On the use of permutation in detecting differential gene expression

Xu Guo and Wei Pan

Division of Biostatistics

School of Public Health

University of Minnesota

Email: weip@biostat.umn.edu

Http: //www.biostat.umn.edu/∼weip

July 2003

Outline

- Introduction
- New Method
- Simulation
- Real Data
- Discussion

1. Introduction

- Goal: detect differential gene expression
- Two-sample comparison $H_{i,0}$: mean expression levels of gene i are the same
- Data: matrix; Features:

  a huge number of genes,

  a small number of arrays,
- A class of nonparametric methods
  - SAM of Tusher et al (2001)
  - EB of Efron et al (2001)
  - MMM of Pan et al (2001)
  - Xu, Olsen and Zhao (2002)

– Key: rank-based; pooled over all genes

- Test stat: Z_i for gene i

  e.g. t-stat or its variants

- Null stat: permuting data and then apply the t-stat to permuted data, $z_i^{(b)}$

- Key assumption: Distr of $z_i^{(b)}$'s is the same as the null distr of Z_i's

  $\Longrightarrow$ Pooling $z_i^{(b)}$'s to estimate the null distr!

- E.g., for any c,

$$\widehat{TP} = \#\{i : |Z_i| > c\}$$

$$\widehat{FP} = \frac{1}{B} \sum_{b=1}^{B} \#\{i : |z_i^{(b)}| > c\}$$

- However, for real data, $H_{i,0}$ holds for some genes, but does not for others!
  $\Longrightarrow$ if not $H_{i,0}$, distr of $z_i^{(b)}$ may differ from the null distr of Z_i!
- Consequence: conservative inference!
  Under-estimate TP or over-estimate FP
  to be shown later
- This problem is known, some methods have appeared
  –Efron et al (2001), Zhao and Pan (2003), Pan (2003)
  –Tricks: take within-sample differences
  –Drawbacks: extra assumptions/conditions,

reduced sample size

- Newton et al (2003): over-estimation of FDR in EB of Efron et al
- A relevant point: a better estimate of FP is $\pi_0 \widehat{FP}$, where π_0 is proportion of non-differentially expressed genes, and π_0 also needs to be estimated

2. New Method

- Trouble: use z_i's of the genes with expression change
- Solution: If know which genes do not have expression change, then use only their z_i's, not others'! –of course, we don't know

- However, we can estimate which genes are likely to have altered expression!

  E.g. EB of Efron et al: p_i=posterior probability of gene i with NO expression change

- Weighting: weight genes proportional to their p_i

- Modify the EM when fitting a finite Normal mixture to $z_i^{(b)}$'s in EB and MMM can have a modified SAM

- A new estimator of FP:

$$\widehat{FP} = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} p_i I(|z_i^{(b)}| > c)$$

3. Simulation

- Set-up: 500+500 genes, 4+4 arrays

  first 500 genes: $Y_{1i}, Y_{2i} \sim N(\mu_i, 1)$, $\mu_i \sim N(0, 5)$

  second half: $Y_{1i} \sim N(\mu_{1i}, 5)$, $Y_{2i} \sim N(\mu_{2i}, 5)$, $\mu_{1i}, \mu_{2i} \sim N(0, 5)$

- $Z_i = \frac{\bar{Y}_{1i} - \bar{Y}_{2i}}{\sqrt{(\frac{1}{J_1} + \frac{1}{J_2})s_i^2 + s_0}}$

  s_0: stablize denominator; chosen to min CV as in SAM; Bayes (Baldi & Long 2001; Lonnstedt & Speed 2002)

- Use MMM (Pan et al, 2003, FIG)

  1. Fit a finite normal mixture f_0 to $z_i^{(b)}$'s

2. For any α, find C s.t.

$$\int_{|z|>C} f_0(z)dz = \alpha$$

3. Gene i significant if $|Z_i| > C$

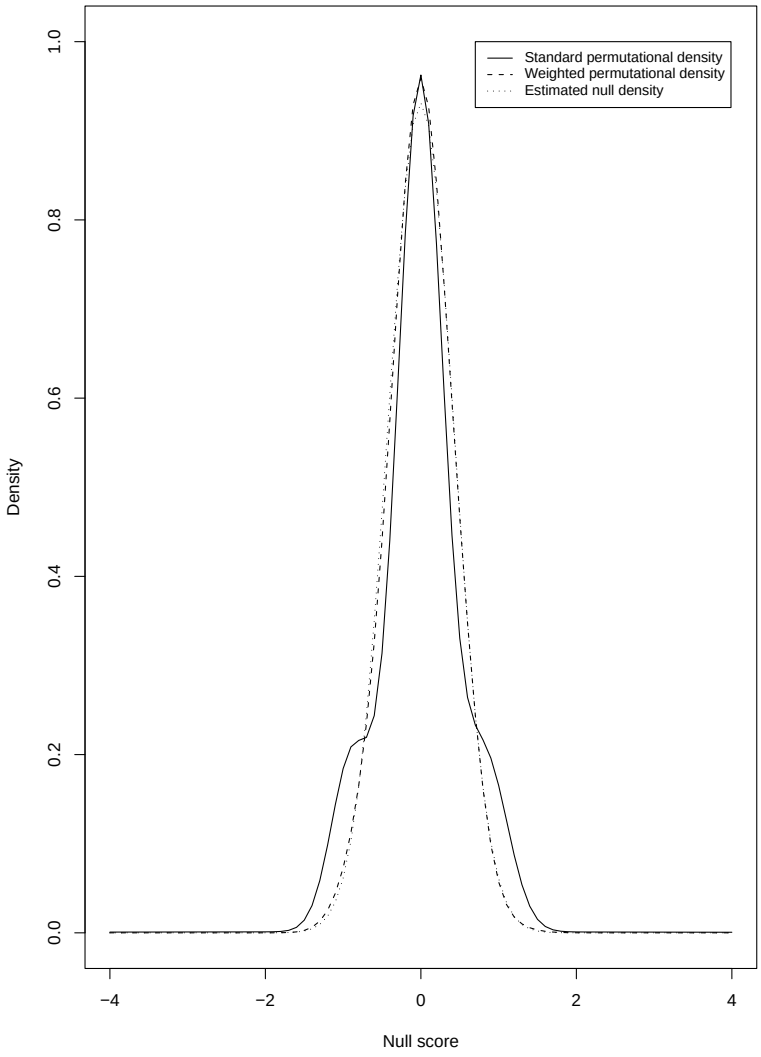
4. Optional: $\widehat{FP} = n\alpha$, or do as before using $z_i^{(b)}$'s

- MMM formalizes ideas in

  Pan et al (2002, GB)

  Broet, Richardson, Radvanyi (2002, JCB)

- Fig 1: estimates of the null distribution of Z_i


Standard permutational density
Weighted permutational density
Estimated null density
Density
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
Null score

	Standard			
α	$\widehat{TP}$	TP	FP	$\widehat{FP}$
0.0001	5	5	0	0
0.0005	20	20	0	0
0.001	39	39	0	1
0.005	207	207	0	5
0.01	356	356	0	10
0.05	392	388	4	54

- Tables: results in MMM

	Weighted				
α	$\widehat{TP}$	TP	FP	$\widehat{FP}$	$\widetilde{FP}$
0.0001	341	341	0	9	0
0.0005	357	357	0	10	0
0.001	363	363	0	12	1
0.005	385	383	2	33	3
0.01	392	388	4	57	6
0.05	435	410	25	148	29

	Estimated				
α	$\widehat{TP}$	TP	FP	$\widehat{FP}$	$\widehat{FP}_0$
0.0001	350	350	0	9	0
0.0005	362	362	0	11	0
0.001	366	365	1	14	1
0.005	387	384	3	41	3
0.01	403	396	7	67	6
0.05	436	410	26	154	27

3. Real Data

- A rare bone marrow cell was identified: mesodermal progenitor cell (MPC) (Reyes et al, 2001).
- MPC can differentiate at single-cell level into mesenchymal cell types such as osteoblasts, chondroblasts and adipocytes, and also into cells of visceral mesodermal origin.
- MPC can be an ideal source of cells to generate osteoblasts to treat bone diseases such as osteoporosis or non-healing fractures, and osteogenesis imperfecta (Hor-

witz et al, 1999).

- Understand the differentiation process of MPC into osteoblasts

  gene regulations of specific signaling proteins and transcription factors (Yamaguchi et al, 2000; Ducy et al, 2000)

- Studied gene expression from undifferentiated MPC (at day 0) to osteoblast lineage-specific differentiation at day 1, day 2 and day 7 by cDNA (Qi et al, PNAS, 2003)

- A key feature: samples taken from the same subject were used to measure gene expression across the seven days.

3 subjects

4132 genes

- Thus, a longitudinal data set with four different time points was generated.
- Q: identify genes differentially expressed over time

  WLOG, only consider days 0, 1, 2
- Test stat: a modified generalized Wald stat (Guo et al, 2003)
- Results:

MMM	Standard		Weighted		
α	$\widehat{TP}$	$\widehat{FP}$	$\widehat{TP}$	$\widehat{FP}$	$\widehat{FP}$
0.0005	8	2	25	7	3
0.001	12	4	40	14	6
0.005	61	23	120	52	30
0.01	108	44	179	81	49

EB	Standard		Weighted		
S	$\widehat{TP}$	$\widehat{FP}$	$\widehat{TP}$	$\widehat{FP}$	$\widehat{FP}$
2	8	2	157	70	29
1.7	15	5	182	83	38
1.4	46	16	213	102	51
1.0	132	56	279	144	79

SAM	Standard		Weighted		
Δ	$\widehat{TP}$	$\widehat{FP}$	$\widehat{TP}$	$\widehat{FP}$	$\widehat{FP}$
50	8	2	11	3	1
35	11	3	24	7	3
20	59	19	71	26	13
12	117	45	143	58	35