Power Calculation for the Overall Test With More Variables than Subjects

Yueh-Yun Chi[1]*, Matthew Gribbin[2], and Keith E. Muller[3]

[1] Department of Biostatistics, University of Florida, Gainesville, FL
[2] Department of Biostatistics, Human Genome Sciences, Rockville, MD
[3] Department of Health Outcomes and Policy, University of Florida , Gainesville, FL
* yychi@ufl.edu

This work is supported primarily by NIH/NIDCR R01-DE020832-01A1 and by NIH/NIDDK R01-DK072398 , NIH/NIDCR U54-DE019261, NIH/NCRR K30-RR022258, NIH/NHLBI R01-HL091005, NIH/NIAAA R01-AA013458-01, and NIH/NIDA R01-DA031017.

Outline

- Motivation
- Overall test
  - Null case
  - Non-null case (power)
- Simulation
- Power analysis
- Discussion

Motivation

- Pathway analysis in microarray analysis
  - Genes are functionally or structurally related
  - System biology-driven analysis
  - Genes in the set are given a priori
  - Also important in metabolomics and proteomics
- An example
  - Wu et al. (2009) analyzed 35 sets of 4-145 genes, N=16 (9 with, 7 without metal particulate exposure)

Overall Hypothesis Testing

- One test on overall significance of the set
- # of genes (p) < sample size (N)
  - MANOVA
- # of genes (p) > sample size (N)
  - Singular sample covariance matrix makes MANOVA statistics undefined
  - Use of regularization (Warton, 2008), generalized inverse (Srivastava, 2007) were proposed to "fix" MANOVA
  - A list of alternative approaches are also available

General Linear Multivariate Model (GLMM)

- Multivariate data can be modeled as follows

$$\underset{(N\times p)}{\boldsymbol{Y}} = \underset{(N\times q)(q\times p)}{\boldsymbol{XB}} + \underset{(N\times p)}{\boldsymbol{E}}$$

- Testing secondary parameters

$$H_0: \underset{(a\times q)(p\times p)(p\times b)}{\boldsymbol{CBU}} = \underset{(a\times b)}{\boldsymbol{\Theta}} = \underset{(a\times b)}{\boldsymbol{\Theta}_0}$$

Overall Test for Null Case

- Our alternative solution (Chi et al., 2012) is a new test for an existing statistic for GLMM
  - Controls the Type I error rate
  - Applies to general design (any GLMM)
  - Applies to data with $p < N$
  - Easy to compute (available in SAS 9.3)

Overall Test for Null Case

- The existing statistic is a ratio of hypothesis to error sums of squares

$$t_u = [\text{tr}(\boldsymbol{S}_h)/a]/[\text{tr}(\boldsymbol{S}_e)/\nu_e]$$

$$\boldsymbol{S}_h = (\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)$$

$$\boldsymbol{S}_e = \nu_e \boldsymbol{U}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}$$

$$\nu_e = N - \text{rank}(\boldsymbol{X})$$

Overall Test for Non-Null Case

- The exact distribution (Theorem 1, Chi et al., in preparation)

$$\Pr\{t_u \le f_0\} = \Pr\left\{\sum_{k=1}^{b} \pi_k y_{kh} - f_0 a \nu_e^{-1} \sum_{k=1}^{b} \pi_k y_{ke} < 0\right\}$$

$$y_{kh} \sim \chi^2(a, \omega_k);\ y_{ke} \sim \chi^2(\nu_e);\ y_{kh} \perp y_{ke};\ \pi_k = \lambda_k \Big/ \left(\sum_{k=1}^{b} \lambda_k\right)$$

  - A minimum set of sufficient parameters
    - Scaled variances of principal components
    - Noncentrality parameters

Overall Test for Non-Null Case

- Noncentrality parameters are 1-1 functions of the squared multiple semi-partial correlations
  - Correlations have better scientific interpretability

$\omega_k = N\rho_k^2/(1-\rho_k^2)$

ρ_k^2: the squared correlation between $\boldsymbol{y}_{uk}$ and the set of predictors tested, with the predictors adjusted for all untested predictors in the model

Overall Test for Non-Null Case

- A convenient non-central F approximation is available (Theorem 3, Chi et al., in preparation)

$$\Pr\{t_u \le f_0\} \approx \Pr\{F(ab\epsilon_n, b\nu_e\epsilon_d, \omega_u) \le f_0\}$$

$$\epsilon_n = \left(a + 2\sum_{k=1}^{b}\pi_k\omega_k\right) \Big/ \left(ab\sum_{k=1}^{b}\pi_k^2 + 2b\sum_{k=1}^{b}\pi_k^2\omega_k\right)$$

$$\epsilon_d = 1\Big/\left(b\sum_{k=1}^{b}\pi_k^2\right) = \epsilon \text{ (Sphericity parameter)}$$

$$\omega_u = \left(\sum_{k=1}^{b}\pi_k\omega_k\right) b\epsilon_n$$

Simulation – One Sample Problem

Number of Outcomes	Number of Conditions	Approximated Power	Empirical, Absolute Bias in Power	
			Max	Mean
64	36	0.20	0.010	0.004
64	36	0.50	0.024	0.004
64	36	0.80	0.005	0.002
64	36	0.90	0.013	0.002
256	36	0.80	0.005	0.002
256	36	0.90	0.005	0.002
1024	36	0.80	0.004	0.001
1024	36	0.90	0.003	0.001

$N \in \{10,\ 20,\ 40\}, \epsilon \in \{0.27, 0.56, 0.76\}$, number of nonzero ρ_k^2 of either 4 or 32, Location of nonzero ρ_k^2 at either the most dominant or middle components

Simulation – Two Sample Problem

Number of Outcomes	Number of Conditions	Approximated Power	Empirical, Absolute Bias in Power	
			Max	Mean
64	36	0.20	0.001	0.006
64	36	0.50	0.034	0.011
64	36	0.80	0.037	0.011
64	36	0.90	0.028	0.009
256	36	0.80	0.031	0.010
256	36	0.90	0.024	0.009
1024	36	0.80	0.027	0.011
1024	36	0.90	0.022	0.009

$N \in \{10,\ 20,\ 40\}, \epsilon \in \{0.27, 0.56, 0.76\}$, number of nonzero ρ_k^2 of either 4 or 32, Location of nonzero ρ_k^2 at either the most dominant or middle components

Power Analysis

- Seven input components
  - Type I error rate
  - Design matrix
  - Between-subject contrast matrix
  - Within-subject contrast matrix
  - Null matrix
  - Primary parameters matrix
  - Error covariance matrix

Power Analysis

- Number of parameters explodes when $p >> N$

Type I error rate	α	1×1
Design matrix	$\boldsymbol{X}$	$N \times q$
Between-subject contrast matrix	$\boldsymbol{C}$	$a \times q$
Within-subject contrast matrix	$\boldsymbol{U}$	$p \times b$
Null matrix	$\boldsymbol{\Theta}_0$	$a \times b$
Primary parameters matrix	$\boldsymbol{B}$	$q \times p$
Error covariance matrix	$\boldsymbol{\Sigma}$	$p \times p$

Power Analysis

- Power equivalence simplifies problem (Thm. 2)

	Hypothesis Testing Scenario	
Feature	S_1	S_2
Model	$\boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{B}_1+\boldsymbol{E}_1$	$\boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{B}_2+\boldsymbol{E}_2$
# of Outcomes	p	b
# of Predictors	q	$r = \text{Rank}(\boldsymbol{X}_1)$
Design Matrix Feature	$\boldsymbol{X}_1$	$\boldsymbol{X}_2'\boldsymbol{X}_2 = \boldsymbol{I}_r$
Error Covariance	$\mathcal{V}(\boldsymbol{E}_1) = \boldsymbol{\Sigma}$	$\mathcal{V}(\boldsymbol{E}_2) = \text{Dg}(\boldsymbol{\pi})$
Between-Subject Contrast	$\boldsymbol{C}_1$	$\boldsymbol{C}_2 = [\,\boldsymbol{I}_a\ \boldsymbol{0}\,]$
Within-Subject Contrast	$\boldsymbol{U}_1$	$\boldsymbol{U}_2 = \boldsymbol{I}_b$
Primary Parameters	$\boldsymbol{B}_1$	$\boldsymbol{B}_2 = [\,\boldsymbol{\beta}_2\ \boldsymbol{0}\,]'$

$\beta_{2k} = (\pi_k\omega_k)^{1/2},\ \pi_k = \lambda_k / \left(\sum_{k=1}^{b}\lambda_k\right)$

Conclusion for $p > N$

- Overall testing is important for pathway analysis
- We have
  - A size α general test for null case
  - Accurate power approximation
  - Software for testing available
  - Updated power software underway
- Power equivalent scenarios help simplify the power analysis

Discussion

- With high outcome dimension, practicing safe computing is particularly essential to ensure numerical accuracy
- Power calculation with random covariate will be explored for future research

Reference

- Chi et al. (2012) Statistics in Medicine, in press
- Srivastava (2007) Journal of Japanese Statistical Society, 37, 53-86
- Warton (2008) JASA, 103, 340-349
- Wu et al. (2009) Bioinformatics, 25, 1145-1151