Conference on Applied Statistics in Defense

Bureau of Labor Statistics
Washington, DC

October 20-24, 2014

Analysis of High Dimensional Biomarker Data for Binary Outcomes

Hongda Zhang[1], Hua Liang[2], Colin Wu[3], Yuanzhang Li[4]

[1]Digital System Inc., Chevy Chase, MD

[2]George Washington University

[3]National Institutes of Health

[4]Walter Reed Army Institute of Research

Disclosure

The opinions or assertions contained herein are the private views of the author(s), and are not to be construed as official, or as reflecting true views of the Department of the Army or the Department of Defense. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this presentation.

Introduction

The effect of individual predictors in a multiple regression may be biased due to multicollinearity.

Multicollinearity often occurs in longitudinal studies, especially when the objective is to study the association between biomarkers and a specific disease.

Biomarker Study Concepts

1. Biological predictors of disease (e.g., schizophrenia)
2. Nested case-control design
3. Binary outcome: disease status
4. Multiple serum specimens per case and control
5. Demographic and medical information available on all subjects

Objectives

1. Identify a biomarker signature that distinguishes individuals with disease from the general population
2. Identify a biomarker that distinguish cases from controls

Methods

1. Decompose the space of X, consisting of all independent variables according to their association, such that all biomarkers in any subspace are independent.
2. Find the gradient, as the linear combination of the biomarkers, that can best separate schizophrenia cases from controls, as well as the perpendicular vectors to the gradient in each subspace.
3. Perform general linear regression (GLR) based on the gradient direction and other significant vectors for dimension reduction and case identification.
4. Propose a sum statistic test used to select biomarkers.

Space Decomposition

- In regression, x_1, x_2, …, x_k are assumed to be independent.
- The multicollinerity in the regression can be solved by space decomposition.
- We divided the whole space into several subspaces.

Space Decomposition

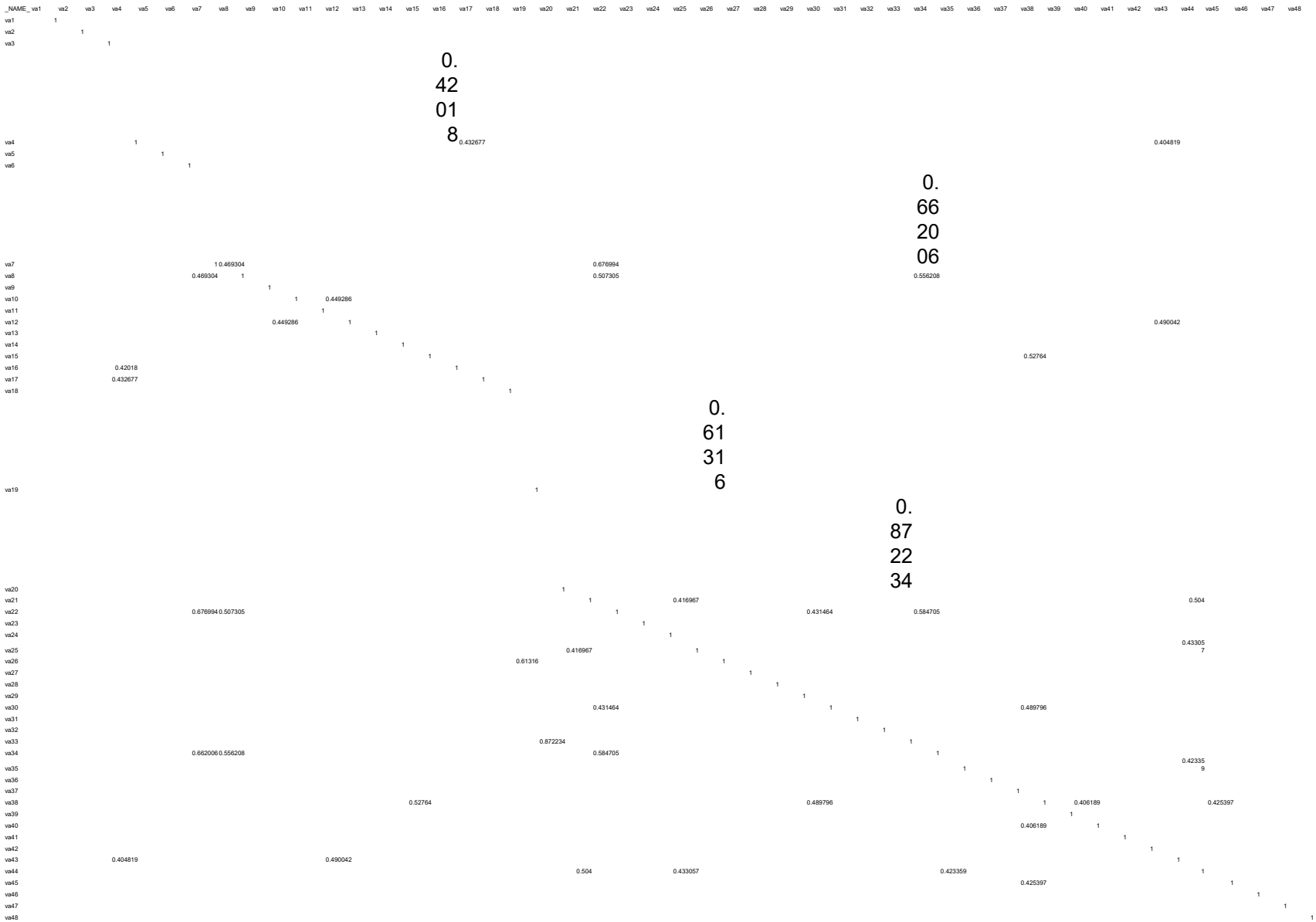
$$\sum = \begin{pmatrix} \Sigma_1 & C_{12} & C_{13} & \dots C_{1k} \\ C_{21} & \Sigma_2 & C_{22} & \dots C_{2k} \\ & & \dots & \\ C_{k1} & C_{k2} & C_{k3} & \dots \Sigma_k \end{pmatrix}$$

Where

$$\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & 0 & 0 & \dots 0 \\ 0 & \sigma_{i2}^2 & 0 & \dots 0 \\ & & \dots & \\ 0 & 0 & \dots 0 & \sigma_{ip_i}^2 \end{pmatrix}$$

for i=1,2,...k

Correlation Example



Observations

1. High correlation:
   - Exists
   - Sparse
2. This makes decomposing data possible

Gradient-Nuisance Direction-Orthogonal Base

$$\mathbf{g(Y)} \approx \mathbf{f(X,\beta)+\varepsilon} \qquad \mathbf{(1)}$$

$$E(g(Y) \mid \mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta}) \qquad (2)$$

$$E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = 0 \qquad (3)$$

The Space X will be decomposed into two parts: U and V. The vectors in Space U will be highly associated with g(y), and the vectors in V will have almost no association with g(y).
Usually in our approach, unlike PCA, the gradient is the only factor in U.

Binary Outcome Study

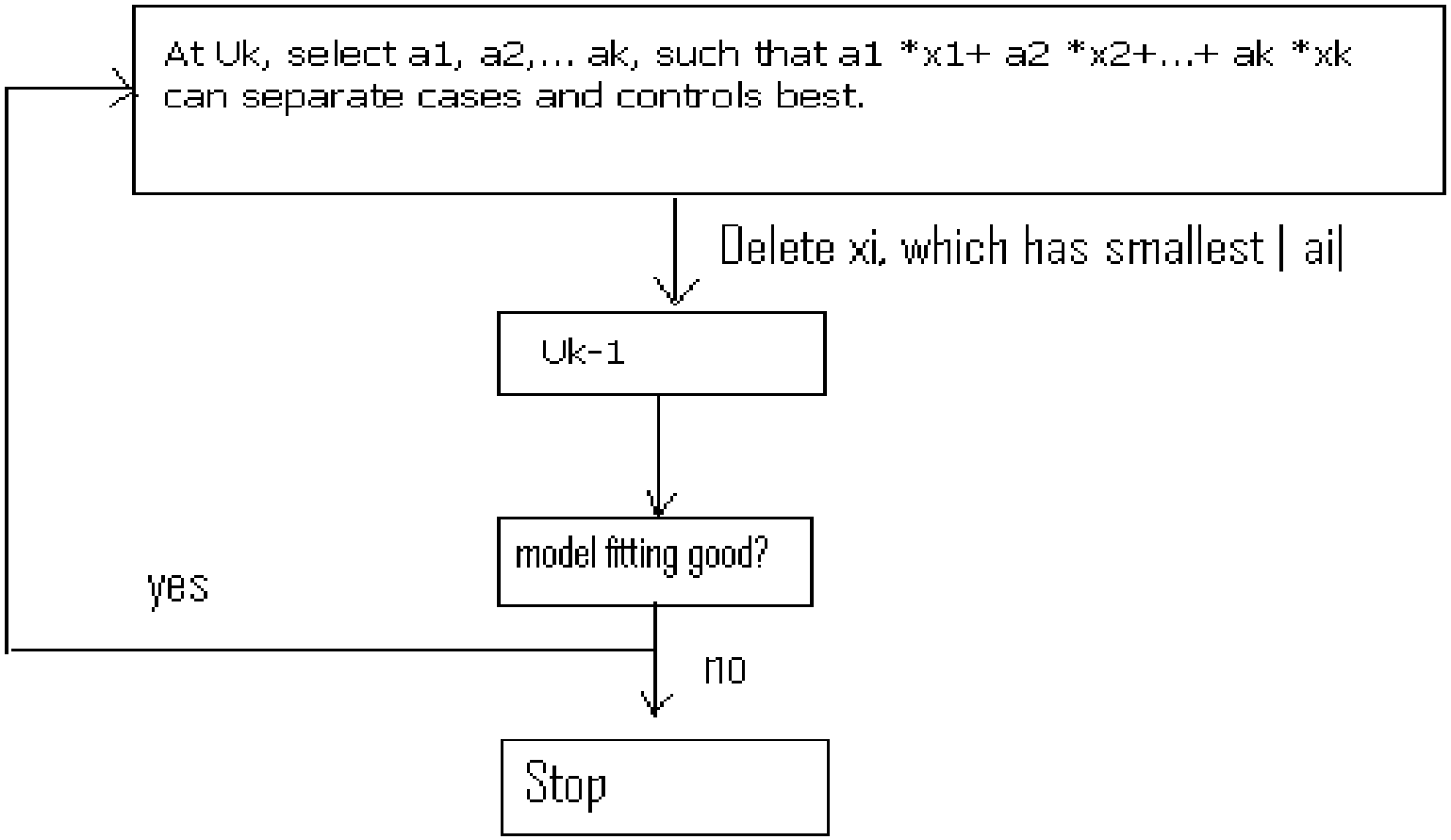
- The probability density functions of **X for** Y=0 or 1, normal $(\mu_0, \sum_0)$ and $(\mu_1, \sum_1)$
- The ratio of the log-likelihoods

$$(X-\mu_0)^T\sum_0^{-1}(X-\mu_0)+ln|\sum_0|-(X-\mu_1)^T\sum_1^{-1}(X-\mu_1)-ln|\sum_1|<T$$

- Solution to get ω_1 such that $\boldsymbol{\omega 1 \cdot X} < c$, where $\boldsymbol{\omega 1 = \Sigma^{-1}(\mu_1 - \mu_0)}$
- ω_j is the one:

Minimize $\{|(\omega_j \cdot \omega_1)| + |(\omega_j \cdot \omega_2)| + \ldots + |(\omega_j \cdot \omega_{j-1})|\}$ j=2, 3,4,…K; (4)

Deduction Approach 1: Ad Hoc. Gradient-Reduction-Sequence



Simulation for Selection

Assume two factors (biomarkers) are associated with binary outcomes

RESULTS

Sample Size=200							Sample Size=200						
		Number of Biomarkers							Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10	Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.89	0.84	0.85	0.85	0.85	(1.5,1.0)	Mean of Sensitivity	0.86	0.85	0.85	0.85	0.85
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.03	0.06	0.05	0.05	0.05		SD of Sensitivity	0.04	0.05	0.05	0.05	0.05
	X1 Remains	100.00%	100.00%	100.00%	100.00%	100.00%		X1 Remains	99.00%	100.00%	100.00%	100.00%	100.00%
	X2 Remains	100.00%	100.00%	100.00%	100.00%	100.00%		X2 Remains	95.00%	95.00%	91.00%	99.00%	90.00%
(2.0,1.0)	Mean of Sensitivity	0.88	0.85	0.85	0.85	0.85	(1.5,0.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.03	0.06	0.05	0.05	0.05		SD of Sensitivity	0.05	0.05	0.05	0.05	0.05
	X1 Remains	100.00%	100.00%	100.00%	100.00%	100.00%		X1 Remains	100.00%	100.00%	100.00%	100.00%	100.00%
	X2 Remains	93.00%	89.00%	85.00%	92.00%	89.00%		X2 Remains	24.00%	35.00%	44.00%	35.00%	34.00%
(2.0,0.5)	Mean of Sensitivity	0.87	0.85	0.85	0.85	0.85	(1.0,1.0)	Mean of Sensitivity	0.83	0.84	0.85	0.85	0.85
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.03	0.05	0.05	0.05	0.05		SD of Sensitivity	0.05	0.05	0.05	0.05	0.05
	X1 Remains	100.00%	100.00%	100.00%	100.00%	100.00%		X1 Remains	91.00%	95.00%	96.00%	99.00%	99.00%
	X2 Remains	22.00%	18.00%	20.00%	17.00%	24.00%		X2 Remains	98.00%	96.00%	95.00%	95.00%	94.00%
(1.5,1.5)	Mean of Sensitivity	0.87	0.85	0.85	0.85	0.85							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.03	0.05	0.05	0.05	0.05							
	X1 Remains	100.00%	100.00%	100.00%	100.00%	100.00%							
	X2 Remains	100.00%	100.00%	100.00%	100.00%	100.00%							

RESULTS

Sample Size=150							Sample Size=150						
		Number of Biomarkers							Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10	Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84	(1.5,1.0)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.05	0.06	0.06	0.07	0.07		SD of Sensitivity	0.05	0.06	0.06	0.07	0.07
	X1 Remains	82.00%	88.00%	88.00%	88.00%	89.00%		X1 Remains	80.00%	81.00%	74.00%	80.00%	77.00%
	X2 Remains	80.00%	81.00%	68.00%	78.00%	78.00%		X2 Remains	50.00%	53.00%	52.00%	51.00%	58.00%
(2.0,1.0)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84	(1.5,0.5)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.05	0.06	0.06	0.07	0.07		SD of Sensitivity	0.06	0.06	0.06	0.07	0.07
	X1 Remains	90.00%	90.00%	86.00%	82.00%	85.00%		X1 Remains	90.00%	85.00%	85.00%	79.00%	84.00%
	X2 Remains	35.00%	48.00%	55.00%	52.00%	57.00%		X2 Remains	10.00%	14.00%	15.00%	14.00%	12.00%
(2.0,0.5)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84	(1.0,1.0)	Mean of Sensitivity	0.84	0.84	0.84	0.84	0.84
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.05	0.06	0.06	0.07	0.07		SD of Sensitivity	0.06	0.06	0.07	0.07	0.07
	X1 Remains	88.00%	88.00%	85.00%	85.00%	89.00%		X1 Remains	57.00%	53.00%	58.00%	66.00%	59.00%
	X2 Remains	12.00%	13.00%	15.00%	19.00%	14.00%		X2 Remains	64.00%	62.00%	50.00%	59.00%	65.00%
(1.5,1.5)	Mean of Sensitivity	0.85	0.84	0.84	0.84	0.84							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.05	0.06	0.06	0.07	0.07							
	X1 Remains	73.00%	62.00%	78.00%	76.00%	66.00%							
	X2 Remains	79.00%	75.00%	76.00%	77.00%	62.00%							

RESULTS

Sample Size=100							Sample Size=100						
		Number of Biomarkers							Number of Biomarkers				
		2	4	6	8	10	(1.5,1.0)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85
Mean Pairs	Statistics							Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
(2.0,1.5)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		X1 Remains	96.00%	99.00%	95.00%	95.00%	97.00%
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		X2 Remains	64.00%	68.00%	69.00%	56.00%	73.00%
	X1 Remains	99.00%	100.00%	100.00%	99.00%	100.00%	(1.5,0.5)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85
	X2 Remains	96.00%	91.00%	86.00%	96.00%	94.00%		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
(2.0,1.0)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		X1 Remains	95.00%	96.00%	98.00%	100.00%	98.00%
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		X2 Remains	8.00%	14.00%	12.00%	11.00%	16.00%
	X1 Remains	99.00%	100.00%	100.00%	100.00%	100.00%	(1.0,1.0)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85
	X2 Remains	51.00%	54.00%	64.00%	53.00%	62.00%		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
(2.0,0.5)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		X1 Remains	68.00%	76.00%	75.00%	65.00%	72.00%
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		X2 Remains	62.00%	69.00%	71.00%	65.00%	64.00%
	X1 Remains	100.00%	99.00%	100.00%	100.00%	98.00%							
	X2 Remains	6.00%	8.00%	9.00%	18.00%	16.00%							
(1.5,1.5)	Mean of Sensitivity	0.84	0.84	0.84	0.85	0.85							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07							
	X1 Remains	100.00%	98.00%	97.00%	98.00%	94.00%							
	X2 Remains	90.00%	95.00%	98.00%	96.00%	95.00%							

RESULTS

Sample Size=80							Sample Size=80						
		Number of Biomarkers							Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10	Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.88	0.84	0.85	0.86	0.87	(1.5,1.0)	Mean of Sensitivity	0.85	0.85	0.86	0.87	0.87
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.05	0.08	0.07	0.07	0.07		SD of Sensitivity	0.06	0.07	0.07	0.07	0.07
	X1 Remains	98.00%	95.00%	99.00%	99.00%	97.00%		X1 Remains	91.00%	92.00%	93.00%	91.00%	86.00%
	X2 Remains	81.00%	89.00%	85.00%	90.00%	83.00%		X2 Remains	45.00%	50.00%	50.00%	56.00%	50.00%
(2.0,1.0)	Mean of Sensitivity	0.88	0.85	0.86	0.86	0.87	(1.5,0.5)	Mean of Sensitivity	0.84	0.85	0.86	0.87	0.87
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.05	0.08	0.07	0.07	0.07		SD of Sensitivity	0.06	0.07	0.07	0.07	0.07
	X1 Remains	98.00%	100.00%	100.00%	98.00%	100.00%		X1 Remains	90.00%	93.00%	91.00%	90.00%	91.00%
	X2 Remains	39.00%	51.00%	45.00%	52.00%	41.00%		X2 Remains	11.00%	10.00%	13.00%	15.00%	14.00%
(2.0,0.5)	Mean of Sensitivity	0.87	0.85	0.86	0.87	0.87	(1.0,1.0)	Mean of Sensitivity	0.83	0.85	0.86	0.87	0.87
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.05	0.07	0.07	0.07	0.07		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	X1 Remains	100.00%	99.00%	97.00%	100.00%	99.00%		X1 Remains	51.00%	70.00%	57.00%	61.00%	62.00%
	X2 Remains	1.00%	6.00%	10.00%	7.00%	9.00%		X2 Remains	60.00%	52.00%	68.00%	66.00%	55.00%
(1.5,1.5)	Mean of Sensitivity	0.87	0.85	0.86	0.87	0.87							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.05	0.07	0.07	0.07	0.07							
	X1 Remains	90.00%	90.00%	89.00%	87.00%	92.00%							
	X2 Remains	87.00%	88.00%	81.00%	93.00%	89.00%							

RESULTS

Sample Size=60							Sample Size=60						
		Number of Biomarkers							Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10	Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.87	0.87	0.86	0.87	0.87	(1.5,1.0)	Mean of Sensitivity	0.87	0.87	0.87	0.87	0.87
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.07	0.07	0.08	0.08	0.08		SD of Sensitivity	0.07	0.07	0.08	0.08	0.08
	X1 Remains	88.00%	86.00%	91.00%	90.00%	92.00%		X1 Remains	60.00%	72.00%	65.00%	73.00%	71.00%
	X2 Remains	56.00%	71.00%	55.00%	59.00%	59.00%		X2 Remains	25.00%	33.00%	38.00%	30.00%	39.00%
(2.0,1.0)	Mean of Sensitivity	0.87	0.87	0.86	0.87	0.87	(1.5,0.5)	Mean of Sensitivity	0.87	0.86	0.87	0.87	0.87
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.07	0.07	0.08	0.08	0.08		SD of Sensitivity	0.07	0.07	0.08	0.08	0.08
	X1 Remains	94.00%	92.00%	92.00%	94.00%	93.00%		X1 Remains	68.00%	69.00%	72.00%	73.00%	68.00%
	X2 Remains	25.00%	22.00%	32.00%	18.00%	27.00%		X2 Remains	2.00%	4.00%	9.00%	3.00%	4.00%
(2.0,0.5)	Mean of Sensitivity	0.87	0.87	0.87	0.87	0.87	(1.0,1.0)	Mean of Sensitivity	0.86	0.86	0.86	0.87	0.87
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.07	0.07	0.08	0.08	0.08		SD of Sensitivity	0.07	0.08	0.08	0.08	0.08
	X1 Remains	93.00%	91.00%	95.00%	95.00%	93.00%		X1 Remains	30.00%	22.00%	25.00%	32.00%	35.00%
	X2 Remains	2.00%	2.00%	4.00%	4.00%	8.00%		X2 Remains	33.00%	31.00%	23.00%	39.00%	43.00%
(1.5,1.5)	Mean of Sensitivity	0.87	0.87	0.87	0.87	0.87							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.07	0.07	0.08	0.08	0.08							
	X1 Remains	66.00%	67.00%	71.00%	67.00%	74.00%							
	X2 Remains	62.00%	60.00%	67.00%	72.00%	59.00%							

RESULTS

Sample Size=50							Sample Size=50						
		Number of Biomarkers							Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10	Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85	(1.5,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89		Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	X1 Remains	69.00%	77.00%	66.00%	71.00%	74.00%		X1 Remains	43.00%	45.00%	53.00%	55.00%	58.00%
	X2 Remains	37.00%	38.00%	42.00%	35.00%	44.00%		X2 Remains	24.00%	18.00%	23.00%	16.00%	28.00%
(2.0,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85	(1.5,0.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87		Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	X1 Remains	71.00%	77.00%	63.00%	67.00%	75.00%		X1 Remains	49.00%	52.00%	44.00%	58.00%	54.00%
	X2 Remains	18.00%	21.00%	11.00%	16.00%	13.00%		X2 Remains	4.00%	5.00%	3.00%	5.00%	4.00%
(2.0,0.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85	(1.0,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85		Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07		SD of Sensitivity	0.07	0.07	0.07	0.07	0.07
	X1 Remains	68.00%	71.00%	78.00%	78.00%	70.00%		X1 Remains	19.00%	24.00%	21.00%	30.00%	23.00%
	X2 Remains	1.00%	3.00%	6.00%	4.00%	3.00%		X2 Remains	15.00%	13.00%	25.00%	18.00%	17.00%
(1.5,1.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.85							
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86							
	SD of Sensitivity	0.07	0.07	0.07	0.07	0.07							
	X1 Remains	49.00%	46.00%	44.00%	40.00%	46.00%							
	X2 Remains	48.00%	46.00%	44.00%	44.00%	41.00%							

RESULTS

Sample Size=30						
		Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10
(2.0,1.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.89	0.89	0.89	0.89	0.89
	SD of Sensitivity	0.07	0.08	0.08	0.08	0.08
	X1 Remains	55.00%	58.00%	51.00%	54.00%	56.00%
	X2 Remains	29.00%	35.00%	22.00%	26.00%	30.00%
(2.0,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.87	0.87	0.87	0.87	0.87
	SD of Sensitivity	0.07	0.08	0.08	0.08	0.08
	X1 Remains	48.00%	53.00%	52.00%	60.00%	58.00%
	X2 Remains	9.00%	10.00%	5.00%	11.00%	9.00%
(2.0,0.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.85	0.85	0.85	0.85	0.85
	SD of Sensitivity	0.07	0.08	0.08	0.08	0.08
	X1 Remains	56.00%	61.00%	62.00%	57.00%	56.00%
	X2 Remains	1.00%	0.00%	1.00%	1.00%	3.00%
(1.5,1.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.86	0.86	0.86	0.86	0.86
	SD of Sensitivity	0.08	0.08	0.08	0.08	0.08
	X1 Remains	30.00%	33.00%	33.00%	34.00%	34.00%
	X2 Remains	25.00%	27.00%	34.00%	35.00%	25.00%

Sample Size=30						
		Number of Biomarkers				
Mean Pairs	Statistics	2	4	6	8	10
(1.5,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.82	0.82	0.82	0.82	0.82
	SD of Sensitivity	0.08	0.08	0.08	0.08	0.08
	X1 Remains	26.00%	27.00%	32.00%	33.00%	32.00%
	X2 Remains	13.00%	9.00%	5.00%	15.00%	11.00%
(1.5,0.5)	Mean of Sensitivity	0.85	0.85	0.85	0.85	0.86
	Theoretical Sensitivity	0.79	0.79	0.79	0.79	0.79
	SD of Sensitivity	0.08	0.08	0.08	0.08	0.08
	X1 Remains	30.00%	36.00%	27.00%	32.00%	30.00%
	X2 Remains	3.00%	2.00%	3.00%	6.00%	3.00%
(1.0,1.0)	Mean of Sensitivity	0.85	0.85	0.85	0.86	0.86
	Theoretical Sensitivity	0.76	0.76	0.76	0.76	0.76
	SD of Sensitivity	0.08	0.08	0.08	0.08	0.08
	X1 Remains	10.00%	12.00%	18.00%	20.00%	17.00%
	X2 Remains	8.00%	14.00%	11.00%	11.00%	13.00%

Regression

In general, for binary outcomes, a logistic model is commonly used

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha + \beta_1 \omega_i + \beta_2 t_i + \beta_3 \omega_i t_i$$

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha + \beta_1 \omega_i + \beta_2 t_{ij} + \beta_3 \omega_i t_{ij}$$

Regression on Gradients

1. Find the plane best separating cases from controls in a d-dimensional space $\mathbf{S}_d$, d=p, p-1, ..., where p is the number of predictors in the space. The new variable, generated by the normal vector of the plane denoted as ω_{d1}, is called "gradient."

2. Find other vectors that are orthogonal to each other as well as to the gradient and generate new variables: ω_{d2}, ω_{d3}, …, ω_{dd}.

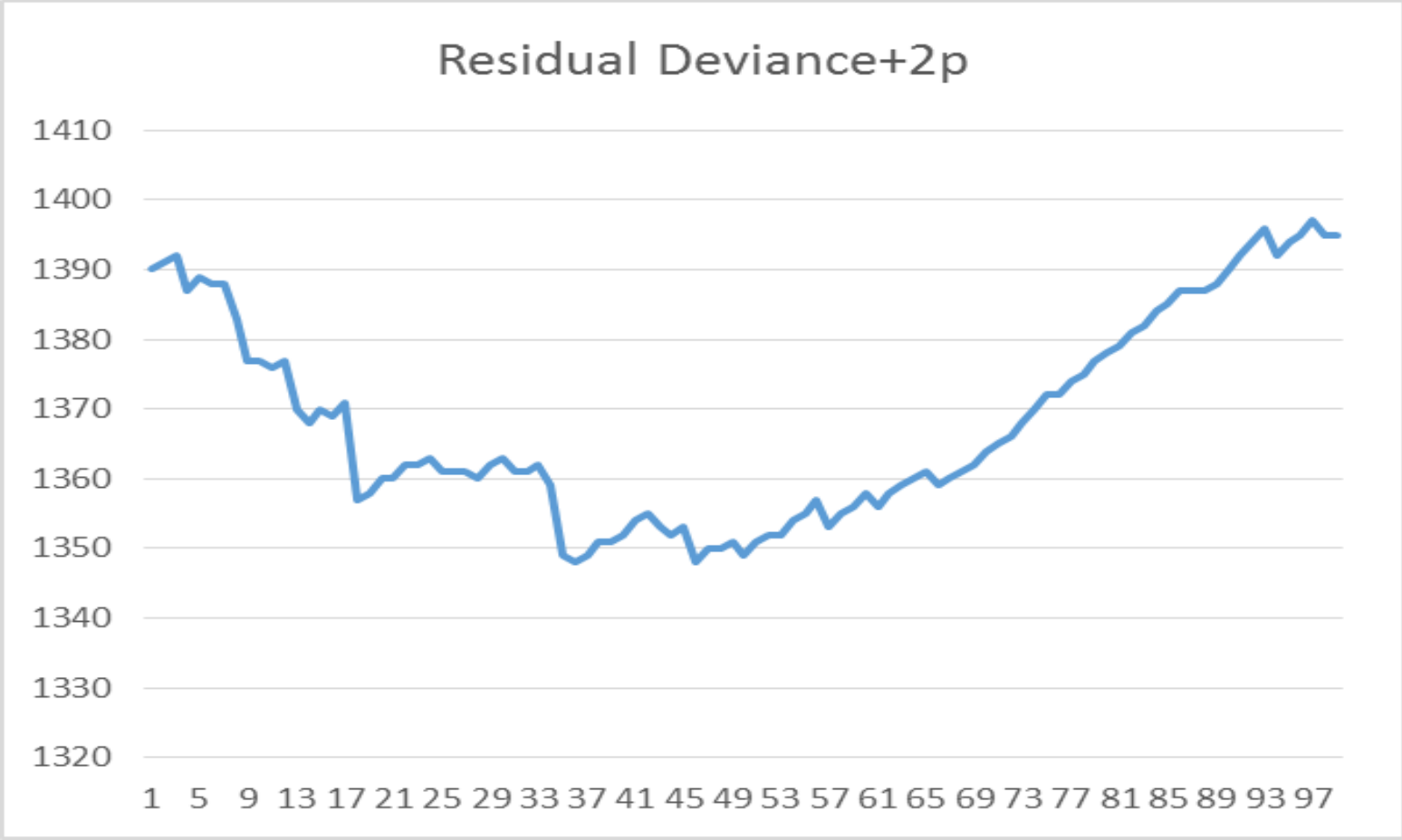
Regression on Gradients Cont.

3. Perform regression modeling on ω_{d1} and other ω_{di} that show significant effect on outcome of $\mathbf{g}(\mathbf{Y})$. We use $\mathbf{U}_d$ to denote the subspace of $\mathbf{S}_d$. Remove the factors with small absolute coefficients in the gradients and other significant vectors. Then repeat Steps 2 and 3 to continue to reduce the number of biomarkers.
4. Use the logarithm of the likelihood G^2 to measure goodness of fit. Simulation results show that the gradient absorbs almost all information from existing biomarkers, hence $\Delta G^2 = G^2(\omega_{p1}) - G^2(\omega_{d1}) \approx G^2(\mathbf{U}_p) - G^2(\mathbf{U}_d)$. If ΔG^2 is large or significant (the degrees of freedom of the ΔG^2 would be approximately p-d), factor elimination may stop. An alternative stop rule would be to examine $\Delta G^2 = G^2(\omega_{d+1}) - G^2(\omega_{d1})$ with 1 degree of freedom. In practice, we may stop eliminating biomarkers based on biological plausibility.

Biomarker Selection by Akaike Information Criterion

- How many biomarkers should be included can be determined by AIC, BIC or QIC.
- The gradient consist of p biomarkers, it has one degree freedom in the regression, but it consist of p biomarkers, we should consider AIC has p degrees of freedom rather than 1.
- A simulation of 100 biomarkers, 50 are assumed to be associated and 50 are not, the AIC should achieve minimum at 50.

Biomarker Selection by Akaike Information Criterion



Sum of Statistic

- Gradient vector $\boldsymbol{\omega} = (u_1, u_2, \ldots, u_P)^{\tau}$, and
- $\sum u_i^2 = 1$, defines the sum statistic as
- $Sumc = \sum |u_i|$,
- Range (1, √P)
- Under the null hypothesis with random results and no effect on outcome, then sumc=√P.
- If the null hypothesis is not true, and one of the biomarkers has the strongest effect while the others have no effect on outcome, then sumc=1.

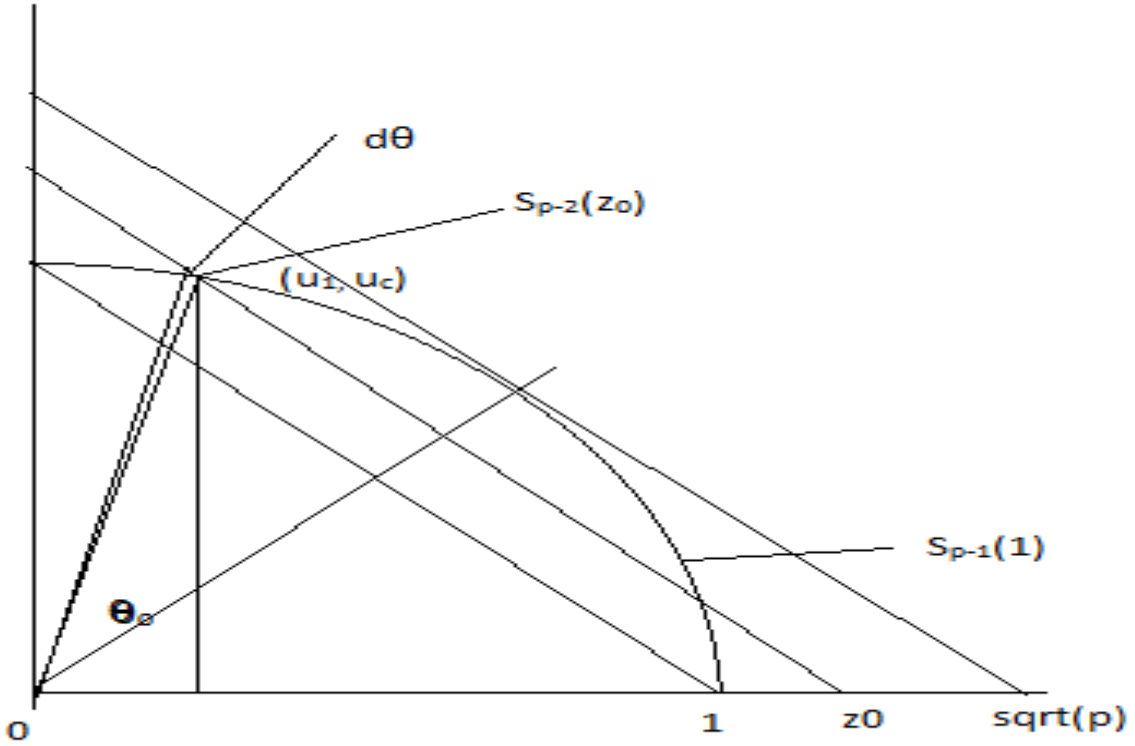
Integration on the Surface of P-Sphere

- For r near 1, the integration domain is a circle, but for r near $1/\sqrt{p}$, the integration domain is part of circle.

$$S_n(R) = \frac{(P-1)\pi^{\frac{P-1}{2}}}{\Gamma(\frac{P}{2}+1)} R^{p-1}$$

- The distribution percentiles can be used to delete non-effect biomarkers simultaneously.
- H_0: all biomarkers are not associated with the binary outcome.
- If gradient score has no significant effect and sumc>selected critical value, say 0.05, or 0.10.

Evaluated Distribution of Sumc



Percentiles of Sumc

P	1%	5%	10%	20%	50%	80%	90%	95%	99%
2	1.008	1.039	1.076	1.144	1.307	1.397	1.410	1.413	1.414
3	1.099	1.211	1.285	1.376	1.515	1.646	1.689	1.710	1.728
4	1.247	1.393	1.469	1.559	1.716	1.847	1.904	1.940	1.979
5	1.399	1.556	1.637	1.731	1.893	2.029	2.090	2.133	2.190
6	1.549	1.710	1.793	1.889	2.055	2.196	2.259	2.305	2.374
7	1.689	1.855	1.939	2.036	2.204	2.349	2.414	2.463	2.539
8	1.823	1.992	2.076	2.174	2.344	2.492	2.559	2.609	2.691
9	1.953	2.121	2.207	2.305	2.477	2.626	2.696	2.748	2.833
10	2.074	2.244	2.330	2.429	2.603	2.754	2.824	2.878	2.967
11	2.191	2.362	2.449	2.548	2.722	2.875	2.947	3.002	3.093
12	2.304	2.476	2.562	2.661	2.836	2.991	3.063	3.119	3.213
13	2.412	2.585	2.671	2.771	2.947	3.102	3.176	3.233	3.329

150	9.276	9.433	9.515	9.613	9.793	9.967	10.055	10.126	10.258
155	9.437	9.594	9.676	9.774	9.955	10.129	10.217	10.288	10.420
160	9.596	9.753	9.835	9.931	10.113	10.287	10.375	10.447	10.578
165	9.752	9.910	9.991	10.089	10.269	10.443	10.531	10.603	10.733
170	9.908	10.064	10.145	10.242	10.423	10.597	10.686	10.757	10.889
175	10.059	10.216	10.297	10.394	10.575	10.749	10.838	10.910	11.041
180	10.208	10.365	10.447	10.544	10.724	10.899	10.987	11.059	11.191
185	10.355	10.512	10.593	10.690	10.871	11.046	11.135	11.207	11.338
190	10.501	10.657	10.739	10.837	11.017	11.191	11.280	11.352	11.484
195	10.644	10.801	10.883	10.980	11.161	11.335	11.424	11.497	11.629
200	10.785	10.942	11.025	11.122	11.302	11.477	11.566	11.638	11.769

Simultaneous Deletion of Biomarkers

- H_0: no biomarker has an effect on the binary outcome
- H_a: some biomarkers have an effect on the binary outcome
- If P(sumc < observed sumc)>α, and the gradient is not significant, then H_0 is not rejected, and all biomarkers can be removed. Otherwise, the biomarkers are divided into two parts, and used the above test is performed separately. In general, only a few biomarkers show an effect on the risk of outcome.

DATA

"Data from the Defense Medical Surveillance System, The Armed Forces Health Surveillance Center, U.S. Department of Defense, Silver Spring, Maryland [Data from 1988 to 2006; released in 2007]"

"Serum specimens from the Department of Defense Serum Repository: The Armed Forced Health Surveillance Center, U.S. Department of Defense, Silver Spring, Maryland [Data from 1990 to 2006; released in 2007]".

Study Design

- Nested case-control
- Multiple serum specimens per case and control
- Cases and controls were matched on their military accession date (±12 months), date of birth (±12 months), sex, race, branch of military service, and the serum specimen draw dates (±90 days)
- All cases had psychiatric evaluations with full clinical evaluations that applied DSM-IV diagnostic criteria

Steps

1. Decompose space X which consists of all independent variables ranked by the significance of their associations starting with Prolactin (PRL) into subspaces devoid of highly correlated variables (X=AUBUC).

   Aim: avoid co-linearity in regression and biomarker selection
2. Find the gradient (e.g., ω_A, ω_B), the linear combination of the biomarkers in each subspace that can best separate the schizophrenia cases from controls in the corresponding subspace as well as the perpendicular vectors to the gradient in each subspace.

   Aim: Reducing dimension
3. Use general linear regression (GLR), which is based on all gradients and other significant vectors, for dimension reduction and case identification. For longitudinal data, the GEE GLR is used.

   Aim: Study association and identification

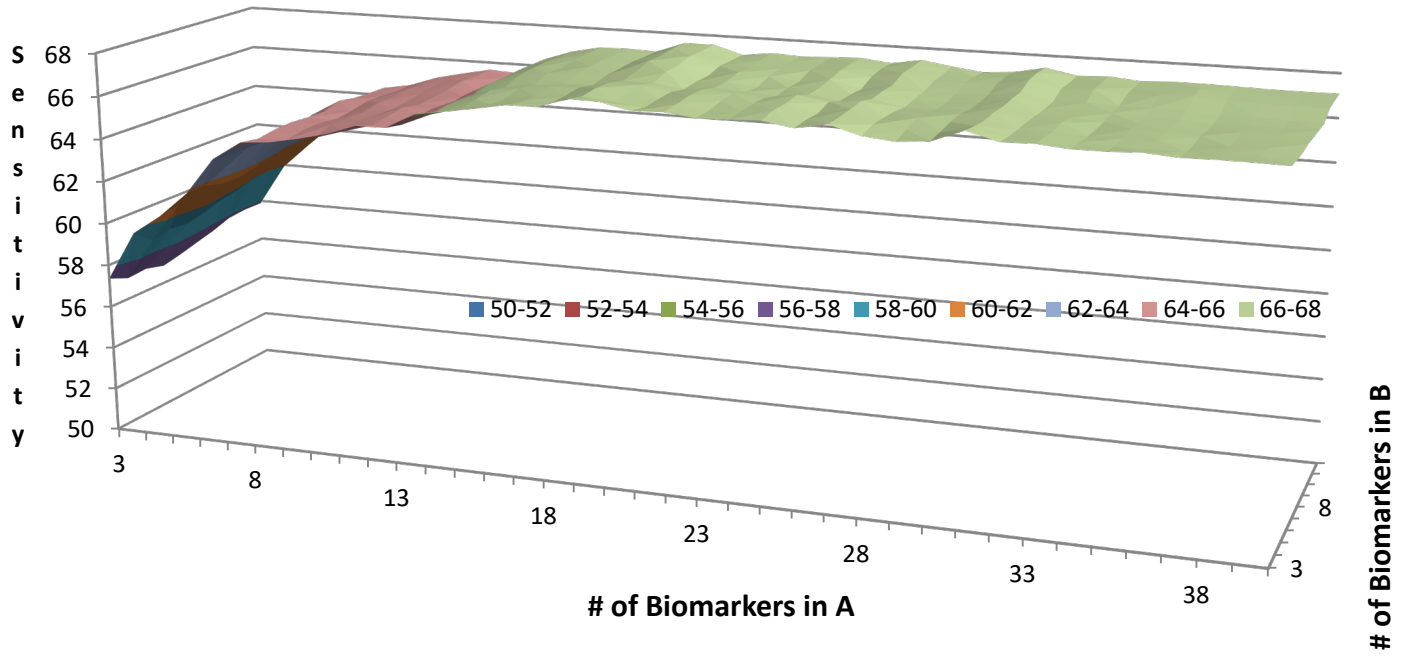
Biomarkers

Space A	Apolipoprotein B (Apo B)
Testosterone, Total	Macrophage-Derived Chemokine (MDC)
Prolactin (PRL)	Cortisol (Cortisol)
Interleukin-6 receptor (IL-6r)	Ferritin (FRTN)
Brain-Derived Neurotrophic Factor (BDNF)	Intercellular Adhesion Molecule 1 (ICAM-1)
Follicle-Stimulating Hormone (FSH)	Betacellulin (BTC)
Fetuin-A	Cancer Antigen 125 (CA-125)
Apolipoprotein A-I (Apo A-I)	Monocyte Chemotactic Protein 2 (MCP-2) ;
Interleukin-7 (IL-7)	
Carcinoembryonic Antigen (CEA)	**Space B**
Beta-2-Microglobulin (B2M)	Apolipoprotein H (Apo H)
Prostatic Acid Phosphatase (PAP)	Tumor Necrosis Factor Receptor 2 (TNFR2)
Peptide YY (PYY)	Connective Tissue Growth Factor (CTGF)
Macrophage Migration Inhibitory Factor (MIF)	Sortilin
Epidermal Growth Factor Receptor (EGFR)	Kidney Injury Molecule-1 (KIM-1)
Serum Amyloid P-Component (SAP)	Macrophage Inflammatory Protein-1 alpha (MIP-1 alpha)
Vascular Endothelial Growth Factor (VEGF)	Serotransferrin (Transferrin)
Immunoglobulin M (IGM)	Thyroid-Stimulating Hormone (TSH)
TNF-Related Apoptosis-Inducing Ligand Receptor 3 (TRAIL-R3)	Apolipoprotein C-I (Apo C-I)
Interleukin-10 (IL-10)	Haptoglobin
Luteinizing Hormone (LH)	Tissue Inhibitor of Metalloproteinases 1 (TIMP-1)
Matrix Metalloproteinase-2 (MMP-2)	Immunoglobulin A (IgA)
Vitronectin	**Space C**
Endothelin-1 (ET-1)	Apolipoprotein A-II (Apo A-II)
CD5 (CD5L)	Complement C3 (C3)
Alpha-1-Antitrypsin (AAT)	Calbindin

Decomposition Result

- In each step k in each space A, B or C, the only significant vector is the gradient. None other vectors approach significance, so the gradient is used to select biomarkers.
- The correlation coefficients between any pairs of gradients in Space A, Space B and Space C decrease as the dimension increases. All were **0.2 or less for this data**. It can be shown that
- $\rho(g_{A,I},g_{B,J}) < \rho(v_k,v_l)$, for $1 \le k \le I$, $1 \le l \le J$, $3 \le I \le 33$, $3 \le J \le 12$
- Where $g_{A,I}$ is the gradient in Space A_I, which consists of one selected biomarker from Space A by GNO; $g_{B,J}$ is the gradient in Space B_J, which consists of J selected biomarkers from Space B by GNO.
- The contribution of Space C did not approach significance; Using the sumc test all biomarkers in Space C were removed from further analysis.
- The collinearity problem is solved.

Figure 1: Effect of the number of biomarkers on sensitivity



QIC Selection

Using QIC:

1. Space A: 12 biomarkers were selected
2. Space B: 4 biomarkers were selected

Figure 2: Quasilikelihood under the Independence model Criterion in Space A

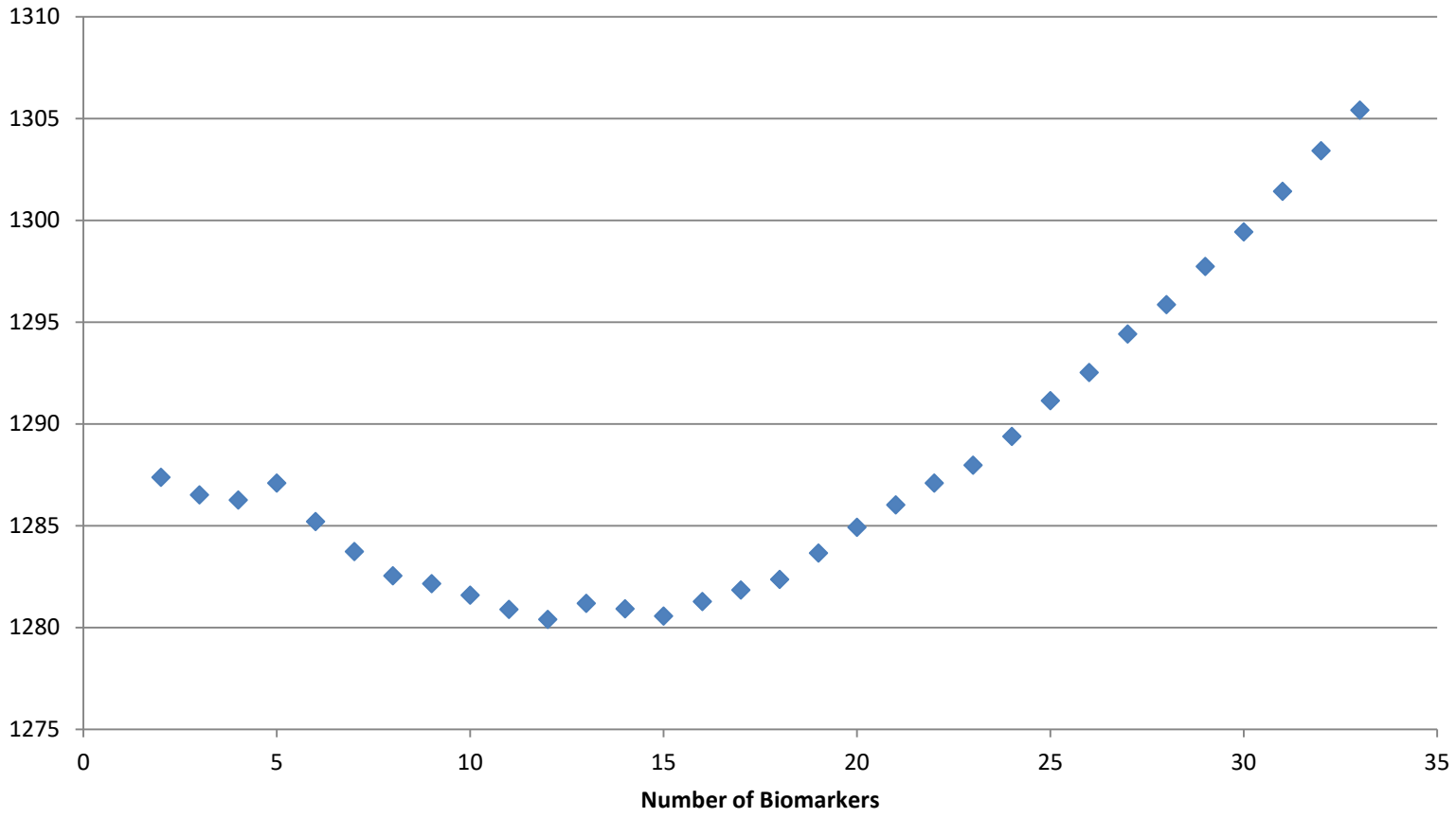
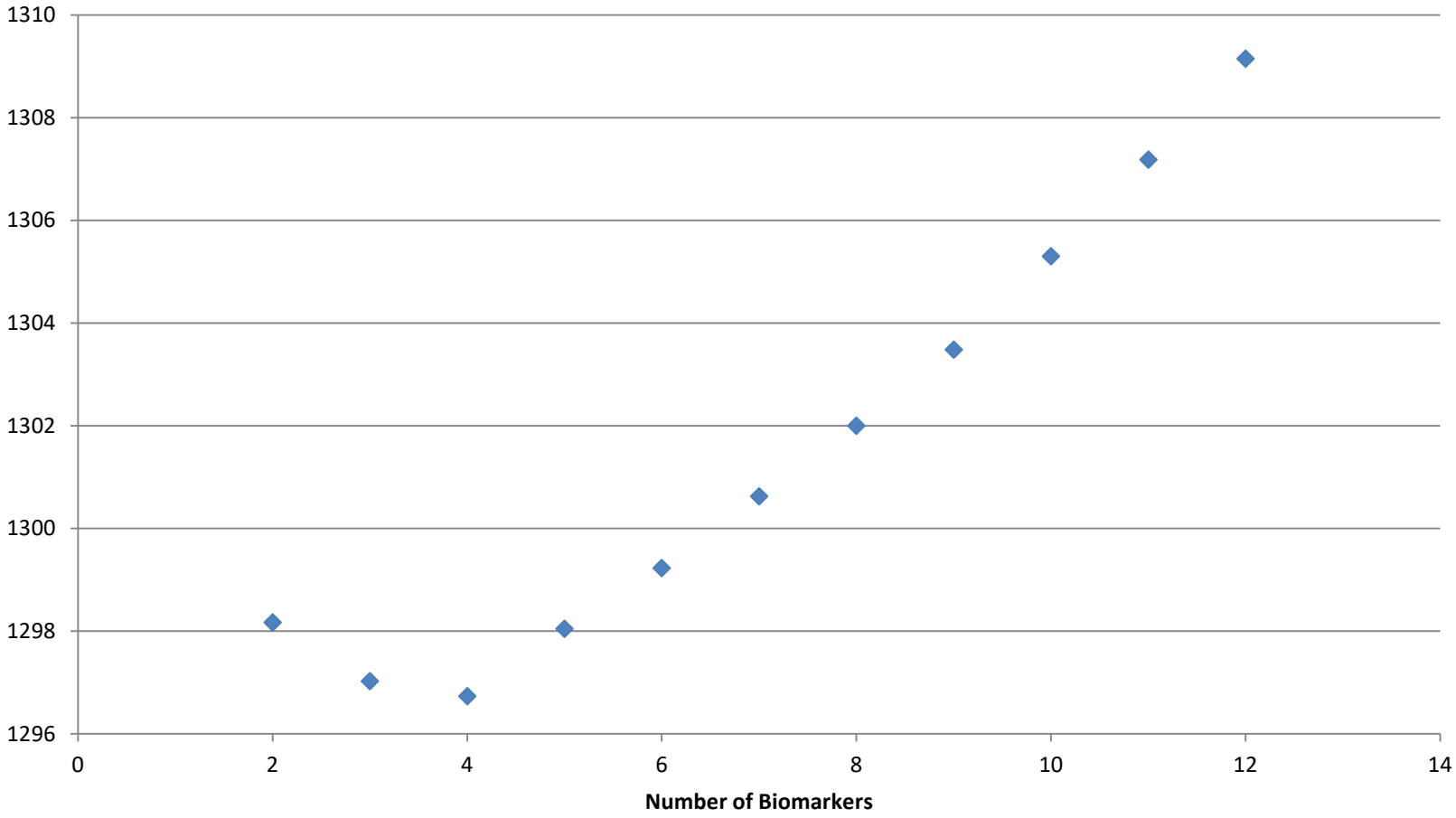


Figure 3: Quasilikelihood under the Independence model Criterion in Space B



Estimate Biomarker Effect

The top 12 or 6 biomarkers from Space A and the top 4 or 3 biomarkers from Space B were used to form the gradient scores in the logistic model with a GEE approach.

Results

The Odds Ratio for One Unit Increased in Gradient Score						
Space	Parameter	OR	ORL	ORU	*p_value in model	Adjusted p-value from Chi-square Wald approach
(6,3)	gradienta	1.5	1.27	1.77	2.00E-06	0.001
	gradientb	1.25	1.05	1.5	0.013	0.11
(12,4)	gradienta	1.78	1.49	2.14	5.16E-10	0.01
	gradientb	1.31	1.09	1.56	0.004	0.09

* Using Bonferoni criterion, all are significant: k=6, $\alpha=0.05/6=0.0083$, k=3; $\alpha=0.05/6=0.0167$

SUMC Test

The six biomarkers are in A12, but not in A6:

1. Matrix Metalloproteinase-2 (MMP-2)
2. Cortisol (Cortisol)
3. Monocyte Chemotactic Protein 2 (MCP-2)
4. CD5 (CD5L)
5. Macrophage Migration Inhibitory Factor (MIF)
6. Epidermal Growth Factor Receptor (EGFR)

The gradient effect by adjusted Chi-square was 0.07

Sumc = 2.36, which was near the 80^{th} percentile for p = 6.

The null hypothesis cannot be rejected, all effects are random, and they can be deleted simultaneously.

Individual Biomarker Effect from Last 6 Biomarkers in A and 3 Biomarkers in B

Space	Male			
	Biomarkers	Coef	Effect %	OR
A	Alpha-1-Antitrypsin (AAT)	0.47	0.22	1.21*
	Apolipoprotein A-I (Apo A-I)	0.38	0.14	1.16
	Immunoglobulin M (IGM)	-0.33	0.11	0.87
	Interleukin-6 receptor (IL-6r)	-0.53	0.28	0.81*
	Prolactin (PRL)	0.36	0.13	1.16
	Serum Amyloid P-Component (SAP)	0.35	0.12	1.15
B	Apolipoprotein H (Apo H)	-0.65	0.19	0.94
	Immunoglobulin A (IgA)	0.51	0.26	1.12
	Connective Tissue Growth Factor (CTGF)	0.56	0.54	1.24*

***significance level < 0.05.**

Observations

1. The gradient contains much more information than noise.
2. The gradient from 12 contains almost the same information as the gradient from the higher dimension space.
3. The surface is smooth, hence we can confidently delete weak-effect biomarkers.
4. 3 biomarkers had higher contributions.

Conclusions

1. The selection of biomarkers is robust and accurate.
2. Individual biomarker effect can be estimated.
3. Correlated biomarkers, if they have effect, all will be selected.
4. Eliminated the collinearity difficulty in regression.
5. Non-significant biomarkers can be tested and eliminated simultaneously by proposed sum statistic.
6. The difference in sensitivity between training and test sets is small.
7. The training group sensitivity is stable, and with a large sample size is large we can expect more reliable selection.

Discussion

- Univariate analysis: Prolactin RR=1.10
- Unispace gradient: RR=1.42
- Split space: RR=1.50
- Split space+time: RR=1.77
- 3 biomarkers were selected

Acknowledgements

- The authors also recognize the contribution of the Armed Forces Health Surveillance Center personnel, particularly Dr. Angelia A. Cost, for providing data, specimens and help with methodological aspects of the study.

- COL Christopher Littell, MAJ Michael Boivin, Natalya Weber, MD, Janice K. Gary, B.S, Walter Reed Army Institute of Research, Preventive Medicine Branch, and COL David Niebuhr, USUHS, for the Project development, administrative support, and preparation of the presentation.

References

- Adragni, K. P. and Cook R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of The Royal Society a: Mathematical, Physical & Engineering Sciences* **367**, 4385-4405.
- Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119-137.
- Boulpaep EL, Boron WF (2005). Medical physiology: a cellular and molecular approach. St. Louis, Mo: Elsevier Saunders. pp. 1125. ISBN 1-4160-2328-3.
- Chakravarti A. (1999). Population genetics-making sense out of sequence. *Nature Genetics* **21**, 56-60.
- Chung, C. P., Avalos, I., Oeser, A., Gebretsadik, T., Shintani, A., Raggi, P. et al. (2007). High prevalence of the metabolic syndrome in patients with systemic lupus erythematosus: association with disease characteristics and cardiovascular risk factors. *Annals of the Rheumatic Diseases* **66,** 208-214.
- Cook, R. D. and Forzani, L. (2009). Likelihood-Based Sufficient Dimension Reduction. *Journal of the American Statistical Association* **104**, 197-208.
- Current Population Survey (CPS) - Definitions and Explanations". US Census Bureau. http://www.census.gov/population/www/cps/cpsdef.html.
- De Hert, M., Van Winkel, R., Van Eyck, D., Hanssens, L., Wampers, M., Scheen, A. et al. (2006). Prevalence of diabetes, metabolic syndrome and metabolic abnormalities in schizophrenia over the course of the illness: a cross-sectional study. *Clinical Practice and Epidemiology in Mental Health* **2**,14.
- Fessel, W. J. and Solomon, G. F. (1960). Psychosis and systemic lupus erythematosus: a review of the literature and case reports. *California Medicine* **92,** 266-270.
- Gini, C. (1912) (Italian: Variabilità e mutabilità (Variability and Mutability), C. Cuppini, Bologna, 156 pages. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).
- Goldberg, R. B. (2009). Cytokine and cytokine-like inflammation markers, endothelial dysfunction, and imbalanced coagulation in development of diabetes and its complications. *The Journal of Clinical Endocrinology & Metabolism* **94,** 3171-3182.
- Good, P.I. and Hardin, J.W. (2009). *Common Errors in Statistics: And How to Avoid Them* (3 edn.), Wiley, New Jersey.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Traning : Data Mining, Inference, and Prediction*. Springer, New York.
- Kuan CT, Wikstrand CJ, Bigner DD (June 2001). "EGF mutant receptor vIII as a molecular target in cancer therapy". *Endocr. Relat. Cancer* **8** (2): 83–96. DOI:10.1677/erc.0.0080083. PMID 11397666.

References

- Li, B. and Dong, Y. X. (2009). Dimension reduction for nonelliptically distributed Predictors. *Annals of Statistics* **37,** 1272-1298.
- Li, K.C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86,** 316-327.
- Niebuhr, D. W., Li, Y, Cowan, D. N., Weber, N. S., Fisher, J. A., Ford, G. M., and Yolken, R. (2011). Association between bovine casein antibody and new onset schizophrenia among US military personnel. *Schizophrenia Research* doi:10.1016/j.schres.2011.02.005.
- Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D., Hallmayer J. et al. (1999). A genomic screen of autism: Evidence for a multilocus etiology. *The American Journal of Human Genetics* **65**, 493-507.
- Scott E (2011). "Cortisol and Stress: How to Stay Healthy". About.com. http://stress.about.com/od/stresshealth/a/cortisol.htm
- Segal, M. R., Dahlquist, K. D, and Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology* **10**, 961-980.
- Shoelson, S. E., Lee, J. and Goldfine, A. B. (2006). Inflammation and insulin resistance. *The Journal of Clinical Investigation* **116**, 1793-1801.
- Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D, et al. (2009). Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.
- Volp, A. C., Alfenas, R. D., Costa, N.M., Minim, V. P., Stringueta, P.C. and Bressan, J. (2008). Inflammation biomarkers capacity in predicting the metabolic syndrome. *Arquivos Brasileiros de Endocrinologia & Metabologia* **52**, 537-549.
- Wajed, J., Ahmad, Y., Durrington, P. N. and Bruce, I. N. (2004). Prevention of cardiovascular disease in systemic lupus erythematosus-proposed guidelines for risk factor management. *Rheumatology (Oxford)* **43**, 7-12.
- Walker F, Abramowitz L, Benabderrahmane D, Duval X, Descatoire V, Hénin D, Lehy T, Aparicio T (November 2009). "Growth factor receptor expression in anal squamous lesions: modifications associated with oncogenic human papillomavirus and human immunodeficiency virus". Hum. Pathol. 40 (11): 1517–27. DOI:10.1016/j.humpath.2009.05.010. PMID 19716155.
- Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society. Series B (Methodological)* **71,** 615-636.
- Yin, X. R. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika* **90** 113-125.

QUESTIONS