Dimension reduction in the study of etiologic heterogeneity

Emily C. Zabor

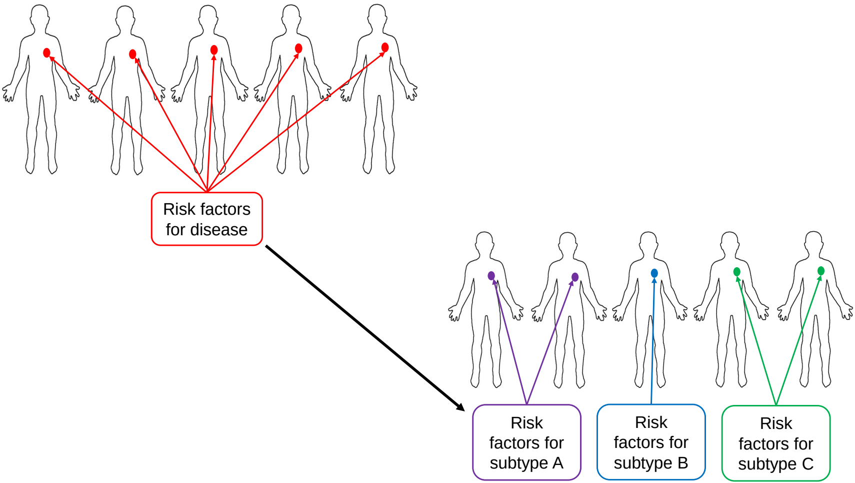
2017 Joint Statistical Meetings, Baltimore, MD

August 1, 2017

Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

The focus of cancer epidemiologic research is shifting from single disease organized by site to disease subtypes



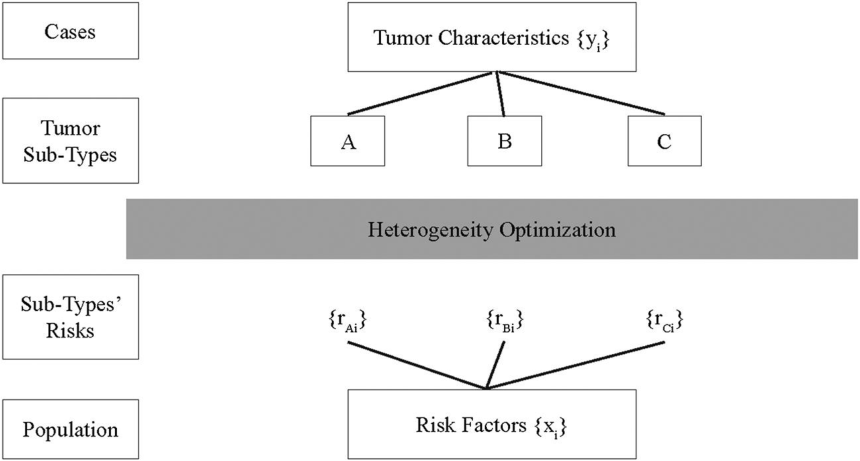
Risk factors for disease

Risk factors for subtype A

Risk factors for subtype B

Risk factors for subtype C

A scalar measure of etiologic heterogeneity is based on risk predictions obtained from a polytomous logistic regression



Begg CB, Zabor EC, Bernstein JL, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013; **32**(29):5039-52

Risk heterogeneity is measured using the coefficient of variation and risk covariance

The total coefficient of variation (CV) for subtypes A, B, C is:

$$K^2 = \pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2 + 2\pi_A \pi_B K_{AB} + 2\pi_A \pi_C K_{AC} + 2\pi_B \pi_C K_{BC}$$

where π_j is the relative frequency and K_j^2 is the CV for subtype j.

Then the incremental explained variation is defined as:

$$D = (\pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2) - K^2$$

Begg CB, Zabor EC, Bernstein JL, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013; **32**(29):5039-52

Risk heterogeneity is measured using the coefficient of variation and risk covariance

The total coefficient of variation (CV) for subtypes A, B, C is:

$$K^2 = \pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2 + 2\pi_A \pi_B K_{AB} + 2\pi_A \pi_C K_{AC} + 2\pi_B \pi_C K_{BC}$$

where π_j is the relative frequency and K_j^2 is the CV for subtype j.

Then the incremental explained variation is defined as:

$$D = (\pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2) - K^2$$

Begg CB, Zabor EC, Bernstein JL, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013; **32**(29):5039-52

Risk heterogeneity is measured using the coefficient of variation and risk covariance

The total coefficient of variation (CV) for subtypes A, B, C is:

$$K^2 = \pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2 + 2\pi_A \pi_B K_{AB} + 2\pi_A \pi_C K_{AC} + 2\pi_B \pi_C K_{BC}$$

where π_j is the relative frequency and K_j^2 is the CV for subtype j.

Then the incremental explained variation is defined as:

$$D = (\pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2) - K^2$$

Begg CB, Zabor EC, Bernstein JL, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013; **32**(29):5039-52

Risk heterogeneity is measured using the coefficient of variation and risk covariance

The total coefficient of variation (CV) for subtypes A, B, C is:

$$K^2 = \pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2 + 2\pi_A \pi_B K_{AB} + 2\pi_A \pi_C K_{AC} + 2\pi_B \pi_C K_{BC}$$

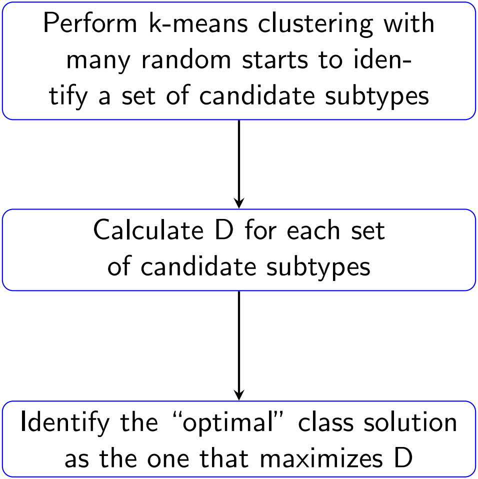
where π_j is the relative frequency and K_j^2 is the CV for subtype j.

Then the incremental explained variation is defined as:

$$D = (\pi_A K_A^2 + \pi_B K_B^2 + \pi_C K_C^2) - K^2$$

Begg CB, Zabor EC, Bernstein JL, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med* 2013; **32**(29):5039-52

An analysis of etiologic heterogeneity in this framework involves three steps



Perform k-means clustering with many random starts to identify a set of candidate subtypes

Calculate D for each set of candidate subtypes

Identify the "optimal" class solution as the one that maximizes D

We use a data example to assess sensitivity of results to different clustering approaches

Hierarchical
vs
K-means

Full gene set
vs
Unsupervised dimension reduction
vs
Supervised dimension reduction

Risk factor data are from the Cancer and Steroid Hormone (CASH) breast cancer case-control study

Risk factor	Controls (N = 2990)	Cases (N = 551)
Age (per 10 years)	4.67 (2, 5.54)	4.73 (2.45, 5.5)
Pre-menopausal BMI (per 20)	1.15 (0.78, 2.74)	1.15 (0.78, 2.21)
Post-menopausal BMI (per 20)	1.2 (0.8, 3.08)	1.2 (0.83, 1.76)
Age at menarche (per 2 years)	6.5 (4, 10)	6 (4, 9)
Parity	3 (1, 13)	3 (1, 9)
Age at first birth (per 5 years)	4.6 (2.2, 8.6)	4.6 (2.6, 8)
Months of breastfeeding (per 6)	0.17 (0, 28)	0.17 (0, 16.33)
Age at menopause (per 5 years)	8.4 (4.2, 10.6)	8.4 (4.6, 10.6)
Non-white race	381 (12.7)	39 (7.1)
Family history of brca	206 (6.9)	73 (13.2)
Benign breast disease	354 (11.8)	100 (18.1)
Nulliparous	405 (13.5)	83 (15.1)
Post-menopausal	1211 (40.5)	204 (37)

Tumor marker data on cases includes 202 gene expression values

Hierarchical clustering results in unbalanced average class size compared to k-means clustering

	Class			
Method	1	2	3	4
k-means	58	110	177	206
hclust complete euclid	26	72	131	322
hclust single euclid	1	1	1	548
hclust avg euclid	1	3	28	519
hclust complete corr	12	28	119	392
hclust single corr	1	1	1	548
hclust avg corr	1	6	23	522

Four approaches to k-means clustering are compared

1. K-means clustering on full gene set
2. K-means clustering on principal components
3. K-means clustering on gene set pre-filtered according to univariate D for each gene
4. K-means clustering on gene set pre-filtered according to F-statistic proposed by Zapala & Schork*

*Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS* 2006; **103**(51):19430-35

Four approaches to k-means clustering are compared

1. K-means clustering on full gene set
2. K-means clustering on principal components
3. K-means clustering on gene set pre-filtered according to univariate D for each gene
4. K-means clustering on gene set pre-filtered according to F-statistic proposed by Zapala & Schork*

*Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS* 2006; **103**(51):19430-35

Four approaches to k-means clustering are compared

1. K-means clustering on full gene set
2. K-means clustering on principal components
3. K-means clustering on gene set pre-filtered according to univariate D for each gene
4. K-means clustering on gene set pre-filtered according to F-statistic proposed by Zapala & Schork*

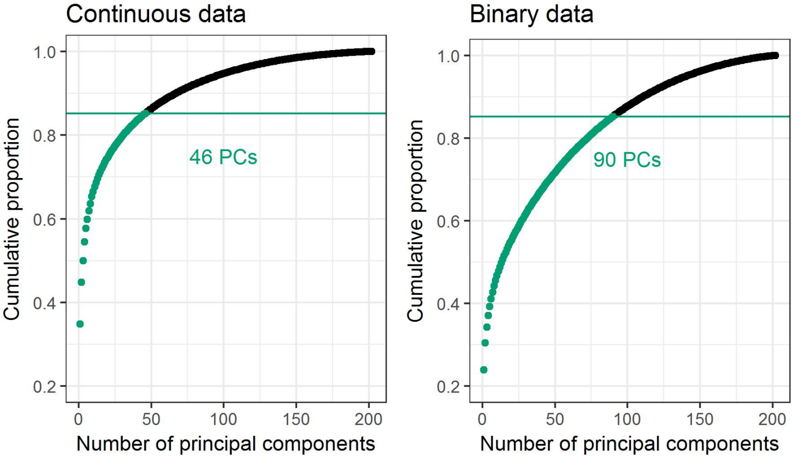
*Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS* 2006; **103**(51):19430-35

Four approaches to k-means clustering are compared

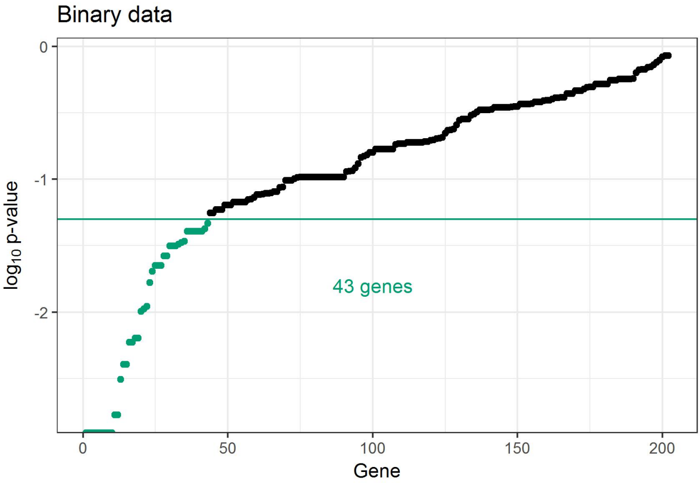
1. K-means clustering on full gene set
2. K-means clustering on principal components
3. K-means clustering on gene set pre-filtered according to univariate D for each gene
4. K-means clustering on gene set pre-filtered according to F-statistic proposed by Zapala & Schork*

*Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS* 2006; **103**(51):19430-35

Different sets of principal components are selected when using continuous vs binary gene expression data



Univariate D p-values, adjusted for multiple comparisons, identify 43 significant genes



Zapala & Schork propose an F-statistic to assess the relationship between risk factors and dissimilarity matrix

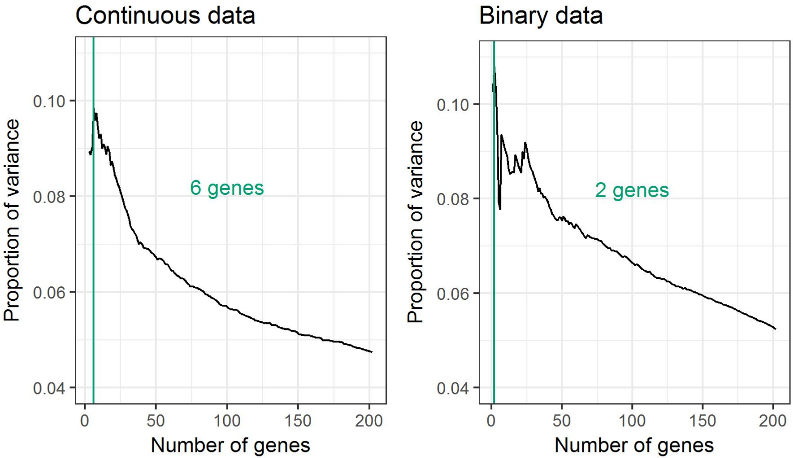
$$F = \frac{tr(\boldsymbol{HGH})/(P-1)}{tr[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{H})]/(N-P)}$$

where:

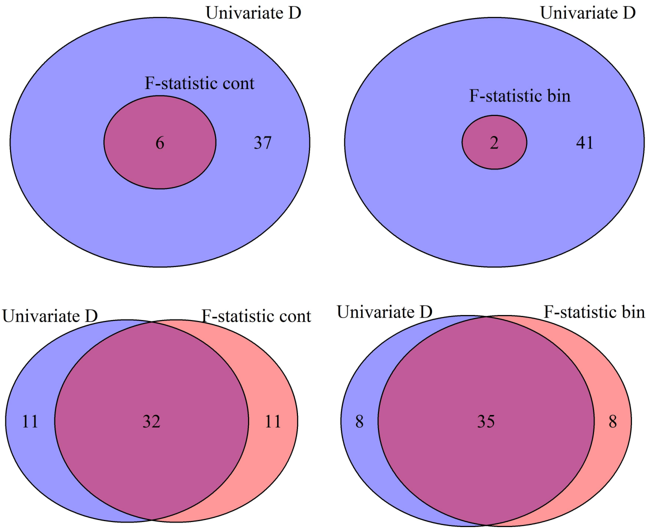
- N indexes cases
- P indexes risk factors
- $\boldsymbol{H}$ an $N \times N$ hat matrix
- $\boldsymbol{G}$ is Gower's centered distance matrix

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *PNAS* 2006; **103(**51):19430-35

The F-statistic identifies different gene sets when using continuous versus binary gene expression data



There is significant overlap in selected genes by the two supervised approaches



The two supervised approaches result in consistently high D-metrics under the various configurations

The full gene set resulted in a D metric of 0.200 for continuous genes

	Number of included elements		
Ranking method	46 (PCA)	43 (Univariate D)	6 (F-statistic)
PCA	**0.196**	0.196	0.191
Univariate D	0.245	**0.248**	0.242
F-statistic	0.233	0.250	**0.330**

The full gene set resulted in a D metric of 0.231 for binary genes

	Number of included elements		
Ranking method	90 (PCA)	43 (Univariate D)	2 (F-statistic)
PCA	**0.213**	0.213	0.226
Univariate D	0.241	**0.283**	0.258
F-statistic	0.250	0.248	**0.198**

There is moderate to substantial alignment of class results based on top 43 continuous genes

	Univariate D				
	1	2	3	4	kappa
PCA					0.443
1	**25 (100)**	102 (48.3)	3 (1.6)	6 (4.9)	
2	0 (0)	**109 (51.7)**	85 (44.3)	15 (12.2)	
3	0 (0)	0 (0)	**104 (54.2)**	14 (11.4)	
4	0 (0)	0 (0)	0 (0)	**88 (71.5)**	
F-stat					0.864
1	**7 (28)**	15 (7.1)	45 (23.4)	35 (28.5)	
2	17 (68)	**186 (88.2)**	12 (6.2)	0 (0)	
3	1 (4)	10 (4.7)	**135 (70.3)**	0 (0)	
4	0 (0)	0 (0)	0 (0)	**88 (71.5)**	

There is moderately good class results based on top 43 binary genes from each ranking method

	Univariate D				
	1	2	3	4	kappa
PCA					0.443
1	**83 (76.9)**	1 (0.7)	16 (11.1)	20 (12.8)	
2	15 (13.9)	**33 (23.1)**	4 (2.8)	55 (35.3)	
3	9 (8.3)	109 (76.2)	**124 (86.1)**	0 (0)	
4	1 (0.9)	0 (0)	0 (0)	**81 (51.9)**	
F-stat					0.681
1	**80 (74.1)**	11 (7.7)	3 (2.1)	47 (30.1)	
2	19 (17.6)	**90 (62.9)**	0 (0)	1 (0.6)	
3	6 (5.6)	42 (29.4)	**141 (97.9)**	0 (0)	
4	3 (2.8)	0 (0)	0 (0)	**108 (69.2)**	

Results of this data example are preliminary in nature and will guide design of future simulation study

- More strongly etiologically distinct subtypes may be discovered after supervised dimension reduction is performed
- The F-statistic is a desirable approach due to its computational simplicity
- Future simulation study will examine properties of these approaches in the context of a gold standard class solution

Acknowledgements:

Colin Begg, Memorial Sloan Kettering Cancer Center
Venkat Seshan, Memorial Sloan Kettering Cancer Center
Shuang Wang, Columbia University

Contact:

E-mail: zabore@mskcc.org
Slides: `https://github.com/zabore/talk-slides`