Large-Scale Prediction Problems

Bradley Efron
Stanford University

http://stat.stanford.edu/~brad/papers

Microarray Predictions

Observe Data Matrix

$$x_{ij} \begin{cases} i = 1, 2, \ldots, N & \text{genes} \\ j = 1, 2, \ldots, n & \text{individuals} \end{cases}$$

Two Classes

$$\begin{cases} j = 1, 2, \ldots, n_1 & \text{healthy} \\ j = n_1 + 1, \ldots, n_1 + n_2 = n & \text{sick} \end{cases}$$

New Array: $X_i, \quad i = 1, 2, \ldots, N$

The Task Use data matrix to predict whether new array is in healthy or sick class.

The Trouble: 1000's of possible predictors, most of which are probably useless.

Prostate Data (Singh et al., 2002)

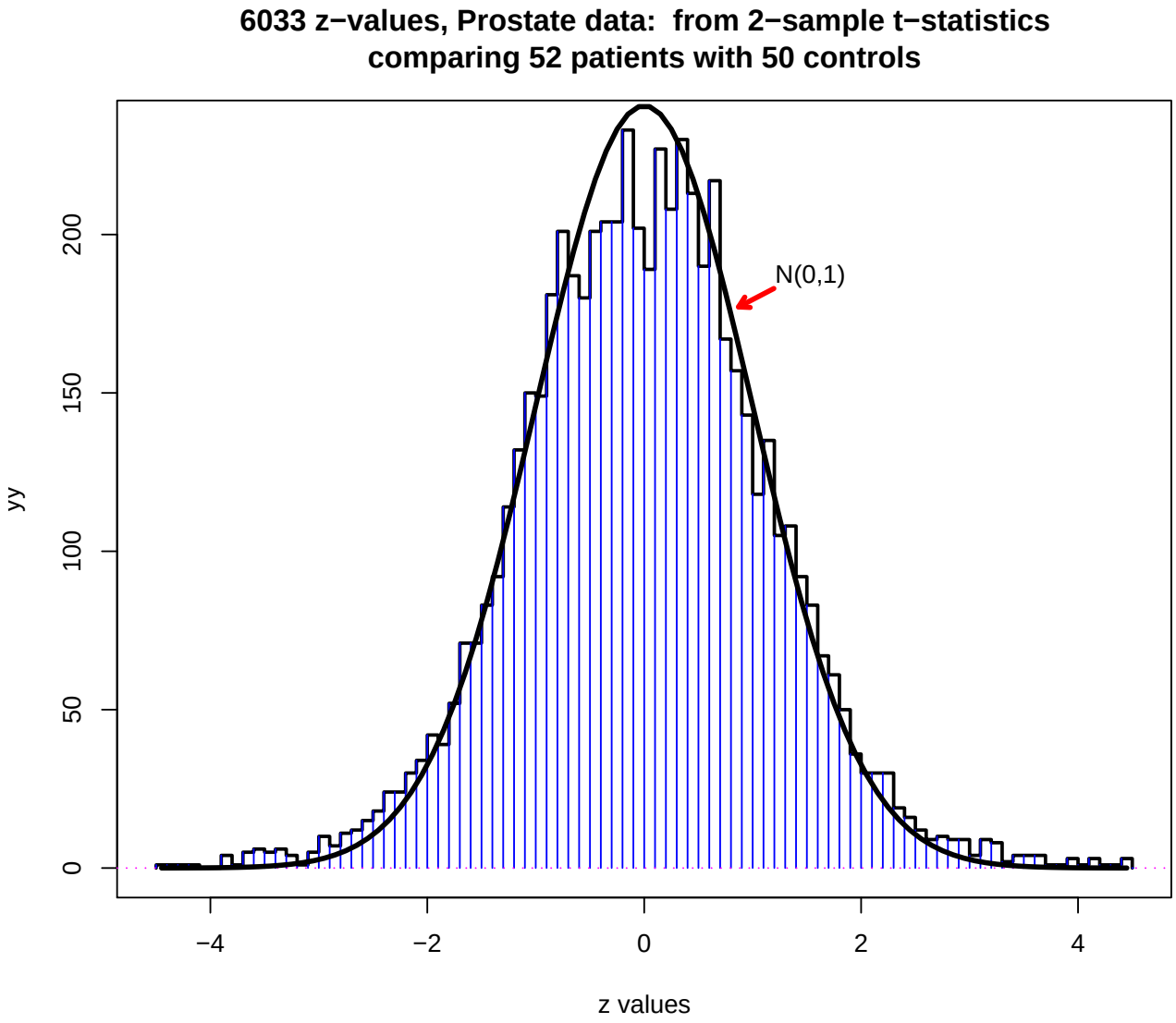
- $N = 6033$ genes, $n = 102$ patients

$$\begin{cases} n_1 = 50 \text{ healthy controls} \\ n_2 = 52 \text{ prostate cancer} \end{cases}$$

- z_i = two-sample t-stat comparing prostate patients with healthy controls for gene i (transformed so that

$$z_i \sim \mathcal{N}(0,1)$$

under null hypothesis of no difference between the two classes)


6033 z−values, Prostate data: from 2−sample t−statistics
comparing 52 patients with 50 controls
N(0,1)
yy
0
50
100
150
200
−4
−2
0
2
4
z values

A Very Simple Model

- For each microarray assume gene measurements $X_1, X_2, \ldots X_N$ are independent normal variables

$$\boxed{\frac{X_i - \mu_i}{\sigma_i} \overset{\text{ind}}{\sim} \mathcal{N}\left(\pm\frac{\delta_i}{2c}, 1\right)} \quad \begin{cases} \text{“}-\text{”} & \text{healthy} \\ \text{“}+\text{”} & \text{sick} \end{cases}$$

- *Constant “c” =*

$$\frac{1}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}} = \sqrt{\frac{n_1 n_2}{n}}$$

$$(= 5.05 \text{ for } n_1 = 50,\ n_2 = 52)$$

- δ_i = “effect size”

Ideal Prediction Rule

- *Compute*

$$\boxed{S = \sum \delta_i \left(\frac{X_i - \mu_i}{\sigma_i} \right) \sim \mathcal{N}\left(\pm \frac{\|\boldsymbol{\delta}\|^2}{2c}, \|\boldsymbol{\delta}\|^2 \right)}$$

where

$$\|\boldsymbol{\delta}\|^2 = \sum \delta_i^2.$$

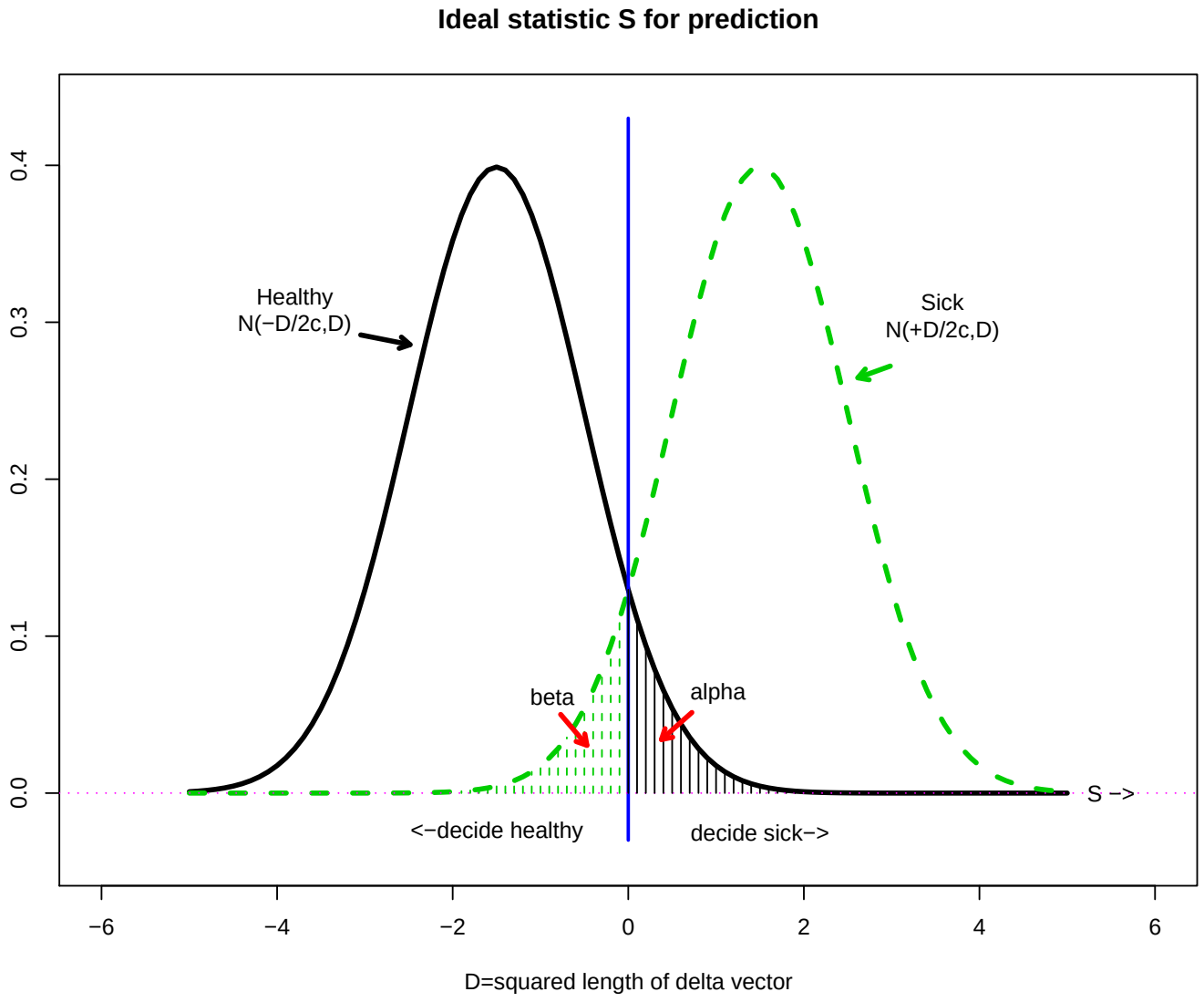
- *Predict*: "healthy" if $S < 0$,

"sick" if $S > 0$.

- *Error Rates*: $\alpha = \beta = \Phi(-\|\boldsymbol{\delta}\|/2c)$

  ◊ accurate prediction if $\|\boldsymbol{\delta}\|$ is large

  ◊ $\alpha =$ "prediction error"


Ideal statistic S for prediction
Healthy
N(−D/2c,D)
Sick
N(+D/2c,D)
beta
alpha
S −>
<−decide healthy
decide sick−>
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
D=squared length of delta vector

Estimating the Ideal Rule

- $S = \sum \delta_i \left(\dfrac{X_i - \mu_i}{\sigma_i} \right)$ depends on $(\delta_i, \mu_i, \sigma_i), \quad i = 1, 2, \ldots, N.$
- Estimate $\mu_i,\ \sigma_i$ in usual way: $\hat{\mu}_i = (\bar{x}_{i1} + \bar{x}_{i2})/2,\ \hat{\sigma}_i^2 = (ss_{i1} + ss_{i2})/(n-2)$
- Obvious estimate of δ_i:

$$z_i = c\,\frac{\bar{x}_{i2} - \bar{x}_{i1}}{\hat{\sigma}_i} \;\dot{\sim}\; \mathcal{N}(\delta_i, 1)$$

Actually "t_i", transformed to $z_i = \Phi F_{n-2}(t_i)$.

Trouble!

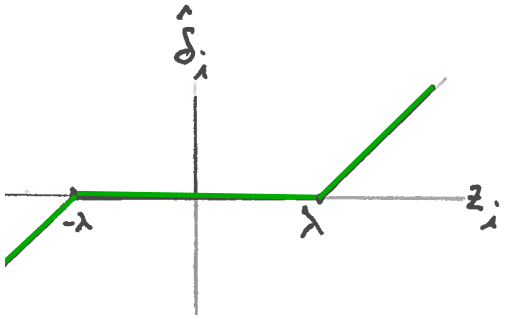
- $\bar{\delta}_i = z_i$ poor estimate of δ_i
- Wildly overestimates $\|\boldsymbol{\delta}\|^2$,
  underestimates prediction error
  $\alpha = \Phi(-\|\boldsymbol{\delta}\|/2c)$
- *Selection Bias*: large z_i's may be "lucky"

Shrunken Centroids (Tibshirani, Hastie, Narasimhan & Chu, 2002)

- *Idea*: shrink estimates $\bar{\delta}_i = z_i$ toward 0 ("soft thresholding")

$$\hat{\delta}_i = \text{sign}(z_i)(|z_i| - \lambda)_+$$

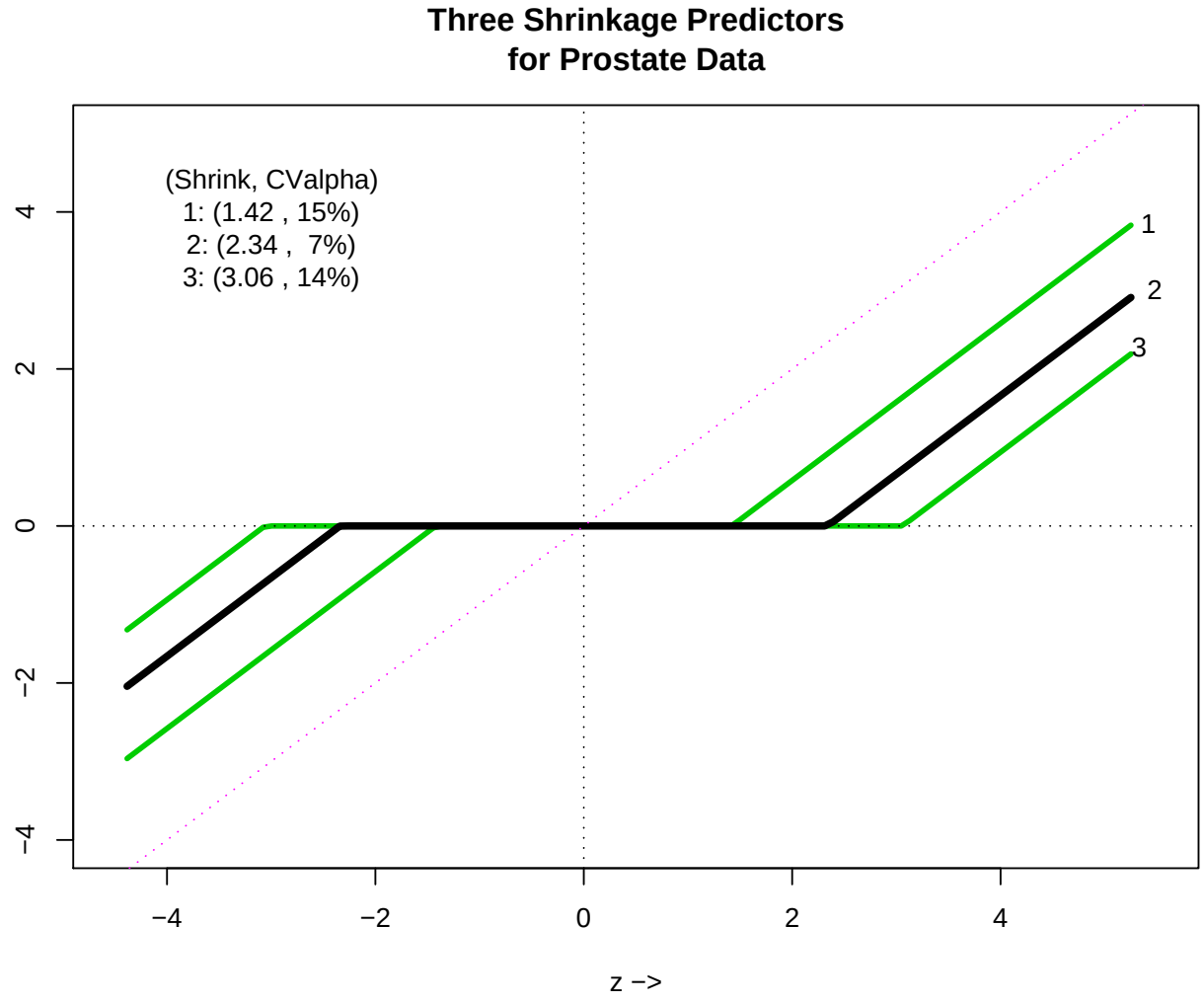
and predict with $\hat{S} = \sum \hat{\delta}_i \left(\frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i} \right)$



- Use cross-validation to estimate prediction error for every choice of shrinkage constant λ (CRAN: "pamr")

Shrunken Centroids for Prostate Data

shrink value	nonzero genes	error rate
0.00	6033	0.34
0.54	3763	0.33
1.08	1931	0.23
1.62	866	0.12
2.16	377	0.09
2.70	172	0.10
3.24	80	0.16
3.78	35	0.30
4.32	4	0.41
4.86	1	0.48
5.29	0	0.52


Three Shrinkage Predictors
for Prostate Data
(Shrink, CValpha)
1: (1.42 , 15%)
2: (2.34 , 7%)
3: (3.06 , 14%)
1
2
3
4
2
0
−2
−4
−4
−2
0
2
4
z −>

Choosing the Shrinkage Constant

- *Irresistible Impulse*: choose the λ that minimizes the cross-validated prediction error
- But this choice itself is not cross-validated
- **Simulation**
  $N = 1000,\ n_1 = n_2 = 10,\ x_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0,1)$
- Minimum cross-validated error rates were $30\% \pm 16\%$

Bayesian Prediction

- Suppose we could calculate Bayes posterior expectations

$$\tilde{\delta}_i = E\{\delta_i | \boldsymbol{z}\}$$

- Could use predictor

$$\tilde{S} = \sum \tilde{\delta}_i \left(\frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i} \right),$$

with approximate error probabilities

$$\tilde{\alpha} = \Phi\left(-\frac{\|\tilde{\boldsymbol{\delta}}\|}{2c} \right)$$

- Bayes estimates immune to selection bias ($z_{610} = 5.29$)

Brown–Stein Model (1971, 1981)

- *Model* $\delta \sim g(\cdot)$ and $z|\delta \sim \mathcal{N}(\delta, 1)$
- *Marginal Density*:

$$f(z) = \int \varphi(z-\delta)\, g(\delta)\, d\delta \qquad \left[\varphi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\right]$$

- **Posterior Density**:

$$\boxed{\begin{array}{c} g(\delta|z) = e^{\delta z - \psi(z)}\ [e^{-\delta^2/2} g(\delta)] \\ \psi(z) = \log\left(f(z)/\varphi(z)\right) \end{array}}$$

- $g(\delta|z)$ exponential family, natural parameter z, cgf $\psi(z)$:

$$E\{\delta|z\} = \psi'(z) \quad \text{Var}\{\delta|z\} = \psi''(z)$$

Empirical Bayes Estimates ("Ebay" program)

- Estimate marginal density $f(z)$ by $\hat{f}(z)$ (from Poisson regression of z-value histogram counts, as natural spline function of z)
- Numerically differentiate $\hat{\psi}(z) = \log\{\hat{f}(z)/\varphi(z)\}$ to give

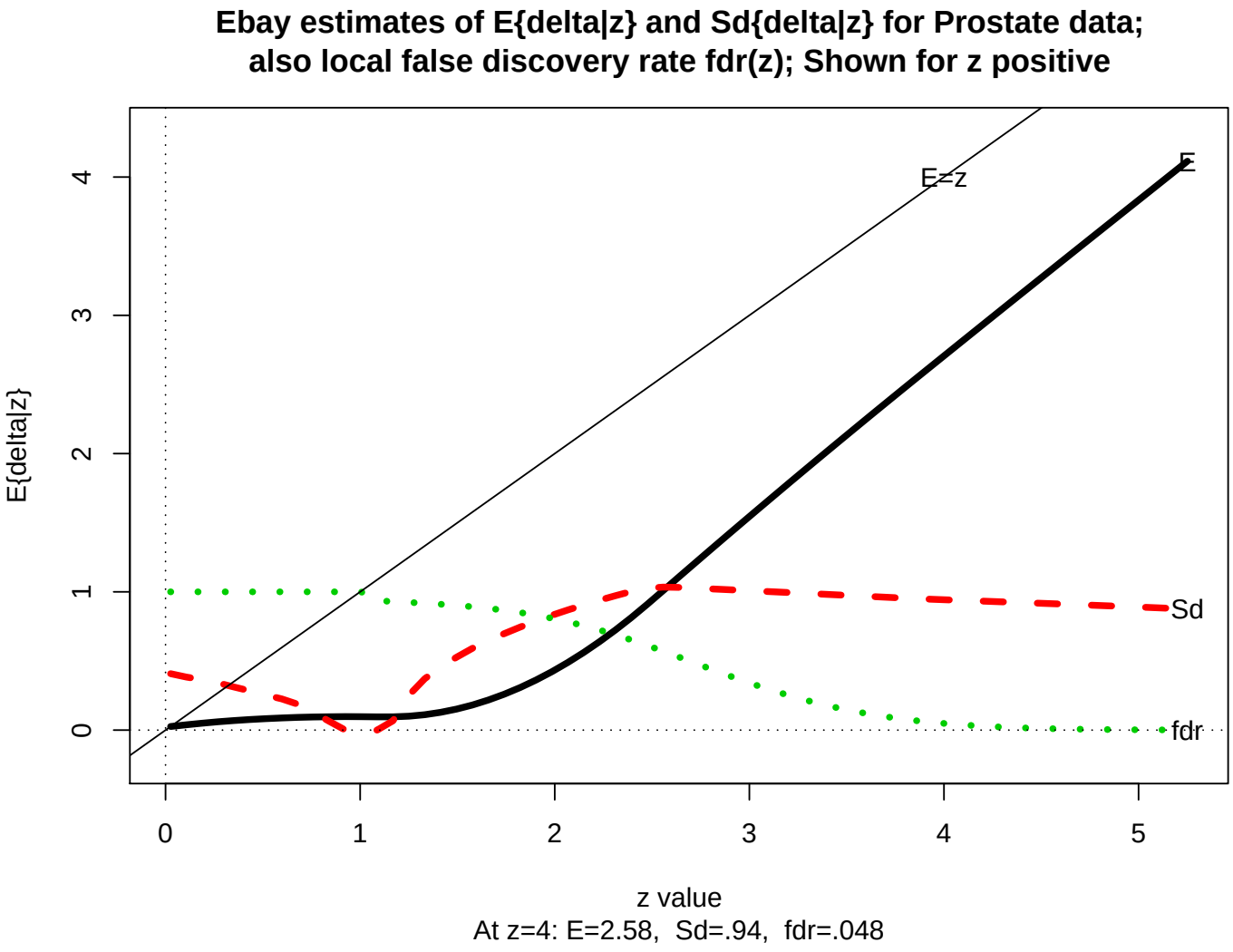
$$\hat{E}\{\delta|z\} = \hat{\psi}'(z), \quad \hat{S}d(z) = \sqrt{\hat{\psi}''(z)}$$

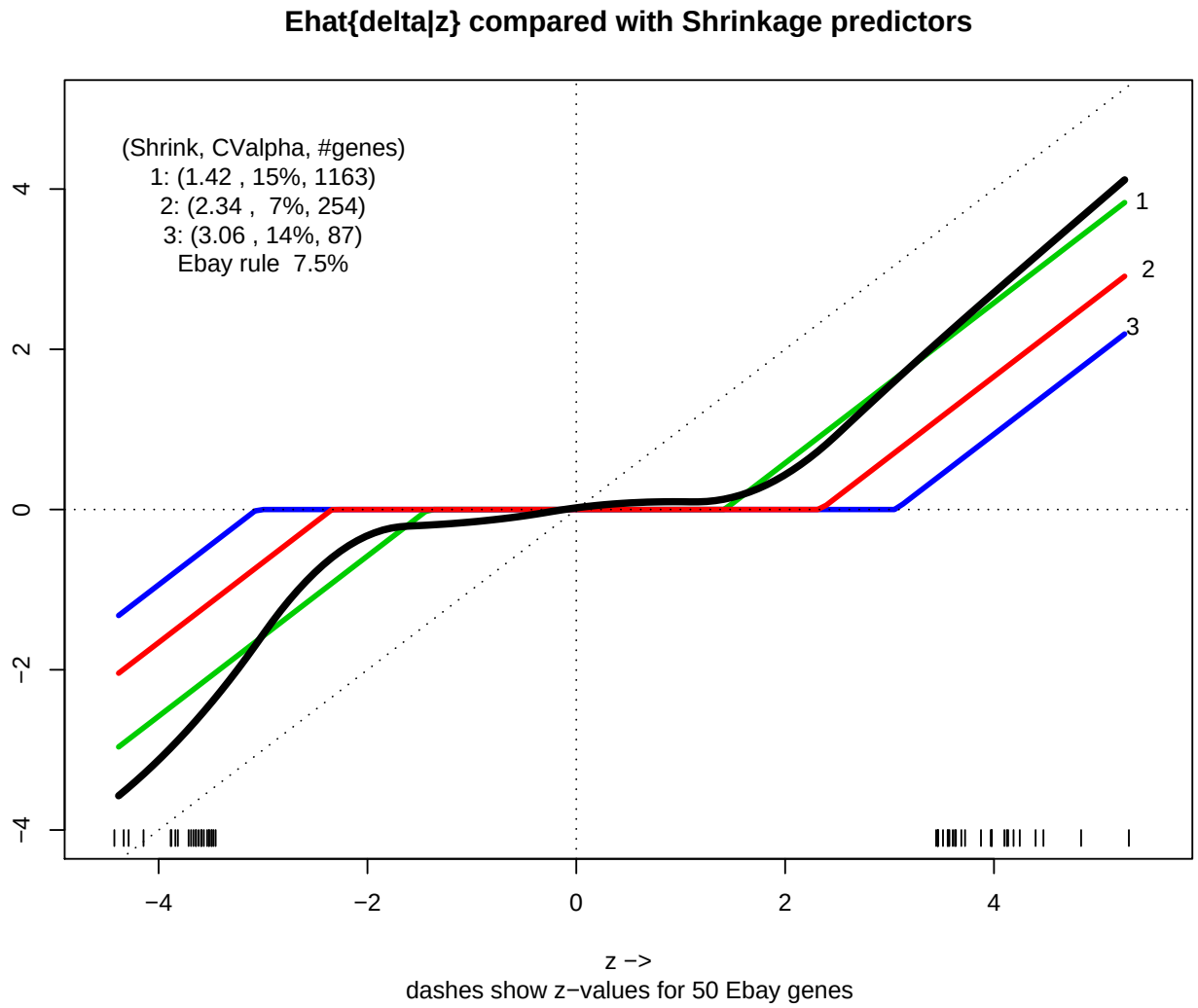
- For a given subset "I" of the genes,

$$\hat{\boldsymbol{\delta}}_I = \left\{\hat{\delta}_i = \hat{E}\left\{\delta|z_i\right\}\right\}, \quad \hat{S} = \sum \hat{\delta}_i \left(\frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i}\right),$$

$$\hat{\alpha}_I = \Phi(-\|\hat{\boldsymbol{\delta}}_I\|/2c)$$

- Can choose "I" to include biggest $\hat{\delta}_i$ values without worrying about selection bias.


Ebay estimates of E{delta|z} and Sd{delta|z} for Prostate data;
also local false discovery rate fdr(z); Shown for z positive
E{delta|z}
0
1
2
3
4
E=z
E
Sd
fdr
0
1
2
3
4
5
z value
At z=4: E=2.58, Sd=.94, fdr=.048


Ehat{delta|z} compared with Shrinkage predictors
(Shrink, CValpha, #genes)
1: (1.42 , 15%, 1163)
2: (2.34 , 7%, 254)
3: (3.06 , 14%, 87)
Ebay rule 7.5%
1
2
3
4
2
0
−2
−4
−4
−2
0
2
4
z −>
dashes show z−values for 50 Ebay genes

"Ebay" Prediction Rule

- Select target error probability α_0
- Program selects predictor genes in order of $|\hat{\delta}_i|$, largest first
- Continues until $\|\hat{\boldsymbol{\delta}}_I\|$ sufficiently large $[\Phi(-\|\hat{\delta}_I\|/2c) \leq \alpha_0]$ or until $\#I = 200$
- Estimates prediction error by 10-fold cross-validation

Note: CV error "honest"

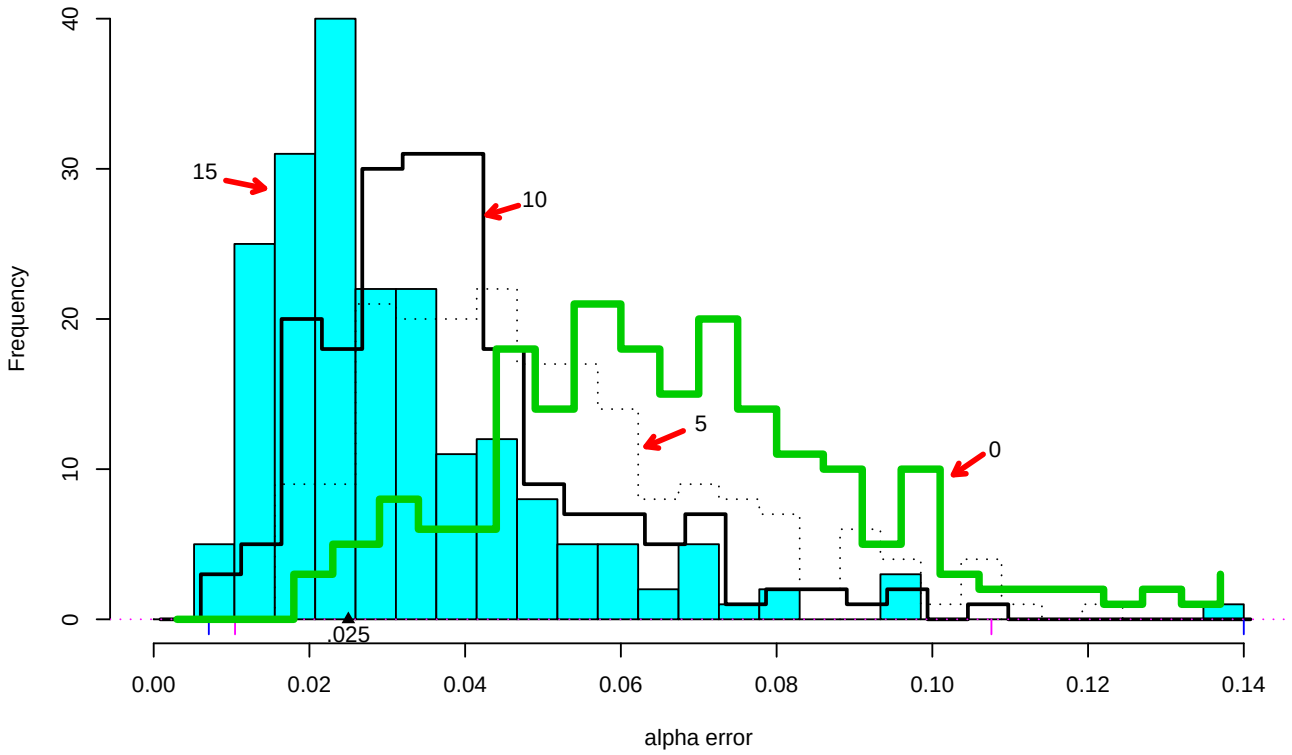
Ebay ($\alpha_0 = .025$) for Prostate Data

step	gene	z	$\hat{\delta}$	$\hat{S}d$	α
1	610	5.29	4.10	1.00	0.341
2	364	−4.66	−3.78	1.01	0.291
3	3940	−4.56	−3.71	1.02	0.255
4	4546	−4.50	−3.66	1.03	0.224
5	1720	5.15	3.61	0.96	0.200
6	4331	−4.34	−3.48	1.08	0.184
7	332	4.72	3.23	0.93	0.169
8	914	4.63	3.15	0.93	0.156
9	4088	−4.05	−3.11	1.16	0.147
10	3991	−4.04	−3.10	1.16	0.136
⋮					
45	2370	3.69	2.33	1.01	0.029
46	3282	3.68	2.32	1.01	0.027
47	905	3.63	2.27	1.02	0.027
48	3260	−3.49	−2.27	1.30	0.026
49	4154	−3.49	−2.26	1.30	0.026
50	4552	3.44	2.22	1.03	0.024

(CV error rate .07)

Simulation Study

- $N = 5000$ genes, $n_1 = 20$ "healthy",
  $n_2 = 20$ "sick"
- *Healthy*: $x_{ij} \sim \mathcal{N}(0,1)$
- *Sick*: $x_{ij} \sim \begin{cases} \mathcal{N}(1.5,2) & i = 1,2,\ldots,250 \\ \mathcal{N}(0,1) & i = 251,\ldots,5000 \end{cases}$
- z_i transformed two-sample t-stat for gene i:
  $z_i \sim \mathcal{N}(0,1),\ i > 250$
- `Ebay` $(z, \alpha_0 = .025)$ gives $\hat{\boldsymbol{\delta}}_I$, with
  $\hat{\alpha} = \Phi(-\|\hat{\boldsymbol{\delta}}_I\|/2c) \le .025$


Frequency
0
10
20
30
40
15
10
5
0
.025
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
alpha error

Connection with False Discovery Rates

- **Two-Groups Model** Assume proportion p_0 of genes "null": $z_i \sim \mathcal{N}(0,1)$;
  remainder $p_1 = 1 - p_0$ "non-null": $z_i \sim f_1(\cdot)$
- *Mixture Density*: $f(z) = p_0\varphi(z) + p_1 f_1(z)$
- *Local False Discovery Rate*:

$$\boxed{\text{fdr}(z) = \text{Prob}\{\text{null}|z\} = p_0\varphi(z)/f(z)}$$

- *Exponential Family Theory*:

$$\psi(z) = \log\{f(z)/\varphi(z)\} = \log\{p_0/\text{fdr}(z)\}$$

so

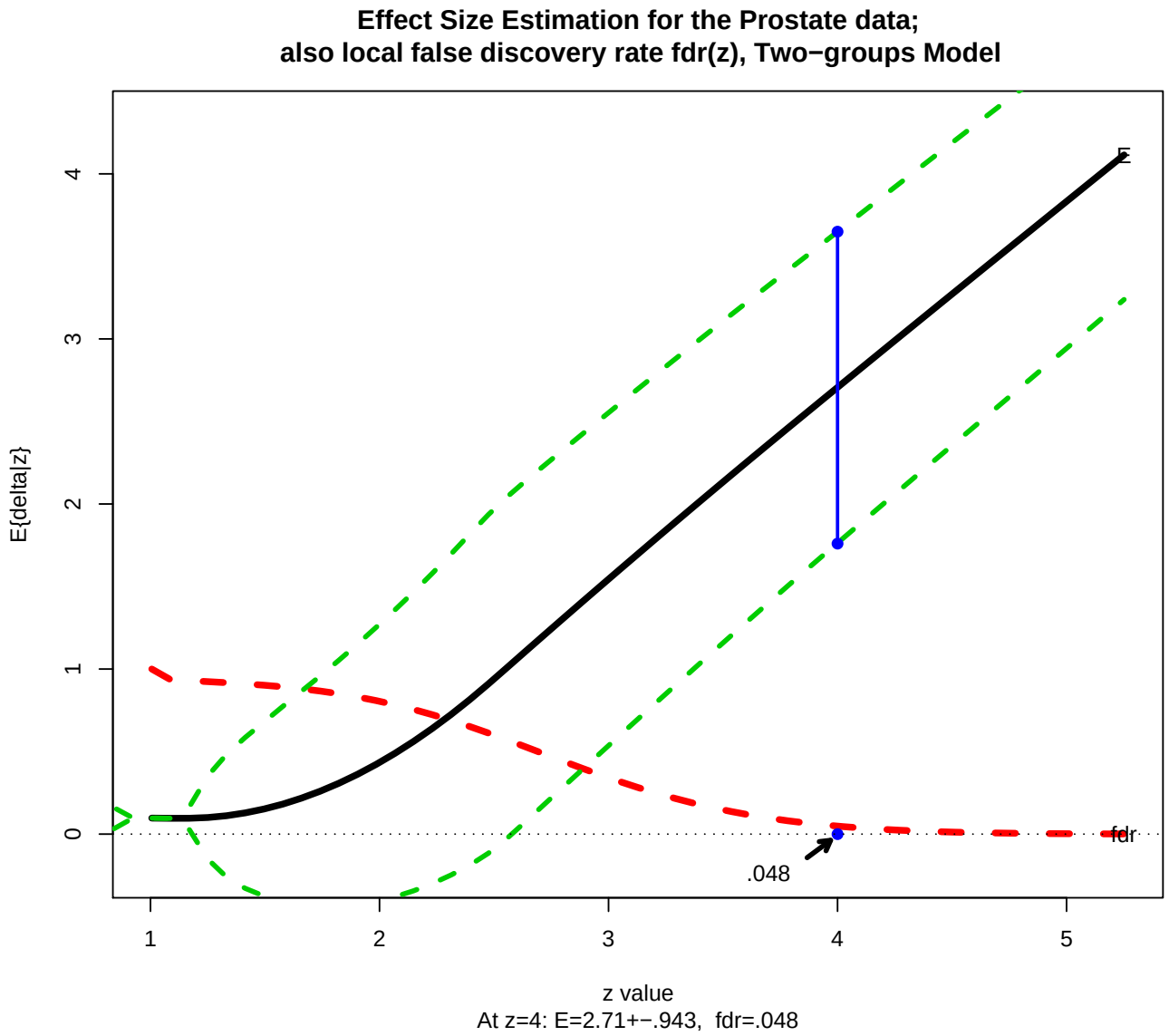
$$\boxed{-\frac{d\log\text{fdr}(z)}{dz} = E\{\delta|z\},} \quad -\frac{d^2\log\text{fdr}(z)}{dz^2} = \text{Var}\{\delta|z\}$$

Non-Null Means and Standard Deviations

- Assume *structural model* $z_i|\delta_i \sim \mathcal{N}(\delta_i, 1)$ and *two-groups model*: null proportion p_0 of genes have $\delta_i = 0$
- **Theorem** $g(\delta|z, \text{non-null}) = e^{\delta z - \psi_1(z)} \left[e^{-\delta^2/2} g(\delta|\text{non-null})\right]$ *with*

$$\psi_1(z) = \log\left\{\frac{p_0}{\text{fdr}(z)} \Big/ \frac{1-p_0}{1-\text{fdr}(z)}\right\}$$

- $E\{\delta_i|z_i, \text{non}\} = \psi_1'(z)$, $Sd\{\delta_i|z_i, \text{non}\} = \psi_1''(z)$
- Now $\hat{E} \pm \hat{S}d$ more appropriate for "effect size" estimation (Benjamini & Yekutieli, 2005; Efron, 2008, Sec. 7)


Effect Size Estimation for the Prostate data;
also local false discovery rate fdr(z), Two−groups Model
E{delta|z}
0
1
2
3
4
E
fdr
.048
1
2
3
4
5
z value
At z=4: E=2.71+−.943, fdr=.048

Correlation

- *Simple Model*: $Y_i \equiv (X_i - \mu_i)/\sigma_i$

  took Y_i's independent

- If $\not{\Sigma} = \text{cor}(\boldsymbol{Y})$ then prediction error of $S = \sum \delta_i Y_i$ is

$$\alpha = \Phi\left(-\frac{\|\boldsymbol{\delta}\|}{2c} \cdot \text{fact}\right), \quad \text{fact} = \left(\frac{\boldsymbol{\delta}'\boldsymbol{\delta}}{\boldsymbol{\delta}'\not{\Sigma}\boldsymbol{\delta}}\right)^{1/2}$$

- Ebay $\rightarrow \hat{\boldsymbol{\delta}}_I$, say $\#I = 50$

- Can use empirical Bayes shrinkage methods for 50×50 estimate $\hat{\not{\Sigma}}$

- Prostate data has, mostly, small correlation between genes, "fact" in range (.94,1.00) for $\#I = 1, 2, \ldots, 50$

step	gene	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\alpha}_{\text{cor}}$
1	610	4.10	0.342	0.34
2	364	−3.78	0.290	0.28
3	3940	−3.72	0.253	0.25
4	4546	−3.66	0.225	0.22
5	1720	3.61	0.202	0.22
6	4331	−3.47	0.183	0.19
7	332	3.23	0.169	0.18
8	914	3.15	0.157	0.17
9	4088	−3.11	0.146	0.15
10	3991	−3.09	0.136	0.15
⋮				
45	2370	2.33	0.029	0.05
46	3282	2.32	0.028	0.05
47	905	2.27	0.027	0.05
48	3260	−2.27	0.026	0.05
49	4154	−2.26	0.025	0.05
50	4552	2.22	0.024	0.05

Some Missing Topics

- Different prior weights for the two classes [change "μ_i" definition]
- Predicting other response variables (e.g., survival time)
- t-statistics and z-values $\left[z_i = \Phi^{-1}F_\nu(t_i) \dot{\sim} \mathcal{N}\left(\delta_i, \sigma^2(\delta_i)\right)\right]$
- How immune are *empirical* Bayes estimates to selection biases?

References

Benjamini and Yekutieli (2005). False discovery rate-adjusted confidence intervals.... *JASA*.

Brown (1971). Admissible estimators.... *Ann. Math. Stat.*, 855–903.

Efron (2008). Microarrays, empirical Bayes, and the two-groups model. Available at `http://stat.stanford.edu/~brad/papers.`

Singh, et al. (2002). Gene expression correlates.... *Cancer Cell*, 302–309.

Stein (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 1135–1151.

Tibshirani, Hastie, Narasimhan and Chu (2002). Diagnosis of multiple cancer types by shrunken centroids.... *PNAS*, 6567–6572.