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

- $z_i = \text{two-sample } t\text{-stat comparing prostate patients with healthy controls for gene } i$ (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
A Very Simple Model

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

$$\frac{X_i - \mu_i}{\sigma_i} \overset{\text{ind}}{\sim} \mathcal{N}\left(\pm \frac{\delta_i}{2c}, 1\right)$$

\[
\begin{cases}
  "-" & \text{healthy} \\
  "+" & \text{sick}
\end{cases}
\]

- Constant "c" =

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

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

- $\delta_i$ = “effect size”
Ideal Prediction Rule

- **Compute**

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

where

\[
||\delta||^2 = \sum \delta_i^2.
\]

- **Predict**: “healthy” if \( S < 0 \),
  “sick” if \( S > 0 \).

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

  ◊ accurate prediction if \( ||\delta|| \) is large
  ◊ \( \alpha = ”prediction error” \)
Ideal statistic $S$ for prediction

Healthy: $N(-D/2c, D)$

Sick: $N(+D/2c, D)$

$S \rightarrow \beta \alpha$

$D = $ squared length of delta vector
Estimating the Ideal Rule

- $S = \sum \delta_i \left( \frac{X_i - \mu_i}{\sigma_i} \right)$ depends on $(\delta_i, \mu_i, \sigma_i), \ i = 1, 2, \ldots, N$.

- Estimate $\mu_i, \sigma_i$ in usual way:
  \[ \hat{\mu}_i = (\bar{x}_{i1} + \bar{x}_{i2})/2, \ \hat{\sigma}^2_i = (ss_{i1} + ss_{i2})/(n-2) \]

- Obvious estimate of $\delta_i$:
  \[
  z_i = c \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\hat{\sigma}_i} \sim \mathcal{N}(\delta_i, 1)
  \]

Actually “$t_i$”, transformed to $z_i = \Phi F_{n-2}(t_i)$. 
Trouble!

- $\bar{\delta}_i = z_i$ poor estimate of $\delta_i$
- Wildly overestimates $||\delta||^2$, underestimates prediction error
  $\alpha = \Phi(-||\delta||/2c)$
- 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 $\lambda$ (CRAN: "pamr")
Shrunken Centroids for Prostate Data

<table>
<thead>
<tr>
<th>shrink value</th>
<th>nonzero genes</th>
<th>error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>6033</td>
<td>0.34</td>
</tr>
<tr>
<td>0.54</td>
<td>3763</td>
<td>0.33</td>
</tr>
<tr>
<td>1.08</td>
<td>1931</td>
<td>0.23</td>
</tr>
<tr>
<td>1.62</td>
<td>866</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>2.16</strong></td>
<td><strong>377</strong></td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>2.70</td>
<td>172</td>
<td>0.10</td>
</tr>
<tr>
<td>3.24</td>
<td>80</td>
<td>0.16</td>
</tr>
<tr>
<td>3.78</td>
<td>35</td>
<td>0.30</td>
</tr>
<tr>
<td>4.32</td>
<td>4</td>
<td>0.41</td>
</tr>
<tr>
<td>4.86</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>5.29</td>
<td>0</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Three Shrinkage Predictors for Prostate Data

(Shrink, CValpha)
1: (1.42, 15%)
2: (2.34, 7%)
3: (3.06, 14%)
Choosing the Shrinkage Constant

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

• Suppose we could calculate Bayes posterior expectations

\[ \tilde{\delta}_i = E\{\delta_i | z\} \]

• Could use predictor

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

with approximate error probabilities

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

• Bayes estimates immune to selection bias \( z_{610} = 5.29 \)

• Model \( \delta \sim g(\cdot) \) and \( z|\delta \sim \mathcal{N}(\delta, 1) \)

• Marginal Density:

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

• Posterior Density:

\[
g(\delta|z) = e^{\delta z - \psi(z)} \left[ e^{-\delta^2/2} g(\delta) \right]
\]
\[
\psi(z) = \log \left( \frac{f(z)}{\varphi(z)} \right)
\]

• \( g(\delta|z) \) 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 
  \( \hat{\psi}(z) = \log\{\hat{f}(z)/\phi(z)\} \) to give 
  \begin{align*}
  \hat{E}\{\delta|z\} &= \hat{\psi}'(z), \\
  \hat{sd}(z) &= \sqrt{\hat{\psi}''(z)}
  \end{align*}

- For a given subset “\( I \)” of the genes,
  \begin{align*}
  \hat{\delta}_I &= \{\hat{\delta}_i = \hat{E}\{\delta|z_i\}\}, \\
  \hat{S} &= \sum \hat{\delta}_i \left( \frac{X_i - \hat{\mu}_i}{\hat{\sigma}_i} \right), \\
  \hat{\alpha}_I &= \Phi(-||\hat{\delta}_I||/2c)
  \end{align*}

- Can choose “\( I \)” to include biggest \( \hat{\delta}_i \) values without worrying about selection bias.
Ebay estimates of $E(\delta|z)$ and $S_d(\delta|z)$ for Prostate data; also local false discovery rate $fdr(z)$; Shown for $z$ positive

At $z=4$: $E=2.58$, $S_d=.94$, $fdr=.048$

E value
$E(\delta|z)$
$S_d$
$fdr$
$E=z$

z value
At $z=4$: $E=2.58$, $S_d=.94$, $fdr=.048$
Large-Scale Prediction Problems

E_{\hat{\delta}|z}$ compared with Shrinkage predictors

$z \rightarrow$

1: (1.42, 15%, 1163)
2: (2.34, 7%, 254)
3: (3.06, 14%, 87)
Ebay rule 7.5%

\begin{itemize}
  \item 1: (1.42, 15%, 1163)
  \item 2: (2.34, 7%, 254)
  \item 3: (3.06, 14%, 87)
\end{itemize}

\begin{itemize}
  \item Ebay rule 7.5%
\end{itemize}
“Ebay” Prediction Rule

- Select target error probability $\alpha_0$
- Program selects predictor genes in order of $|\hat{\delta}_i|$, largest first
- Continues until $\|\hat{\delta}_I\|$ 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

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

(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 $\hat{\delta}_I$, with
  $\hat{\alpha} = \Phi(-||\hat{\delta}_I||/2c) \leq .025$
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**:
  
  $$fdr(z) = \text{Prob}\{\text{null}|z\} = \frac{p_0 \varphi(z)}{f(z)}$$

- **Exponential Family Theory**:

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

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

\[
\psi_1(z) = \log \left\{ \frac{p_0}{\text{fdr}(z)} \left/ \frac{1 - p_0}{1 - \text{fdr}(z)} \right. \right\}
\]

- $E\{\delta_i|z_i, \text{non}\} = \psi_1'(z)$, $Sd\{\delta_i|z_i, \text{non}\} = \psi_1''(z)$

- Now $\hat{E} \pm \hat{Sd}$ 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

At $z=4$: $E=2.71\pm0.943$, $fdr=0.048$
Correlation

- **Simple Model**: \( Y_i \equiv (X_i - \mu_i)/\sigma_i \)
  took \( Y_i \)'s independent

- If \( \Sigma = \text{cor}(Y) \) then prediction error of \( S = \sum \delta_i Y_i \) is

\[
\alpha = \Phi\left(-\frac{\|\delta\| \cdot \text{fact}}{2c}\right), \quad \text{fact} = \left(\frac{\delta' \delta}{\delta' \Sigma \delta}\right)^{1/2}
\]

- Ebay \( \rightarrow \hat{\delta}_I \), say \#\( I = 50 \)

- Can use empirical Bayes shrinkage methods for \( 50 \times 50 \) estimate \( \hat{\Sigma} \)

- Prostate data has, mostly, small correlation between genes, “fact” in range \((.94,1.00)\) for \#\( I = 1, 2, \ldots, 50 \)
## Large-Scale Prediction Problems

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

$$\begin{align*}
  z_i &= \Phi^{-1}F_v(t_i) \sim N(\delta_i, \sigma^2(\delta_i))
\end{align*}$$

• How immune are empirical Bayes estimates to selection biases?
References

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


