

Some Statistical Issues Related to Feature Detection Using Random Forests

Grant Izmirlian

DHHS/NIH/NCI

Division of Cancer Prevention

Biometry Research Group

Bethesda, MD

`izmirlian@nih.gov`

Proteomics: Why?

- transcription translation
- gene $\xrightarrow{\hspace{1.5cm}}$ RNA $\xrightarrow{\hspace{1.5cm}}$ protein
 - direction of increasing complexity
 - isoforms, folding
 - study of RNA expression
 - misses “post-translational modifications”
 - study of protein products
 - the true substance of cellular processes

Proteomics: Innovation

- SELDI-TOF
 - surface enhanced laser desorption time of flight
 - much faster than previous methods
- before SELDI-TOF
 - studies limited to purified samples within limited mass ranges
 - electrophoresis gels

SELDI-TOF: Procedural Overview

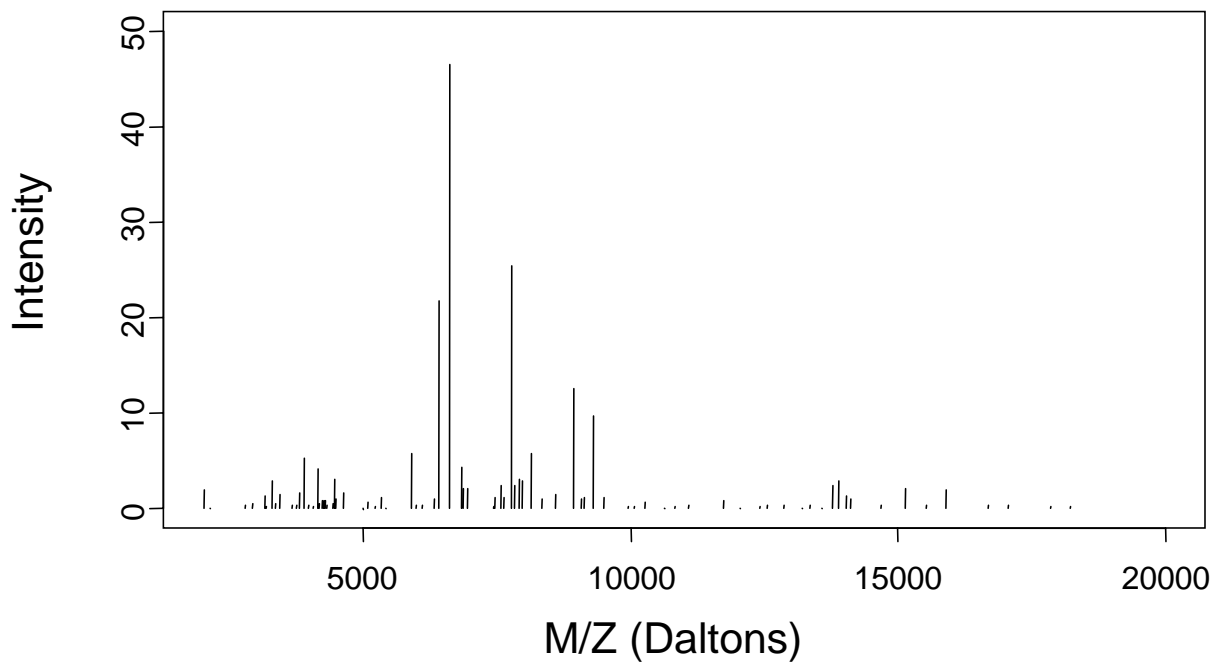
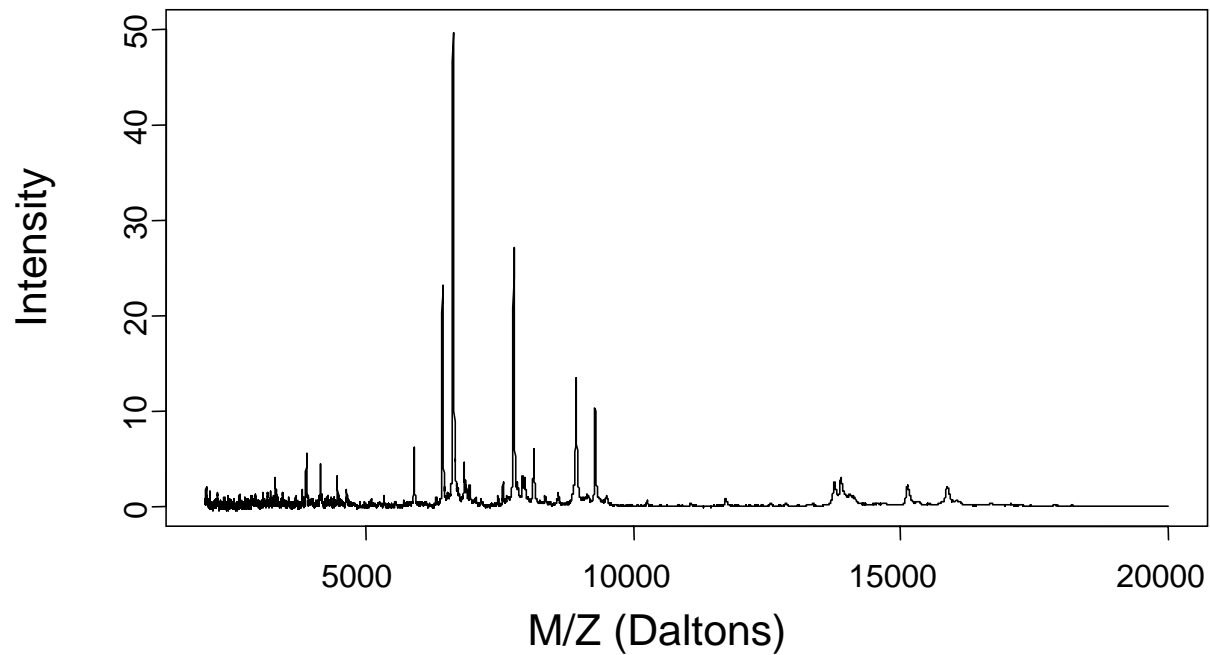
- samples processed with a “matrix” of compounds
- placed onto a metal surface
- surface and matrix combination give specific affinity e.g.:
 - hydrophilic strong anion exchange
- ionized via a cool laser
 - matrix absorbs energy—prevents vaporization
 - proteins separated according to mass via time of flight down a magnetic column

SELDI-TOF: Caveats

- mass to charge ratio
 - two ions with masses in ratio 2:1 and charges in ratio 1:2 are indistinguishable
 - most likely prevalent charge is +1, non-ignorable prevalence of +2.
- many proteins are fragmented
 - original target mass represented by multiple smaller masses

Signal Compression: Yay or Nay

- Nay
 - compression results in loss of information
 - requires choosing a method
 - if your classifier can handle under-specified problems then just feed it raw spectra
- Yay
 - goals are classification *and* important feature identification with emphasis on the latter
 - SELDI-TOF manufacturer's method: θ, w, r



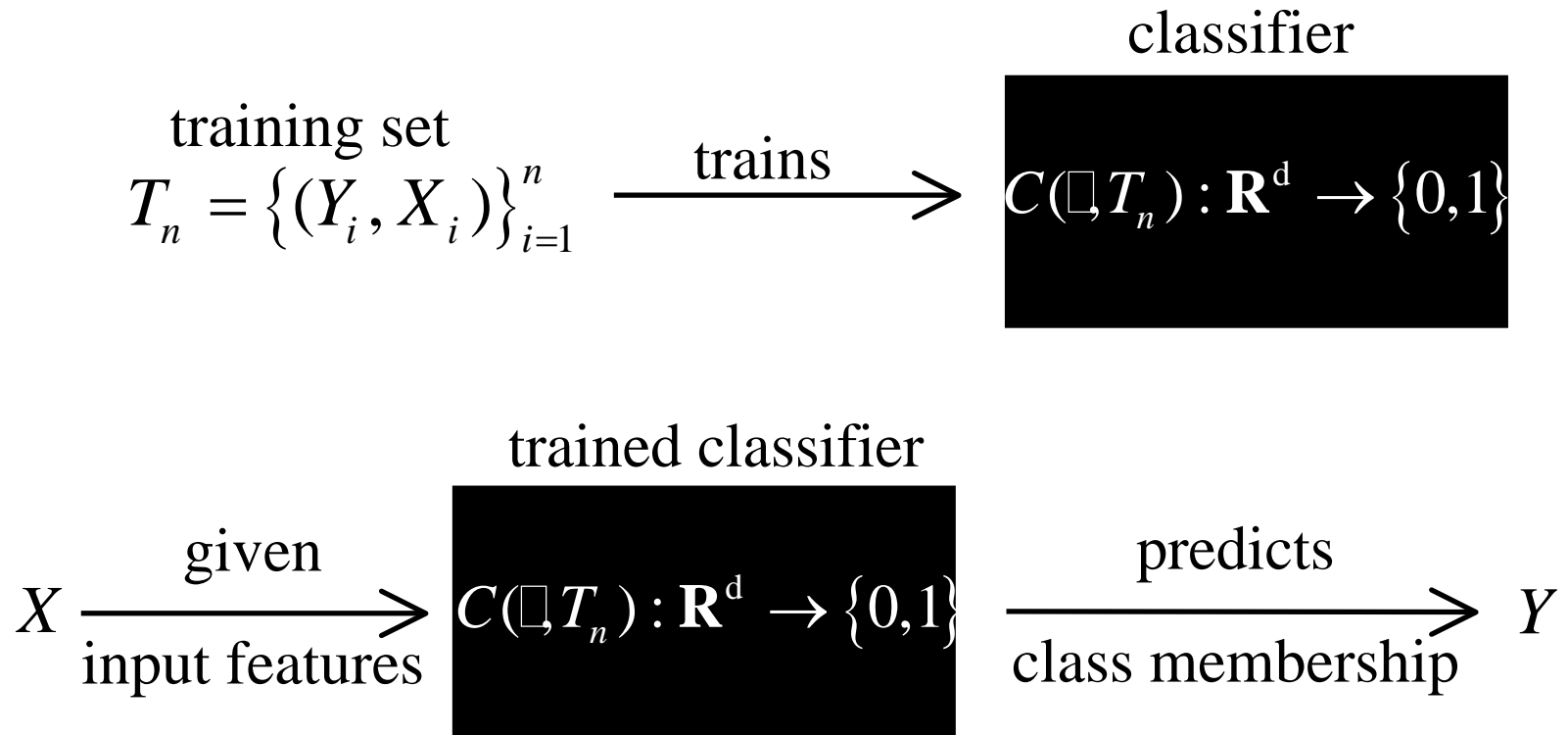
Data

- subjects $i=1,2,\dots,n$ each with data on
 - class membership, $Y_i \in \{0,1\}$ e.g. cancer/no-cancer or respondent/non-respondent
 - intensities at each peak in the compressed spectrum, $X_i \in \mathbb{P}^d$
 - $X_i(k)$ is subject i 's intensity at feature k

$$T_n = \{(Y_i, X_i) : i = 1, 2, \dots, n\}, \quad Y_i \in \{0, 1\}, \quad X_i \in \mathbf{R}^d \text{ and } d > n$$

$$F(x) = \mathbb{P} \{X_i(k) \leq x(k) : 1 \leq k \leq d\} \text{ and } \pi(x) = \mathbb{P} \{Y_i = 1 | X_i = x\}$$

Classification



- conditional upon $X = x$, the *expected value* of the output of a trained *unbiased* classifier is equal to

$$\pi(x) = \mathbf{P} \{Y_i = 1 | X_i = x\}$$

Validation

- split the sample into training and validation portions

$$T_n = \tilde{T}_m \cup V_{n-m}$$

- sensitivity, specificity and classification error

$$Se = \frac{\sum_{i \in V_{n-m}} Y_i C(X_i, \tilde{T}_m)}{\sum_{i \in V_{n-m}} Y_i} \quad Sp = \frac{\sum_{i \in V_{n-m}} (1 - Y_i) (1 - C(X_i, \tilde{T}_m))}{\sum_{i \in V_{n-m}} (1 - Y_i)}$$

$$Err = \frac{1}{n - m} \sum_{i \in V_{n-m}} I(C(X_i, \tilde{T}_m) \neq Y_i)$$

The Classification Tree (CT)

- the root node: all of P^d , entire sample
- A parent node is split into two child nodes according to a rule based upon the value of a single feature: $I(X_i(k) \geq x)$
- Feature, k , and threshold, x , chosen to optimize
 - between generation increase in node class purity
 - “purity” measured using the Gini criterion
- Node splitting continues until entire node belongs to a single class
- predicted class assigned to input, $x \in P^d$:
 - starting at the root node
 - following all splitting rules on x
 - until terminal node is reached

Over-fit, Bias and Variance

- CT perfectly classifies T_n
- Y given X not deterministic
 - “noise” in the data
 - perfect classification on data with noise
 - \longleftrightarrow the “fit” is “fitting” noise
- the CT’s are dense in $L_2(\mathbf{R}^d, \{0,1\})$
- CT is an unbiased classifier $\mathbf{E}[C(x, T_n)] = \pi(x)$
 - but due to over-fit it has high variance

Aggregating a Classifier

- imagine having a sequence of i.i.d. replicates of the training set $\{T_n^1, T_n^2, \dots, T_n^M\}$
- train a classifier on each of the training sets
- form aggregate classifier by voting

$$C\left(x, \{T_n^r\}_{r=1}^M\right) = I\left(\frac{1}{M} \sum_{r=1}^M C(x, T_n^r) > \frac{1}{2}\right)$$

- Aggregation
 - Smooths out over-fitting
 - Reduces variability

“Bagging” a Classifier

- In practice we have only a single training set, T_n
- But can form a sequence of bootstrap replicates, $\{\tilde{T}_n^{(1)}, \tilde{T}_n^{(2)}, \dots, \tilde{T}_n^{(B)}\}$ e.g. each is a random sample of size, n , drawn with replacement from T_n
- Form the aggregated classifier from the sequence of bootstrapped replicated training sets

$$C(x, T_n) = I\left(\frac{1}{B} \sum_{b=1}^B C(x, \tilde{T}_n^{(b)}) > \frac{1}{2}\right)$$

“632 Cross-Validation”

- Bagging allows a novel type of validation
- Let $O_b = \{i : 1 \leq i \leq n, (Y_i, X_i) \notin \tilde{T}_n^{(b)}\}$ for $b \in \{1, 2, \dots, B\}$
- called the portion of the sample that is “out-of-bag” relative to the bootstrap sample, $\tilde{T}_n^{(b)}$
- Let $O_i^{\text{a}} = \{b : 1 \leq b \leq B, i \in O_b\}$ for $i \in \{1, 2, \dots, n\}$
- The indices of bootstrap replicates for which sample i is out-of-bag

“632”-cont'd: Sens. Spec. & Err.

$$Se = \frac{\sum_{i=1}^n Y_i I\left(\frac{1}{|\mathcal{O}_i^{\hat{a}}|} \sum_{b \in \mathcal{O}_i^{\hat{a}}} C(X_i, \tilde{T}_n^{(b)}) > \frac{1}{2}\right)}{\sum_{i=1}^n Y_i}$$

$$Sp = \frac{\sum_{i=1}^n (1 - Y_i) I\left(\frac{1}{|\mathcal{O}_i^{\hat{a}}|} \sum_{b \in \mathcal{O}_i^{\hat{a}}} C(X_i, \tilde{T}_n^{(b)}) < \frac{1}{2}\right)}{\sum_{i=1}^n (1 - Y_i)}$$

$$Err_{n,B} = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i I\left(\frac{1}{|\mathcal{O}_i^{\hat{a}}|} \sum_{b \in \mathcal{O}_i^{\hat{a}}} C(X_i, \tilde{T}_n^{(b)}) < \frac{1}{2}\right) + (1 - Y_i) I\left(\frac{1}{|\mathcal{O}_i^{\hat{a}}|} \sum_{b \in \mathcal{O}_i^{\hat{a}}} C(X_i, \tilde{T}_n^{(b)}) > \frac{1}{2}\right) \right\}$$

Random Forest (RF) Algorithm

- Random Forest (RF) algorithm is a bagged ensemble of CT's
- Combined with 632 cross-validation estimates of sensitivity, specificity and error rate
- Additional Randomness Property:
 - During training when a node is split search for best feature limited to randomly selected subset of all features, size $\log_2(d)$

$$C(x, T_n, \xi) = I\left(\frac{1}{B} \sum_{b=1}^B C(x, \tilde{T}_n^{(b)}, \xi^{(b)}) > \frac{1}{2}\right)$$

- Helps to reduce between tree correlation

Importance Measure

- RF computes a feature importance measure
- For a given feature, j , define:

$$\Delta_{j,b,i} = I\left(C\left(X_i, \tilde{T}_n^{(b)}, \tilde{\xi}\right) = Y_i\right) - I\left(C\left(\tilde{X}_i^{(j)}, \tilde{T}_n^{(b)}, \tilde{\xi}\right) = Y_i\right)$$

- where $\tilde{X}_i^{(j)}(k) = X_i(k)$, $k \neq j$ and $\tilde{X}_i^{(j)}(j) = X_{i'}(j)$, $i' \sim \text{UNIF}[\mathcal{O}_b]$,

$$b \in \{1, 2, \dots, B\}, \text{ and } i \in \mathcal{O}_b$$

- The measure of feature j importance is:

$$\Delta_j = \frac{1}{B} \sum_{b=1}^B \frac{1}{|\mathcal{O}_b|} \sum_{i \in \mathcal{O}_b} \Delta_{j,b,i} \xrightarrow{\text{a.s.}} \mathbf{E}\left[\Delta_{j,b,i} \mid T_n\right] \text{ as } B \rightarrow \infty$$

Variance of Importance Measure

$$\begin{aligned}
 \text{var}[\Delta_j] &= \frac{1}{B} \mathbf{E} |O_b|^{-1} \text{var}[\Delta_{j,b,i}] + \frac{1}{B} \mathbf{E} \frac{|O_b|^{-1}}{|O_b|} \text{cov}[\Delta_{j,b,i}, \Delta_{j,b,i'}] \\
 &+ \frac{B-1}{B} \mathbf{E} \frac{|O_b \cap O_{b'}|}{|O_b| |O_{b'}|} \text{cov}[\Delta_{j,b,i}, \Delta_{j,b',i} | O_b \cap O_{b'} \neq \emptyset] \mathbf{P}\{O_b \cap O_{b'} \neq \emptyset\} \\
 &+ 2 \frac{B-1}{B} \mathbf{E} \frac{|O_b \setminus O_{b'}|}{|O_b| |O_{b'}|} \text{cov}[\Delta_{j,b,i}, \Delta_{j,b',i''} | O_b \setminus O_{b'} \neq \emptyset] \mathbf{P}\{O_b \setminus O_{b'} \neq \emptyset\} \\
 &= O\left(\frac{1}{nB}\right) + O\left(\frac{1}{B}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right)
 \end{aligned}$$

Variance cont'd

- Since n is fixed but B can be made arbitrarily large, the sum of the last two terms is the dominating quantity
- Note: the dominating quantity arises from
 - within individual, between tree correlation of Δ
 - between individual, between tree correlation of Δ
- The variance supplied in the RF algorithm is $O(1/B)$ and therefore incorrect

Inference on Importance Measures

- Existing controversy even in the medical literature surrounding reproducibility of findings in proteomics profiling studies
- Reproducibility should be an important part of any scientific investigation
- Goal for importance measures:
 - Normalization by correct $O\left(n^{-1/2}\right)$ standard error this will give the correct ranking
- Correct p -values are also crucial
 - so you aren't surprised when your findings aren't reproducible

Simulation Studies: Outline

- Part I: monte carlo investigation under H0
- Part II: monte carlo investigation under an H1
- In both cases
 - Sample size $n = 100$, $d = 138$ features, 1000 monte carlo replicates
 - the importance measures, $\{\Delta_j\}_{j=1}^d$, are normalized by the monte carlo standard errors
 - conduct Benjamini Hochberg FDR step-down procedure controlling FDR at 10%

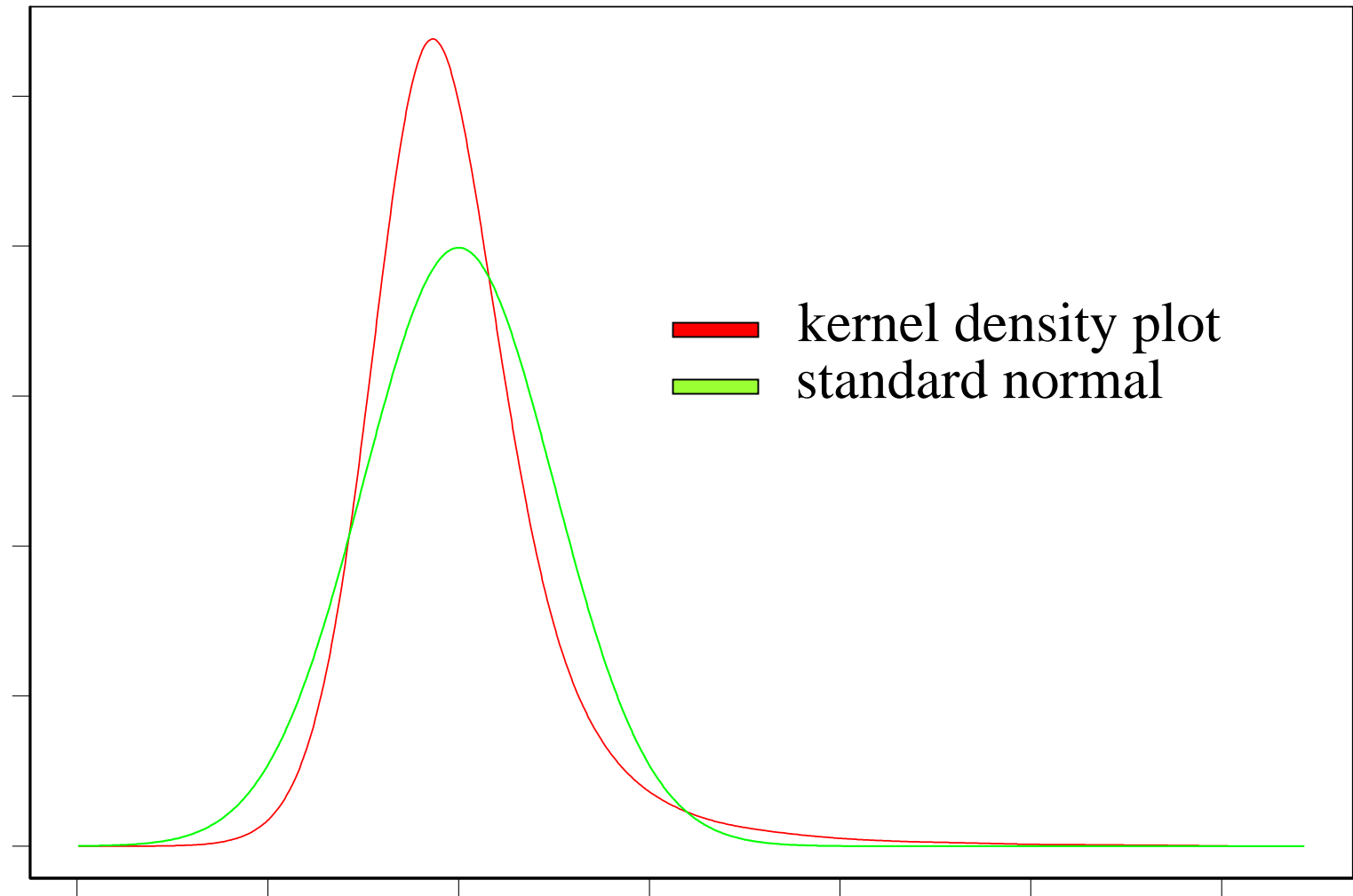
Outline-cont'd

- H0 and H1 differ in specification of $\pi(x) = \mathbf{P}\{Y_i = 1 | X_i = x\}$
- Both are based upon the same distribution of spectra, $F(x) = \mathbf{P}\{X_i(k) \leq x(k) : 1 \leq k \leq d\}$
- Spectrum distribution, $F(x)$, specified as $\log(X_i) \sim \text{MVN}(\mu, \Sigma)$, where log is component-wise
- Means and diagonal variances estimated from human serum proteomics data
- Correlation matrix: not enough data, truncated values < 0.9 to zero, values ≥ 0.9 to 0.9

Part I: monte carlo under H0

- balanced study: $\pi(x) = \mathbf{P}\{Y_i = 1 | X_i = x\} \equiv \frac{1}{2}$
- normalized importance measures, $\{\Delta_j\}_{j=1}^d$,
by monte carlo variance giving nominal t-statistics
- assumed that the per feature nominal t-statistics are identically distributed
- Formed kernel density plot from the aggregate sample of 138,000 nominal t-statistics
- mean was very nearly zero, variances forced to be 1

Importance measure nominal t-statistics null distribution



Ramifications for FDR

Mass #	T-stat	True p	Nominal p	BH
M.062	3.924365	0.0087609	0.000205870	0.000725
M.022	3.582749	0.0110805	0.000535262	0.00145
M.037	3.010289	0.0169607	0.002473143	0.00217
M.026	2.195273	0.0337820	0.017605429	0.0029
M.060	1.931609	0.0438409	0.030973049	0.00362
M.044	1.905829	0.0449734	0.032663204	0.00435
M.055	1.897222	0.0453673	0.033244945	0.00507
M.029	1.851478	0.0476202	0.036488500	0.0058
M.036	1.781664	0.0516695	0.041958674	0.00652
M.009	1.626041	0.0610650	0.056679714	0.00725
M.002	1.398013	0.0802351	0.085671609	0.00797
M.020	1.218667	0.0992894	0.115762714	0.0087
M.004	1.178971	0.1043369	0.123378482	0.00942
M.068	1.165515	0.1061538	0.126041473	0.0101

Empirical FDR

- For each monte carlo rep
 - Table of per feature nominal t-statistics sorted
 - Computed nominal p-value (under t -33df)
 - Computed “true p-value” as percentile
 - Applied B-H step-down procedure at FDR=10%
 - Counted discoveries under “nominal” and “true”

Classifier Error Rate and Importance Measure FDR

- o.o.b. error rate (95% mc CI)

50% (40%, 62%)

- Empirical FDR = $\frac{1}{\#m.c. \text{ reps}} \sum_{m.c. \text{ reps}} I(1 \text{ or more discoveries})$

– under nominal p-value: 79%

– under true p-value: 8.5%

Part II: monte carlo under an H1 specification of $\pi(x)$

- arbitrarily picked 3 features
 - split at respective medians
- formed 7 dummy variables for 8 categories
- chose 8 coefficients,
 - using inv logit to assign prob. to each of 8 cat's

specification of $\pi(x)$

#22	#88	#96	π_k
0	0	0	0.05
1	0	0	0.15
0	1	0	0.70
1	1	0	0.70
0	0	1	0.10
1	0	1	0.30
0	1	1	0.10
1	1	1	0.99

- Overall class-1 prevalence
$$\sum_{k=0}^7 \pi_k p_k = 0.34$$
 - where $\pi(x) = \pi_k$ for $x \in \{\text{cell } k\}$
 - and $p_k = \int_{\{\text{cell } k\}} dF(x)$was estimated from a dataset of size 10,000

Comparison w/ simple 2x2 table

- If first row is $j=0$, then
 - logged odds ratio for “ j even to j odd” is
$$\text{logit}\left(\sum_{j \text{ odd}} \pi_j p_j\right) - \text{logit}\left(\sum_{j \text{ even}} \pi_j p_j\right) = 2.23,$$
- The simple 2x2 table test for association has power 90% to detect $\log(\text{OR}) = 2.23$ if class-1 prevalence is 34% with a sample size of $n = 70$
- Keep in mind the 2x2 table test
 - knows what to look at “omnipotent 2x2 test”

Power under H1

- Effective sample size for importance measure is $n/3$ = size of out-of-bag sample
- So to compare the power of RF with the “omnipotent 2x2 test” the RF sample size should be $n = 210$

Classifier Error Rate and Importance Measure FDR

- o.o.b. error rate (95% mc CI)
31.1% (25.3%, 36.3%)
- Strongest feature nominal t -statistic mean: 7.0
- Strongest feature detected using FDR 10%
 - under nominal p-value: 93% of the mc reps
 - under true p-value: 2.5% of the mc reps

Conclusions

- Recommend normalization by correct $O(n^{-1/2})$ - standard errors
- Under H_0 , even w/ correct std err normalization, FDR's under nominal p-value seriously inflated
 - Recommend use of the correct p-value via bootstrap
 - Current work to derive reasonable tail probs

Conclusions: H1

- Under a specified H1,
 - the B-H step-down at FDR=10% under nominal p-values gives a power of 90%.
 - appears that RF pays \$0 to sift through garbage
- The same B-H procedure under correct p-values paints an entirely different picture