A Prediction Interval for the Misclassification Rate

E.B. Laber

&

S.A. Murphy

MICHIGAN STATISTICS

Institute for Social Research
Survey Research Center
SRC

Outline

– Review
– Three challenges in constructing PIs
– Combining a statistical approach with a learning theory approach to constructing PIs
– Relevance to confidence measures for the value of a dynamic treatment regime.

Review

– *X* is the vector of features in R^q, *Y* is the binary label in *{-1,1}*

– Misclassification Rate: $err(f) = E[1\{Y \neq f(X)\}]$

– Data: *N* iid observations of *(Y,X)*

– Given a space of classifiers, $\mathcal{F}$, and the data, use some method to construct a classifier, $\widehat{f}$

– The goal is to provide a PI for $err(\widehat{f})$

Review

– Since the loss function $1\{Y \neq f(X)\}$ is not smooth, one commonly uses a smooth surrogate loss to estimate the classifier

– Surrogate Loss: *L(Y,f(X))*

– $\widehat{f} \in \min_{f \in \mathcal{F}} E_N[L(Y, f(X))]$

(E_N denotes expectation with respect to empirical distribution)

Review

General approach to providing a PI:

- We estimate $err(\hat{f})$ using the data, resulting in $\widehat{err(\hat{f})}$
- Derive approximate distribution for

$$\left(\widehat{err(\hat{f})} - err(\hat{f})\right)$$

- Use this approximate distribution to construct a prediction interval for $err(\hat{f})$

Review

A common choice for $\widehat{err(\hat{f}\,)}$ is the resubstitution error or training error:

$$\widehat{err}_{rs}(f) = E_n[1\{Y \neq f(X)\}]$$

evaluated at $f = \hat{f}$ e.g. if $f(x) = sign(x^T\beta))$ then

$$\widehat{err(\hat{f})} = E_n[1\{YX^T\hat{\beta} < 0\}]$$

Three challenges

1) $\mathcal{F}$ is too large leading to over-fitting and
$E\left[\widehat{err(\hat{f}\,)} - err(\hat{f}\,)\right] < 0$ (negative bias)

2) $err(f) = E[1\{Y \neq f(X)\}]$ is a non-smooth function of f.

3) $\widehat{err(\hat{f}\,)}$ may behave like an extreme quantity

No assumption that $\hat{f}$ is close to optimal.

A Challenge

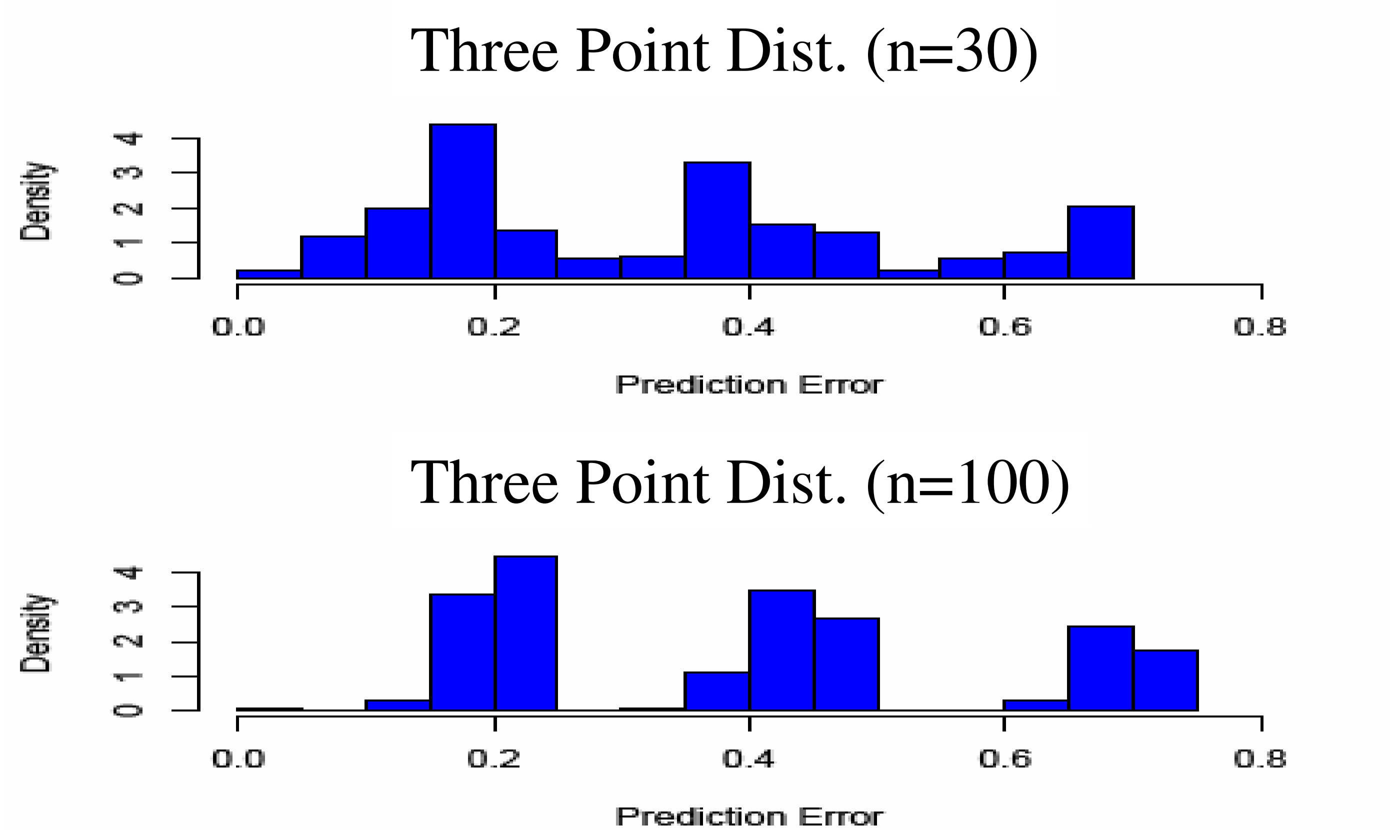
2) $err(f) = E[1\{Y \neq f(X)\}]$ is non-smooth.

Example: The unknown optimal classifier has quadratic decision boundary. We fit, by least squares, a linear decision boundary

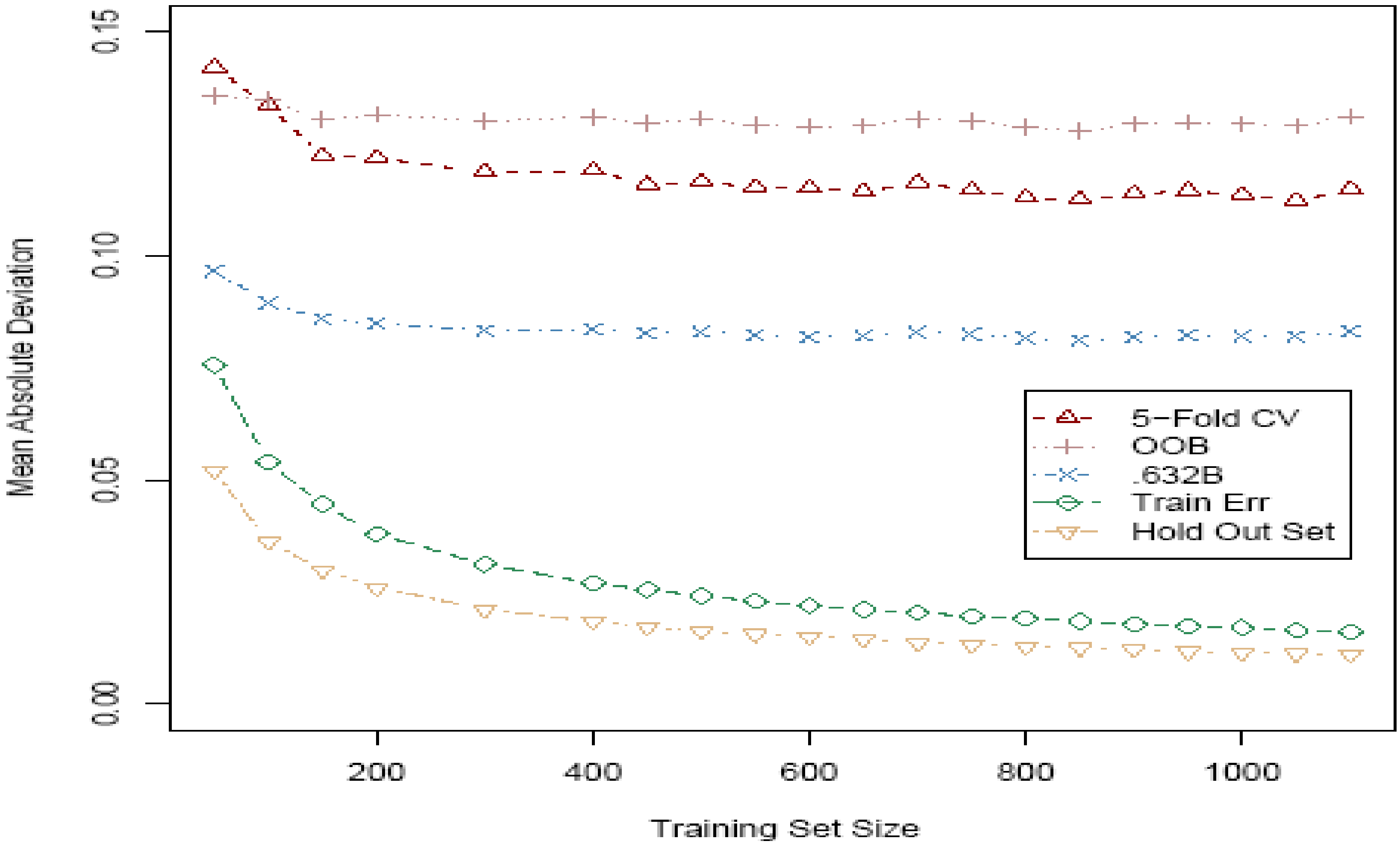
$$f(x) = sign(\beta_0 + \beta_1 x)$$

$$err(f) = E[1\{Y(\beta_0 + \beta_1 X) < 0)\}]$$

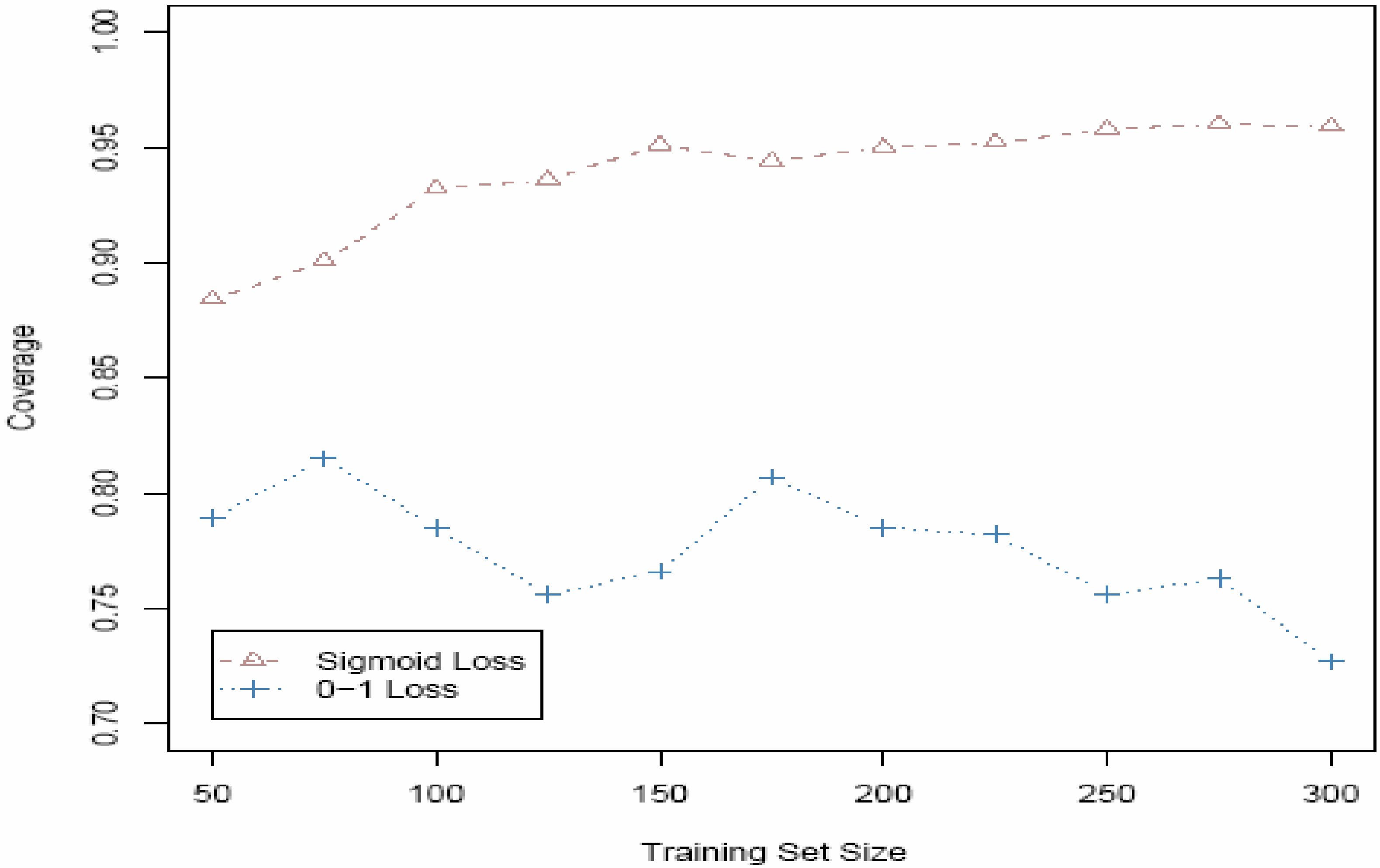
Density of $err(\widehat{f})$



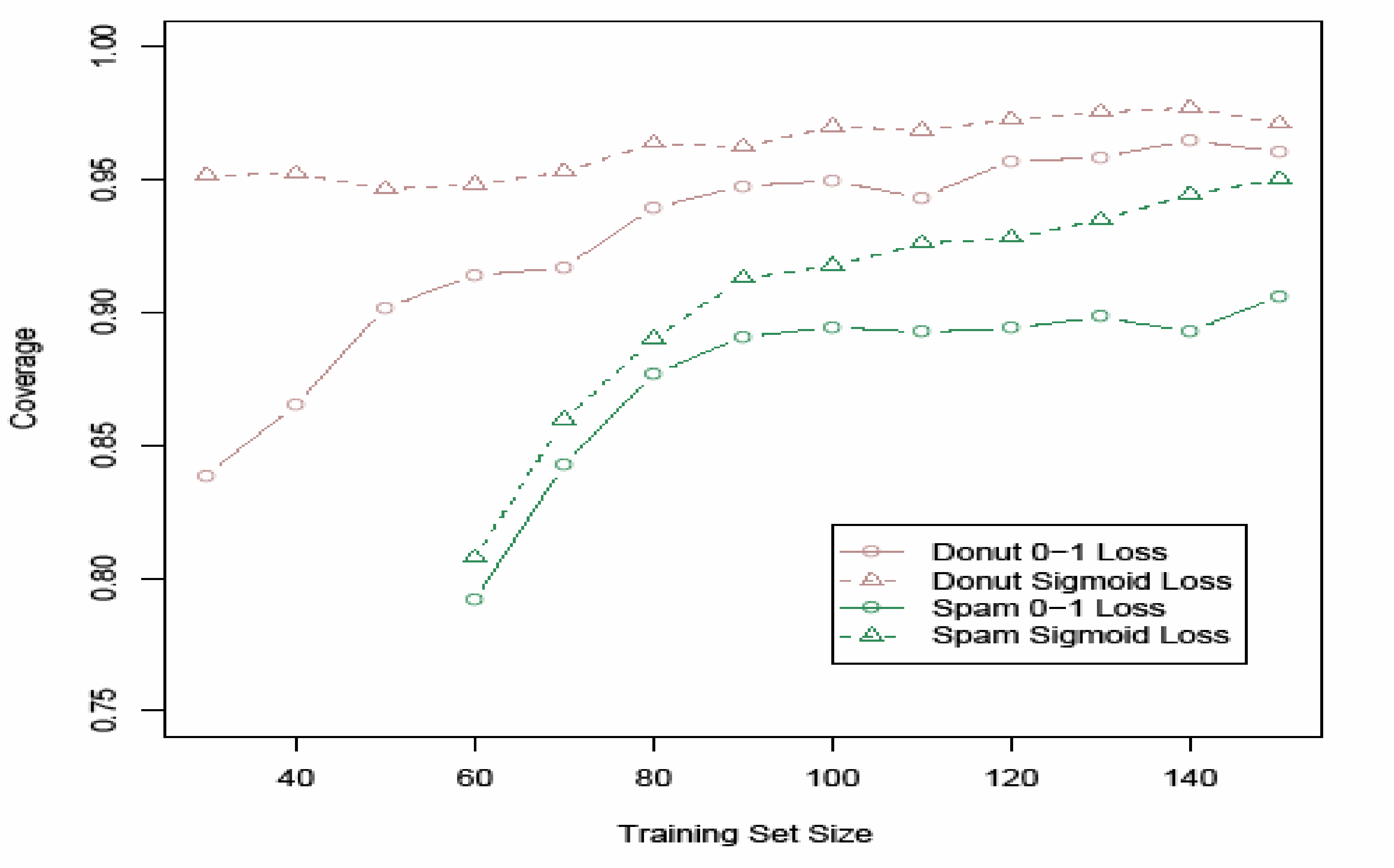
Bias of Common $\widehat{err(\hat{f})}$ on Three Point Example



Coverage of Bootstrap PI in Three Point Example (goal =95%)



Coverage of Correctly Centered Bootstrap PI (goal= 95%)



Coverage of 95% PI (Three Point Example)

Sample Size	Bootstrap Percentile	Yang CV	CUD-Bound
30	.79	.75	.97
50	.79	.62	.97
100	.78	.46	.96
200	.78	.35	.96

Non-smooth

In general the distribution of

$$\sqrt{N}(\widehat{err(\hat{f}\,)} - err(\hat{f}\,))$$

may not converge as the training set increases (variance never settles down).

Intuition

Consider the large sample variance of

$$\sqrt{N}\left(E_N[1\{YX^T\beta < 0] - E[1\{YX^T\beta < 0]\right)$$

Variance is $p(1-p),\ p = P[YX^T\beta < 0]$

if in place of β we put $\widehat{\beta}$ where $\widehat{\beta}$ is close to 0 then due to the non-smoothness in $p = P(YX^T\beta < 0)$ at $\beta = 0$ we can get jittering.

PIs from Learning Theory

Given a result of the form: for all *N*

$$P\left[\sup_{f\in\mathcal{G}_N}\left|\widehat{err}_{rs}(f)-err(f)\right|<B_{N,\delta}\right]>1-\delta$$

where $\hat{f}$ is known to belong to $\mathcal{G}_N$ and

$\widehat{err}_{rs}(f)=E_N[1\{Y\neq f(X)\}]$

forms a conservative *1*-δ PI:

$$\widehat{err}_{rs}(\hat{f})-B_{N,\delta}<err(\hat{f})<\widehat{err}_{rs}(\hat{f})+B_{N,\delta}$$

Combine statistical ideas with learning theory ideas

Construct a prediction interval for

$$\sup_{f \in \mathcal{G}_N} |\widehat{err}_{rs}(f) - err(f)|$$

where $\mathcal{G}_N$ is chosen to be small yet contain $\hat{f}$

---from this PI deduce a conservative PI for

$$err(\hat{f})$$

---use the surrogate loss to perform estimation and to construct $\mathcal{G}_N$

Construct a prediction interval for

$$\sup_{f \in \mathcal{G}_N} |\widehat{err}_{rs}(f) - err(f)|$$

--- $\mathcal{G}_N$ should contain all f that are close to $\hat{f}$

--- all f for which

$$E_N[L(Y, \tilde{f}(X)) - E_N[L(Y, f(X)] > 0$$

--- $\tilde{f}$ is the "limiting value" of $\hat{f}$;

$$\tilde{f} = \arg\max_{f \in \mathcal{F}} E[L(Y, f(X)]$$

Prediction Interval

Construct a prediction interval for

$$\sup_{f \in \mathcal{F}} \left\{ \begin{array}{l} \left| \widehat{err}_{rs}(f) - err(f) \right| \times \\ \qquad g\left(N(E_N[L(Y, \tilde{f}(X)) - E_N[L(Y, f(X)]) \right) \end{array} \right\}$$

$$g(u) = (1 + u)1\{-1 \leq u \leq 0\} + 1\{u > 0\}$$

Prediction Interval

$$\left|\widehat{err(\hat{f}\,)} - err(\hat{f}\,)\right| \lesssim \left|\widehat{err}_{rs}(\hat{f}\,) - err(\hat{f}\,)\right|$$

$$=$$

$$\left|\widehat{err}_{rs}(\hat{f}\,) - err(\hat{f}\,)\right| \times$$
$$g\left(N(E_N[L(Y,\tilde{f}(X)) - E_N[L(Y,\hat{f}(X)])\right)$$

$$\leq$$

$$\sup_{f\in\mathcal{F}}\left\{\begin{array}{l}|\widehat{err}_{rs}(f) - err(f)| \times \\ \qquad g\left(N(E_N[L(Y,\tilde{f}(X)) - E_N[L(Y,f(X)])\right)\end{array}\right\}$$

Bootstrap

We use bootstrap to obtain an estimate of an upper percentile of the distribution of

$$\sup_{f \in \mathcal{F}} \left\{ \begin{array}{l} (\widehat{err}_{rs}(f) - err(f)) \times \\ \qquad g\left(N(E_N[L(Y, \tilde{f}(X)) - E_N[L(Y, f(X)])\right) \end{array} \right\}$$

to obtain b_U. The PI is then

$$\widehat{err(\hat{f}\,)} - b_L \leq err(\hat{f}\,) \leq \widehat{err(\hat{f}\,)} + b_U$$

Implementation

- Approximation space for the classifier is linear:

$$\mathcal{F} = \{f(x) = sign(x^T\beta) : \beta \in R^p\}$$

- Surrogate loss is least squares:

$$L(y, f(x)) = (y - x^T\beta)^2$$

- $\widehat{err(\hat{f})} = err_{rs}(\hat{f})$ (resubstitution error)

Implementation

$$\sup_{f \in \mathcal{F}} \left\{ \begin{array}{l} (\widehat{err}_{rs}(f) - err(f)) \times \\ \qquad g\left(N(E_N[L(Y, \tilde{f}(X)) - E_N[L(Y, f(X)])\right) \end{array} \right\}$$

becomes

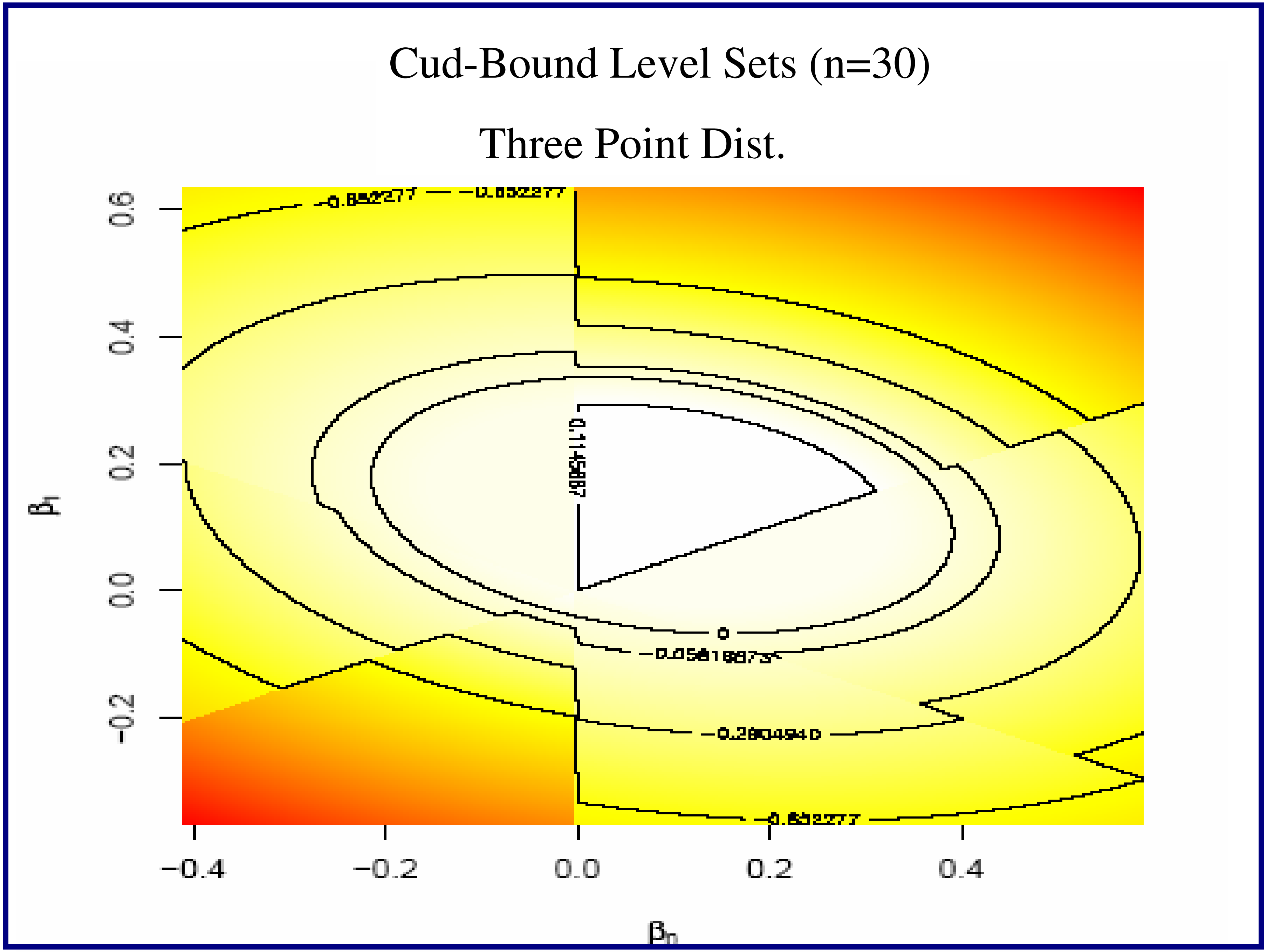
$$\sup_{\beta \in R^q} \left\{ \begin{array}{l} \left(E_N[1\{YX^T\beta < 0] - E[1\{YX^T\beta < 0]\right) \times \\ \qquad g\left(N(E_N[(Y - X^T\tilde{\beta})^2 - E_N[(Y - X^T\beta)^2])\right) \end{array} \right\}$$

Implementation

- Bootstrap version:

$$\sup_{\beta \in R^q} \left\{ \begin{array}{c} \left[E_N^*[1\{YX^T\beta < 0] - E_N[1\{YX^T\beta < 0] \right] \times \\ g\left(N(E_N^*[(Y - X^T\widehat{\beta})^2 - E_N^*[(Y - X^T\beta)^2]) \right) \end{array} \right\}$$

- E_N^* denotes the expectation for the bootstrap distribution


Cud-Bound Level Sets (n=30)
Three Point Dist.
−0.852277
−0.852277
0
−0.852277
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
β_1
β_0

Computational Issues

$$\sup_{\beta \in R^q} \left\{ \begin{array}{r} \left[E_N^*[1\{YX^T\beta < 0] - E_N[1\{YX^T\beta < 0]\right] \times \\ g\left(N(E_N^*[(Y - X^T\widehat{\beta})^2 - E_N^*[(Y - X^T\beta)^2])\right) \end{array} \right\}$$

- Partition R^q into equivalence classes defined by the 2N possible values of the first term.
- Each equivalence class, $\mathcal{M}_i$ can be written as a set of β satisfying linear constraints.
- The first term is constant on $\mathcal{M}_i$

Computational Issues

$$\sup_{\beta \in R^q} \left\{ \begin{array}{l} \left[E_N^*[1\{YX^T\beta < 0\}] - E_N[1\{YX^T\beta < 0\}] \right] \times \\ \qquad g\left(N(E_N^*[(Y - X^T\widehat{\beta})^2 - E_N^*[(Y - X^T\beta)^2]) \right) \end{array} \right\}$$

can be written as

$$\max_i \left\{ \begin{array}{l} C(\mathcal{M}_i) \times \\ \qquad g\left(N(E_N^*[(Y - X^T\widehat{\beta})^2 - \inf_{\beta \in \mathcal{M}_i} E_N^*[(Y - X^T\beta)^2]) \right) \end{array} \right\}$$

since g is non-decreasing.

Computational Issues

- Reduced the problem to the computation of at most 2N mixed integer quadratic programming problems.

- Using commercial solvers (e.g. CPLEX) the CUD bound can be computed for moderately sized data sets in a few minutes on a standard desktop (2.8 GHz processor 2GB RAM).

Comparisons, 95% PI

Data	CUD	BS	M	Y
Magic	1.0	.92	.98	.99
Mamm.	1.0	.68	.43	.98
Ion.	1.0	.61	.76	.99
Donut	1.0	.88	.63	.94
3-Pt	.97	.83	.90	.75
Balance	.95	.91	.61	.99
Liver	1.0	.96	1.0	1.0

Sample size = 30 (1000 data sets)

Comparisons, Length of PI

Data	CUD	BS	M	Y
Magic	.60	.31	.28	.46
Mamm.	.46	.53	.32	.42
Ion.	.42	.43	.30	.50
Donut	.47	.59	.32	.41
3-Pt	.38	.48	.32	.46
Balance	.38	.09	.29	.48
Liver	.62	.37	.33	.49

Sample size=30 (1000 data sets)

Intuition

In large samples

$$\sup_{\beta \in R^q} \left\{ \begin{array}{r} \sqrt{N}\left(E_N[1\{YX^T\beta < 0] - E[1\{YX^T\beta < 0]\right) \times \\ g\left(N(E_N[(Y - X^T\tilde{\beta})^2 - E_N[(Y - X^T\beta)^2])\right) \end{array} \right\}$$

behaves like

$$\sup_{\gamma \in R^q}[X(\gamma)]g\left(Z^T\gamma - \frac{1}{2}\gamma^T\Sigma\gamma\right)$$

$$\gamma = \sqrt{N}(\beta - \tilde{\beta})$$

Intuition

The large sample distribution is the same as the distribution of

$$\sup_{\gamma \in R^q} [X(\gamma)] g\left(Z^T\gamma - \frac{1}{2}\gamma^T \Sigma \gamma\right)$$

where

$$\Sigma = E\left[XX^T\right],\ Z \sim N(0, \sigma^2\Sigma),$$
$$X(\gamma) \sim N(0, p_\gamma(1 - p_\gamma))$$

$$p_\gamma = P[X^T\tilde{\beta} < 0] + P[X^T\tilde{\beta} = 0,\ YX^T\gamma < 0]$$

Intuition

If $P[X^T\tilde{\beta} \neq 0] = 1$

then the distribution is approximately that of a

$$N(0, p(1-p)),\ p = P[X^T\tilde{\beta} < 0]$$

(limiting distribution for binomial, as expected).

Intuition

If $P[X^T\tilde{\beta} = 0] = 1$

the distribution is approximately that of

$$\sup_{\gamma \in \mathcal{G}} N(0, P[YX^T\gamma < 0]P[YX^T\gamma \geq 0])$$

where

$$\mathcal{G} = \{\gamma : (\gamma - \Sigma^{-1}Z)^T\Sigma(\gamma - \Sigma^{-1}Z) \leq B\}$$

$$\sqrt{N}(\hat{\beta} - \tilde{\beta}) = \Sigma_n^{-1}Z_n$$

Discussion

- Further reduce the conservatism of the CUD-bound.
  - Replace $\tilde{\beta}$ by other quantities.
  - Other surrogates (exponential, logit)
- Construct a principle for minimizing the length of the conservative PI?
- The real goal is to produce PIs for the Value of a policy.

The simplest Dynamic treatment regime (e.g. policy) is a decision rule if there is only one stage of treatment

1 Stage for each individual

$$X_1,\ A_1,\ X_2$$

X_j: Observation available at j[th] stage

A_j : Action at j[th] stage (usually a treatment)

Primary Outcome:

$$Y = r(X_1, X_2)$$

Goal:

Construct decision rules that input patient information and output a recommended action; these decision rules should lead to a maximal mean Y.

In future one selects action: $a_1 = d(X_1)$

Single Stage

- Find a confidence interval for the mean outcome if a particular estimated policy (here one decision rule) is employed.
- Treatment A is randomized in {-1,1}.
- Suppose the decision rule is of form

$$\hat{d}(X_1) = sign(\hat{\beta}^T X_1)$$

- *We do not assume the optimal decision boundary is linear.*

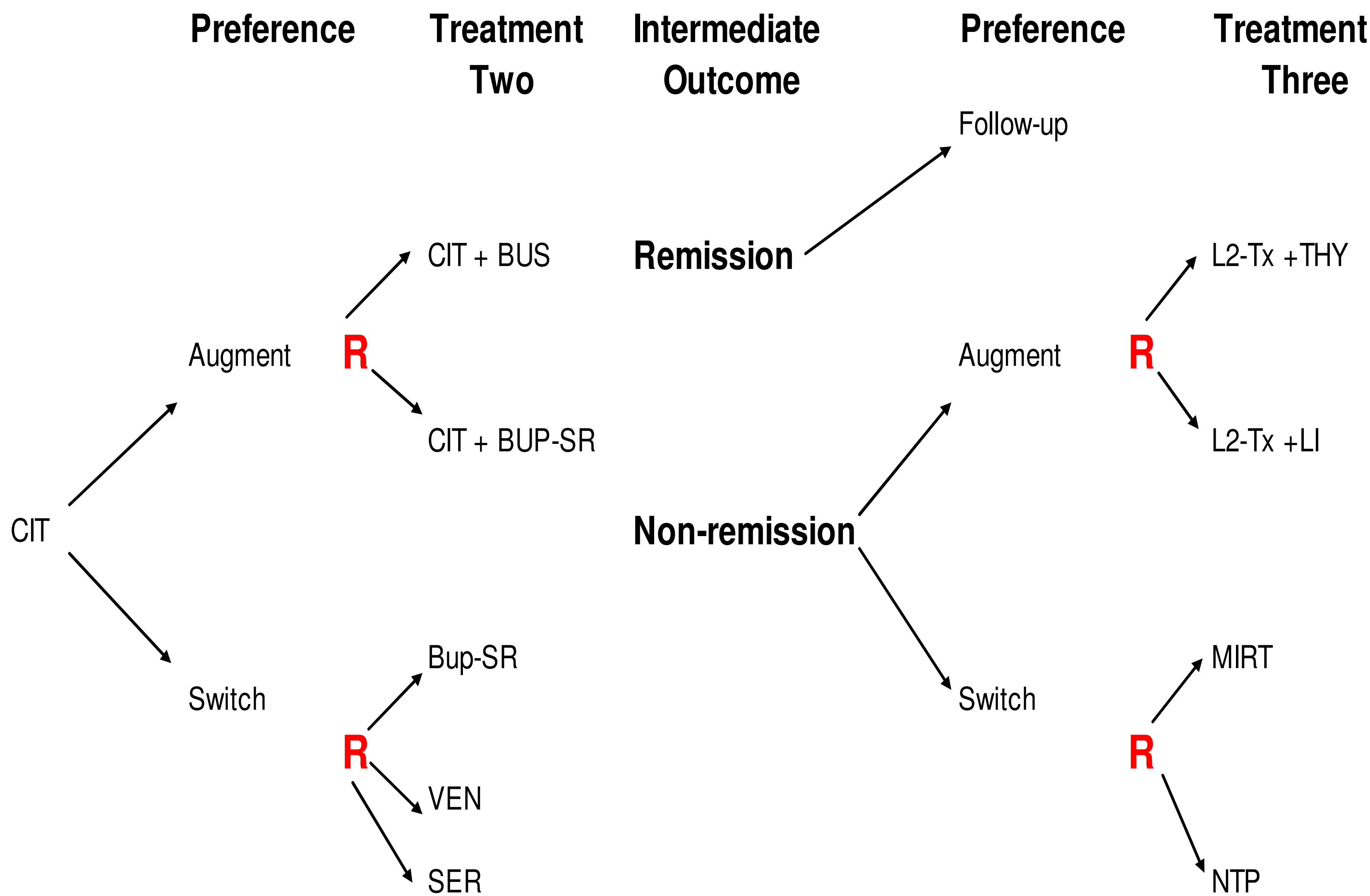
Single Stage

Mean outcome following this policy is $V(\widehat{\beta})$

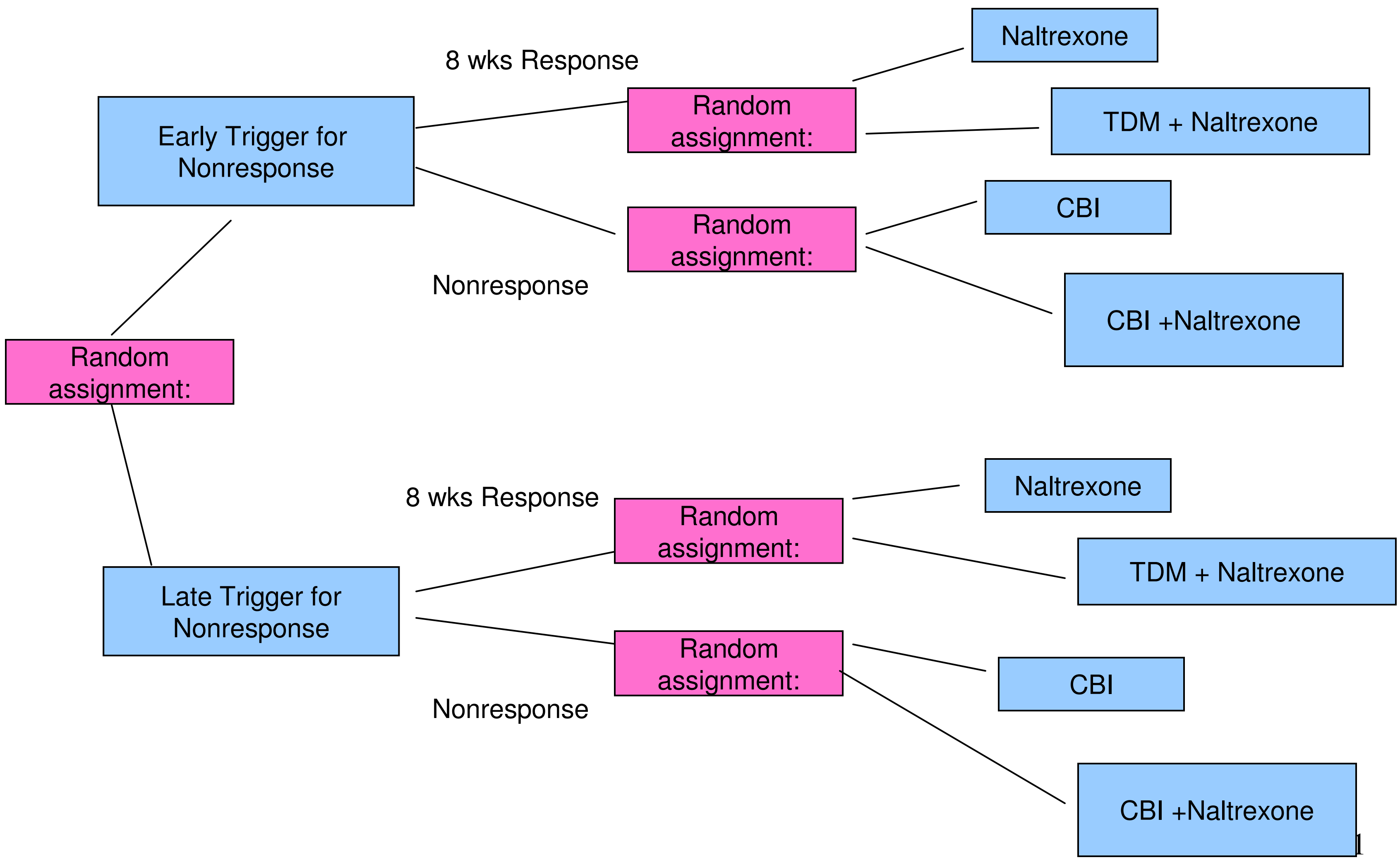
$$
\begin{aligned}
V(\beta) &= E\left[E[Y|X_1, A = sign(X_1^T\beta)]\right] \\
&= E\left[\frac{Y}{p(A|X_1)} I\{AX_1^T\beta > 0\}\right]
\end{aligned}
$$

$p(A_1|X_1)$ is the randomization probability

STAR*D "Sequenced Treatment to Relieve Depression"



Oslin ExTENd



This seminar can be found at:

http://www.stat.lsa.umich.edu/~samurphy/ seminars/UFlorida01.09.09.ppt

Email Eric or me with questions or if you would like a copy of the associated paper:

laber@umich.edu or samurphy@umich.edu

Bias of Common $\widehat{err(\hat{f})}$ on Three Point Example