Adaptive inference for non-regular functionals

Eric B. Laber
Department of Statistics
Institute for Social Research
University of Michigan, Ann Arbor

UNC, January 16, 2011
Introduction

- Modern statistical analysis is rife with non-regularity
  1. Test error of a learned classifier
  2. Parameters in an estimated optimal treatment policy
  3. Inference after model selection
  4. ...

- Ignoring or assuming away this non-regularity can lead to poor small sample performance under many realistic generative models

- An asymptotic framework that faithfully represents small sample behavior is needed for the development and evaluation of inferential procedures
Roadmap

1. Test error in classification
2. Inference for estimated optimal dynamic treatment policies
3. The Lasso and a general framework for adaptive inference
Inference for the test error in classification
Setup

Simple beginnings:

1. Observe iid training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$
   - inputs $X \in \mathbb{R}^p$
   - outputs $Y \in \{-1, 1\}$

2. Construct classifier $\hat{c}_\mathcal{D}(X) : \mathbb{R}^p \mapsto \{-1, 1\}$

3. Use classifier for prediction at new inputs

Questions:

- How well will my classifier perform its task?
  - Point estimate: for test error $\tau(\hat{c}_\mathcal{D}) \equiv P_{Y \neq \hat{c}_\mathcal{D}(X)}$

- How confident am I in the above estimate?
  - Interval estimate: for test error $\tau(\hat{c}_\mathcal{D})$
Simple beginnings:

1. Observe iid training data \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \)
   - inputs \( X \in \mathbb{R}^p \)
   - outputs \( Y \in \{-1, 1\} \)

2. Construct classifier \( \hat{c}_{\mathcal{D}}(X) : \mathbb{R}^p \rightarrow \{-1, 1\} \)

3. Use classifier for prediction at new inputs

Questions:

- How well will my classifier perform its task?
  
  **Point estimate:** for test error \( \tau(\hat{c}_{\mathcal{D}}) \triangleq P1_{Y \neq \hat{c}_{\mathcal{D}}(X)} \)

- How confident am I in the above estimate?
  
  **Interval estimate:** for test error \( \tau(\hat{c}_{\mathcal{D}}) \)
Background

- Long standing problem
- Primary focus has been point estimation
  - CV methods: Krzanowski and Hand [1986], Langford [2005], Yang [2006]
  - Hybrid methods: Fu [2005], Kim [2009]
  - More than 200 references!
Historically interval estimation of secondary interest

- **Point estimation paradigm:** Efron [1983], Efron & Tibshirani [1997], Yang [2002]
  - **Step 1:** Develop best point estimator
  - **Step 2:** Estimate standard error

- **Assume regularity:** Polenik [1995], Mamman & Tsybakov [1999], Tsybakov [2004], Song et al. [2010]
  - Assume model space is correct (e.g. Bayes classifier is linear)
  - Assume a margin condition (e.g. require rates on $P(|X^T \beta^*| \leq \epsilon)$)
Why is this problem still open?

- Non-regular
- Previous methods either
  - ad-hoc
  - assume regularity
- Point estimation paradigm
The problem

- Construct a classifier using surrogate loss $L(X, Y, \beta)$
  1. $\hat{\beta}_n \triangleq \arg \min_{\beta \in \mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$
  2. $\hat{c}_D(X) = \text{sign} \left( X^T \hat{\beta}_n \right)$
The problem

- Construct a classifier using surrogate loss $L(X, Y, \beta)$
  1. $\hat{\beta}_n \triangleq \arg\min_{\beta \in \mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$
  2. $\hat{c}_D(X) = \text{sign}(X^T \hat{\beta}_n)$

- Review: surrogate loss function $L(X, Y, \beta)$
  - like to minimize error rate $\mathbb{P}_n 1_{Y \neq \text{sign}(X^T \beta)}$
  - non-smoothness $\Rightarrow$ computational difficulty
  - replace $1_{Y \neq \text{sign}(X^T \beta)} = 1_{YX^T \beta < 0}$ with smooth surrogate

  - Support Vector Machines:
    $L(X, Y, \beta) = (1 - YX^T \beta)_+ + \gamma \|\beta\|^2$
  - Binomial Deviance:
    $L(X, Y, \beta) = \log(1 + e^{-YX^T \beta})$
  - Squared Error:
    $L(X, Y, \beta) = (1 - YX^T \beta)^2$
Test error \( \tau(\hat{\beta}_n) \triangleq P_{Y\mid X}\tau \hat{\beta}_n < 0 \)

The problem cont’d
The problem cont’d

- Test error $\tau(\hat{\beta}_n) \triangleq P_{Y|X} \{\hat{\beta}_n < 0\}$
- **Goal:** given $\alpha \in (0, 1)$ construct $\hat{l}$ and $\hat{u}$ so that
  
  \[
P_D \{\hat{l} \leq \tau(\hat{\beta}_n) \leq \hat{u}\} \geq 1 - \alpha
  \]
The problem cont’d

- Test error $\tau(\hat{\beta}_n) \triangleq P_{1_{Y^X\tau\hat{\beta}_n<0}}$
- Goal: given $\alpha \in (0, 1)$ construct $\hat{u}$ and $\hat{l}$ so that
  \[
P_D \left\{ \hat{l} \leq \tau(\hat{\beta}_n) \leq \hat{u} \right\} \geq 1 - \alpha
  \]

Context
- Model space may not be correct
- $p << n$
- Cannot afford a test set
Non-regularity

- Simple estimate of $\tau(\hat{\beta}_n)$ is $\hat{\tau}(\hat{\beta}_n) \triangleq \mathbb{P}_n 1_{YX^T \hat{\beta}_n < 0}$
- Natural starting point for inference:

\[
\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \triangleq \sqrt{n}(\mathbb{P}_n - P) 1_{YX^T \hat{\beta}_n < 0} = \sqrt{n}(\mathbb{P}_n - P) 1_{X^T \beta^* = 0} 1_{YX^T \sqrt{n}(\hat{\beta}_n - \beta^*) < 0} + \sqrt{n}(\mathbb{P}_n - P) 1_{X^T \beta^* \neq 0} 1_{YX^T \hat{\beta}_n < 0}
\]

- $P1_{X^T \beta^* = 0} > 0$ implies $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$ has non-regular limit
  - points near the boundary cause jittering
  - $P1_{YX^T \hat{\beta}_n < 0}$ need not concentrate about its mean
  - bootstrap and normal approximations are invalid
Simple example

Suppose

- \((X_1, X_2) \sim \text{Unif}[0, 5]^2\)
- \(\epsilon \sim \mathcal{N}(0, 1/4)\)
- \(Y = \text{sign}(X_2 - (4/25)X_1^2 - 1 + \epsilon)\)

Properties of this example

- \(P_1 x_\top \beta^* = 0 = 0\)
- Linear classifier is a good fit
- E.g. if \(n = 30\)
  - \(\mathbb{E}(\tau(\hat{\beta}_n)) \approx .11\)
  - Bayes error \(\approx .09\)
Simple example cont’d

Under “regular” framework

- Centered bootstrap $\sqrt{n}(\hat{P}_n^{(b)} - P_n)1_{YX^T} \hat{\beta}_n^{(b)} < 0$

- Normal approximation $\hat{\tau} (\hat{\beta}_n) \pm z_{1-\gamma/2} \sqrt{\frac{\hat{\tau} (\hat{\beta}_n)(1-\hat{\tau} (\hat{\beta}_n))}{n}}$

are both asymptotically valid
Simple example cont’d

Under “regular” framework

- Centered bootstrap $\sqrt{n}(\hat{P}_n^{(b)} - P_n)1_{Y^TX^T\hat{\beta}^{(b)}_n < 0}$
- Normal approximation $\hat{\tau}(\hat{\beta}_n) \pm z_{1-\gamma/2} \sqrt{\hat{\tau}(\hat{\beta}_n)(1-\hat{\tau}(\hat{\beta}_n)) / n}$

are both asymptotically valid

Coverage estimated using 1000 Monte Carlo iterations

- Below nominal coverage even for $n = 250$
- Coverage especially poor for small samples
Simple example cont’d

Why do these methods fail?
Simple example cont’d

Why do these methods fail?

- Non-smoothness $\Rightarrow$ non-regularity
- Performance inversely proportional to smoothness

Continuing our example

$\tau(\hat{\beta}_n)$ consider $\tau_{\text{smooth}}(\hat{\beta}_n) \equiv P\left(1 + \exp(aYX^\top\hat{\beta}_n)\right) - 1$

$\tau_{\text{smooth}}(\hat{\beta}_n)$ is smooth for fixed $a > 0$

If $a \to \infty$ then $\tau_{\text{smooth}}(\hat{\beta}_n) \to \tau(\hat{\beta}_n)$

Conjecture: Bootstrap coverage should deteriorate as $a$ grows
Simple example cont’d

Why do these methods fail?
▷ Non-smoothness $\Rightarrow$ non-regularity
▷ Performance inversely proportional to smoothness

Continuing our example
▷ Instead of test error $\tau(\hat{\beta}_n)$ consider

$$\tau_{\text{smooth}}(\hat{\beta}_n) \triangleq P \left( 1 + \exp(aYX^T\hat{\beta}_n) \right)^{-1}$$

▷ $\tau_{\text{smooth}}(\hat{\beta}_n)$ is smooth for fixed $a > 0$
▷ If $a \to \infty$ then $\tau_{\text{smooth}}(\hat{\beta}_n) \to \tau(\hat{\beta}_n)$
Simple example cont’d

Why do these methods fail?
▶ Non-smoothness ⇒ non-regularity
▶ Performance inversely proportional to smoothness

Continuing our example
▶ Instead of test error \( \tau(\hat{\beta}_n) \) consider

\[
\tau_{\text{smooth}}(\hat{\beta}_n) \triangleq P \left( 1 + \exp(aYX^T\hat{\beta}_n) \right)^{-1}
\]

▶ \( \tau_{\text{smooth}}(\hat{\beta}_n) \) is smooth for fixed \( a > 0 \)
▶ If \( a \to \infty \) then \( \tau_{\text{smooth}}(\hat{\beta}_n) \to \tau(\hat{\beta}_n) \)
▶ Conjecture: Bootstrap coverage should deteriorate as \( a \) grows
Simple example cont’d

Smoothed Loss Functions

Estimated Coverage Quadratic Example: Smoothed
Simple example cont’d

Smoothed Loss Functions

Estimated Coverage Quadratic Example: Smoothed
An approach

- Confidence interval the primary focus
- Asymptotic framework that permits non-regularity
- Construct data-driven upper and lower bounds on the test error
  - Bounds are smooth functionals of the data
  - If generative model induces regularity, then bounds collapse to the test error
- Confidence intervals are formed by bootstrapping these bounds
Adaptive confidence interval

Basic idea: data adaptive bound on $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$

- Recall $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$ is equal to

$$\sqrt{n}(\mathbb{P}_n - P)1_{X^T\beta^*=0}1_{YX^T\hat{\beta}_n<0} + \sqrt{n}(\mathbb{P}_n - P)1_{X^T\beta^*\neq0}1_{YX^T\hat{\beta}_n<0}$$

- Take supremum/infimum only when $X$ is in a region near the decision boundary $X^T\beta^* = 0$

$$\mathbb{C}_n \triangleq \sup_{u \in \mathbb{R}^p} \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^T\hat{\beta}_n)^2}{X^T\Sigma X} \leq \lambda_n}1_{YX^T u < 0}$$

$$+ \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^T\hat{\beta}_n)^2}{X^T\Sigma X} > \lambda_n}1_{YX^T\hat{\beta}_n<0}$$

where $\lambda_n \to \infty$, $\lambda_n, = o(n)$ and $\Sigma = n\text{Cov}(\hat{\beta}_n)$
Adaptive confidence interval

Basic idea: data adaptive bound on $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$

- Recall $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$ is equal to

  $$\sqrt{n}(\mathbb{P}_n - P)1_{X^T \hat{\beta}^* = 0} 1_{YX^T \hat{\beta}_n < 0} + \sqrt{n}(\mathbb{P}_n - P)1_{X^T \hat{\beta}^* \neq 0} 1_{YX^T \hat{\beta}_n < 0}$$

- Take supremum/infimum only when $X$ is in a region near the decision boundary $X^T \hat{\beta}^* = 0$

$$\mathbb{C}_n \triangleq \sup_{u \in \mathbb{R}^p} \sqrt{n}(\mathbb{P}_n - P)1_{n(\frac{X^T \hat{\beta}_n)^2}{X^T \Sigma X} \leq \lambda_n} 1_{YX^T u < 0}$$

  $$\quad + \sqrt{n}(\mathbb{P}_n - P)1_{n(\frac{X^T \hat{\beta}_n)^2}{X^T \Sigma X} > \lambda_n} 1_{YX^T \hat{\beta}_n < 0}$$

where $\lambda_n \to \infty$, $\lambda_n = o(n)$ and $\Sigma = n\text{Cov}({\hat{\beta}_n})$
Assumptions

Some technical assumptions:

(A1) $L(X, Y, \beta)$ is convex with respect to $\beta$ for each $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$

(A2) $Q(\beta) \triangleq PL(X, Y, \beta)$ exists and is finite for all $\beta \in \mathbb{R}^p$

(A3) $\beta^* \triangleq \arg \min_{\beta \in \mathbb{R}^p} Q(\beta)$ exists and is unique

(A4) Let $g(X, Y, \beta)$ be a sub-gradient of $L(X, Y, \beta)$. Then $P\|g(X, Y, \beta)\|^2 < \infty$ for all $\beta$ in a neighborhood of $\beta^*$.

(A5) $Q(\beta)$ is twice continuously differentiable at $\beta^*$ and $H \triangleq \nabla^2 Q(\beta^*)$ is positive definite.
Adaptive prediction interval

Theorem (Convergence)

If (A1)-(A5) hold:

1. $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \rightsquigarrow \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*)$
2. $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \leq C_n$ for all $n$
3. $C_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$.

where $\mathbb{V}$ and $\mathbb{B}$ are Gaussian processes and $z_\infty$ is a $p$-dim normal with covariance $\Sigma$. 

Theorem (Adaptation)

Assuming (A1)-(A5) hold then if either the Bayes decision boundary is linear or $P(X^\top \beta^* = 0) = 0$ then $C_n$ and $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n))$ have the same limiting distribution.
Adaptive prediction interval

Theorem (Convergence)

If (A1)-(A5) hold:

1. \( \sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \Rightarrow \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*) \)
2. \( \sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \leq C_n \) for all \( n \)
3. \( C_n \Rightarrow \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*) \).

where \( \mathbb{V} \) and \( \mathbb{B} \) are Gaussian processes and \( z_\infty \) is a \( p \)-dim normal with covariance \( \Sigma \).

Theorem (Adaptation)

Assuming (A1)-(A5) hold then if either the Bayes decision boundary is linear or \( P(X^T \beta^* = 0) = 0 \) then \( C_n \) and \( \sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \) have the same limiting distribution.
Adaptive confidence interval

- Another asymptotic framework for studying non-regularity is local alternatives
- For each $n$ assume training data $\mathcal{D} = \{(X_{n,i}, Y_{n,i})\}_{i=1}^{n}$ drawn iid from $P_n$
  - (B1) $P_n$ is a sequence of contiguous alternatives to $P$ in the sense of van der Vaart and Wellner (1996)
  - (B2) $\beta^*_n \triangleq \arg \min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta)$ satisfies $\beta^*_n = \beta^* + u/\sqrt{n}$.
Adaptive confidence interval

- Another asymptotic framework for studying non-regularity is local alternatives
- For each $n$ assume training data $D = \{(X_{n,i}, Y_{n,i})\}_{i=1}^n$ drawn iid from $P_n$
  - (B1) $P_n$ is a sequence of contiguous alternatives to $P$ in the sense of van der Vaart and Wellner (1996)
  - (B2) $\beta^*_n \triangleq \arg \min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta)$ satisfies $\beta^*_n = \beta^* + u/\sqrt{n}$.

Theorem (Convergence under local alternatives)

If (A1)-(A5) and (B1)-(B2) hold, then under $P_n$
1. $\sqrt{n}(\hat{\tau}(\hat{\beta}_n) - \tau(\hat{\beta}_n)) \rightsquigarrow \mathbb{V}(z_\infty + u) + \mathbb{B}(\beta^*)$
2. $C_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$. 
Adaptive confidence interval

Recall

\[ C_n \triangleq \sup_{u \in \mathbb{R}^p} \sqrt{n} (\mathbb{P}_n - P) 1 \frac{n(x^T \hat{\beta}_n)^2}{x^T \Sigma x} \leq \lambda_n 1_{Yx^T u < 0} + \sqrt{n} (\mathbb{P}_n - P) 1 \frac{n(x^T \hat{\beta}_n)^2}{x^T \Sigma x} > \lambda_n 1_{Yx^T \hat{\beta}_n < 0} \]

Bootstrap analog

\[ \hat{C}_n^{(b)} \triangleq \sup_{u \in \mathbb{R}^p} \sqrt{n} (\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n) 1 \frac{n(x^T \hat{\beta}_n^{(b)})^2}{x^T \hat{\Sigma}_n x} \leq \lambda_n 1_{Yx^T u < 0} + \sqrt{n} (\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n) 1 \frac{n(x^T \hat{\beta}_n^{(b)})^2}{x^T \hat{\Sigma}_n x} > \lambda_n 1_{Yx^T \hat{\beta}_n^{(b)} < 0} \]
Adaptive prediction interval

Let $\hat{u}$ be the $1 - \alpha/2$ of percentile $\hat{C}_n^{(b)}$ then:

$$P \left\{ \sqrt{n}(\hat{P}_n - P)1_{YX^T \hat{\beta}_n < 0} \leq \hat{u} \right\} \geq P \left\{ C_n \leq \hat{u} \right\} \approx P \left\{ \hat{C}_n^{(b)} \leq \hat{u} \right\} = 1 - \alpha/2$$

so that

$$P \left\{ \hat{\mathbb{P}}_n 1_{YX^T \hat{\beta}_n < 0} - \hat{u}/\sqrt{n} \leq P 1_{YX^T \hat{\beta}_n < 0} \right\} = P \left\{ \hat{\tau}(\hat{\beta}_n) - \hat{u}/\sqrt{n} \leq \tau(\hat{\beta}_n) \right\} \geq 1 - \delta/2$$

and $[\hat{\tau}(\hat{\beta}) - \hat{u}/\sqrt{n}, 1]$ is an approximate confidence interval.
Adaptive prediction interval

Theorem
Suppose that (A1)-(A5) hold then \( C_n \) and \( \hat{C}_n^{(b)} \) converge to the same limiting distribution in probability.
Computation

\[ \hat{C}_n(b) \triangleq \sup_{u \in \mathbb{R}^p} \sqrt{n}(\hat{P}_n(b) - P_n) \mathbf{1}_{n(\mathbf{X}^\top \hat{\beta}_n(b))^2 \leq \lambda_n \mathbf{1}_{\mathbf{Y}^\top \mathbf{X}} u < 0} + \sqrt{n}(\hat{P}_n(b) - P_n) \mathbf{1}_{n(\mathbf{X}^\top \hat{\beta}_n(b))^2 > \lambda_n \mathbf{1}_{\mathbf{Y}^\top \hat{\beta}_n(b)} < 0} \]

- Computing \( \hat{C}_n(b) \) is a Mixed Integer Programming problem
  - requires specialized software (e.g. CPLEX)
  - computational cost excessive for large problems
- Admits convex relaxation
  - solved in polynomial time
  - negligible loss in solution quality
  - no specialized software required (can be solved in R)
Computation

Let \((m_{n1}, m_{n2}, \ldots, m_{nn})\) denote a single instance of bootstrap multinomial weights.

Computing the ACI requires computing

\[
\inf_{u \in \mathbb{R}^p} \sum_{i \in N_n^{(b)}} (m_{ni} - 1)1_{y_i x_i^T u < 0}
\]

where \(N_n^{(b)} \triangleq \left\{ i : \frac{(x_i^T \hat{\beta}_n^{(b)})^2}{x_i^T \Sigma x_i} \leq \lambda_n \right\}\)

Note that this is a weighted classification problem.
Computation: convex relaxation

- Write
  \[
  \sum_{i \in N_n^{(b)}} (m_{ni} - 1)1_{y_i x_i^T u < 0} = \sum_{i \in N_n^{(b)}} m_{ni}1_{y_i x_i^T u < 0} + \sum_{i \in N_n^{(b)}} (-1)_{y_i x_i^T u < 0}
  \]

- Replace \(1_{z<0}\) and \(-1_{z<0}\) with convex upper bounds
Computation: convex relaxation

- Write

\[
\sum_{i \in N_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^T u < 0} = \sum_{i \in N_n^{(b)}} m_{ni} 1_{y_i x_i^T u < 0} + \sum_{i \in N_n^{(b)}} (-1) 1_{y_i x_i^T u < 0}
\]

- Replace \( 1_{z < 0} \) and \(-1_{z < 0}\) with convex upper bounds
Computation: convex relaxation

Write

\[ \sum_{i \in \mathcal{N}_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^T u < 0} = \sum_{i \in \mathcal{N}_n^{(b)}} m_{ni} 1_{y_i x_i^T u < 0} + \sum_{i \in \mathcal{N}_n^{(b)}} (-1) 1_{y_i x_i^T u < 0} \]

Replace \(1_{z < 0}\) and \((-1)_{z < 0}\) with convex upper bounds.

**Surrogate Loss Function Piece One**

**Surrogate Loss Function Piece Two**
### Computation: convex relaxation

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Coverage Relaxed</th>
<th>MIP</th>
<th>(p_{.99})</th>
<th>(p_{.95})</th>
<th>(p_{.75})</th>
<th>(p_{.5})</th>
<th>Computation Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreePt</td>
<td>.948</td>
<td>.948</td>
<td>.0334</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>4.23×</td>
</tr>
<tr>
<td>Magic</td>
<td>.944</td>
<td>.945</td>
<td>.0334</td>
<td>.0334</td>
<td>0.00</td>
<td>0.00</td>
<td>1.56×</td>
</tr>
<tr>
<td>Mam.</td>
<td>.957</td>
<td>.958</td>
<td>.0334</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.07×</td>
</tr>
<tr>
<td>Ion.</td>
<td>.954</td>
<td>.954</td>
<td>.0334</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.30×</td>
</tr>
<tr>
<td>Donut.</td>
<td>.967</td>
<td>.968</td>
<td>.0667</td>
<td>.0334</td>
<td>.0334</td>
<td>0.00</td>
<td>3.21×</td>
</tr>
<tr>
<td>Bal.</td>
<td>.969</td>
<td>.969</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.72×</td>
</tr>
<tr>
<td>Liver</td>
<td>.956</td>
<td>.956</td>
<td>.0333</td>
<td>.0333</td>
<td>0.00</td>
<td>0.00</td>
<td>1.55×</td>
</tr>
<tr>
<td>Spam</td>
<td>.984</td>
<td>.987</td>
<td>.0333</td>
<td>.0333</td>
<td>0.00</td>
<td>0.00</td>
<td>1.95×</td>
</tr>
<tr>
<td>Quad</td>
<td>.959</td>
<td>.962</td>
<td>.0333</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.38×</td>
</tr>
<tr>
<td>Heart</td>
<td>.960</td>
<td>.961</td>
<td>.0333</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>3.08×</td>
</tr>
</tbody>
</table>

- Notes: sample size \(n = 30\), 1000 Monte Carlo iterations, 1000 bootstrap resamples
Computation: data-augmentation black box

- Write

\[
\sum_{i \in N_n^{(b)}} (m_{ni} - 1) \mathbf{1}_{y_i x_i^T u < 0} = c + \sum_{i \in N_n^{(b)}} m_{ni} \mathbf{1}_{y_i x_i^T u < 0} + \sum_{i \in N_n^{(b)}} \mathbf{1}_{-y_i x_i^T u < 0}
\]

- Create augmented data set \( \{ (\tilde{y}_i, \tilde{x}_i) \}_{i=1}^M \) with

\( M = \# \{ i \in N_n^{(b)} : m_{ni} \neq 1 \} \)

1. \( m_{ni} - 1 \) copies of \((y_i, x_i)\) if \( m_{ni} > 1 \)
2. One copy of \(( -y_i, x_i)\) if \( m_{ni} = 0 \)
Computation: data-augmentation black box

- Write

\[
\sum_{i \in \mathcal{N}_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^T u < 0} = c + \sum_{i \in \mathcal{N}_n^{(b)}} m_{ni} 1_{y_i x_i^T u < 0} + \sum_{i \in \mathcal{N}_n^{(b)}} 1_{-y_i x_i^T u < 0}
\]

- Create augmented data set \( \{(\tilde{y}_i, \tilde{x}_i)\}_{i=1}^M \) with

  - \( M = \#\{i \in \mathcal{N}_n^{(b)} : m_{ni} \neq 1\} \)
  1. \( m_{ni} - 1 \) copies of \((y_i, x_i)\) if \( m_{ni} > 1 \)
  2. One copy of \((-y_i, x_i)\) if \( m_{ni} = 0 \)

- Approximate

\[
\inf_{u \in \mathbb{R}^p} \sum_{i \in \mathcal{N}_n^{(b)}} (m_{ni} - 1) 1_{y_i x_i^T u < 0}
\]

with plug-in estimate of test error using a classifier fit to augmented data \( \{(\tilde{y}_i, \tilde{x}_i)\}_{i=1}^M \)
Experiments

Compare performance of
- Adaptive confidence interval (ACI)
- CV-Normal approximation [Yang 2006]
- BCCVP-BR approximation [Jiang 2008]
Experiments

Compare performance of
- Adaptive confidence interval (ACI)
- CV-Normal approximation [Yang 2006]
- BCCVP-BR approximation [Jiang 2008]

Details
- 1000 Monte Carlo iterations
- 10 data sets
- Compare estimated coverage and width
- \( \lambda_n \triangleq \max(\sqrt{n}, \chi^2_{.995}) \)
Results

Target coverage .950, loss function $L(X, Y, \beta) = (1 - YX^T\beta)^2$, $n = 30$

<table>
<thead>
<tr>
<th>Data Set/Method</th>
<th>ACI</th>
<th>CV-Normal</th>
<th>BCCVP-BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreePt</td>
<td>.948</td>
<td>.930</td>
<td>.863</td>
</tr>
<tr>
<td>Magic</td>
<td>.944</td>
<td>.996</td>
<td>.979</td>
</tr>
<tr>
<td>Mam.</td>
<td>.957</td>
<td>.989</td>
<td>.966</td>
</tr>
<tr>
<td>Ion.</td>
<td>.941</td>
<td>.989</td>
<td>.972</td>
</tr>
<tr>
<td>Donut</td>
<td>.965</td>
<td>.967</td>
<td>.908</td>
</tr>
<tr>
<td>Bal.</td>
<td>.976</td>
<td>.989</td>
<td>.966</td>
</tr>
<tr>
<td>Liver</td>
<td>.956</td>
<td>.997</td>
<td>.970</td>
</tr>
<tr>
<td>Spam</td>
<td>.984</td>
<td>.998</td>
<td>.975</td>
</tr>
<tr>
<td>Quad</td>
<td>.959</td>
<td>.983</td>
<td>.945</td>
</tr>
<tr>
<td>Heart</td>
<td>.960</td>
<td>.995</td>
<td>.976</td>
</tr>
</tbody>
</table>

Table: Estimated coverage of competing confidence procedures. Coverage is highlighted if not different from .950 at the .01 level.
Results

Target coverage .950, loss function $L(X, Y, \beta) = (1 - YX^T\beta)^2$, $n = 30$

<table>
<thead>
<tr>
<th>Data Set/Method</th>
<th>ACI</th>
<th>CV-Normal</th>
<th>BCCVP-BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreePt</td>
<td>.385</td>
<td>.548</td>
<td>.720</td>
</tr>
<tr>
<td>Magic</td>
<td>.498</td>
<td>.548</td>
<td>.500</td>
</tr>
<tr>
<td>Mam.</td>
<td>.374</td>
<td>.456</td>
<td>.384</td>
</tr>
<tr>
<td>Ion.</td>
<td>.313</td>
<td>.466</td>
<td>.388</td>
</tr>
<tr>
<td>Donut</td>
<td>.424</td>
<td>.483</td>
<td>.485</td>
</tr>
<tr>
<td>Bal.</td>
<td>.217</td>
<td>.350</td>
<td>.232</td>
</tr>
<tr>
<td>Liver</td>
<td>.534</td>
<td>.527</td>
<td>.500</td>
</tr>
<tr>
<td>Spam</td>
<td>.428</td>
<td>.496</td>
<td>.418</td>
</tr>
<tr>
<td>Quad</td>
<td>.246</td>
<td>.360</td>
<td>.267</td>
</tr>
<tr>
<td>Heart</td>
<td>.367</td>
<td>.476</td>
<td>.404</td>
</tr>
</tbody>
</table>

Table: Estimated width of competing confidence procedures. Width is highlighted if coverage is at least .950 and the interval is smallest.
Results

Target coverage .950, loss function $L(X, Y, \beta) = \log(1 + e^{-Y^T X \beta})$, $n = 30$

<table>
<thead>
<tr>
<th>Data Set/Method</th>
<th>ACI</th>
<th>CV-Normal</th>
<th>BCCVP-BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreePt</td>
<td>.976</td>
<td>.893</td>
<td>.914</td>
</tr>
<tr>
<td>Magic</td>
<td>.955</td>
<td>.999</td>
<td>.983</td>
</tr>
<tr>
<td>Mam.</td>
<td>.951</td>
<td>.993</td>
<td>.974</td>
</tr>
<tr>
<td>Ion.</td>
<td>.947</td>
<td>.995</td>
<td>.985</td>
</tr>
<tr>
<td>Donut</td>
<td>.968</td>
<td>.966</td>
<td>.908</td>
</tr>
<tr>
<td>Bal.</td>
<td>.979</td>
<td>.996</td>
<td>.972</td>
</tr>
<tr>
<td>Liver</td>
<td>.946</td>
<td>.997</td>
<td>.972</td>
</tr>
<tr>
<td>Spam</td>
<td>.985</td>
<td>.999</td>
<td>.981</td>
</tr>
<tr>
<td>Quad</td>
<td>.978</td>
<td>.997</td>
<td>.945</td>
</tr>
<tr>
<td>Heart</td>
<td>.960</td>
<td>.995</td>
<td>.976</td>
</tr>
</tbody>
</table>

Table: Estimated coverage of competing confidence procedures. Coverage is highlighted if not different from .950 at the .01 level.
Results

Target coverage .950, loss function \( L(X, Y, \beta) = (1 - YX^\top \beta)^2 \), \( n = 30 \)

<table>
<thead>
<tr>
<th>Data Set/Method</th>
<th>ACI</th>
<th>CV-Normal</th>
<th>BCCVP-BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThreePt</td>
<td>.374</td>
<td>.551</td>
<td>.742</td>
</tr>
<tr>
<td>Magic</td>
<td>.466</td>
<td>.526</td>
<td>.504</td>
</tr>
<tr>
<td>Mam.</td>
<td>.373</td>
<td>.448</td>
<td>.387</td>
</tr>
<tr>
<td>Ion.</td>
<td>.305</td>
<td>.459</td>
<td>.401</td>
</tr>
<tr>
<td>Donut</td>
<td>.434</td>
<td>.485</td>
<td>.494</td>
</tr>
<tr>
<td>Bal.</td>
<td>.262</td>
<td>.349</td>
<td>.257</td>
</tr>
<tr>
<td>Liver</td>
<td>.533</td>
<td>.526</td>
<td>.518</td>
</tr>
<tr>
<td>Spam</td>
<td>.454</td>
<td>.494</td>
<td>.423</td>
</tr>
<tr>
<td>Quad</td>
<td>.310</td>
<td>.372</td>
<td>.267</td>
</tr>
<tr>
<td>Heart</td>
<td>.367</td>
<td>.476</td>
<td>.404</td>
</tr>
</tbody>
</table>

Table: Estimated width of competing confidence procedures. Width is highlighted if coverage is at least .950 and the interval is smallest.
Conclusions for test error

ACI

- Provides nominal coverage
- Non-trivial width
- Consistent under non-regular setting
- Computationally efficient

Notes

- Consistent under local alternatives
- \( p \gg n \) is an important extension
- Robust to the choice of \( \lambda \)
Conclusions for test error

ACI

- Provides nominal coverage
- Non-trivial width
- Consistent under non-regular setting
- Computationally efficient

Notes

- Consistent under local alternatives
- $p \gg n$ important extension
- Robust to the choice $\lambda_n$
Inference for estimated optimal dynamic treatment policies
Dynamic treatment policies in mental health

- Motivation: treatment of chronic illness
  - Some examples: HIV/AIDS, cancer, depression, schizophrenia, drug and alcohol addiction, ADHD, etc.
  - Treatment must adapt to the evolving health status of each patient
- Multistage decision making problem
- Long-term treatment requires long-term evaluation of treatment regimes (e.g. cumulative not myopic evaluation)

- Dynamic treatment regimes (DTRs)
  - Operationalize multistage decision making via as sequence of decision rules
    - One decision rule for each time (decision) point
    - A decision rule is a function inputs patient history and outputs a recommended treatment
  - Aim to optimize some cumulative clinical outcome
Beyond DTRs

- Dynamic treatment policies have applications beyond informing clinical practice
  1. Robotics (autonomous helicopter, drones, etc. Ng 2003)
  2. Marketing (Simester, Sun and Tsitsiklis, 2003)
  3. Active labor market policies (Lechner and Smith 2003)
  4. ...
Inference in DTRs

- Basic inferential tasks bursting with non-regularity
  - Non-regularity increases with number of stages and number of treatments
  - Inference must be done in a data impoverished setting (clinical trials ⇒ expensive data, massive missingness, etc.)
  - In practice (e.g. under realistic generative models) standard methods like the bootstrap can lead to poor small sample performance

In Laber et al. (2010) we offer confidence intervals for various functionals of interest in DTRs

- Data-driven upper and lower bounds
- Consistent under fixed and local alternatives
- Good empirical performance
Inference in DTRs

- Basic inferential tasks bursting with non-regularity
  - Non-regularity increases with number of stages and number of treatments
  - Inference must be done in a data impoverished setting (clinical trials $\Rightarrow$ expensive data, massive missingness, etc.)
  - In practice (e.g. under realistic generative models) standard methods like the bootstrap can lead to poor small sample performance
- In Laber et al. (2010) we offer confidence intervals for various functionals of interest in DTRs
  - Data-driven upper and lower bounds
  - Consistent under fixed and local alternatives
  - Good empirical performance
The Lasso and a general framework for non-regular inference
The Lasso

Consider the following generative model

\[ Y_i = x_i^T \beta^* + \epsilon_i, \]

where \( \epsilon_1, \epsilon_2, \ldots, \epsilon_n \) are iid with mean 0 and variance \( \sigma^2 \)
The Lasso

Consider the following generative model

\[ Y_i = x_i^T \beta^* + \epsilon_i, \]

where \( \epsilon_1, \epsilon_2, \ldots, \epsilon_n \) are iid with mean 0 and variance \( \sigma^2 \).

Estimate \( \beta^* \) using \( \hat{\beta}_n \) where

\[
\hat{\beta}_n \triangleq \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|
\]

is the well known Lasso estimator (Tibshirani 1996)
The Lasso

- **Goal:** construct a confidence interval for $c^T \beta^*$
- $c \in \mathbb{R}^p$ fixed
  1. Contrast of scientific interest
  2. Feature vector of a future patient (prediction interval)
  3. Gradient of some smooth non-linear function

$c^T \sqrt{n} (\hat{\beta}_n - \beta^*)$ is non-regular when $\beta^*_j = 0$ for at least one $j$ (see Knight and Fu 2000)

Degree of non-regularity is proportional to $p - s \equiv \# \{j: \beta^*_j = 0\}$
The Lasso

▶ **Goal:** construct a confidence interval for $c^T \beta^*$

▶ $c \in \mathbb{R}^p$ fixed
  1. Contrast of scientific interest
  2. Feature vector of a future patient (prediction interval)
  3. Gradient of some smooth non-linear function

▶ $c^T \sqrt{n}(\hat{\beta}_n - \beta^*)$ is non-regular when $\beta_j^* = 0$ for at least one $j$ (see Knight and Fu 2000)

▶ Degree of non-regularity is proportional to sparsity
$p - s \overset{\Delta}{=} \# \left\{ j : \beta_j^* = 0 \right\}$
Lasso and non-regularity

- Recall our generative model
  - $\epsilon_i \sim iid \ N(0, .25)$
  - $X_i \sim iid \ N_p(0, I_p)$
  - $Y_i = x_i^T \beta^* + \epsilon_i$
  - $\beta_j^* = 1_{j \leq 5}$
- Keep training set size $n$ fixed at 100 and vary $p$
- Record coverage of residual bootstrap and adaptive confidence interval
Lasso and non-regularity

- Recall our generative model:
  - $\epsilon_i \sim_{iid} \mathcal{N}(0, .25)$
  - $X_i \sim_{iid} \mathcal{N}_p(0, I_p)$
  - $Y_i = x_i^T \beta^* + \epsilon_i$
  - $\beta^*_j = 1_{j \leq 5}$
- Keep training set size $n$ fixed at 100 and vary $p$
- Record coverage of residual bootstrap and adaptive confidence interval

![Estimated Coverage: Toy Example](image)
Conclusions

- Constructed adaptive confidence intervals using smooth data-dependent bounds
- Consistency under fixed and local alternatives
- Wide range of applications
  1. Classification
  2. Dynamic treatment policies
  3. Inference after model selection
  4. . . .
- Many directions for future work...
Thank you for your attention.

Questions and comments: Eric Laber, laber@umich.edu
Adaptive confidence intervals and the Lasso

To apply same methodology as test error we want smooth data-dependent upper and lower bounds on $c^T \sqrt{n}(\hat{\beta}_n - \beta^*)$.

- No closed formula for $c^T \sqrt{n}(\hat{\beta}_n - \beta^*)$.
- Use implicit minimization definition instead.

Define

$$V_n(u, t) \triangleq -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i x_i^T u + u^T \left( \frac{\sum_{i=1}^{n} x_i x_i^T}{n} \right) u$$

$$+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \left[ |\sqrt{n}\beta^*_j - u_j| - |\sqrt{n}\beta^*_j| \right] \frac{1}{\hat{\sigma}_j^2} > \lambda_n$$

$$+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \left[ |t_j - u_j| - |t_j| \right] \frac{1}{\hat{\sigma}_j^2} \leq \lambda_n$$

then $\sqrt{n}(\hat{\beta}_n - \beta^*) = \arg \min_u V_n(u, \sqrt{n}\beta^*)$. 
Define upper bound on $c^T \sqrt{n}(\hat{\beta}_n - \beta^*)$ to be

$$
\mathcal{U}(c) \triangleq \sup_{t \in \mathcal{C}_n} c^T \left[ \arg \min_u V_n(u, t) \right]
$$

$\mathcal{C}_n$ is the constraint set for $t$ given by

$$
\mathcal{C}_n \triangleq \left\{ t_j : \frac{(\sqrt{n}(\hat{\beta}_{n,j} - \beta_j^*) - t_j)^2}{\hat{\sigma}_j^2} \leq \lambda_n \right\}
$$
Assumptions

(A1) $\Omega_n \triangleq \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \rightarrow \Omega$ positive definite

(A2) $\frac{1}{n} \max_{i \leq n} x_i^T x_i \rightarrow 0$

(A3) $\lambda_n \rightarrow \infty$ and $\lambda_n = o(n)$
Theoretical results

Lemma (Limiting objective function)

Assume (A1)-(A3) and that data are generated under \( \beta_n \triangleq \beta^* + t/\sqrt{n} \) then:

\[
\arg\min_u V_n(u, t) \xrightarrow{\sim} \arg\min_u V_\infty(u, t)
\]

where convergence is uniform over \( t \) on compact sets, and \( V_\infty(u, t) \) equals

\[
-2u^T w + u^T \Omega u + \lambda_0 \sum_{j=1}^{p} \left[ |u_j + t_j| - |t_j| \right] 1_{\beta_j^* = 0} + \lambda_0 \sum_{j=1}^{p} \text{sgn}(\beta_j^*) u_j 1_{\beta_j^* \neq 0}.
\]
Theoretical results

Theorem (Consistency)

Assume (A1)-(A3) and that the data are generated under 
\( \beta_n^* \equiv \beta^* + t/\sqrt{n} \) then:

1. \( c^T \sqrt{n} (\hat{\beta}_n - \beta^*) \sim c^T \arg \min_u V_\infty(u, t) \)
2. \( \mathcal{U}(c) \sim \sup_{t \in \mathbb{R}^p} c^T [\arg \min_u V_\infty(u, t)] \).
Theoretical results

Theorem (Consistency)

Assume (A1)-(A3) and that the data are generated under 
\[ \beta_n^* \triangleq \beta^* + t / \sqrt{n} \]
then:

1. \( c^T \sqrt{n}(\hat{\beta}_n - \beta^*) \rightsquigarrow c^T \arg \min_u V_\infty(u, t) \)
2. \( \mathcal{U}(c) \rightsquigarrow \sup_{t \in \mathbb{R}^p} c^T [\arg \min_u V_\infty(u, t)] \).

Corollary (Regularity of bounds)

The upper bound \( \mathcal{U}(c) \) is regular.
A note on computation

- Bootstrap distribution of bounds are used to construct a confidence interval.
- For every bootstrap, need to compute
  \[ \sup_{t \in C_n} c^T \text{arg min}_u V_n(u, t) \]
A note on computation

- Bootstrap distribution of bounds are used to construct a confidence interval.
- For every bootstrap, need to compute:
  \[ \sup_{t \in C_n} c^T \arg \min_u V_n(u, t) \]
  - For each \( t \) computing \( \arg \min_u V_n(u, t) \) is quadratic program.
A note on computation

- Bootstrap distribution of bounds are used to construct a confidence interval
- For every bootstrap need to compute $\sup_{t \in C_n} c^T \arg\min_u V_n(u, t)$
  - For each $t$ computing $\arg\min_u V_n(u, t)$ is quadratic program
  - For each fixed $t$ let $KKT(t, u, \theta)$ denote the Karush-Kuhn-Tucker conditions for the corresponding quadratic program
A note on computation

- Bootstrap distribution of bounds are used to construct a confidence interval
- For every bootstrap need to compute
  \[ \sup_{t \in C_n} c^T \arg\min_u V_n(u, t) \]
  - For each \( t \) computing \( \arg\min_u V_n(u, t) \) is quadratic program
  - For each fixed \( t \) let \( KKT(t, u, \theta) \) denote the Karush-Kuhn-Tucker conditions for the corresponding quadratic program
  - New optimization problem

\[
\sup_{t, u, \theta} c^T u \\
\text{s.t.} \quad KKT(t, u, \theta) \\
\quad t \in C_n
\]
A note on computation

- Bootstrap distribution of bounds are used to construct a confidence interval.
- For every bootstrap need to compute
  \[ \sup_{t \in C_n} c^T \arg \min_u V_n(u, t) \]
  - For each \( t \) computing \( \arg \min_u V_n(u, t) \) is quadratic program.
  - For each fixed \( t \) let \( KKT(t, u, \theta) \) denote the Karush-Kuhn-Tucker conditions for the corresponding quadratic program.
- New optimization problem
  \[
  \sup_{t, u, \theta} c^T u \\
  \text{s.t.} \quad KKT(t, u, \theta) \\
  t \in C_n
  \]
- New optimization problem is a Mixed Integer Program.
Lasso Summary

- Preliminary work shows promise
  - Consistent under fixed and local alternatives
  - Bootstrap also consistent (not shown here)
  - Computationally very interesting!
- Fixed $p$ framework
- Need to extend to a “large $p$ small $n$” asymptotic framework
A general framework

- Lasso problem suggests a general approach for M-estimators
  1. Derive limiting process of objective function indexed by local alternatives
  2. Approximate the supremum of this resultant process
  3. Efficiency? Optimality?