Adaptive confidence intervals for nonregular parameters

Eric B. Laber\textsuperscript{1} & Susan A. Murphy\textsuperscript{2}

\textsuperscript{1}North Carolina State University, Raleigh
\textsuperscript{2}University of Michigan, Ann Arbor

University of Alberta
February 28, 2013
Introduction

- Modern statistical analysis is rife with non-regularity
  1. Test error of a learned classifier
  2. Parameters in a treatment policy
  3. Inference based on thresholded estimators
  4. ... 

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

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

1. Confidence intervals for the test error of a classifier
2. Confidence intervals for parameters in optimal treatment policies
Example I: Classification

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 $\hat{c}_D(X) : \mathbb{R}^p \mapsto \{-1, 1\}$
3. Use classifier for prediction at new inputs
Example I: Classification

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

Goal:
- **Interval estimator**: for test error $\tau(\hat{c}_\mathcal{D}) \triangleq P1_{Y \neq \hat{c}_\mathcal{D}(X)}$
**The problem**

- Focus on linear approximations to the Bayes decision boundary
  - We do not assume the approximation space is correct
The problem

- Focus on linear approximations to the Bayes decision boundary
  - We do not assume the approximation space is correct
- Construct a classifier using surrogate loss $L(X, Y, \beta)$
  1. $\hat{\beta} \triangleq \arg\min_{\beta \in \mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$
  2. $\hat{c}_D(X) = \text{sign}(X^T \hat{\beta})$
The problem

- Focus on linear approximations to the Bayes decision boundary
  - We do not assume the approximation space is correct
- Construct a classifier using surrogate loss $L(X, Y, \beta)$
  1. $\hat{\beta} \triangleq \arg \min_{\beta \in \mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$
  2. $\hat{c}_D(X) = \text{sign} (X^T \hat{\beta})$
- 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:
      $L(X, Y, \beta) = (1 - YX^T \beta)^2$
The problem cont’d

- Test error

\[ \tau(\hat{\beta}) \triangleq P_{1_{YX^T\hat{\beta} < 0}} = \int 1_{yX^T\hat{\beta} < 0} dP(x, y) \]
The problem cont’d

- Test error

\[ \tau(\hat{\beta}) \triangleq P1_{YX^T\hat{\beta}<0} = \int 1_{y\times\hat{\beta}<0}dP(x,y) \]

- Averages over new input-output pair \((X, Y)\) but *not* training data—evaluates the performance of the learned classifier
Test error

\[ \tau(\hat{\beta}) \triangleq P1_{YX^T \hat{\beta} < 0} = \int 1_{yX^T \hat{\beta} < 0} dP(x, y) \]

- Averages over new input-output pair \((X, Y)\) but not training data—evaluates the performance of the learned classifier
- The test error \(\tau(\hat{\beta})\) is random quantity
  - Data-dependent parameter (Dawid 1994)
The problem cont’d

- Test error

\[ \tau(\hat{\beta}) \triangleq P_{YX^\top \hat{\beta} < 0} = \int 1_{yX^\top \hat{\beta} < 0} dP(x, y) \]

- Averages over new input-output pair \((X, Y)\) but not training data—evaluates the performance of the learned classifier

- The test error \(\tau(\hat{\beta})\) is random quantity
  - Data-dependent parameter (Dawid 1994)

- Contrast with expected test error which averages over training data—evaluates performance of the algorithm used to construct the classifier
The problem cont’d

- **Goal**: given $\alpha \in (0, 1)$ construct $\hat{u}$ and $\hat{l}$ so that

$$P_D \{ \hat{l} \leq \tau(\hat{\beta}) \leq \hat{u} \} \geq 1 - \alpha$$
The problem cont’d

- **Goal:** given $\alpha \in (0, 1)$ construct $\hat{u}$ and $\hat{l}$ so that

$$P_D \left\{ \hat{l} \leq \tau(\hat{y}) \leq \hat{u} \right\} \geq 1 - \alpha$$

**Context**
- Model space may not be correct
- Low dimensional setting ($p$ fixed)
- Cannot afford a test set
Non-regularity

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

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

- $P_1_{X^T \beta^* = 0} > 0$ implies $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$ has non-regular limit
  - points near the boundary cause jittering
  - $P_1_{YX^T \hat{\beta} < 0}$ need not concentrate about its mean
  - bootstrap and normal approximations are invalid
Suppose

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

Properties of this example

- $P_1\mathbf{x}^\top \beta^* = 0$ (seemingly regular)
- Linear classifier is a good fit
- E.g. if $n = 30$
  - $\mathbb{E}(\tau(\hat{\beta})) \approx .11$
  - Bayes error $\approx .09$
Illustration cont’d

Under “regular” framework

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

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

are both asymptotically valid
Under “regular” framework
- Centered bootstrap $\sqrt{n}(\hat{P}_n^{(b)} - P_n)1_{YX^T\hat{\beta}^{(b)} < 0}$
- Normal approximation $\hat{\tau}(\hat{\beta}) \pm z_{1-\gamma/2} \sqrt{\frac{\hat{\tau}(\hat{\beta})(1-\hat{\tau}(\hat{\beta}))}{n}}$

are both asymptotically valid

Coverage estimated using 1000 Monte Carlo data sets
Below nominal coverage even for $n = 250$
Coverage especially poor for small samples
Illustration cont’d

Why do these methods fail?
Illustration cont’d

Why do these methods fail?

- Non-smoothness $\Rightarrow$ non-regularity
- Performance inversely proportional to smoothness
Illustration cont’d

Why do these methods fail?

- Non-smoothness ⇒ non-regularity
- Performance inversely proportional to smoothness

Continuing our example

- Instead of test error \( \tau(\hat{\beta}) \) consider

\[
\tau_a(\hat{\beta}) \triangleq P \left( 1 + \exp(a Y X^T \hat{\beta}) \right)^{-1}
\]

- \( \tau_a(\hat{\beta}) \) is smooth for fixed \( a > 0 \)
- If \( a \to \infty \) then \( \tau_a(\hat{\beta}) \to \tau(\hat{\beta}) \)
Illustration cont’d

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

Continuing our example
▶ Instead of test error $\tau(\hat{\beta})$ consider
  
  $\tau_a(\hat{\beta}) \triangleq P \left( 1 + \exp(aYX^T\hat{\beta}) \right)^{-1}$

▶ $\tau_a(\hat{\beta})$ is smooth for fixed $a > 0$
▶ If $a \to \infty$ then $\tau_a(\hat{\beta}) \to \tau(\hat{\beta})$
▶ Conjecture: Bootstrap coverage should deteriorate as $a$ grows
Illustration cont’d

Smoothed Loss Functions

Margin

Smoothed Loss

a = 0.1

a = 1.0

a = 10

a = Inf

Estimated Coverage Quadratic Example: Smoothed

Sample Size

Estimated Coverage
Illustration cont’d

Smoothed Loss Functions

Estimated Coverage Quadratic Example: Smoothed
Two Examples

1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies
Example II: Treatment Policies

- Motivation: treatment of chronic illness
  - Some examples: HIV/AIDS, cancer, depression, schizophrenia, drug and alcohol addiction, ADHD, etc.
  - Multistage decision making problem
  - Longer-term treatment requires balancing present and future benefit as opposed to focusing only on present benefit.

- Treatment Policies
  - Operationalize multistage decision making via a 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
Construction and inference for policies have applications beyond medicine

1. Artificial Intelligence and Reinforcement Learning (autonomous helicopter, drones, etc., Ng 2003)
2. Marketing (Simester, Sun and Tsitsiklis, 2003)
3. Active labor market policies (Lechner and Miquel, 2010)
4. Recruitment and retention policies in Survey Research (Wagner, 2012?)
5. …
An Example Policy for ADHD

Prior medication?  
Yes → Low dose MEDS
No → Low dose BMOD

Adequate response?  
Yes → Continue MEDS  
No → Intensify MEDS

High adherence?  
Yes → Intensify BMOD  
No → Add BMOD

Adequate response?  
Yes → Continue BMOD  
No → No

High adherence?  
Yes → Intensify BMOD  
No → Add MEDS
ADHD Trial (Pelham, PI)

Treatment A
Low Intensity BMOD

Response?

Yes → Continue
Low Intensity BMOD

No → Treatment AA
Augment with MEDS

Treatment AB
Intensify BMOD

No → Continue
Low Intensity MEDS

Treatment BA
Augment with BMOD

Yes → Intensify MEDS

Treatment BB

Response?

Yes → Continue
Low Intensity BMOD

No → Intensify BMOD
Data

- \((X_1, A_1, X_2, A_2, Y)\) for each individual
  - \(X_j\): Observations available at stage \(j\)
  - \(A_j\): Treatment at stage \(j\), with known distribution (usually uniform)
  - \(Y\): Primary outcome (larger is better)
  - \(H_j\): History at stage \(j\), \(H_1 = X_1\), \(H_2 = (X_1, A_1, X_2)\)

- The policy, \(\pi = \{\pi_1, \pi_2\}\), \(\pi_j : H_j \rightarrow A_j\), should have high Value: \(V^\pi = E^\pi(Y)\)
Constructing a policy from data: Q-Learning

- Generalization of regression to multiple treatment stages
- Backwards induction like dynamic programming
- Approximate conditional expectation with regression
Constructing a policy from data: Q-Learning

- Generalization of regression to multiple treatment stages
- Backwards induction like dynamic programming
- Approximate conditional expectation with regression

- In computer science there are many variations; almost always presented as part of a stochastic approximation algorithm for solving a problem with an infinite number of stages (infinite horizon) Watkins (1989), Sutton & Barto (1998)
- In statistics there are a few variations, with a finite number of stages, appearing in Murphy (2003), Robins (2004), Henderson et al. (2009) + more
Simple Version of Q-Learning

Two stages; linear regressions; \( A_j \in \{0, 1\} \), \( H_{j1}, H_{j2} \) features of patient history, \( H_j \):

- Stage 2 regression: Regress \( Y \) on \( H_{21}, H_{22} \) to obtain
  \[
  \hat{Q}_2(H_2, A_2) = \hat{\alpha}_1^T H_{21} + \hat{\alpha}_2^T H_{22} A_2
  \]
  \[
  \hat{\pi}_2(H_2) = \arg \max_{a_2} \hat{Q}_2(H_2, a_2) = \arg \max_{a_2} \hat{\alpha}_2^T H_{22} a_2
  \]
Simple Version of Q-Learning

Two stages; linear regressions; \( A_j \in \{0, 1\}, \ H_{j1}, H_{j2} \) features of patient history, \( H_j \):

- **Stage 2 regression:** Regress \( Y \) on \( H_{21}, H_{22} \) to obtain
  \[
  \hat{Q}_2(H_2, A_2) = \hat{\alpha}_1^T H_{21} + \hat{\alpha}_2^T H_{22} A_2
  \]
  \( \hat{\pi}_2(H_2) = \arg \max_{a_2} \hat{Q}_2(H_2, a_2) = \arg \max_{a_2} \hat{\alpha}_2^T H_{22} a_2 \)

- \( \tilde{Y} = \hat{\alpha}_1^T H_{21} + \max_{a_2} \hat{\alpha}_2^T H_{22} a_2 \) (\( \tilde{Y} \) is a predictor of \( \max_{a_2} Q_2(H_2, a_2) \))
Simple Version of Q-Learning

Two stages; linear regressions; $A_j \in \{0, 1\}$, $H_{j1}$, $H_{j2}$ features of patient history, $H_j$:

- **Stage 2 regression**: Regress $Y$ on $H_{21}, H_{22}$ to obtain
  \[ \hat{Q}_2(H_2, A_2) = \hat{\alpha}_1^T H_{21} + \hat{\alpha}_2^T H_{22} A_2 \]
  - $\hat{\pi}_2(H_2) = \arg \max_{a_2} \hat{Q}_2(H_2, a_2) = \arg \max_{a_2} \hat{\alpha}_2^T H_{22} a_2$

- $\tilde{Y} = \hat{\alpha}_1^T H_{21} + \max_{a_2} \hat{\alpha}_2^T H_{22} a_2$ (\(\tilde{Y}\) is a predictor of $\max_{a_2} Q_2(H_2, a_2)$)

- **Stage 1 regression**: Regress $\tilde{Y}$ on $H_{11}, H_{12}$ to obtain
  \[ \hat{Q}_1(H_1, A_1) = \hat{\beta}_1^T H_{11} + \hat{\beta}_2^T H_{12} A_1 \]
  - $\hat{\pi}_1(H_{12}) = \arg \max_{a_1} \hat{Q}_1(H_1, a_1) = \arg \max_{a_1} \hat{\beta}_2^T H_{12} a_1$
GOAL: confidence interval for a contrast of stage 1 parameters: \( c^T \beta^* \)

- Non-regular due to non-differentiable max operator used in Q-learning; recall
  \[ \hat{Y} = \hat{\alpha}_1^T H_{21} + \max_{a_2} \hat{\alpha}_2^T H_{22} a_2 \]

- In this setting the centered percentile bootstrap confidence interval for stage 1 \( \beta \) parameters can be anticonservative,
  (95% confidence interval covers 90%-93% in two stages, each with two treatments; 84%-93% for two stages, each with three treatments)
Limiting Distribution of centered $c^T \sqrt{n} \hat{\beta}$

- Local Alternative:
  - $\alpha_{2,n}^* = \alpha_2^* + u/\sqrt{n}$
  - The limiting distribution of $c^T \sqrt{n}(\hat{\beta} - \beta_n^*)$ is the distribution of $c^T \Sigma_1^{-1} (\mathbb{W} + f(\mathbb{V}, u))$

where

$$f(v, u) = E \left[ B_1 ( [H_{22}^T v + H_{22}^T u]_+ - [H_{22}^T u]_+ ) 1_{H_{22}^T \alpha_2^* = 0} \right]$$

and $B_1 = (H_{11}^T, H_{12}^T A_1)^T$ (e.g. the design matrix) and $\mathbb{W}, \mathbb{V}$ are jointly normal vectors

- The fact that the limiting distribution depends on the direction, $u$, means that $\hat{\beta}$ is a nonregular estimator (unless $P[H_{22}^T \alpha_2^* = 0] = 0$)
Ideas
Ideas

This work builds on ideas from

- Generalization error bounds
  - Construct smooth data-based upper and lower bounds on a centered estimator:
    - $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$ (centered estimator of test error)
    - $\sqrt{n}(c^T\hat{\beta} - c^T\beta)$ (centered stage 1 regression coefficient)
  - If generative model induces regularity, then bounds collapse to centered parameter

- Pretests (e.g. hypothesis tests) for use in inference concerning weakly identified parameters in econometrics (Andrews 2001, Andrews and Soares 2007; Cheng 2008). We use the pretest idea to test if the parameter is near a “bad” parameter value.
Ideas

- Confidence interval is the primary focus

- Construct smooth data-based upper and lower bounds on a centered estimator:
  - $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$ (centered estimator of test error)
  - $\sqrt{n}(c^T\hat{\beta}_1 - c^T\beta_1)$ (centered stage 1 regression coefficient)

- Confidence intervals are formed by bootstrapping these bounds

- Evaluate using an asymptotic framework that permits non-regularity
The Adaptive Confidence Intervals

1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies
Adaptive CI for the test error

Idea: construct smooth upper and lower bounds on $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$

- Recall $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$ is equal to
  $$\sqrt{n}(\mathbb{P}_n - P)1_{YX\tau\beta<0}$$

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

$$\text{UB}_n \triangleq \sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$$

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

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

where $\hat{\Sigma} = n\text{Cov}(\hat{\beta})$
Adaptive CI for the test error

Idea: construct smooth upper and lower bounds on $\sqrt{n}(\hat{\tau} - \tau)$

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

  $$\sqrt{n}(P_n - P)1_{YX\hat{\tau}<0}$$

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

$$UB_n \triangleq \sqrt{n}(\hat{\tau}) - \tau - \sqrt{n}(P_n - P)1_{n(X^\top \hat{\beta})^2 < \lambda_n} \leq \lambda_n 1_{YX^\top \hat{\beta} < 0}$$

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

where $\hat{\Sigma} = n\text{Cov}(\hat{\beta})$

(Replace supremum with infimum to obtain lower bound.)
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.

(A6) The sequence \( \lambda_n \) tends to infinity and satisfies \( \lambda_n = o(n) \).
Properties

Theorem (Convergence)

1. $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) \rightsquigarrow W + V(z_{\infty})$
2. $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) \leq UB_n$ for all $n$
3. $UB_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} W + V(u)$
4. $UB_n^{(b)} \rightsquigarrow \sup_{u \in \mathbb{R}^p} W + V(u)$ in probability.

where $(V, W, z_{\infty})$ is zero mean Gaussian; $V$ is a Gaussian process, $W$ is a normal random variable and $z_{\infty}$ is $p$-dim normal.
Properties

Theorem (Convergence)

1. $\sqrt{n}(\hat{\beta}(\hat{\beta}) - \tau(\hat{\beta})) \rightsquigarrow W + \mathbb{V}(z_\infty)$
2. $\sqrt{n}(\hat{\beta}(\hat{\beta}) - \tau(\hat{\beta})) \leq UB_n$ for all $n$
3. $UB_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} W + \mathbb{V}(u)$
4. $UB_n^{(b)} \rightsquigarrow \sup_{u \in \mathbb{R}^p} W + \mathbb{V}(u)$ in probability.

where $(\mathbb{V}, W, z_\infty)$ is zero mean Gaussian; $\mathbb{V}$ is a Gaussian process, $W$ is a normal random variable and $z_\infty$ is $p$-dim normal.

Theorem (Adaptation)

If either the Bayes decision boundary is linear or $P(X^T \beta^* = 0) = 0$ then $UB_n$ and $\sqrt{n}(\hat{\beta}(\hat{\beta}) - \tau(\hat{\beta}))$ have the same limiting distribution.
Properties

The supremum in the upper bound $\mathbb{UB}_n$ can be viewed as a supremum over local alternatives:

Theorem (Convergence under local alternatives)

Under $P_n$

1. $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) \rightsquigarrow \mathbb{W} + \mathbb{V}(z_\infty + u)$
2. $\mathbb{UB}_n \rightsquigarrow \sup_{\mathbb{R}^p} \mathbb{W} + \mathbb{V}(u)$.

where $P_n$ is a sequence of local alternatives contiguous to $P$ for which $\beta_n^* \triangleq \arg \min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta)$ satisfies $\beta_n^* = \beta^* + u/\sqrt{n}$. 
The Adaptive Confidence Intervals

1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies
Adaptive CI for the treatment effect

Idea: construct smooth upper and lower bounds on $c^T \sqrt{n}(\hat{\beta} - \beta^*)$.

$$ UB_n \triangleq c^T \sqrt{n}(\hat{\beta} - \beta^*) $$

$$ - c^T \hat{\Sigma}_{11}^{-1} P_n B_1 \left( [H_{22}^T V_n + H_{22}^T u]_+ - [H_{22}^T u]_+ \right) 1_{\frac{n(H_{22}^T \hat{\alpha}_2)^2}{H_{22}^T \hat{\Sigma} H_{22}} \leq \lambda_n} \left| u = \sqrt{n} \alpha_2^* \right| $$

$$ + \sup_u c^T \hat{\Sigma}_{11}^{-1} P_n B_1 \left( [H_{22}^T V_n + H_{22}^T u]_+ - [H_{22}^T u]_+ \right) 1_{\frac{n(H_{22}^T \hat{\alpha}_2)^2}{H_{22}^T \hat{\Sigma} H_{22}} \leq \lambda_n} $$

where the supremum is taken only when $H_{22}$ is in a region near the decision boundary $H_{22}^T \alpha_2^* = 0$

- $B_1 = (H_{11}^T, H_{12}^T A_1)^T$
- $V_n = \sqrt{n}(\hat{\alpha}_2 - \alpha_2^*)$
- $\hat{\Sigma} = n \widehat{\text{Cov}}(\hat{\alpha}_2)$
Adaptive CI for the treatment effect

Idea: construct smooth upper and lower bounds on $c^T \sqrt{n}(\hat{\beta} - \beta^*)$.

$$\text{UB}_n \triangleq c^T \sqrt{n}(\hat{\beta} - \beta^*)$$

$$- c^T \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left( \left[ H_{22}^T \mathbb{V}_n + H_{22}^T u \right]_+ - \left[ H_{22}^T u \right]_+ \right) 1_{\frac{n(H_{22}^T \hat{\alpha}_2)^2}{H_{22}^T \hat{\Sigma} H_{22}} \leq \lambda} \bigg| u = \sqrt{n\alpha^*_2}$$

$$+ \sup_u c^T \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left( \left[ H_{22}^T \mathbb{V}_n + H_{22}^T u \right]_+ - \left[ H_{22}^T u \right]_+ \right) 1_{\frac{n(H_{22}^T \hat{\alpha}_2)^2}{H_{22}^T \hat{\Sigma} H_{22}} \leq \lambda}$$

where the supremum is taken only when $H_{22}$ is in a region near the decision boundary $H_{22}^T \alpha_2^* = 0$

- $B_1 = (H_{11}^T, H_{12}^T A_1)^T$
- $\mathbb{V}_n = \sqrt{n}(\hat{\alpha}_2 - \alpha_2^*)$
- $\hat{\Sigma} = n \hat{\text{Cov}}(\hat{\alpha}_2)$

(Replace supremum with infimum to obtain lower bound.)
Assumptions

(A1) The histories $H_j$ with $B_j = (H_{j1}^T, H_{j2}^T, A_j)^T$, $j = 1, 2$ and primary outcome $Y$, satisfy the moment inequalities $P\|H_2\|^2 \|B_1\|^2 < \infty$ and $PY^2\|B_j\|^2 < \infty$.

(A2) Define:

1. $\Sigma_j \triangleq PB_j^T B_j$ for $j = 1, 2$;
2. $g_2(B_2, Y_2; \alpha^*) \triangleq B_2^T (Y_2 - B_2 \alpha^*)$;
3. $g_1(B_1, Y_1, H_2; \beta^*, \alpha^*) \triangleq B_1^T (H_{21}^T \alpha_1^* + |H_{22}^T \alpha_2^*| - B_1 \beta^*)$;

assume the matrices $\Sigma_j$ and $\Omega \triangleq \text{Var-cov} (g_1, g_2)$ are strictly positive definite.

(A3) The sequence $\lambda_n$ tends to infinity and satisfies $\lambda_n = o(n)$. 
Properties

Theorem (Convergence)

1. \( c^T \sqrt{n}(\hat{\beta} - \beta^*) \sim c^T \Sigma_1^{-1} (W + f(V, 0)) \)
2. \( c^T \sqrt{n}(\hat{\beta} - \beta^*) \leq UB_n \) for all \( n \)
3. \( UB_n \sim \sup_{u \in \mathbb{R}^p} c^T \Sigma_1^{-1} (W + f(V, u)) \)
4. \( UB_n^{(b)} \sim \sup_{u \in \mathbb{R}^p} c^T \Sigma_1^{-1} (W + f(V, u)) \) in probability.

where

\[
f(v, u) = E \left[ B_1 \left( [H_{22}^T v + H_{22}^T u]^+ - [H_{22}^T u]^+ \right) 1_{H_{22}^T \alpha^*_2 = 0} \right]
\]

and \( B_1 = (H_{11}^T, H_{12}^T A_1)^T \) (e.g. row of the design matrix) and \( W, V \) are jointly normal vectors.
Properties

Theorem (Adaptation)

If \( P(H_{22}^\top \alpha_2^* = 0) = 0 \) then \( UB_n \) and \( c^\top \sqrt{n}(\hat{\beta} - \beta^*) \) have the same limiting distribution.
Properties

Theorem (Adaptation)
If \( P(H_{22}^\top \alpha_2^* = 0) = 0 \) then \( UB_n \) and \( c^\top \sqrt{n}(\hat{\beta} - \beta^*) \) have the same limiting distribution.

The supremum in the upper bound \( UB_n \) can be viewed as a supremum over local alternatives:

Theorem (Convergence under local alternatives)
Under \( P_n \) for which \( \alpha_{2,n}^* = \alpha_{2}^* + u/\sqrt{n} \),

1. \( c^\top \sqrt{n}(\hat{\beta} - \beta_{n}^*) \rightsquigarrow c^\top \Sigma_1^{-1} (W + f(V, u)) \)
2. \( UB_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} c^\top \Sigma_1^{-1} (W + f(V, u)) \).
Simulation Experiments

1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies
Experiments

Compare performance of

- Adaptive confidence interval (ACI)
- CV-Normal approximation [Yang 2006]
- BCCVP-BR approximation [Jiang 2008]
- ACI uses $\lambda_n \triangleq \max(\sqrt{n}, \chi_{0.995}^2)$
Experiments

Compare performance of

- Adaptive confidence interval (ACI)
- CV-Normal approximation [Yang 2006]
- BCCVP-BR approximation [Jiang 2008]
- ACI uses $\lambda_n \triangleq \max(\sqrt{n}, \chi^2_{.995})$

Details

- 1000 Monte Carlo replications
- 10 data sets
Results

Target coverage $0.950$, loss function $L(X, Y, \beta) = \log(1 + e^{-YX^T\beta})$, $n = 30$

<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>0.976</td>
<td>0.893</td>
<td>0.914</td>
</tr>
<tr>
<td>Magic</td>
<td>0.955</td>
<td>0.999</td>
<td>0.983</td>
</tr>
<tr>
<td>Mam.</td>
<td>0.951</td>
<td>0.993</td>
<td>0.974</td>
</tr>
<tr>
<td>Ion.</td>
<td>0.947</td>
<td>0.995</td>
<td>0.985</td>
</tr>
<tr>
<td>Donut</td>
<td>0.968</td>
<td>0.966</td>
<td>0.908</td>
</tr>
<tr>
<td>Bal.</td>
<td>0.979</td>
<td>0.996</td>
<td>0.972</td>
</tr>
<tr>
<td>Liver</td>
<td>0.946</td>
<td>0.997</td>
<td>0.972</td>
</tr>
<tr>
<td>Spam</td>
<td>0.985</td>
<td>0.999</td>
<td>0.981</td>
</tr>
<tr>
<td>Quad</td>
<td>0.978</td>
<td>0.997</td>
<td>0.945</td>
</tr>
<tr>
<td>Heart</td>
<td>0.960</td>
<td>0.995</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Table: Estimated coverage of competing confidence procedures. Coverage is highlighted if not different from $0.950$ at the $0.01$ level.
Results

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

<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>.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

- ACI achieves nominal coverage
- Non-trivial width
- Computationally efficient
- Robust to choice of $\lambda_n$
Simulation Experiments

1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies
Empirical study

- Compare performance of
  - Soft-thresholding (ST) (Chakraborty et al., 2009)
  - Centered percentile bootstrap (CPB)
  - Plug-in pretesting estimator (PPE) (uses idea of Chatterjee and Lahiri, 2011)
  - ACI uses $\lambda_n = \log \log n$
Empirical study

- Compare performance of
  - Soft-thresholding (ST) (Chakraborty et al., 2009)
  - Centered percentile bootstrap (CPB)
  - Plug-in pretesting estimator (PPE) (uses idea of Chatterjee and Lahiri, 2011)
  - ACI uses $\lambda_n = \log \log n$

- Generative models
  1. Non-regular (NR): $P(H_{22}^T \beta_{22}^* = 0) > 0$
  2. Nearly non-regular (NNR): $P(H_{22}^T \beta_{22}^* \approx 0) > 0$
  3. Regular (R): $P(H_{22}^T \beta_{22}^* \approx 0) = 0$

- 1000 Monte Carlo replications
Results

Target coverage .950 for coefficient of stage 1 treatment, $n = 150$

<table>
<thead>
<tr>
<th>2 stages</th>
<th>2 txts</th>
<th>Ex1</th>
<th>Ex2</th>
<th>Ex3</th>
<th>Ex4</th>
<th>Ex5</th>
<th>Ex6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPB</td>
<td>NR</td>
<td>0.934</td>
<td>0.935</td>
<td>0.930</td>
<td>0.939</td>
<td>0.925</td>
<td>0.928</td>
</tr>
<tr>
<td>ST</td>
<td>NNR</td>
<td>0.948</td>
<td>0.945</td>
<td>0.938</td>
<td>0.919</td>
<td>0.759</td>
<td>0.762</td>
</tr>
<tr>
<td>PPE</td>
<td>0.931</td>
<td>0.940</td>
<td>0.938</td>
<td>0.931</td>
<td>0.904</td>
<td>0.903</td>
<td></td>
</tr>
<tr>
<td>ACI</td>
<td>0.992</td>
<td>0.992</td>
<td>0.968</td>
<td>0.950</td>
<td>0.964</td>
<td>0.965</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 stages</th>
<th>3 txts</th>
<th>Ex1</th>
<th>Ex2</th>
<th>Ex3</th>
<th>Ex4</th>
<th>Ex5</th>
<th>Ex6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPB</td>
<td>NR</td>
<td>0.933</td>
<td>0.938</td>
<td>0.915</td>
<td>0.940</td>
<td>0.885</td>
<td>0.895</td>
</tr>
<tr>
<td>PPE</td>
<td>0.931</td>
<td>0.932</td>
<td>0.927</td>
<td>0.918</td>
<td>0.858</td>
<td>0.856</td>
<td></td>
</tr>
<tr>
<td>ACI</td>
<td>0.999</td>
<td>0.999</td>
<td>0.968</td>
<td>0.964</td>
<td>0.970</td>
<td>0.971</td>
<td></td>
</tr>
</tbody>
</table>

Table: Coverage is NOT highlighted if significantly below .95 at the .05 level.
Conclusion

- ACI achieved nominal or improved coverage on all examples
- ACI is conservative when there is no stage 2 treatment effect.
- Relative performance of ACI improves on examples with increasing numbers of stages and/or treatments
- Robust to choice of $\lambda_n$
Many modern statistical problems involve nonregular estimators. Most frequently these occur in $p$ large ($p < n$) or $p >> n$ problems. Examples:

- Inference based on estimators that involve the estimation of a matrix with eigenvalues that may be near zero,
- Prediction intervals after using lasso or other variable selection methods,
- Evaluation of the misclassification rate of a learned classifier
- Constrained estimation

Principled approaches to forming confidence intervals and hypothesis tests are currently lacking.
Questions: laber@stat.ncsu.edu, samurphy@umich.edu
A copy of this talk can found at:
www.stat.lsa.umich.edu/~samurphy

Acknowledgements:
National Institute of Health grants RO1 MH080015 and P50 DA10075
ADHD Trial (Pelham, PI)

**Treatment A**
Low Intensity BMOD → Response?

- **Response?**
  - Yes → Continue
    - Low Intensity BMOD
  - No → Intensify BMOD → Treatment AB

**Treatment B**
Low Intensity MEDS → Response?

- **Response?**
  - Yes → Continue
    - Low Intensity MEDS
  - No → Intensify MEDS → Treatment BB
ADHD Dynamic Treatment Regime

Prior medication?
- Yes: Low dose MEDS
- No: Low dose BMOD

Adequate response?
- Yes: Continue MEDS
- No: Intensify MEDS

High adherence?
- Yes: Add BMOD
- No: Add MEDS

Adequate response?
- Yes: Continue BMOD
- No: Intensify BMOD
## Inference for ADHD Treatment Effects

<table>
<thead>
<tr>
<th>Stage</th>
<th>History</th>
<th>Lower (5%)</th>
<th>Upper (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Had prior med.</td>
<td>-0.51</td>
<td>0.14</td>
</tr>
<tr>
<td>1</td>
<td>No prior med.</td>
<td>-0.05</td>
<td>0.39</td>
</tr>
<tr>
<td>2</td>
<td>High adherence and BMOD</td>
<td>-0.08</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>Low adherence and BMOD</td>
<td>-1.10</td>
<td>-0.28</td>
</tr>
<tr>
<td>2</td>
<td>High adherence and MEDS</td>
<td>-0.18</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>Low adherence and MEDS</td>
<td>-1.25</td>
<td>-0.29</td>
</tr>
</tbody>
</table>

- Positive stage 1 effect favors BMOD \( A_1 = 1 \) if BMOD; \( A_1 = -1 \) if MED
- Positive stage 2 effect favors Intensify \( A_2 = 1 \) if Intensify; \( A_2 = -1 \) if Augment
ADHD Dynamic Treatment Regime

Prior medication? Yes → Low dose MEDS ~OR~ BMOD

No → Low dose BMOD

Adequate response? Yes → Intensify SAME

No → Intensify SAME

High adherence? Yes → Add MEDS ~OR~ Intensify BMOD

No → Add MEDS

Adequate response? Yes → Continue SAME

No → Add OTHER ~OR~ Intensify SAME