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

Porcelli Lectures, Louisiana State University
April 28, 2014
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 in many realistic settings

- An asymptotic framework that faithfully represents small sample behavior is needed for the development and evaluation of inferential procedures
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 $\hat{c}_D(X) : \mathbb{R}^p \mapsto \{-1, 1\}$
   - Classifier should minimize test error,
     \[ \tau(\hat{c}_D) = \int 1_{y\hat{c}_D(x)<0} dP(x, y) \]
     where $P$ is unknown probability distribution of $(X, Y)$

3. Use classifier for prediction at new inputs
Classification

1. Observe \textit{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 \( \hat{c}_D(X) : \mathbb{R}^p \mapsto \{-1, 1\} \)
   - Classifier should minimize test error, \( \tau(\hat{c}_D) = \int 1_{y \hat{c}_D(x) < 0} dP(x, y) \) where \( P \) is unknown probability distribution of \( (X, Y) \)

3. Use classifier for prediction at new inputs

Goal:
   - \textit{Interval estimator}: for test error \( \tau(\hat{c}_D) \)
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.
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. $\hat{c}_D(X) = \text{sign} \left( X^T \hat{\beta} \right)$
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

\[ \tau(\hat{\beta}) \triangleq P1_{YX^T\hat{\beta}<0} = \int 1_{yX^T\hat{\beta}<0} dP(x, y) \]
The problem cont’d

- Test error

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

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

- 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

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

- 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)
  - \(\tau(\cdot) : \mathbb{R}^p \rightarrow [0, 1]\) can be discontinuous
Goal: given $\alpha \in (0, 1)$ construct $\hat{u}$ and $\hat{l}$ so that

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

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

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

**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
Illustration

Suppose

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

Properties of this example

- $P1_{X^\top\beta^* = 0} = 0$ (seemingly regular)
- Linear classifier is a good fit
- E.g. if $n = 30$
  - $\mathbb{E}(\tau(\hat{\beta})) \approx 0.11$
  - Bayes error $\approx 0.09$
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 
  \[ \hat{\tau}(\hat{\beta}) \pm Z_{1-\gamma/2} \sqrt{\alpha(1-\alpha) / n} \]

are both asymptotically valid
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

![Estimated Coverage Quadratic Example: 0–1](image)

- 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?

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

Continuing our example

- Instead of test error $\tau(\hat{\beta})$ consider $\tau_a(\hat{\beta}) ≜ P\left(1 + \exp(aYX^{\top}\hat{\beta})\right) - 1$

- $\tau_a(\hat{\beta})$ 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
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 $\Rightarrow$ non-regularity
- Performance inversely proportional to smoothness

Continuing our example

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

$$
\tau_a(\hat{\beta}) \triangleq P \left( 1 + \exp(a YX^T \hat{\beta}) \right)^{-1}
$$

- $\tau_a(\hat{\beta})$ is smooth for fixed $a > 0$
- If $a \to \infty$ then $\tau_a(\hat{\beta}) \to \tau(\hat{\beta})$
Why do these methods fail?

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

Continuing our example

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

Estimated Coverage Quadratic Example: Smoothed

- Sample Size
- Estimated Coverage

- Various values of \( a \):
  - \( a = 0.1 \)
  - \( a = 1.0 \)
  - \( a = 10 \)
  - \( a = \infty \)
Illustration cont’d
Interlude: 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 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
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

1. **Prior medication?**
   - Yes: Low dose MEDS
     - Yes: Adequate response?
       - Yes: Continue MEDS
       - No: High adherence?
         - No: Add BMOD
         - Yes: Intensify MEDS
     - No: Adequate response?
       - Yes: Continue BMOD
       - No: High adherence?
         - No: Add MEDS
         - Yes: 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
  - $\tilde{Y} = \hat{\alpha}_1^T H_{21} + \max_{a_2} \hat{\alpha}_2^T H_{22} a_2$

- 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)
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)
  - 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)

- Confidence intervals are formed by bootstrapping these bounds

- Evaluate using an asymptotic framework that permits non-regularity
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

\[ \sqrt{n}(\mathbb{P}_n - P)1_{YX^\top \hat{\beta} < 0} \]

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

\[ \text{UB}_n \triangleq \sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) - \sqrt{n}(\mathbb{P}_n - P)1_{YX^\top \hat{\beta} < 0} \]

\[ + \sup_{u \in \mathbb{R}^p} \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\top \hat{\beta})^2}{X^\top \hat{\Sigma} X} \leq \lambda_n} \]

\[ 1_{YX^\top u < 0} \]

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

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

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

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

\[ \text{UB}_n \triangleq \sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) - \sqrt{n}(\mathbb{P}_n - P)1_{YX^\top \hat{\beta} < 0} \]
\[ + \sup_{u \in \mathbb{R}^p} \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\top \hat{\beta})^2}{X^\top \hat{\Sigma} X} \leq \lambda n} 1_{YX^\top u < 0} \]

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

Theorem (Adaptation)

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

Compare performance of

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

Compare performance of

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

Details

- 1000 Monte Carlo replications
- 10 data sets
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>.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 $L(X, Y, \beta) = \log(1 + e^{-YX^\top\beta})$, $n = 30$

<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 $\lambda_n$
Many modern statistical problems involve nonregular estimators. Most frequently these occur in $p$ large ($p < n$) or $p \gg n$ 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 be 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?

No

Treatment AB
Intensify BMOD

Yes

Continue
Low Intensity BMOD

Response?

No

Treatment AA
Augment with MEDS

Yes

Continue
Low Intensity MEDS

Treatment AB
Intensify BMOD

No

Treatment BA
Augment with BMOD

Response?

No

Treatment BB
Intensify MEDS
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
- Intensify BMOD

No
- Add BMOD

Adequate response?

Yes
- Continue BMOD

No
- High adherence?
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 → Adequate response?

No → High adherence?

Yes → Intensify SAME

No → Add OTHER OR Itensify SAME

Adequate response? Yes → Continue SAME

No → Add MEDS

High adherence? Yes → Add MEDS OR Intensify BMOD

No → Add MEDS

High adherence? Yes → Continue SAME

No → Add MEDS

Intensify SAME