Adaptive confidence intervals for nonregular parameters

Eric B. Laber[1] **& Susan A. Murphy**[2]

[1]Department of Statistics
[1]North Carolina State University, Raleigh
[2]Department of Statistics
[2]University of Michigan, Ann Arbor

High Dimensional Problems in Statistics
Sept. 22, 2011

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 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

Two Examples

1. Confidence intervals for the test error in classification
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 $\hat{c}_{\mathcal{D}}(X) : \mathbb{R}^p \mapsto \{-1, 1\}$
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. $\hat{c}_{\mathcal{D}}(X) = sign\left(X^{\intercal}\hat{\beta}\right)$

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. $\hat{c}_{\mathcal{D}}(X) = sign\left(X^\intercal \hat{\beta}\right)$
- Review: surrogate loss function $L(X, Y, \beta)$
  - like to minimize error rate $\mathbb{P}_n 1_{Y \neq sign(X^\intercal \beta)}$
  - non-smoothness $\Rightarrow$ computational difficulty
  - replace $1_{Y \neq sign(X^\intercal \beta)} = 1_{YX^\intercal \beta < 0}$ with smooth surrogate
    - Support Vector Machines : $L(X, Y, \beta) = (1 - YX^\intercal \beta)_+ + \gamma ||\beta||^2$
    - Binomial Deviance : $L(X, Y, \beta) = log(1 + e^{-YX^\intercal \beta})$
    - Squared Error: $L(X, Y, \beta) = (1 - YX^\intercal \beta)^2$

The problem cont'd

- Test error

$$\tau(\hat{\beta}) \triangleq P1_{YX^\intercal\hat{\beta}<0} = \int 1_{yx^\intercal\hat{\beta}<0} dP(x,y)$$

The problem cont'd

- Test error

$$\tau(\hat{\beta}) \triangleq P1_{YX^\intercal\hat{\beta}<0} = \int 1_{yx^\intercal\hat{\beta}<0}\, dP(x,y)$$

- 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

$$\tau(\hat{\beta}) \triangleq P1_{YX^\intercal\hat{\beta}<0} = \int 1_{yx^\intercal\hat{\beta}<0} dP(x, y)$$

- 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

$$\tau(\hat{\beta}) \triangleq P1_{YX^\intercal\hat{\beta}<0} = \int 1_{yx^\intercal\hat{\beta}<0} dP(x,y)$$

- 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

$$P_{\mathcal{D}}\left\{\hat{l} \leq \tau(\hat{\beta}) \leq \hat{u}\right\} \geq 1-\alpha$$

The problem cont'd

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

$$P_{\mathcal{D}}\left\{\hat{l} \leq \tau(\hat{\beta}) \leq \hat{u}\right\} \geq 1-\alpha$$

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

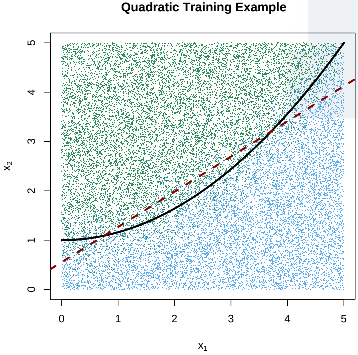
$$\begin{aligned}
\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) &\triangleq \sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}<0} \\
&= \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*=0}1_{YX^\intercal\sqrt{n}(\hat{\beta}-\beta^*)<0} \\
&\quad + \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*\neq 0}1_{YX^\intercal\hat{\beta}<0}
\end{aligned}$$

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

Illustration

Suppose

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



Properties of this example

- $P1_{X^\intercal\beta^*=0} = 0$ (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 $\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\hat{\beta}^{(b)}<0}$
- Normal approximation $\hat{\tau}(\hat{\beta}) \pm z_{1-\gamma/2}\sqrt{\frac{\hat{\tau}(\hat{\beta})(1-\hat{\tau}(\hat{\beta}))}{n}}$

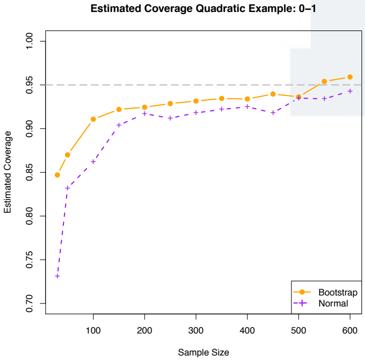
are both asymptotically valid

Illustration cont'd

Under "regular" framework

- Centered bootstrap $\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\hat{\beta}^{(b)}<0}$
- 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 $\Rightarrow$ non-regularity
- Performance inversely proportional to smoothness

Continuing our example

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

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

- $\tau_{\text{smooth}}(\hat{\beta})$ is smooth for fixed $a > 0$
- If $a \to \infty$ then $\tau_{\text{smooth}}(\hat{\beta}) \to \tau(\hat{\beta})$

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

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

- $\tau_{\text{smooth}}(\hat{\beta})$ is smooth for fixed $a > 0$
- If $a \to \infty$ then $\tau_{\text{smooth}}(\hat{\beta}) \to \tau(\hat{\beta})$
- Conjecture: Bootstrap coverage should deteriorate as a grows

Illustration cont'd

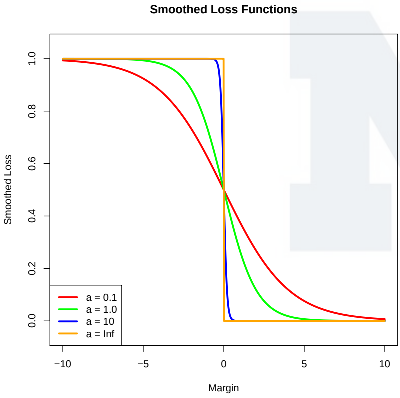
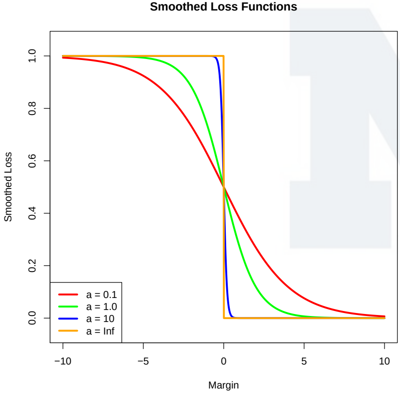
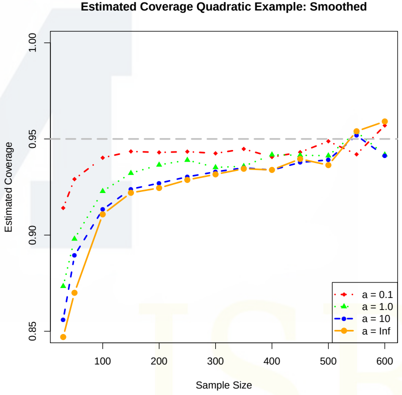


Illustration cont'd





Two Examples

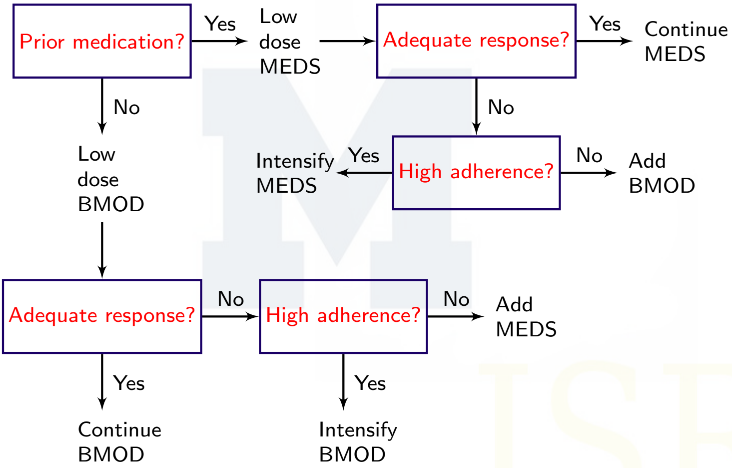
1. Confidence intervals for the test error in classification
2. Confidence intervals for parameters in optimal treatment policies

Example II: Treatment Policies

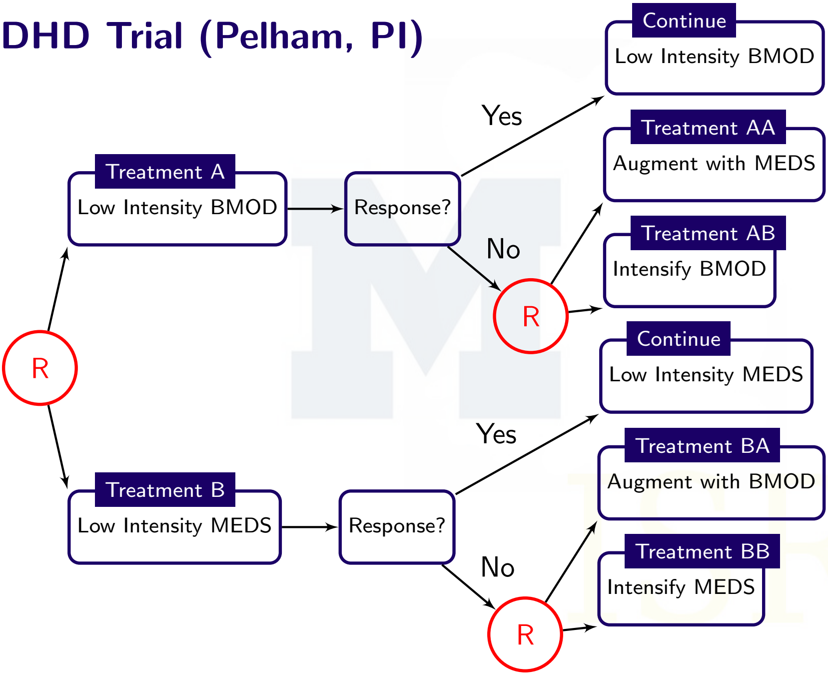
- Motivation : treatment of chronic illness
  - Some examples: HIV/AIDS, cancer, depression, schizophrenia, drug and alcohol addiction, ADHD, etc.
  - Multistage decision making problem
  - Longer-term treatment requires cumulative as opposed to myopic evaluation.
- Treatment Policies
  - Operationalize 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. ...

An Example Policy for ADHD



ADHD Trial (Pelham, PI)



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
  Y: Primary outcome (larger is better)
  H_j: History at stage j, $H_1 = X_1$, $H_2 = (X_1, A_1, X_2)$
  –Known randomization probability at stage j (usually uniform)–
- The policy, $\pi = \{\pi_1, \pi_2\}$, $\pi_j : \mathcal{H}_j \to \mathcal{A}_j$, 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 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 $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^T H_{21} + \hat{\beta}_{22}^T H_{22} A_2$
  - $\hat{\pi}_2(H_2) = \arg\max_{a_2} \hat{Q}_2(H_2, a_2) = \arg\max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$

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 $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^T H_{21} + \hat{\beta}_{22}^T H_{22} A_2$
  - $\hat{\pi}_2(H_2) = \arg\max_{a_2} \hat{Q}_2(H_2, a_2) = \arg\max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$
- $\tilde{Y} = \hat{\beta}_{21}^T H_{21} + \max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$ ($\tilde{Y}$ 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 $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^T H_{21} + \hat{\beta}_{22}^T H_{22} A_2$
  - $\hat{\pi}_2(H_2) = \arg\max_{a_2} \hat{Q}_2(H_2, a_2) = \arg\max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$
- $\tilde{Y} = \hat{\beta}_{21}^T H_{21} + \max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$ ($\tilde{Y}$ 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 $\hat{Q}_1(H_1, A_1) = \hat{\beta}_{11}^T H_{11} + \hat{\beta}_{12}^T H_{12} A_1$
  - $\hat{\pi}_1(H_{12}) = \arg\max_{a_1} \hat{Q}_1(H_1, a_1) = \arg\max_{a_1} \hat{\beta}_{12}^T H_{12} a_1$

GOAL: confidence interval for a contrast of stage 1 parameters: $c^{\intercal}\beta_1^*$

- Non-regular due to non-differentiable max operator used in Q-learning; recall
  - $\tilde{Y} = \hat{\beta}_{21}^T H_{21} + \max_{a_2} \hat{\beta}_{22}^T H_{22} a_2$
- In this setting the centered percentile bootstrap confidence interval for $c^{\intercal}\beta_1^*$ 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 $c^\intercal\sqrt{n}\hat{\beta}_1$

- Local Alternative:
  - $\beta^*_{22,n} = \beta^*_{22} + u/\sqrt{n}$
  - The limiting distribution of $c^\intercal\sqrt{n}(\hat{\beta}_1 - \beta^*_{1,n})$ is the distribution of

$$c^\intercal\Sigma_1^{-1}\left(\mathbb{W} + f(\mathbb{V}, u)\right)$$

  where

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

  and $B_1 = (H_{11}^\intercal, H_{12}^\intercal A_1)^\intercal$ (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 $\hat{\beta}_1$ is a *nonregular* estimator (unless $P\left[H_{22}^\intercal\beta^*_{22} = 0\right] = 0$)

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)
  - $\sqrt{n}(c^\intercal \hat{\hat{\beta}}_1 - c^\intercal \beta_1)$ (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)
  - $\sqrt{n}(c^{\intercal}\hat{\hat{\beta}}_1 - c^{\intercal}\beta_1)$ (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

$$\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}<0}$$

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

$$\begin{aligned}\mathbb{UB}_n \triangleq\ & \sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) \\ & - \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\intercal\hat{\beta})^2}{X^\intercal\hat{\Sigma}X}\le\lambda_n}1_{YX^\intercal\hat{\beta}<0} \\ & + \sup_{u\in\mathbb{R}^p}\sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\intercal\hat{\beta})^2}{X^\intercal\hat{\Sigma}X}\le\lambda_n}1_{YX^\intercal u<0}\end{aligned}$$

where $\hat{\Sigma} = nCov(\hat{\beta})$

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

$$\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}<0}$$

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

$$\begin{aligned}\mathbb{UB}_n \triangleq \sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta})) \\ - \sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\intercal\hat{\beta})^2}{X^\intercal\hat{\Sigma}X}\leq\lambda_n}1_{YX^\intercal\hat{\beta}<0} \\ + \sup_{u\in\mathbb{R}^p}\sqrt{n}(\mathbb{P}_n - P)1_{\frac{n(X^\intercal\hat{\beta})^2}{X^\intercal\hat{\Sigma}X}\leq\lambda_n}1_{YX^\intercal u<0}\end{aligned}$$

where $\hat{\Sigma} = nCov(\hat{\beta})$

(Replace supremum with infimum to obtain lower bound.)

Assumptions

Some technical assumptions:

(A1) $L(X, Y, \beta)$ is convex with respect to β 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 β in a neighborhood of β^*.

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

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

Properties

Theorem (Convergence)

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

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

Properties

Theorem (Convergence)

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

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

Theorem (Adaptation)

If either the Bayes decision boundary is linear or $P(X^\intercal \beta^* = 0) = 0$ *then* $\mathbb{UB}_n$ *and* $\sqrt{n}(\hat{\tau}(\hat{\beta}) - \tau(\hat{\beta}))$ *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. $\mathbb{UB}_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} \mathbb{W} + \mathbb{V}(u)$.

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 $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$.

$$
\begin{aligned}
\mathbb{UB}_n &\triangleq c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*) \\
&- c^\intercal \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left(\left[H_{22}^\intercal \mathbb{V}_n + H_{22}^\intercal u \right]_+ - \left[H_{22}^\intercal u \right]_+ \right) 1_{\frac{n(H_{22}^\intercal \hat{\beta}_{22})^2}{H_{22}^\intercal \hat{\Sigma} H_{22}} \le \lambda_n} \Big|_{u = \sqrt{n}\beta_1^*} \\
&+ \sup_u c^\intercal \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left(\left[H_{22}^\intercal \mathbb{V}_n + H_{22}^\intercal u \right]_+ - \left[H_{22}^\intercal u \right]_+ \right) 1_{\frac{n(H_{22}^\intercal \hat{\beta}_{22})^2}{H_{22}^\intercal \hat{\Sigma} H_{22}} \le \lambda_n}
\end{aligned}
$$

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

- $B_1 = (H_{11}^\intercal, H_{12}^\intercal A_1)^\intercal$
- $\mathbb{V}_n = \sqrt{n}(\hat{\beta}_{22} - \beta_{22}^*)$
- $\hat{\Sigma} = nCov(\hat{\beta}_{22})$

Adaptive CI for the treatment effect

Idea: construct smooth upper and lower bounds on $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$.

$$
\begin{aligned}
&\mathbb{UB}_n \triangleq c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*) \\
&-c^\intercal \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left(\left[H_{22}^\intercal \mathbb{V}_n + H_{22}^\intercal u \right]_+ - \left[H_{22}^\intercal u \right]_+ \right) 1_{\frac{n(H_{22}^\intercal \hat{\beta}_{22})^2}{H_{22}^\intercal \hat{\Sigma} H_{22}} \le \lambda_n} \Big|_{u=\sqrt{n}\beta_1^*} \\
&+ \sup_u c^\intercal \hat{\Sigma}_{11}^{-1} \mathbb{P}_n B_1 \left(\left[H_{22}^\intercal \mathbb{V}_n + H_{22}^\intercal u \right]_+ - \left[H_{22}^\intercal u \right]_+ \right) 1_{\frac{n(H_{22}^\intercal \hat{\beta}_{22})^2}{H_{22}^\intercal \hat{\Sigma} H_{22}} \le \lambda_n}
\end{aligned}
$$

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

- $B_1 = (H_{11}^\intercal, H_{12}^\intercal A_1)^\intercal$
- $\mathbb{V}_n = \sqrt{n}(\hat{\beta}_{22} - \beta_{22}^*)$
- $\hat{\Sigma} = nCov(\hat{\beta}_{22})$

(Replace supremum with infimum to obtain lower bound.)

Assumptions

(A1) The histories H_j with $B_j = (H_{j1}^{\intercal}, H_{j2}^{\intercal} A_j)$, $j = 1, 2$ 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. $\Sigma_j \triangleq PB_j^{\intercal} B_j$ for $j = 1, 2$;
2. $g_2(B_2, Y_2; \beta_2^*) \triangleq B_2^{\intercal}(Y_2 - B_2\beta_2^*)$;
3. $g_1\left(B_1, Y_1, H_2; \beta_1^*, \beta_2^*\right) \triangleq B_1^{\intercal}\left(H_{21}^{\intercal}\beta_{21}^* + |H_{22}^{\intercal}\beta_{22}^*| - B_1\beta_1^*\right)$;

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

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

Properties

Theorem (Convergence)

1. $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow c^\intercal \Sigma_1^{-1} \left(\mathbb{W} + f(\mathbb{V}, 0)\right)$
2. $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*) \leq \mathbb{UB}_n$ *for all n*
3. $\mathbb{UB}_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} c^\intercal \Sigma_1^{-1} \left(\mathbb{W} + f(\mathbb{V}, u)\right)$
4. $\mathbb{UB}_n^{(b)} \rightsquigarrow \sup_{u \in \mathbb{R}^p} c^\intercal \Sigma_1^{-1} \left(\mathbb{W} + f(\mathbb{V}, u)\right)$ *in probability.*

where

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

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

Properties

Theorem (Adaptation)

If $P(H_{22}^{\intercal}\beta_{22}^{*}=0)=0$ *then* $\mathbb{UB}_n$ *and* $c^{\intercal}\sqrt{n}(\hat{\beta}_1-\beta_1^{*})$ *have the same limiting distribution.*

Properties

Theorem (Adaptation)

If $P(H_{22}^{\mathsf{T}}\beta_{22}^{*} = 0) = 0$ then $\mathbb{UB}_n$ and $c^{\mathsf{T}}\sqrt{n}(\hat{\beta}_1 - \beta_1^{*})$ have the same limiting distribution.

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 for which $\beta_{22,n}^{*} = \beta_{22}^{*} + u/\sqrt{n}$,

1. $c^{\mathsf{T}}\sqrt{n}(\hat{\beta}_1 - \beta_{1n}^{*}) \rightsquigarrow c^{\mathsf{T}}\Sigma_1^{-1}(\mathbb{W} + f(\mathbb{V}, u))$
2. $\mathbb{UB}_n \rightsquigarrow \sup_{u \in \mathbb{R}^p} c^{\mathsf{T}}\Sigma_1^{-1}(\mathbb{W} + f(\mathbb{V}, u))$.

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 $\lambda_n \triangleq max(\sqrt{n}, \chi^2_{.995})$

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

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

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 $L(X, Y, \beta) = (1 - YX^{\intercal}\beta)^2$, $n = 30$

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

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 $L(X, Y, \beta) = log(1 + e^{-YX^{\intercal}\beta})$, $n = 30$

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

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 $L(X, Y, \beta) = log(1 + e^{-YX^{\intercal}\beta})$, $n = 30$

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

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 λ_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): $P(H_{22}^{\intercal}\beta_{22}^{*} = 0) > 0$
  2. Nearly non-regular (NNR) : $P(H_{22}^{\intercal}\beta_{22}^{*} \approx 0) > 0$
  3. Regular (R) : $P(H_{22}^{\intercal}\beta_{22}^{*} \approx 0) = 0$
- 1000 Monte Carlo replicatons

Results

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

2 stages 2 txts	Ex1 NR	Ex2 NNR	Ex3 NR	Ex4 R	Ex5 NR	Ex6 NNR
CPB	0.934	0.935	0.930	0.939	0.925	0.928
ST	0.948	0.945	0.938	0.919	0.759	0.762
PPE	0.931	0.940	0.938	0.931	0.904	0.903
ACI	0.992	0.992	0.968	0.950	0.964	0.965

2 stages 3 txts	Ex1 NR	Ex2 NNR	Ex3 NR	Ex4 R	Ex5 NR	Ex6 NNR
CPB	0.933	0.938	0.915	0.940	0.885	0.895
PPE	0.931	0.932	0.927	0.918	0.858	0.856
ACI	0.999	0.999	0.968	0.964	0.970	0.971

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 ACE improves on examples with increasing numbers of stages and/or treatments
- Robust to choice of λ_n

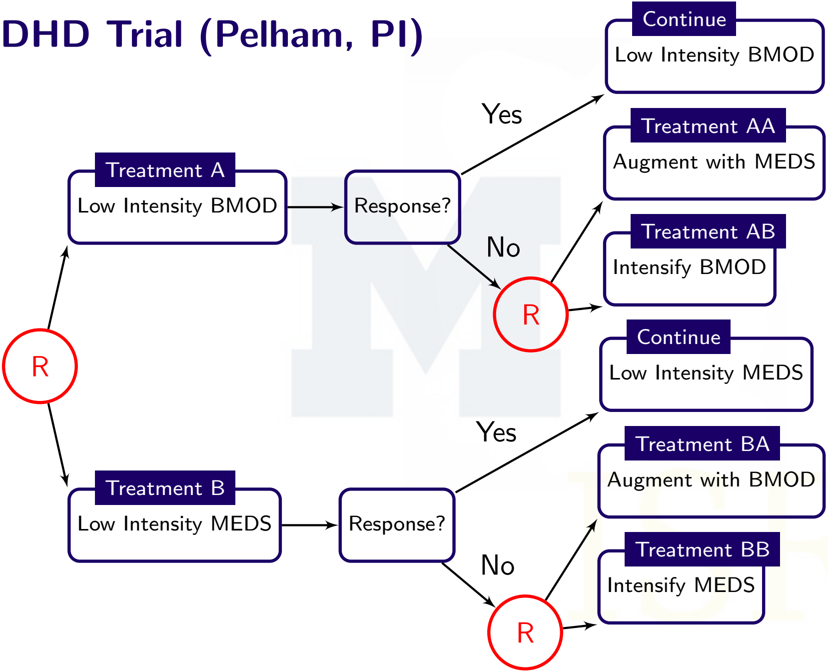
Discussion

- 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.

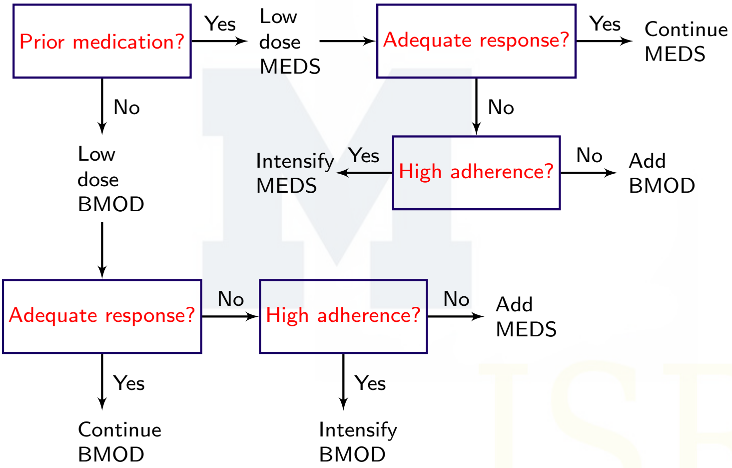
Questions: `laber@umich.edu`, `samurphy@umich.edu`
A copy of this talk can found at:
`www.stat.lsa.umich.edu/~samurphy`

Acknowledgements:
National Institute of Health grants RO1 MH080015 and P50 DA10075

ADHD Trial (Pelham, PI)



ADHD Dynamic Treatment Regime



Inference for ADHD Treatment Effects

Stage	History	Lower (5%)	Upper (95%)
1	Had prior med.	-0.51	0.14
1	No prior med.	-0.05	0.39
2	High adherence and BMOD	-0.08	0.69
2	Low adherence and BMOD	-1.10	-0.28
2	High adherence and MEDS	-0.18	0.62
2	Low adherence and MEDS	-1.25	-0.29

- 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