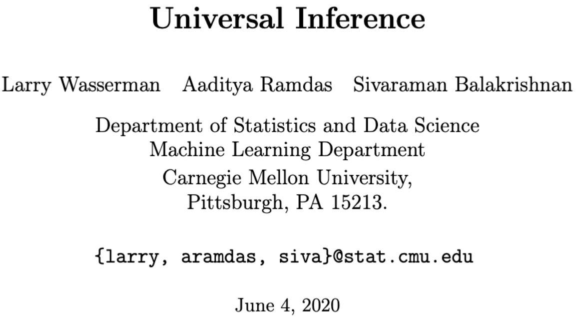
Universal Inference

Behrad Moniri

Dept. of Electrical and Systems Engineering
University of Pennsylvania
bemoniri@seas.upenn.edu

April 12, 2022

What is this presentation about?



Universal Inference

Larry Wasserman Aaditya Ramdas Sivaraman Balakrishnan

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University,
Pittsburgh, PA 15213.

{larry, aramdas, siva}@stat.cmu.edu

June 4, 2020

Published in the Proceedings of the National Academy of Sciences (PNAS).

Introduction

- **Pillars of classical statistics**: Likelihood ratio test, and confidence intervals obtained from asymptotically pivotal estimators.

Introduction

- **Pillars of classical statistics**: Likelihood ratio test, and confidence intervals obtained from asymptotically pivotal estimators.
- These methods rely on large sample asymptotic theory and this often need regularity conditions.

- **Pillars of classical statistics**: Likelihood ratio test, and confidence intervals obtained from asymptotically pivotal estimators.
- These methods rely on large sample asymptotic theory and this often need regularity conditions.
- When these conditions do not hold, there is no general method for statistical inference, with provable guarantees and these settings are typically considered in an *ad-hoc* manner.

- **One-sentence summary**:

- **One-sentence summary**:
  They propose a general method for constructing confidence sets and hypothesis tests that have **finite-sample** guarantees **without** regularity conditions.

Summary

- **One-sentence summary**:
  They propose a general method for constructing confidence sets and hypothesis tests that have **finite-sample** guarantees **without** regularity conditions. $\rightsquigarrow$ *Universal* Inference.

- **One-sentence summary**:
  They propose a general method for constructing confidence sets and hypothesis tests that have **finite-sample** guarantees **without** regularity conditions. $\rightsquigarrow$ *Universal* Inference.
- Based on a modified version of the usual likelihood ratio statistic, called "the split likelihood ratio statistics".

Summary

- **One-sentence summary**:
  They propose a general method for constructing confidence sets and hypothesis tests that have **finite-sample** guarantees **without** regularity conditions. $\rightsquigarrow$ *Universal* Inference.
- Based on a modified version of the usual likelihood ratio statistic, called "the split likelihood ratio statistics".
- They also develop various extensions of this basic methods.

Notation

- Consider a parametric family $\{P_\theta : \theta \in \Theta\}$, for some set Θ.

Notation

- Consider a parametric family $\{P_\theta : \theta \in \Theta\}$, for some set Θ.
- Assume that each distribution has density with respect to some fixed measure μ. Let the corresponding densities be p_θ.

Notation

- Consider a parametric family $\{P_\theta : \theta \in \Theta\}$, for some set Θ.
- Assume that each distribution has density with respect to some fixed measure μ. Let the corresponding densities be p_θ.
- We are given $Y_1, \ldots, Y_{2n} \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.

Notation

- Consider a parametric family $\{P_\theta : \theta \in \Theta\}$, for some set Θ.
- Assume that each distribution has density with respect to some fixed measure μ. Let the corresponding densities be p_θ.
- We are given $Y_1, \ldots, Y_{2n} \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- We want to construct confidence intervals for θ^*.

Recap: Regular Models

For regular models, we proceed as follows:

For regular models, we proceed as follows:

- If $\Theta = \mathbb{R}^d$, set

$$A_n = \left\{ \theta : 2 \log \frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\theta)} \leq c_{\alpha,d} \right\},$$

  - $c_{\alpha,d}$ is the α-quantile of a χ^2_d distribution.
  - $\mathcal{L}(\cdot)$ is the likelihood function.
  - $\hat{\theta}$ is the MLE.

For regular models, we proceed as follows:

- If $\Theta = \mathbb{R}^d$, set

$$A_n = \left\{ \theta : 2 \log \frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\theta)} \leq c_{\alpha,d} \right\},$$

  - $c_{\alpha,d}$ is the α-quantile of a χ^2_d distribution.
  - $\mathcal{L}(\cdot)$ is the likelihood function.
  - $\hat{\theta}$ is the MLE.

Wilks' Theorem (Wilks, 1938)

For regular models,

$$P_{\theta^*}\left(\theta^* \in A_n\right) \to 1 - \alpha.$$

Universal Confidence Intervals

Confidence Intervals with Split Likelihood-Ratio Statistics

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.
- Let $\hat{\theta}_1$ be **any** estimator constructed from D_1.

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.
- Let $\hat{\theta}_1$ be **any** estimator constructed from D_1.
  This can be the MLE, a Bayes estimator that utilizes prior knowledge, a robust estimator, etc.

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.
- Let $\hat{\theta}_1$ be **any** estimator constructed from D_1.
  This can be the MLE, a Bayes estimator that utilizes prior knowledge, a robust estimator, etc.
- The likelihood function based on D_0 is $\mathcal{L}_0(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$

Universal: Split Likelihood-Ratio Statistics

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.
- Let $\hat{\theta}_1$ be **any** estimator constructed from D_1.
  This can be the MLE, a Bayes estimator that utilizes prior knowledge, a robust estimator, etc.
- The likelihood function based on D_0 is $\mathcal{L}_0(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$
- Define the split likelihood ratio statistic as

$$T_n(\theta) = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)}$$

Universal: Split Likelihood-Ratio Statistics

Confidence Intervals with Split Likelihood-Ratio Statistics

- Split data into two sets D_0, D_1 randomly.
- Let $\hat{\theta}_1$ be **any** estimator constructed from D_1.
  This can be the MLE, a Bayes estimator that utilizes prior knowledge, a robust estimator, etc.
- The likelihood function based on D_0 is $\mathcal{L}_0(\theta) = \prod_{i \in D_0} p_\theta(Y_i)$
- Define the split likelihood ratio statistic as

$$T_n(\theta) = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)}$$

- The universal confidence set is

$$\mathcal{C}_n = \left\{ \theta \in \Theta : T_n(\theta) \leq \frac{1}{\alpha} \right\}$$

- If we did not split the data and $\hat{\theta}_1$ was the MLE, then $T_n(\theta)$ would have been the usual likelihood ratio statistic.

Discussion

- If we did not split the data and $\hat{\theta}_1$ was the MLE, then $T_n(\theta)$ would have been the usual likelihood ratio statistic.
- Can we prove an analog of Wilks' theorem here?

Discussion

- If we did not split the data and $\hat{\theta}_1$ was the MLE, then $T_n(\theta)$ would have been the usual likelihood ratio statistic.
- Can we prove an analog of Wilks' theorem here? The answer is yes.

Discussion

- If we did not split the data and $\hat{\theta}_1$ was the MLE, then $T_n(\theta)$ would have been the usual likelihood ratio statistic.
- Can we prove an analog of Wilks' theorem here? The answer is yes.
- Finding or approximating the distribution of the likelihood ratio statistic is highly nontrivial in irregular models. The split LRS avoids these complications.

Theorem

$\mathcal{C}_n$ is a **finite-sample** *valid $1-\alpha$ confidence set for θ^*, meaning that*

$$P_{\theta^*}(\theta^* \in \mathcal{C}_n) \geq 1-\alpha.$$

The proof is extremely simple.

Proof.

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

Proof

Proof.
Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0\left(\theta^*\right)}\right]=\mathbb{E}_{\theta^*}\left[\frac{\prod_{i \in D_0} p_\psi\left(Y_i\right)}{\prod_{i \in D_0} p_{\theta^*}\left(Y_i\right)}\right]$$

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} p_\psi(Y_i)}{\prod_{i\in D_0} p_{\theta^*}(Y_i)}\right] \\ &= \int_A \frac{\prod_{i\in D_0} p_\psi(y_i)}{\prod_{i\in D_0} p_{\theta^*}(y_i)} \prod_{i\in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n\end{aligned}$$

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$
\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} p_\psi(Y_i)}{\prod_{i\in D_0} p_{\theta^*}(Y_i)}\right] \\
&= \int_A \frac{\prod_{i\in D_0} p_\psi(y_i)}{\prod_{i\in D_0} p_{\theta^*}(y_i)} \prod_{i\in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n \\
&= \int_A \prod_{i\in D_0} p_\psi(y_i)\, dy_1 \cdots dy_n
\end{aligned}
$$

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] = \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} p_\psi(Y_i)}{\prod_{i\in D_0} p_{\theta^*}(Y_i)}\right]$$
$$= \int_A \frac{\prod_{i\in D_0} p_\psi(y_i)}{\prod_{i\in D_0} p_{\theta^*}(y_i)} \prod_{i\in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n$$
$$= \int_A \prod_{i\in D_0} p_\psi(y_i)\, dy_1 \cdots dy_n \le \prod_{i\in D_0}\left[\int p_\psi(y_i)\, dy_i\right]$$

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$
\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} p_\psi(Y_i)}{\prod_{i\in D_0} p_{\theta^*}(Y_i)}\right] \\
&= \int_A \frac{\prod_{i\in D_0} p_\psi(y_i)}{\prod_{i\in D_0} p_{\theta^*}(y_i)} \prod_{i\in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n \\
&= \int_A \prod_{i\in D_0} p_\psi(y_i)\, dy_1 \cdots dy_n \le \prod_{i\in D_0}\left[\int p_\psi(y_i)\, dy_i\right] = 1
\end{aligned}
$$

Proof.

Consider any fixed $\psi \in \Theta$ and let A denote the support of P_{θ^*}.

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\psi)}{\mathcal{L}_0(\theta^*)}\right] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i \in D_0} p_\psi(Y_i)}{\prod_{i \in D_0} p_{\theta^*}(Y_i)}\right] \\
&= \int_A \frac{\prod_{i \in D_0} p_\psi(y_i)}{\prod_{i \in D_0} p_{\theta^*}(y_i)} \prod_{i \in D_0} p_{\theta^*}(y_i)\, dy_1 \cdots dy_n \\
&= \int_A \prod_{i \in D_0} p_\psi(y_i)\, dy_1 \cdots dy_n \leq \prod_{i \in D_0}\left[\int p_\psi(y_i)\, dy_i\right] = 1
\end{aligned}$$

$\hat{\theta}_1$ is fixed when we condition on D_1. So we have

$$\mathbb{E}_{\theta^*}\left[T_n(\theta^*) \mid D_1\right] = \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{\mathcal{L}_0(\theta^*)}\middle| D_1\right] \leq 1.$$

Proof

Now, using Markov's inequality,

$$P_{\theta^*}\left(\theta^* \notin \mathcal{C}_n\right) = P_{\theta^*}\left(T_n\left(\theta^*\right) > \frac{1}{\alpha}\right) \leq \alpha \mathbb{E}_{\theta^*}\left[T_n\left(\theta^*\right)\right]$$

Now, using Markov's inequality,

$$P_{\theta^*}\left(\theta^* \notin \mathcal{C}_n\right) = P_{\theta^*}\left(T_n\left(\theta^*\right) > \frac{1}{\alpha}\right) \leq \alpha \mathbb{E}_{\theta^*}\left[T_n\left(\theta^*\right)\right]$$

$$= \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0\left(\hat{\theta}_1\right)}{\mathcal{L}_0\left(\theta^*\right)}\right] = \alpha \mathbb{E}_{\theta^*}\left(\mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{\mathcal{L}_0\left(\theta^*\right)}\middle| D_1\right]\right) \leq \alpha$$

This completes the proof.

Non-parametric Settings

- The parametric setup adopted above generalizes easily to nonparametric settings as long as we can calculate a likelihood.

- The parametric setup adopted above generalizes easily to nonparametric settings as long as we can calculate a likelihood.
- For a collection of densities $\mathcal{P}$, and a true density $p^* \in \mathcal{P}$, suppose we use D_1 to identify $\hat{p}_1 \in \mathcal{P}$, and D_0 to calculate

$$T_n(p) = \prod_{i \in D_0} \frac{\widehat{p}_1(Y_i)}{p(Y_i)}.$$

- We then define, $\mathcal{C}_n = \{p \in \mathcal{P} : T_n(p) \leq \frac{1}{\alpha}\}$, and our previous argument ensures that

$$P_{p^*}(p^* \in \mathcal{C}_n) \geq 1 - \alpha.$$

Universal Hypothesis Testing

- Let $\Theta_0 \subset \Theta$ be a null-set and consider testing

$$H_0 : \theta^* \in \Theta_0 \quad \text{versus} \quad \theta^* \notin \Theta_0$$

- **Using the duality between hypothesis testing and confidence intervals**:
  We simply reject the null hypothesis if $\mathcal{C}_n \cap \Theta_0 = \emptyset$. The type I error of this test is clearly at most α.

- Let $\Theta_0 \subset \Theta$ be a null-set and consider testing

$$H_0 : \theta^* \in \Theta_0 \quad \text{versus} \quad \theta^* \notin \Theta_0$$

- **Using the duality between hypothesis testing and confidence intervals**:
  We simply reject the null hypothesis if $\mathcal{C}_n \cap \Theta_0 = \emptyset$. The type I error of this test is clearly at most α.
- Can we find a computationally efficient way?

- Let $\hat{\theta}_1$ be any estimator constructed from D_1.

Universal Hypothesis Testing

- Let $\hat{\theta}_1$ be any estimator constructed from D_1.
- Let $\hat{\theta}_0 := \underset{\theta \in \Theta_0}{\operatorname{argmax}} \mathcal{L}_0(\theta)$ be the MLE under null from D_0.

Universal Hypothesis Testing

- Let $\hat{\theta}_1$ be any estimator constructed from D_1.
- Let $\hat{\theta}_0 := \underset{\theta \in \Theta_0}{\operatorname{argmax}} \mathcal{L}_0(\theta)$ be the MLE under null from D_0.
- Reject H_0 if

$$\frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} > \frac{1}{\alpha}.$$

- Let $\hat{\theta}_1$ be any estimator constructed from D_1.
- Let $\hat{\theta}_0 := \underset{\theta \in \Theta_0}{\text{argmax}} \mathcal{L}_0(\theta)$ be the MLE under null from D_0.
- Reject H_0 if

$$\frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} > \frac{1}{\alpha}.$$

Theorem

This test controls the type I error at level α.

Universal Hypothesis Testing

- Let $\hat{\theta}_1$ be any estimator constructed from D_1.
- Let $\hat{\theta}_0 := \underset{\theta \in \Theta_0}{\operatorname{argmax}} \mathcal{L}_0(\theta)$ be the MLE under null from D_0.
- Reject H_0 if

$$\frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)} > \frac{1}{\alpha}.$$

Theorem

This test controls the type I error at level α.

Proof.

The proof is one line.

$$P_{\theta^*}\left(\mathcal{L}_0\left(\widehat{\theta}_1\right) / \mathcal{L}_0\left(\widehat{\theta}_0\right) > 1/\alpha\right) \leq \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{\mathcal{L}_0\left(\widehat{\theta}_0\right)}\right] \leq \alpha \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{\mathcal{L}_0\left(\theta^*\right)}\right] \leq \alpha$$

Some Discussions

- **Regular models**:
  Compare the log-likelihood ratio to the $(1-\alpha)$-quantile of a χ^2 distribution (dof = dimension of null - dimension of alternative)

What are we doing?

- **Regular models**:
  Compare the log-likelihood ratio to the $(1-\alpha)$-quantile of a χ^2 distribution (dof $=$ dimension of null - dimension of alternative)
- **This paper**:
  Compare the **split**-log-split-likelihood ratio to $\log(1/\alpha) \rightsquigarrow (1-\alpha)$-quantile of a χ^2 distribution with **one** degree of freedom.

- You are only using Markov?! This isn't tight enough!

- You are only using Markov?! This isn't tight enough!
  Yes and No!

Poor Man's Chernoff Bound

- You are only using Markov?! This isn't tight enough!
  Yes and No!
- We are really using the fact that $\log \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$ has an exponential tail, just as an asymptotic argument would.

Poor Man's Chernoff Bound

- You are only using Markov?! This isn't tight enough! Yes and No!
- We are really using the fact that $\log \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$ has an exponential tail, just as an asymptotic argument would.
- In true Chernoff bounds:

$$\mathbb{E}_{\theta^*}\left[\exp\left(a \log \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}\right)\right] \leq \text{ MGF of } \chi^2, \mathcal{N}, \ldots$$

- One should view this proof as a **poor man's Chernoff bound**:

$$\mathbb{E}_{\theta^*}\left[\exp\left(\log \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}\right)\right] \leq 1$$

Sanity Check: Regular Models

Sanity Check: Gaussian Models

- Suppose that $Y_1, \ldots, Y_n \sim \mathcal{N}_d(\theta, I)$ where $\theta \in \mathbb{R}^d$.

Sanity Check: Gaussian Models

- Suppose that $Y_1, \dots, Y_n \sim \mathcal{N}_d(\theta, I)$ where $\theta \in \mathbb{R}^d$.
- Let $c_{\alpha,d}$ and z_α denote the upper α quantiles of the χ^2_d and standard Gaussian respectively.

Sanity Check: Gaussian Models

- Suppose that $Y_1, \ldots, Y_n \sim \mathcal{N}_d(\theta, I)$ where $\theta \in \mathbb{R}^d$.
- Let $c_{\alpha,d}$ and z_α denote the upper α quantiles of the χ^2_d and standard Gaussian respectively.
- The usual confidence set for θ based on the LRT can be computed as follows:
  - The likelihood function and MLE:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(Y_i - \mu)^2}{2}\right), \qquad \hat{\theta}_{MLE} = \bar{Y}$$

$$A_n = \left\{\theta : \; \|\theta - \overline{Y}\|^2 \leq \frac{c_{\alpha,d}}{n}\right\}$$
$$= \left\{\theta : \; \|\theta - \overline{Y}\|^2 \leq \frac{d + \sqrt{2d}z_\alpha + o(\sqrt{d})}{n}\right\}.$$

Sanity Check: Gaussian Models

- Denoting the sample means $\overline{Y}_1$ and $\overline{Y}_0$ we see that:

$$\log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) = -\left(\frac{n}{2}\right) \frac{\|\overline{Y}_0 - \overline{Y}_1\|^2}{2} + \left(\frac{n}{2}\right) \frac{\|\theta - \overline{Y}_0\|^2}{2}.$$

Sanity Check: Gaussian Models

- Denoting the sample means $\overline{Y}_1$ and $\overline{Y}_0$ we see that:

$$\log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) = -\left(\frac{n}{2}\right)\frac{\|\overline{Y}_0 - \overline{Y}_1\|^2}{2} + \left(\frac{n}{2}\right)\frac{\|\theta - \overline{Y}_0\|^2}{2}.$$

- The universal confidence set is

$$\begin{aligned} C_n &= \{\theta : \ \log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) \leq \log(1/\alpha)\} \\ &= \left\{\theta : \ \|\theta - \overline{Y}_0\|^2 \leq \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}. \end{aligned}$$

Sanity Check: Gaussian Models

- Denoting the sample means $\overline{Y}_1$ and $\overline{Y}_0$ we see that:

$$\log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) = -\left(\frac{n}{2}\right)\frac{\|\overline{Y}_0 - \overline{Y}_1\|^2}{2} + \left(\frac{n}{2}\right)\frac{\|\theta - \overline{Y}_0\|^2}{2}.$$

- The universal confidence set is

$$C_n = \left\{\theta : \log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) \le \log(1/\alpha)\right\}$$
$$= \left\{\theta : \|\theta - \overline{Y}_0\|^2 \le \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}.$$

- Note that $\|\overline{Y}_0 - \overline{Y}_1\|^2 = O_p(d/n)$, so both sets have radii $O_p(d/n)$.

Sanity Check: Gaussian Models

- Denoting the sample means $\overline{Y}_1$ and $\overline{Y}_0$ we see that:

$$\log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) = -\left(\frac{n}{2}\right) \frac{\|\overline{Y}_0 - \overline{Y}_1\|^2}{2} + \left(\frac{n}{2}\right) \frac{\|\theta - \overline{Y}_0\|^2}{2}.$$

- The universal confidence set is

$$\begin{aligned} C_n &= \{\theta : \ \log \mathcal{L}_0(\overline{Y}_1) - \log \mathcal{L}_0(\theta) \leq \log(1/\alpha)\} \\ &= \left\{\theta : \ \|\theta - \overline{Y}_0\|^2 \leq \frac{4}{n} \log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2 \right\}. \end{aligned}$$

- Note that $\|\overline{Y}_0 - \overline{Y}_1\|^2 = O_p(d/n)$, so both sets have radii $O_p(d/n)$.
- For constant α, the radius is four times larger.

Sanity Check: Regular Models

1. **Identifiable**: any $\theta \neq \theta^*$ it is the case that $P_\theta \neq P_{\theta^*}$.
2. Differentiable in quadratic mean **(DQM)** at θ^*: there exists a function s_{θ^*} such that:

$$\int \left[\sqrt{p_\theta} - \sqrt{p_{\theta^*}} - \frac{1}{2}(\theta - \theta^*)^T s_{\theta^*} \sqrt{p_{\theta^*}} \right]^2 d\mu = \quad o(\|\theta - \theta^*\|^2), \text{ as } \theta \to \theta^*.$$

3. The parameter space $\Theta \subset \mathbb{R}^d$ is **compact**.
4. **Smoothness**: There is a function ℓ with $\sup_\theta \mathbb{E}_{x \sim P_\theta} \ell^2(X) < \infty$ s.t.

$$\forall \theta_1, \theta_2 \in \Theta : |\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \ell(x) \|\theta_1 - \theta_2\|.$$

5. A consequence of the DQM condition is that the Fisher information matrix is well-defined, and we assume it is **non-degenerate**.

Theorem

Under the regularity conditions in the previous slide, and $||\hat{\theta}_1 - \theta^*|| = O_p(1/\sqrt{n})$*, the split LRT has diameter* $O_p(\sqrt{\log(1/\delta)/n})$

Sanity Check: Regular Models

Theorem

Under the regularity conditions in the previous slide, and $||\hat{\theta}_1 - \theta^*|| = O_p(1/\sqrt{n})$*, the split LRT has diameter* $O_p(\sqrt{\log(1/\delta)/n})$

Proof.

The high level idea: it suffices to show that for all θ sufficiently far from θ^*, we have

$$\frac{\mathcal{L}_0(\theta)}{\mathcal{L}_0(\hat{\theta}_1)} \leq \alpha.$$

□

Example of an Irregular Model

Example: Mixture Models

- Let $Y_1, \ldots, Y_{2n} \sim P$ where $Y_i \in \mathbb{R}$.
- We want to test

$$H_0 : \quad P \in \mathcal{M}_1 \;\text{ versus }\; H_1 : \quad P \in \mathcal{M}_2,$$

where $\mathcal{M}_k$ denotes the set of mixtures of k Gaussians, with an appropriately restricted parameter space Θ.

- Let $Y_1, \ldots, Y_{2n} \sim P$ where $Y_i \in \mathbb{R}$.
- We want to test

$$H_0 : \quad P \in \mathcal{M}_1 \quad \text{versus} \quad H_1 : \quad P \in \mathcal{M}_2,$$

  where $\mathcal{M}_k$ denotes the set of mixtures of k Gaussians, with an appropriately restricted parameter space Θ.
- LRT has an intractable limiting distribution. There is no known confidence set for mixture problems with guaranteed coverage properties.

Example: Mixture Models

- The true model is assumed to be $\frac{1}{2}\phi(y; -\mu, 1) + \frac{1}{2}\phi(y; \mu, 1)$
- The null: $\mu = 0$. We set $\alpha = 0.1$ and $n = 200$.
- Let $\hat{\theta}_1$ be the MLE under $\mathcal{M}_2$.
- This MLE is calculated using the EM algorithm (does it converge? IDK!)

Example: Mixture Models

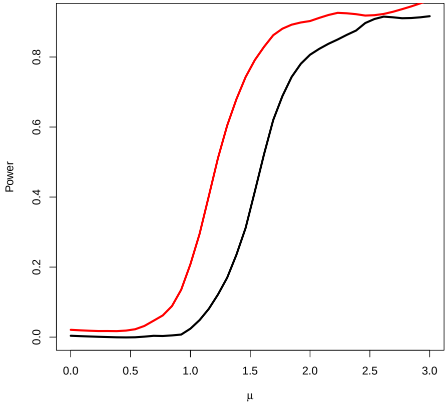


Figure: *Black = Universal / Red = Bootstrap*

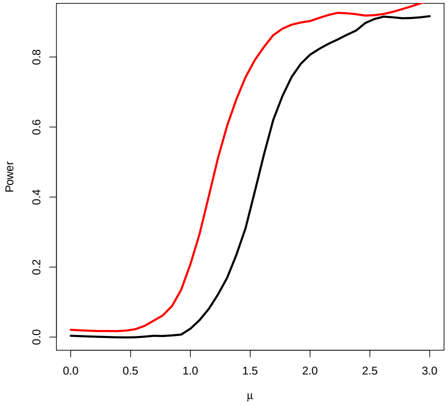


Figure: *Black* = *Universal* / *Red* = *Bootstrap*

The bootstrap test does not have any guarantee on the type I error.

Extensions

De-randomization

- The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split.

De-randomization

- The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split.
- For the test to work, we needed $\mathbb{E}_{\theta^*}[T_n] \leq 1$ where $T_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$.

De-randomization

- The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split.
- For the test to work, we needed $\mathbb{E}_{\theta^*}[T_n] \leq 1$ where $T_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$.
- Imagine that we obtained B such statistics $T_{n,1}...,\ T_{n,B}$ with the same property. Let

$$\bar{T}_n = B^{-1} \sum_{j=1}^{B} T_{n,j}.$$

Then we still have that $\mathbb{E}_{\theta^*}[\bar{T}_n]$.

- The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split.
- For the test to work, we needed $\mathbb{E}_{\theta^*}[T_n] \leq 1$ where $T_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$.
- Imagine that we obtained B such statistics $T_{n,1}$..., $T_{n,B}$ with the same property. Let

$$\bar{T}_n = B^{-1} \sum_{j=1}^{B} T_{n,j}.$$

  Then we still have that $\mathbb{E}_{\theta^*}[\bar{T}_n]$.
- K-fold and All split.

De-randomization

- The universal method involves randomly splitting the data and the final inferences will depend on the randomness of the split.
- For the test to work, we needed $\mathbb{E}_{\theta^*}[T_n] \leq 1$ where $T_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\hat{\theta}_0)}$.
- Imagine that we obtained B such statistics $T_{n,1}$..., $T_{n,B}$ with the same property. Let

$$\bar{T}_n = B^{-1} \sum_{j=1}^{B} T_{n,j}.$$

  Then we still have that $\mathbb{E}_{\theta^*}[\bar{T}_n]$.
- K-fold and All split.
- **Broader Impact**:
  These methods will potentially lead to cherry-picking :)

Upper Bounding the Likelihood

- Computing the maximum likelihood (under the null) is sometimes computationally hard.

Upper Bounding the Likelihood

- Computing the maximum likelihood (under the null) is sometimes computationally hard.
- Suppose one could come up with a relaxation F_0 of the null likelihood $\mathcal{L}_0$:

$$\max_{\theta} F_0(\theta) \geq \max_{\theta} \mathcal{L}_0(\theta).$$

Upper Bounding the Likelihood

- Computing the maximum likelihood (under the null) is sometimes computationally hard.
- Suppose one could come up with a relaxation F_0 of the null likelihood $\mathcal{L}_0$:

$$\max_{\theta} F_0(\theta) \geq \max_{\theta} \mathcal{L}_0(\theta).$$

- Define $\widehat{\theta}_0^F := \underset{\theta}{\text{argmax}} F_0(\theta)$, and consider the statistics

$$T_n' := \frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{F_0\left(\widehat{\theta}_0^F\right)}$$

Upper Bounding the Likelihood

- Computing the maximum likelihood (under the null) is sometimes computationally hard.
- Suppose one could come up with a relaxation F_0 of the null likelihood $\mathcal{L}_0$:

$$\max_\theta F_0(\theta) \geq \max_\theta \mathcal{L}_0(\theta).$$

- Define $\widehat{\theta}_0^F := \underset{\theta}{\text{argmax}} F_0(\theta)$, and consider the statistics

$$T_n' := \frac{\mathcal{L}_0\left(\widehat{\theta}_1\right)}{F_0\left(\widehat{\theta}_0^F\right)}$$

- then the split LRT may proceed using T' instead of T. This is because $F(\widehat{\theta}_0^F) \geq \mathcal{L}(\hat{\theta})$, and hence $T_n' \leq T_n$.

Smoothed Likelihood

- Sometimes the MLE may not exist since the likelihood function is unbounded. A smoothed likelihood has been proposed as an alternative.

Smoothed Likelihood

- Sometimes the MLE may not exist since the likelihood function is unbounded. A smoothed likelihood has been proposed as an alternative.
- Consider a kernel $k(x, y)$ such that $\int k(x, y)dy = 1$ for any x.

$$\widetilde{p}_\theta(y) := \int k(x, y) p_\theta(x) dx.$$

Smoothed Likelihood

- Sometimes the MLE may not exist since the likelihood function is unbounded. A smoothed likelihood has been proposed as an alternative.
- Consider a kernel $k(x, y)$ such that $\int k(x, y)dy = 1$ for any x.

$$\widetilde{p}_\theta(y) := \int k(x, y) p_\theta(x) dx.$$

- Denote the smoothed empirical density based on D_0 as

$$\widetilde{p}_n := \frac{1}{|D_0|} \sum_{i \in D_0} k(X_i, \cdot).$$

Smoothed Likelihood

- Sometimes the MLE may not exist since the likelihood function is unbounded. A smoothed likelihood has been proposed as an alternative.
- Consider a kernel $k(x, y)$ such that $\int k(x, y)dy = 1$ for any x.

$$\widetilde{p}_\theta(y) := \int k(x, y) p_\theta(x) dx.$$

- Denote the smoothed empirical density based on D_0 as

$$\widetilde{p}_n := \frac{1}{|D_0|} \sum_{i \in D_0} k(X_i, \cdot).$$

- Define the smoothed likelihood on D_0 as

$$\widetilde{\mathcal{L}}_0(\theta) := \prod_{i \in D_0} \exp \int k(X_i, y) \log \widetilde{p}_\theta(y) dy \rightsquigarrow \widetilde{\theta}_0 := \arg\min_{\theta \in \Theta_0} KL(\widetilde{p}_n, \widetilde{p}_\theta)$$

- As before, let $\widehat{\theta}_1 \in \Theta$ be any estimator based on D_1. The smoothed split LRT:

$$\text{reject } H_0 \text{ if } \widetilde{U}_n > 1/\alpha, \text{ where } \widetilde{U}_n = \frac{\widetilde{\mathcal{L}}_0(\widehat{\theta}_1)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}.$$

- As before, let $\widehat{\theta}_1 \in \Theta$ be any estimator based on D_1. The smoothed split LRT:

$$\text{reject } H_0 \text{ if } \widetilde{U}_n > 1/\alpha, \text{ where } \widetilde{U}_n = \frac{\widetilde{\mathcal{L}}_0(\widehat{\theta}_1)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}.$$

Fix $\psi \in \Theta$, we have

$$\mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\widetilde{\theta}_0)}\right] \overset{\text{(i)}}{\leq} \mathbb{E}_{\theta^*}\left[\frac{\widetilde{\mathcal{L}}_0(\psi)}{\widetilde{\mathcal{L}}_0(\theta^*)}\right] = \mathbb{E}_{\theta^*}\left[\frac{\prod_{i\in D_0} \exp \int k(X_i, y) \log \widetilde{p}_\psi(y)dy}{\prod_{i\in D_0} \exp \int k(X_i, y) \log \widetilde{p}_{\theta^*}(y)dy}\right]$$

$$= \prod_{i\in D_0} \int \exp\left(\int k(x, y) \log \frac{\widetilde{p}_\psi(y)}{\widetilde{p}_{\theta^*}(y)} dy\right) p_{\theta^*}(x)dx \leq \cdots \leq 1.$$

Sequential Testing

Sequential Split Likelihood Ratio Test

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.
- Let $\widehat{\theta}_{1,t-1}$ be any *non-anticipating* estimator based on the first $t-1$ samples.

Sequential Split Likelihood Ratio Test

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.
- Let $\widehat{\theta}_{1,t-1}$ be any *non-anticipating* estimator based on the first $t-1$ samples.
- Denote the null MLE as $\widehat{\theta}_{0,t} = \arg\max_{\theta \in \Theta_0} \prod_{i=1}^{t} p_\theta(Y_i)$.

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.
- Let $\widehat{\theta}_{1,t-1}$ be any *non-anticipating* estimator based on the first $t-1$ samples.
- Denote the null MLE as $\widehat{\theta}_{0,t} = \arg\max_{\theta \in \Theta_0} \prod_{i=1}^{t} p_\theta(Y_i)$.
- At any time t, reject the null and stop if

$$M_t := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\widehat{\theta}_{0,t}}(Y_i)} > 1/\alpha.$$

Sequential Split Likelihood Ratio Test

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.
- Let $\widehat{\theta}_{1,t-1}$ be any *non-anticipating* estimator based on the first $t-1$ samples.
- Denote the null MLE as $\widehat{\theta}_{0,t} = \arg\max_{\theta \in \Theta_0} \prod_{i=1}^{t} p_\theta(Y_i)$.
- At any time t, reject the null and stop if

$$M_t := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\widehat{\theta}_{0,t}}(Y_i)} > 1/\alpha.$$

- Let τ_θ denote the stopping time when the data is drawn from P_θ.

- Consider the following, standard, sequential testing/estimation setup:
- We observe an i.i.d. sequence $Y_1, Y_2, \ldots$ from P_{θ^*}.
- Let $\widehat{\theta}_{1,t-1}$ be any *non-anticipating* estimator based on the first $t-1$ samples.
- Denote the null MLE as $\widehat{\theta}_{0,t} = \arg\max_{\theta \in \Theta_0} \prod_{i=1}^{t} p_\theta(Y_i)$.
- At any time t, reject the null and stop if

$$M_t := \frac{\prod_{i=1}^{t} p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^{t} p_{\widehat{\theta}_{0,t}}(Y_i)} > 1/\alpha.$$

- Let τ_θ denote the stopping time when the data is drawn from P_θ.

Theorem

The running MLE LRT has type I error at most α, meaning that $\sup_{\theta^* \in \Theta_0} P_{\theta^*}(\tau_{\theta^*} < \infty) \leq \alpha$.

Proof

- For M_t we can write:

$$M_t := \frac{\prod_{i=1}^t p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^t p_{\widehat{\theta}_{0,t}}(Y_i)} \leq \underbrace{\frac{\prod_{i=1}^t p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^t p_{\theta^*}(Y_i)}}_{L_t} = L_{t-1} \frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)}.$$

Proof

- For M_t we can write:

$$M_t := \frac{\prod_{i=1}^t p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^t p_{\widehat{\theta}_{0,t}}(Y_i)} \leq \underbrace{\frac{\prod_{i=1}^t p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^t p_{\theta^*}(Y_i)}}_{L_t} = L_{t-1}\frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)}.$$

- It is easy to verify that L_t is a nonnegative super-martingale with respect to the natural filtration $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$:

$$\mathbb{E}_{\theta^*}[L_t|\mathcal{F}_{t-1}] = \mathbb{E}_{\theta^*}\left[\frac{\prod_{i=1}^t p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^t p_{\theta^*}(Y_i)} \,\middle|\, \mathcal{F}_{t-1}\right]$$

$$= L_{t-1}\mathbb{E}_{\theta^*}\left[\frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)} \,\middle|\, \mathcal{F}_{t-1}\right] \leq L_{t-1} \rightsquigarrow \text{Super-Martingale}$$

Proof

- For M_t we can write:

$$M_t := \frac{\prod_{i=1}^t p_{\widehat{\theta}_{1,i-1}}(Y_i)}{\prod_{i=1}^t p_{\widehat{\theta}_{0,t}}(Y_i)} \leq \underbrace{\frac{\prod_{i=1}^t p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^t p_{\theta^*}(Y_i)}}_{L_t} = L_{t-1} \frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)}.$$

- It is easy to verify that L_t is a nonnegative super-martingale with respect to the natural filtration $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$:

$$\begin{aligned}
\mathbb{E}_{\theta^*}[L_t|\mathcal{F}_{t-1}] &= \mathbb{E}_{\theta^*}\left[\frac{\prod_{i=1}^t p_{\widehat{\theta}_{i-1}}(Y_i)}{\prod_{i=1}^t p_{\theta^*}(Y_i)} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= L_{t-1}\mathbb{E}_{\theta^*}\left[\frac{p_{\widehat{\theta}_{t-1}}(Y_t)}{p_{\theta^*}(Y_t)} \,\middle|\, \mathcal{F}_{t-1}\right] \leq L_{t-1} \rightsquigarrow \text{Super-Martingale}
\end{aligned}$$

- Now we proceed as follows:

$$\begin{aligned}
P_{\theta^*}(\exists t \in \mathbb{N} : M_t > 1/\alpha) &\leq P_{\theta^*}(\exists t \in \mathbb{N} : L_t > 1/\alpha) \\
&\overset{(\star)}{\leq} \mathbb{E}_{\theta^*}[L_0] \cdot \alpha = \alpha,
\end{aligned}$$

(⋆) Ville's Inequality

Theorem [Ville (1939)]

For any nonnegative supermartingale L_t and any $x > 1$, we have

$$\mathbb{P}[\exists t : L_t \geq x] \leq \frac{\mathbb{E}[L_0]}{x}$$

Proof.

The idea is to consider the following stopping time

$$N = \inf\{t \geq 1 : L_t \geq x\},$$

and use the optional stopping time theorem. □

Conclusion

Conclusion

- Inference based on the split likelihood ratio statistic (and variants) leads to simple tests and confidence sets with finite-sample guarantees.

Conclusion

- Inference based on the split likelihood ratio statistic (and variants) leads to simple tests and confidence sets with finite-sample guarantees.
- These methods are most useful in problems where standard asymptotic methods are difficult/impossible to apply.

- **Going forward:** Optimality? Power of the Test?
  How does the choice of $\hat{\theta}_1$ affect the power of the test?

Thank You!