High-Dimensional Inference With No Assumptions

Texas A & M, 2016

Larry Wasserman
Carnegie Mellon University

With: Jing Lei, Max G’Sell, Alessandro Rinaldo and Ryan Tibshirani
In the olden days, we made lots of assumptions.
In the olden days, we made lots of assumptions. These assumptions were checkable.
Assumptions

In the olden days, we made lots of assumptions. These assumptions were checkable. We derived lots of strong results.
Assumptions

In the olden days, we made lots of assumptions. These assumptions were checkable. We derived lots of strong results.

Today, with high-dimensional problems, we have it backwards.
Assumptions

In the olden days, we made lots of assumptions. These assumptions were checkable. We derived lots of strong results.

Today, with high-dimensional problems, we have it backwards. We derive lots of strong results.
Assumptions

In the olden days, we made lots of assumptions. These assumptions were checkable. We derived lots of strong results.

Today, with high-dimensional problems, we have it backwards. We derive lots of strong results. Then we add strong (uncheckable) assumptions to prove the results.
High Dimensional Regression

Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
where \(Y_i \in \mathbb{R}\) and

\[X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d.\]
Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
where \(Y_i \in \mathbb{R}\) and

\[X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d.\]

We let \(d = d_n\) increase with \(n\).
Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
where \(Y_i \in \mathbb{R}\) and
\[
X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d.
\]

We let \(d = d_n\) increase with \(n\).
New pair \((X, Y)\).
High Dimensional Regression

Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
where \(Y_i \in \mathbb{R}\) and

\[ X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d. \]

We let \(d = d_n\) increase with \(n\).
New pair \((X, Y)\).
Predict \(Y\) from \(X\).
Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
where \(Y_i \in \mathbb{R}\) and
\[
X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d.
\]

We let \(d = d_n\) increase with \(n\).
New pair \((X, Y)\).
Predict \(Y\) from \(X\).
Inference: how important is each covariate (feature)?
Usual Assumptions for High Dimensional Regression

1. \( Y = X \beta + \epsilon \)
2. \( \beta \) is sparse
3. \( \text{Var}(Y|X) \) is constant.
4. The design matrix is nice (incoherent/incompatible/restricted eigenvalue etc)
5. Minimum non-zero beta \( \text{something} \)
6. \( \epsilon \) is Normal
7. \( X \) is fixed.

We call these the Kool-Aid Assumptions.
Usual Assumptions for High Dimensional Regression

1. \( Y = X\beta + \epsilon \)
2. \( \beta \) is sparse
3. \( \text{Var}(Y|X) \) is constant.
4. The design matrix is nice (incoherent/incompatible/restricted eigenvalue etc)
5. Minimum non-zero beta \( > \) something
6. \( \epsilon \) is Normal
7. \( X \) is fixed.
Usual Assumptions for High Dimensional Regression

1. $Y = X\beta + \epsilon$
2. $\beta$ is sparse
3. $\text{Var}(Y|X)$ is constant.
4. The design matrix is nice (incoherent/incompatible/restricted eigenvalue etc)
5. Minimum non-zero beta $> \text{something}$
6. $\epsilon$ is Normal
7. $X$ is fixed.

We call these the Kool-Aid Assumptions.
These assumptions are unrealistic.
Usual Assumptions

These assumptions are unrealistic.
They came from trying to import our intuition and results from low dimensions to high dimensions.
Usual Assumptions

These assumptions are unrealistic. They came from trying to import our intuition and results from low dimensions to high dimensions.

1. They never hold (except in signal processing).
Usual Assumptions

These assumptions are unrealistic. They came from trying to import our intuition and results from low dimensions to high dimensions.

1. They never hold (except in signal processing).
2. They are untestable.
Usual Assumptions

These assumptions are unrealistic. They came from trying to import our intuition and results from low dimensions to high dimensions.

1. They never hold (except in signal processing).
2. They are untestable.
3. Procedures based on these assumptions are fragile.
Usual Assumptions

These assumptions are unrealistic. They came from trying to import our intuition and results from low dimensions to high dimensions.

1. They never hold (except in signal processing).
2. They are untestable.
3. Procedures based on these assumptions are fragile.

Luckily, we do not need to make any of these assumptions.
There are many approaches that are assumption free. We’ll consider two:

1. Conformal prediction.
2. Sample splitting.
There are many approaches that are assumption free. We’ll consider two:

1. Conformal prediction.
2. Sample splitting.

These are not new. We are taking a modern view of (somewhat) old ideas.
There are many approaches that are assumption free. We’ll consider two:

1. Conformal prediction.
2. Sample splitting.

These are not new. We are taking a modern view of (somewhat) old ideas.

An extension: IDA (interactive data analysis), a.k.a. adaptive data analysis. Many looks at the data. Same problem (but harder).
Outline

1. Conformal prediction Probability = 1
2. LOCO (Leave-Out-One-Covariate) Probability = .9
3. Inference Via Sample Splitting Probability = .3
4. Fragility Probability = .1
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

1. Fix $y$. We will test:

   $H_0: Y_{n+1} = y$.

2. Form augmented data $(Y_1, \ldots, Y_n, Y_{n+1})$ where $Y_{n+1} = y$.

3. Compute scores $R_1, \ldots, R_{n+1}$ where $R_i = R_i(Y_1, \ldots, Y_n, Y_{n+1})$.

   Example: $|Y_i - y|$.

4. Test $H_0$: $Y_{n+1} = y$.

   The p-value is $p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.

   Under $H_0$: $Y_{n+1} = y$, this is (discrete) Uniform (0,1).

5. Invert: $C_n(y) = \{y: p(y) \geq \alpha\}$.
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe \( Y_1, \ldots, Y_n \)

Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).

\[
\begin{align*}
1. \text{Fix } y. \text{ We will test: } H_0 : Y_{n+1} = y.
\end{align*}
\]
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).
Basic idea: Observe
\[ Y_1, \ldots, Y_n \]
Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).
2. Form augmented data \( (Y_1, \ldots, Y_n, Y_{n+1}) \) where \( Y_{n+1} = y \).
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe $Y_1, \ldots, Y_n$

Predict new $Y_{n+1}$.

1. Fix $y$. We will test: $H_0 : Y_{n+1} = y$.
2. Form augmented data $(Y_1, \ldots, Y_n, Y_{n+1})$ where $Y_{n+1} = y$.
3. Compute scores $R_1, \ldots, R_{n+1}$ where $R_i = R_i(Y_1, \ldots, Y_{n+1})$.
   Example: $|Y_i - \bar{Y}_y|$.
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe

\[ Y_1, \ldots, Y_n \]

Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).

2. Form augmented data \((Y_1, \ldots, Y_n, Y_{n+1})\) where \( Y_{n+1} = y \).

3. Compute scores \( R_1, \ldots, R_{n+1} \) where \( R_i = R_i(Y_1, \ldots, Y_{n+1}) \).
   Example:
   \[ |Y_i - \bar{Y}_y|. \]

4. Test \( H_0 : Y_{n+1} = y \).
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe

\[ Y_1, \ldots, Y_n \]

Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).
2. Form augmented data \( (Y_1, \ldots, Y_n, Y_{n+1}) \) where \( Y_{n+1} = y \).
3. Compute scores \( R_1, \ldots, R_{n+1} \) where \( R_i = R_i(Y_1, \ldots, Y_{n+1}) \).
   Example:
   \[ |Y_i - \bar{Y}_y|. \]
4. Test \( H_0 : Y_{n+1} = y \).
   The p-value is \( p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}) \).
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe

\[ Y_1, \ldots, Y_n \]

Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).
2. Form augmented data \((Y_1, \ldots, Y_n, Y_{n+1})\) where \( Y_{n+1} = y \).
3. Compute scores \( R_1, \ldots, R_{n+1} \) where \( R_i = R_i(Y_1, \ldots, Y_{n+1}) \).
   Example:
   \[ |Y_i - \bar{Y}_y|. \]
4. Test \( H_0 : Y_{n+1} = y \).
   The p-value is \( p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}) \).
   Under \( H_0 : Y_{n+1} = y \), this is (discrete) Uniform \((0,1)\).
Conformal Inference

Conformal inference was invented by Vovk et al (1990’s).

Basic idea: Observe

\[ Y_1, \ldots, Y_n \]

Predict new \( Y_{n+1} \).

1. Fix \( y \). We will test: \( H_0 : Y_{n+1} = y \).

2. Form augmented data \( (Y_1, \ldots, Y_n, Y_{n+1}) \) where \( Y_{n+1} = y \).

3. Compute scores \( R_1, \ldots, R_{n+1} \) where \( R_i = R_i(Y_1, \ldots, Y_{n+1}) \).
Example:

\[ |Y_i - \bar{Y}_y|. \]

4. Test \( H_0 : Y_{n+1} = y \).

The p-value is \( p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}) \).
Under \( H_0 : Y_{n+1} = y \), this is (discrete) Uniform (0,1).

5. Invert:

\[ C_n(y) = \{ y : p(y) \geq \alpha \}. \]
For any $P$ and any $n$,

$$P(Y_{n+1} \in C_n) \geq 1 - \alpha.$$ 

Distribution-free, finite-sample.
Validity

For any $P$ and any $n$,

$$P(Y_{n+1} \in C_n) \geq 1 - \alpha.$$  

Distribution-free, finite-sample.

Vovk and his colleagues have many papers with different versions and interesting applications.
Density Estimation

Lei, Robins and Wasserman (2013), JASA.
Let $\hat{p}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K \left( \frac{\|y - Y_i\|}{h} \right)$.

Let $R_i = 1_{\hat{p}_h(y) \leq 1}$.

Let $C_n = \{y : \pi(y) \geq \alpha\}$ where $\pi(y) = \frac{1}{n + 1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.

Then $P(Y_{n+1} \in C_n) \geq 1 - \alpha$. 

Lei, Robins and Wasserman (2013), JASA.
Let \( \hat{p}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K \left( \frac{||y - Y_i||}{h} \right) \).

Let \( R_i = \frac{1}{\hat{p}_{h,y}(Y_i)} \). (Using the augmented data.)
Lei, Robins and Wasserman (2013), JASA.

Let \( \hat{p}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K \left( \frac{||y - Y_i||}{h} \right) \).

Let \( R_i = \frac{1}{\hat{p}_{h,y}(Y_i)} \). (Using the augmented data.)

Let \( C_n = \{ y : \pi(y) \geq \alpha \} \) where

\[
\pi(y) = \frac{1}{n + 1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}).
\]
Let $\hat{p}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K \left( \frac{||y - Y_i||}{h} \right)$.

Let $R_i = \frac{1}{\hat{p}_{h, y}(Y_i)}$. (Using the augmented data.)

Let $C_n = \{ y : \pi(y) \geq \alpha \}$ where

$$\pi(y) = \frac{1}{n + 1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1}).$$

Then

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha.$$
A Speedup

Let $Z_i = \hat{p}(Y_i)$ and

$$Z(1) \leq \cdots Z(n).$$
A Speedup

Let $Z_i = \hat{p}(Y_i)$ and

$$Z(1) \leq \cdots Z(n).$$

$$t = Z(r) - \frac{K(0)}{nh^d}$$

where $r = \lceil \alpha (n + 1) \rceil$. 

No augmentation involved.
A Speedup

Let \( Z_i = \hat{p}(Y_i) \) and

\[
Z(1) \leq \cdots Z(n).
\]

\[
t = Z(r) - \frac{K(0)}{nh^d}
\]

where \( r = \lfloor \alpha(n + 1) \rfloor \).

\[
C_n^+ = \{ y : \hat{p}(y) \geq t \}.
\]
A Speedup

Let $Z_i = \hat{p}(Y_i)$ and

$$Z(1) \leq \cdots Z(n).$$

$$t = Z(r) - \frac{K(0)}{nh^d}$$

where $r = \lfloor \alpha(n + 1) \rfloor$.

$$C_n^+ = \{y : \hat{p}(y) \geq t\}.$$  

Then $C_n \subset C_n^+$ and so

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha.$$
A Speedup

Let $Z_i = \hat{p}(Y_i)$ and $Z(1) \leq \cdots Z(n)$.

$$t = Z(r) - \frac{K(0)}{nh^d}$$

where $r = \lfloor \alpha(n + 1) \rfloor$.

$$C_n^+ = \{ y : \hat{p}(y) \geq t \}.$$

Then $C_n \subset C_n^+$ and so

$$\mathbb{P}(Y_{n+1} \in C_n) \geq 1 - \alpha.$$

No augmentation involved.
Minimaxity

Conformalized methods inherit the minimax properties of the estimator.
Conformalized methods inherit the minimax properties of the estimator.

If \( p \) smooth: \( C_n \) is minimax. If \( p \) is not smooth: \( C_n \) still has valid coverage.
Conformalized methods inherit the minimax properties of the estimator.

If $p$ smooth: $C_n$ is minimax. If $p$ is not smooth: $C_n$ still has valid coverage.

Let $C_\alpha$ be the oracle, i.e. the set with smallest Lebesgue measure such that $P(Y_{n+1} \in C_\alpha) \geq 1 - \alpha$. Then $C_\alpha = \{y : p(y) \geq t_\alpha\}$. 
Minimaxity

Conformalized methods inherit the minimax properties of the estimator.

If $p$ smooth: $C_n$ is minimax. If $p$ is not smooth: $C_n$ still has valid coverage.

Let $C_\alpha$ be the oracle, i.e. the set with smallest Lebesgue measure such that $P(Y_{n+1} \in C_\alpha) \geq 1 - \alpha$. Then $C_\alpha = \{y : p(y) \geq t_\alpha\}$.

We want $\mu(C_n \Delta C_\alpha)$ to be as small as possible where $\mu$ is Lebesgue measure. Here, $\Delta$ denotes the set difference.
Minimaxity

(A1) Suppose $p$ is $\beta$-Holder smooth.
(A2) Suppose

$$c_1|\epsilon|^\gamma \leq |P(\{y : p(y) \leq t_\alpha + \epsilon\}) - \alpha| \leq c_2|\epsilon|^\gamma.$$
Minimaxity

(A1) Suppose $p$ is $\beta$-Holder smooth.

(A2) Suppose

$$c_1|\epsilon|^\gamma \leq |P\{y : p(y) \leq t_\alpha + \epsilon\} - \alpha| \leq c_2|\epsilon|^\gamma.$$ 

Then, for all $\lambda > 0$,

$$\mathbb{P}\left(\mu(C_n \Delta C_\alpha) \geq \left(\frac{\log n}{n}\right)^{\beta \gamma/(2\beta + d)} + \left(\frac{\log n}{n}\right)^{1/2}\right) \leq \left(\frac{1}{n}\right)^\lambda.$$
Minimaxity

(A1) Suppose $p$ is $\beta$-Holder smooth.

(A2) Suppose

$$c_1|\epsilon|^{\gamma} \leq |P(\{y : p(y) \leq t_\alpha + \epsilon\}) - \alpha| \leq c_2|\epsilon|^{\gamma}.$$ 

Then, for all $\lambda > 0$,

$$\mathbb{P} \left( \mu(C_n \Delta C_\alpha) \geq \left( \frac{\log n}{n} \right)^{\beta \gamma/(2\beta+d)} + \left( \frac{\log n}{n} \right)^{1/2} \right) \leq \left( \frac{1}{n} \right)^{\lambda}.$$ 

Same for outer approximation.
Minimaxity

(A1) Suppose $p$ is $\beta$-Holder smooth.

(A2) Suppose

$$c_1|\epsilon|^\gamma \leq |P(\{y : p(y) \leq t_\alpha + \epsilon\}) - \alpha| \leq c_2|\epsilon|^\gamma.$$

Then, for all $\lambda > 0$,

$$\mathbb{P}\left(\mu(C_n \Delta C_\alpha) \geq \left(\frac{\log n}{n}\right)^{\beta\gamma/(2\beta+d)} + \left(\frac{\log n}{n}\right)^{1/2}\right) \leq \left(\frac{1}{n}\right)^{\lambda}.$$

Same for outer approximation.

But ... even with no conditions, we still have coverage (even if $P$ does not have a density).
Minimaxity

(A1) Suppose $p$ is $\beta$-Holder smooth.
(A2) Suppose

$$c_1|\epsilon|^\gamma \leq |P\{y : p(y) \leq t_\alpha + \epsilon\} - \alpha| \leq c_2|\epsilon|^\gamma.$$

Then, for all $\lambda > 0$,

$$\mathbb{P}\left(\mu(C_n \Delta C_\alpha) \geq \left(\frac{\log n}{n}\right)^{\beta \gamma/(2\beta + d)} + \left(\frac{\log n}{n}\right)^{1/2}\right) \leq \left(\frac{1}{n}\right)^\lambda.$$

Same for outer approximation.
But ... even with no conditions, we still have coverage (even if $P$ does not have a density).
Can also use size of $C_n$ to choose the bandwidth.
Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).

Want \(P(Y \in C_n(x, y)) \geq 1 - \alpha\) for all \(P\) and all \(n\).

Construct nonparametric joint region \(C_n(X, Y)\) using previous method.

Take \(C_n(x) = \{ y : (x, y) \in C_n(x, y) \}\). Again, there is a fast approximation.
Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).
Nonparametric Regression

Lei and Wasserman 2014, JRSS B.

Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).

Want

\[
P(Y \in C_n(X)) \geq 1 - \alpha
\]

for all \(P\) and all \(n\).
Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).

Want

\[
P(Y \in C_n(X)) \geq 1 - \alpha
\]

for all \(P\) and all \(n\).

Construct nonparametric joint region \(C_n(X, Y)\) using previous method.
Nonparametric Regression

Lei and Wasserman 2014, JRSS B.

Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).

Want

\[
\mathbb{P}(Y \in C_n(X)) \geq 1 - \alpha
\]

for all \(P\) and all \(n\).

Construct nonparametric joint region \(C_n(X, Y)\) using previous method.

Take

\[
C_n(x) = \{y : (x, y) \in C_n(x, y)\}.
\]
Nonparametric Regression

Lei and Wasserman 2014, JRSS B.

Observe \((X_1, Y_1), \ldots, (X_n, Y_n)\) and a new \(X\). Predict \(Y\).

Want

\[
P(Y \in C_n(X)) \geq 1 - \alpha
\]

for all \(P\) and all \(n\).

Construct nonparametric joint region \(C_n(X, Y)\) using previous method.

Take

\[
C_n(x) = \{ y : (x, y) \in C_n(x, y) \}.
\]

Again, there is a fast approximation.
Minimaxity

\[ \mathbb{P} \left( \sup_x \mu \left( C_n(x) \Delta C_\alpha(x) \right) > C r_n^{\gamma} \right) \leq \left( \frac{1}{n} \right)^{\lambda} \]

where

\[ r_n = \left( \frac{\log n}{n} \right)^{\frac{\beta}{\beta(d+2)+1}}. \]
Minimaxity

\[ \mathbb{P}\left( \sup_{x} \mu(C_n(x) \Delta C_{\alpha}(x)) > C r_n^\gamma \right) \leq \left( \frac{1}{n} \right)^\lambda \]

where

\[ r_n = \left( \frac{\log n}{n} \right)^{\frac{\beta}{\beta(d+2)+1}}. \]

This is minimax.
(The rate is non-standard because we require different smoothness for \( x \) and \( y \) to get local validity for \( x \)).
Bandwidth

Lebesgue Measure
As before: \((X_1, Y_1), \ldots, (X_n, Y_n)\) but now \(X_i = (X_i^{(1)}, \ldots, X_i^{(d)})\) where \(d\) can be large (increasing with \(n\)). No assumptions except iid.

Predictor: \(\hat{\mu}(x)\).

Linear case: \(\hat{\mu}(x) = \hat{\beta}^T x\).
With: Max G'Sell, Jing Lei, Alessandro Rinaldo and Ryan Tibshirani.

As before: \((X_1, Y_1), \ldots, (X_n, Y_n)\) but now

\[ X_i = (X_i(1), \ldots, X_i(d)) \]

where \(d\) can be large (increasing with \(n\)).
With: Max G'Sell, Jing Lei, Alessandro Rinaldo and Ryan Tibshirani.

As before: \((X_1, Y_1), \ldots, (X_n, Y_n)\) but now

\[ X_i = (X_i(1), \ldots, X_i(d)) \]

where \(d\) can be large (increasing with \(n\)).

No assumptions except iid.
High Dimensional Regression: (Work In Progress)

With: Max G'Sell, Jing Lei, Alessandro Rinaldo and Ryan Tibshirani.

As before: \((X_1, Y_1), \ldots, (X_n, Y_n)\) but now

\[ X_i = (X_i(1), \ldots, X_i(d)) \]

where \(d\) can be large (increasing with \(n\)).

No assumptions except iid.

Predictor: \(\hat{\mu}(x)\).
With: Max G'Sell, Jing Lei, Alessandro Rinaldo and Ryan Tibshirani.

As before: \((X_1, Y_1), \ldots, (X_n, Y_n)\) but now

\[ X_i = (X_i(1), \ldots, X_i(d)) \]

where \(d\) can be large (increasing with \(n\)).

No assumptions except iid.

Predictor: \(\hat{\mu}(x)\).

Linear case: \(\hat{\mu}(x) = \hat{\beta}^T x\).
High Dimensional Regression

Goals:

(1) accurate prediction: $\|Y - \hat{\mu}(X)\|^2$.

(2) valid predictive inference: $P(Y \in C_n(X)) \geq 1 - \alpha$ for all $P$ and $n$, even when we use model selection (lasso, stepwise, etc).

(3) Variable importance: having selected $S \subset \{1, \ldots, d\}$, how important is $X(j)$? (i.e. in the linear case, this is $\beta_j$.)
High Dimensional Regression

Goals:

(1) accurate prediction: $\mathbb{E} | Y - \hat{\mu}(X) |^2$. 

(2) valid predictive inference: $P( Y \in C_n(X)) \geq 1 - \alpha$ for all $P$ and $n$, even when we use model selection (lasso, stepwise, etc).

(3) Variable importance: having selected $S \subset \{1, \ldots, d\}$, how important is $X(j)$? (i.e. in the linear case, this is $\beta_j$).
High Dimensional Regression

Goals:

(1) accurate prediction: \( \mathbb{E}|Y - \hat{\mu}(X)|^2 \).

(2) valid predictive inference:

\[
P(Y \in C_n(X)) \geq 1 - \alpha
\]

for all \( P \) and \( n \), even when we use model selection (lasso, stepwise, etc).
Goals:
(1) accurate prediction: $\mathbb{E}|Y - \hat{\mu}(X)|^2$.
(2) valid predictive inference:

$$P(Y \in C_n(X)) \geq 1 - \alpha$$

for all $P$ and $n$, even when we use model selection (lasso, stepwise, etc).

(3) Variable importance: having selected $S \subset \{1, \ldots, d\}$, how important is $X(j)$? (i.e. in the linear case, this is $\beta_j$.)
Linear Regression (with model selection)

Full conformal:

- Data $\Rightarrow$ model selection $\Rightarrow \hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n)$, $(x, y) = \Rightarrow \hat{\mu}(x, y) = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}(x, y)(X_i)|$
- $\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$
- Repeat for every $(x, y)$.
- $C_n(x) = \{y : \pi(x, y) \geq \alpha\}$
- Without any assumptions: $P(Y \in C_n(X)) \geq 1 - \alpha$. 
Linear Regression (with model selection)

Full conformal:

- Data $\implies$ model selection $\implies \hat{\beta}$
Linear Regression (with model selection)

Full conformal:
- Data $\implies$ model selection $\implies \hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \implies \hat{\mu}_{(x,y)} = \hat{\beta}^T x$

Residuals:
$R_i = |Y_i - \hat{\mu}_{T(x,y)}(X_i)|$

$\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$

Repeat for every $(x, y)$.

$C_n(x) = \{y : \pi(x, y) \geq \alpha\}$

Without any assumptions:
$P(Y \in C_n(X)) \geq 1 - \alpha$
Linear Regression (with model selection)

Full conformal:

- Data $\implies$ model selection $\implies \hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \implies \hat{\mu}_{(x,y)} = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}_{(x,y)}(X_i)|$. 

$\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.

$C_n(x) = \{y : \pi(x, y) \geq \alpha \}$.

Without any assumptions:

$P(Y \in C_n(X)) \geq 1 - \alpha$. 

Linear Regression (with model selection)

Full conformal:
- Data $\implies$ model selection $\implies\hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \implies \hat{\mu}_{(x,y)} = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}^T_{(x,y)}(X_i)|$
- $\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$. 


Linear Regression (with model selection)

Full conformal:

- Data $\implies$ model selection $\implies \hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \implies \hat{\mu}_{(x,y)} = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}_{(x,y)}(X_i)|$.
- $\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.
- Repeat for every $(x, y)$. 
Linear Regression (with model selection)

Full conformal:

- Data $\rightarrow$ model selection $\rightarrow$ $\hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \rightarrow \hat{\mu}_{(x,y)} = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}_{(x,y)}(X_i)|$.
- $\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.
- Repeat for every $(x, y)$.
- $C_n(x) = \{y : \pi(x, y) \geq \alpha\}$. 
Linear Regression (with model selection)

Full conformal:

- Data $\implies$ model selection $\implies \hat{\beta}$
- Augment: $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y) \implies \hat{\mu}_{(x,y)} = \hat{\beta}^T x$
- Residuals: $R_i = |Y_i - \hat{\mu}_{(x,y)}(X_i)|$
- $\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$
- Repeat for every $(x, y)$.
- $C_n(x) = \{y : \pi(x, y) \geq \alpha\}$
- Without any assumptions: $\mathbb{P}(Y \in C_n(X)) \geq 1 - \alpha$. 
The previous algorithm is very expensive. Here is a cheaper version:

1. Split data: $D_1$ and $D_2$.
2. Get $\hat{\mu}$ from $D_1$.
3. Residuals from $D_2$. Empirical cdf $\hat{F}_n$.
4. $C_n(x) = \hat{\mu}(x) \pm \hat{F}_{n}(1 - \alpha)$.

Then $C_n$ is valid (but less efficient than full conformal).

(This is not the 'sample splitting' we will discuss later.)
The previous algorithm is very expensive. Here is a cheaper version:

1. Split data: $D_1$ and $D_2$.
2. Get $\hat{\mu}$ from $D_1$.
3. Residuals from $D_2$. Empirical cdf $\hat{F}_n$.
4. $C_n(x) = \hat{\mu}(x) \pm \hat{F}_n^{-1}(1 - \alpha)$. 

Then $C_n$ is valid (but less efficient than full conformal).
The previous algorithm is very expensive. Here is a cheaper version:

1. Split data: $D_1$ and $D_2$.
2. Get $\hat{\mu}$ from $D_1$.
3. Residuals from $D_2$. Empirical cdf $\hat{F}_n$.
4. $C_n(x) = \hat{\mu}(x) \pm \hat{F}_n^{-1}(1 - \alpha)$.

Then $C_n$ is valid (but less efficient than full conformal).
The previous algorithm is very expensive. Here is a cheaper version:

1. Split data: $D_1$ and $D_2$.
2. Get $\hat{\mu}$ from $D_1$.
3. Residuals from $D_2$. Empirical cdf $\hat{F}_n$.
4. $C_n(x) = \hat{\mu}(x) \pm \hat{F}_n^{-1}(1 - \alpha)$.

Then $C_n$ is valid (but less efficient than full conformal).

This is very fast.

(This is not the ‘sample splitting’ we will discuss later.)
Multisplit

Split $N$ times.

Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$.

Set $C(x) = \bigcap_{j=1}^{N} C_j(x)$.

This is valid and reduces the randomness.

What happens to the length?

$\alpha \rightarrow \alpha/N$ increases length but taking the intersection reduces the length.
Split $N$ times.
Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$. 

\[ C(x) = \bigcap_{j=1}^N C_j(x). \]
This is valid and reduces the randomness.
What happens to the length?
$\alpha \rightarrow \alpha/N$ increases length BUT taking the intersection reduces the length.
Split $N$ times.
Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$.
Set $C(x) = \bigcap_{j=1}^{N} C_j(x)$. 
Split $N$ times.

Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$.

Set $C(x) = \bigcap_{j=1}^{N} C_j(x)$.

This is valid and reduces the randomness.
Split $N$ times.
Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$.
Set $C(x) = \bigcap_{j=1}^{N} C_j(x)$.
This is valid and reduces the randomness.
What happens to the length?
Split $N$ times.

Get $C_1, \ldots, C_N$ each at level $1 - \alpha/N$.

Set $C(x) = \bigcap_{j=1}^{N} C_j(x)$.

This is valid and reduces the randomness.

What happens to the length?

$\alpha \rightarrow \alpha/N$ increases length BUT taking the intersection reduces the length.
Length

Let $C$ be from one split. Let $C^N$ be from multi-split.
Let $C$ be from one split. Let $C^N$ be from multi-split. Suppose there exists some function $g$ such that

$$P(\|\hat{\mu} - g\|_\infty > \eta_n) \leq \rho_n$$

where $\eta_n = o(1)$ and $\rho_n = o(1)$. Note that $g$ need not be $g$ i.e we don’t require consistency.
Let $C$ be from one split. Let $C^N$ be from multi-split. Suppose there exists some function $g$ such that

$$P(||\hat{\mu} - g||_\infty > \eta_n) \leq \rho_n$$

where $\eta_n = o(1)$ and $\rho_n = o(1)$. Note that $g$ need not be $g$ i.e we don’t require consistency.

Also suppose that

$$N = o(\rho_n^{-1}) \quad \text{and} \quad N = o(e^{n\alpha^2/32}).$$
Length

Let $C$ be from one split. Let $C^N$ be from multi-split.

Suppose there exists some function $g$ such that

$$P(||\hat{\mu} - g||_\infty > \eta_n) \leq \rho_n$$

where $\eta_n = o(1)$ and $\rho_n = o(1)$. Note that $g$ need not be $g$ i.e we don’t require consistency.

Also suppose that

$$N = o\left(\rho_n^{-1}\right) \quad \text{and} \quad N = o\left(e^{n\alpha^2/32}\right).$$

Then

$$P(\text{length}(C^N) > \text{length}(C)) \to 1.$$
Length

Let $C$ be from one split. Let $C^N$ be from multi-split.
Suppose there exists some function $g$ such that

$$P(||\hat{\mu} - g||_\infty > \eta_n) \leq \rho_n$$

where $\eta_n = o(1)$ and $\rho_n = o(1)$. Note that $g$ need not be $g$ i.e we don’t require consistency.
Also suppose that

$$N = o(\rho_n^{-1}) \quad \text{and} \quad N = o(e^{n\alpha^2/32}).$$

Then

$$P(\text{length}(C^N) > \text{length}(C)) \to 1.$$ 

We suggest: take $N = 1$ and live with the randomness or do full conformal.
THEOREM: If the Kool-Aid assumptions do hold, then the conformal interval has optimal rate. (Nothing is lost by conformalization.)
THEOREM: If the Kool-Aid assumptions do hold, then the conformal interval has optimal rate. (Nothing is lost by conformalization.)

But what can say in general?
Efficiency

THEOREM: If the Kool-Aid assumptions do hold, then the conformal interval has optimal rate. (Nothing is lost by conformalization.)

But what can say in general?

We will assess the properties of $C_n$ by comparing it to some oracles.
Oracles

Let $F_n(t) = P(|Y - \mu_n(X)| \leq t)$ where $\hat{\mu}$ is based on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X, Y)$ is a new pair.

Define the oracle $C_{\text{oracle}}(x) = \hat{\mu}(X) \pm F_{n-1}(1 - \alpha)$.

This is an exact but unobtainable interval.

Define the uber-oracle $C_{\text{uber}}(x) = \mu(X) \pm F_{n-1}(1 - \alpha)$ where $F(t) = P(|\epsilon| \leq t)$ and $Y = \mu(X) + \epsilon$. Exact and even more unobtainable.
Oracles

Let

\[ F_n(t) = \mathbb{P}(|Y - \hat{\mu}_n(X)| \leq t) \]

where \( \hat{\mu} \) is based on \((X_1, Y_1), \ldots, (X_n, Y_n)\) and \((X, Y)\) is a new pair.
Let

\[ F_n(t) = \mathbb{P}(|Y - \hat{\mu}_n(X)| \leq t) \]

where \( \hat{\mu} \) is based on \((X_1, Y_1), \ldots, (X_n, Y_n)\) and \((X, Y)\) is a new pair.

Define the oracle

\[ C_{\text{oracle}}(x) = \hat{\mu}(X) \pm F_n^{-1}(1 - \alpha). \]

This is an exact but unobtainable interval.
Let
\[ F_n(t) = \mathbb{P}(|Y - \hat{\mu}_n(X)| \leq t) \]
where \( \hat{\mu} \) is based on \((X_1, Y_1), \ldots, (X_n, Y_n)\) and \((X, Y)\) is a new pair.

Define the oracle
\[ C_{\text{oracle}}(x) = \hat{\mu}(X) \pm F_n^{-1}(1 - \alpha). \]

This is an exact but unobtainable interval.

Define the uber-oracle
\[ C_{\text{uber}}(x) = \mu(X) \pm F^{-1}(1 - \alpha) \]

where \( F(t) = \mathbb{P}(|\epsilon| \leq t) \) and \( Y = \mu(X) + \epsilon \). Exact and even more unobtainable.
Oracles

Let $\nu$ denote Lebesgue measure.
Let $\nu$ denote Lebesgue measure.

Let $\Delta = \hat{\mu}(X) - \mu(X)$. 

The split-conformal interval is a better approximation to a slightly worse oracle.
Let $\nu$ denote Lebesgue measure.

Let $\Delta = \hat{\mu}(X) - \mu(X)$.

Then:

$$\nu(C_{\text{oracle}}) - \nu(C_{\text{uber}}) \leq \mathbb{E}[\Delta^2]$$

$$\nu(C_n) - \nu(C_{\text{oracle}}) \leq O_P(\sqrt{ns_n^2})$$ where $s_n$ is a stability measure

$$\nu(C_{\text{split}}) - \nu(C_{\text{oracle},n/2}) \leq O_P(n^{-1/2})$$
Let $\nu$ denote Lebesgue measure.

Let $\Delta = \hat{\mu}(X) - \mu(X)$.

Then:

$$\nu(C_{\text{oracle}}) - \nu(C_{\text{uber}}) \leq C \mathbb{E}[\Delta^2]$$

$$\nu(C_n) - \nu(C_{\text{oracle}}) \leq O_P\left(\sqrt{ns_n^2}\right) \quad \text{where } s_n \text{ is a stability measure}$$

$$\nu(C_{\text{split}}) - \nu(C_{\text{oracle}, n/2}) \leq O_P\left(n^{-1/2}\right)$$

The split-conformal interval is a better approximation to a slightly worse oracle.
Overfitting

Under extreme overfitting, $C_n$ is still valid but $\nu(C_n) \to \infty$.  

For example, for $k$-nn regression:  

$$\nu(C_n) = C_n + \text{bias}^2/n + \text{var} \sqrt{k^2 - 1}.$$  

When $k \to 1$, $\nu(C_n) \to \infty$.  

In practice, the minimizer of $\nu(C_n)$ underfits relative to minimizer of MSE.
Under extreme overfitting, $C_n$ is still valid but $\nu(C_n) \to \infty$.

For example, for $k$-nn regression:

$$\nu(C_n) = \frac{C + \text{bias}^2_n + \text{var}_n}{\sqrt{k^2 - 1}}.$$ 

When $k \to 1$, $\nu(C_n) \to \infty$. 
Under extreme overfitting, $C_n$ is still valid but $\nu(C_n) \rightarrow \infty$.

For example, for $k$-nn regression:

$$\nu(C_n) = \frac{C + \text{bias}_n^2 + \text{var}_n}{\sqrt{k^2 - 1}}.$$ 

When $k \rightarrow 1$, $\nu(C_n) \rightarrow \infty$.

In practice, the minimizer of $\nu(C_n)$ underfits relative to minimizer of MSE.
The Tradeoff

![Graph showing the tradeoff between underfit and overfit with different models: Oracle, Conformal, Split-Oracle, Split-Conformal. The x-axis represents 1/h, and the y-axis represents length. The graph illustrates the progression from underfit to overfit, with each model showing distinct behaviors.](image-url)
High-dimensional
Example

High-dimensional
Highly correlated
Example

High-dimensional
Highly correlated
Non-constant variance
Example

High-dimensional
Highly correlated
Non-constant variance
Thick tailed errors
The diagram illustrates the coverage of different models as a function of relative optimism. The models compared are Stepwise, Lasso, Elastic net, and Ridge. The x-axis represents the relative optimism, while the y-axis represents coverage. The lines on the graph show the performance of each model across various optimism levels.
Variable Importance: LOCO

We want a predictive measure of variable importance that does not require specifying the “true model.” The usual approach is the projection parameter (best linear predictor):

$$\beta = \arg\min_{\beta} \mathbb{E} \left( Y - X^T \beta \right)^2.$$  

Now create confidence interval for each \( \beta(j) \).

We want to infer \( \beta \) because we are used to it. But it is not a good target for inference.

1. When \( m(x) \) is not linear, it is not answering the right question.
2. When we get to sample splitting (if we get to sample splitting) we’ll see that the accuracy for \( \beta \) is poor (slow rate).
3. Not even defined for nonparametric regression (random forest).
4. If you do want to infer \( \beta \), use sample splitting (later).
Variable Importance: LOCO

We want a predictive measure of variable importance that does not require specifying the “true model.”
Variable Importance: LOCO

We want a predictive measure of variable importance that does not require specifying the "true model."
The usual approach is the projection parameter (best linear predictor):

$$\beta = \arg\min \mathbb{E}(Y - X^T\beta)^2.$$ 

Now create confidence interval for each $\beta(j)$. 
Variable Importance: LOCO

We want a predictive measure of variable importance that does not require specifying the “true model.”
The usual approach is the projection parameter (best linear predictor):

$$\beta = \text{argmin } \mathbb{E}(Y - X^T \beta)^2.$$  

Now create confidence interval for each $\beta(j)$.
We want to infer $\beta$ because we are used to it. But it is not a good target for inference.
We want a predictive measure of variable importance that does not require specifying the “true model.”

The usual approach is the projection parameter (best linear predictor):

\[ \beta = \arg\min_{\beta} \mathbb{E}(Y - X^T\beta)^2. \]

Now create confidence interval for each \( \beta(j) \).

We want to infer \( \beta \) because we are used to it. But it is not a good target for inference.

1. When \( m(x) \) is not linear, it is not answering the right question.
2. When we get to sample splitting (if we get to sample splitting) we’ll see that the accuracy for \( \beta \) is poor (slow rate).
3. Not even defined for nonparametric regression (random forest).
4. If you do want to infer \( \beta \), use sample splitting (later).
LOCO (Leave-One-Covariate-Out):

Let $\hat{\mu}_j$ be the estimator when $X(j)$ is deleted from the analysis. Either re-fit without $X(j)$ or set $\hat{\beta}(j) = 0$.

Let $W_j(X) = |Y - \hat{\mu}_j(X)| - |Y - \hat{\mu}(X)|$.

Quantities of interest: $W_j(x)$, $G_j(t) = P(W_j \leq t | \hat{\mu})$, $\theta_j$, $\phi_j$, etc.
Variable Importance: LOCO

LOCO (Leave-One-Covariate-Out):
Let $\hat{\mu}_j$ be the estimator when $X(j)$ is deleted from the analysis.
LOCO (Leave-One-Covariate-Out):
Let $\hat{\mu}_j$ be the estimator when $X(j)$ is deleted from the analysis.
Either re-fit without $X(j)$ or set $\hat{\beta}(j) = 0$. 

Quantities of interest:
$W_j(X) = |Y - \hat{\mu}_j(X)| - |Y - \hat{\mu}(X)|$. 

$G_j(t) = P(W_j \leq t | \hat{\mu})$

$\theta_j = \text{mean of } W_j$

$\phi_j = \text{median of } W_j$ 

etc.
Variable Importance: LOCO

LOCO (Leave-One-Covariate-Out): Let $\hat{\mu}_j$ be the estimator when $X(j)$ is deleted from the analysis. Either re-fit without $X(j)$ or set $\hat{\beta}(j) = 0$. Let

$$W_j(X) = |Y - \hat{\mu}_j(X)| - |Y - \hat{\mu}(X)|.$$  

Quantities of interest:

- $W_j(x)$
- $G_j(t) = P(W_j \leq t | \hat{\mu})$
- $\theta_j = \text{mean of } W_j$
- $\phi_j = \text{median of } W_j$

etc.
Local importance: for a given $x$:

$$V_j(x) = \left\{ |y - \hat{\mu}_{(-j)}(x)| - |y - \hat{\mu}(x)| : y \in C_n(x) \right\}.$$
Local importance: for a given $x$:

$$V_j(x) = \left\{ |y - \hat{\mu}_{(-j)}(x)| - |y - \hat{\mu}(x)| : y \in C_n(x) \right\}.$$

Then

$$\mathbb{P}(W_j(x) \in V_j \text{ for all } j \text{ and all } x) \geq 1 - \alpha.$$
In the split conformal approach, we can use the residuals from $\mathcal{D}_2$ to define

$$W_{ji} = |Y_i - \hat{\mu}_j(X_i)| - |Y_i - \hat{\mu}(X_i)|$$

to get various distribution-free tests and confidence intervals.
In the split conformal approach, we can use the residuals from $D_2$ to define

$$W_{ji} = |Y_i - \hat{\mu}_j(X_i)| - |Y_i - \hat{\mu}(X_i)|$$

to get various distribution-free tests and confidence intervals. For example we can test:
In the split conformal approach, we can use the residuals from $\mathcal{D}_2$ to define

$$W_{ji} = |Y_i - \hat{\mu}_j(X_i)| - |Y_i - \hat{\mu}(X_i)|$$

to get various distribution-free tests and confidence intervals. For example we can test:

$H_0 : F_j = F$ versus $H_1 : F_j \succeq F$

$F_j(t) = \mathbb{P}(|Y - \hat{\mu}_j(X)| \leq t), \ F(t) = \mathbb{P}(|Y - \hat{\mu}(X)| \leq t).$
Suppose $Y \in \{1, \ldots, k\}$. (See Lei 2015 for the case $k = 2$.)

Three types of guarantee:

$P(Y \in C(X)) \geq 1 - \alpha$

$P(Y \in C(X) | Y = j) \geq 1 - \alpha$

$P(Y \in C(X) | X = x) \geq 1 - \alpha$.

The latter cannot be done in a distribution-free way.
Suppose $Y \in \{1, \ldots, k\}$. (See Lei 2015 for the case $k = 2$.)
(with Jing Lei and Mauricio Sadinle).

Suppose \( Y \in \{1, \ldots, k\} \). (See Lei 2015 for the case \( k = 2 \).)

Same method works. Three types of guarantee:

\[
\Pr(Y \in C(X)) \geq 1 - \alpha \\
\Pr(Y \in C(X) | Y = j) \geq 1 - \alpha \\
\Pr(Y \in C(X) | X = x) \geq 1 - \alpha.
\]
Beyond Regression: Multiclass Prediction

(with Jing Lei and Mauricio Sadinle).

Suppose $Y \in \{1, \ldots, k\}$. (See Lei 2015 for the case $k = 2$.)

Same method works. Three types of guarantee:

\[
\begin{align*}
P(Y \in C(X)) & \geq 1 - \alpha \\
P(Y \in C(X)|Y = j) & \geq 1 - \alpha \\
P(Y \in C(X)|X = x) & \geq 1 - \alpha.
\end{align*}
\]

The latter cannot be done in a distribution-free way.
Multiclass Prediction

In this case we want to minimize ambiguity $E|C_n(X)$. This can be done optimally. $C_n = \{j: \hat{P}(Y = j | X = x) > \hat{t}\}$. Interesting twist: the optimal sets can be empty: $C_n(X) = \emptyset$. Solution: expand $C_n(x)$ (optimally).
In this case we want to minimize ambiguity

\[ \mathbb{E}|C_n(X)|. \]
Multiclass Prediction

In this case we want to minimize ambiguity

\[ \mathbb{E}|C_n(X)|. \]

This can be done optimally.

\[ C_n = \{ j : \hat{P}(Y = j|X = x) > \hat{t} \}. \]
In this case we want to minimize ambiguity

\[ \mathbb{E} | C_n(X) |. \]

This can be done optimally. 

\[ C_n = \{ j : \hat{P}(Y = j|X = x) > \hat{t} \}. \]

Interesting twist: the optimal sets can be empty: \( C_n(X) = \emptyset. \)
Multiclass Prediction

In this case we want to minimize ambiguity

$$\mathbb{E}|C_n(X)|.$$ 

This can be done optimally. 

$$C_n = \{j : \hat{P}(Y = j|X = x) > \hat{t}\}.$$ 

Interesting twist: the optimal sets can be empty: 

$$C_n(X) = \emptyset.$$ 

Solution: expand $$C_n(x)$$ (Optimally).
Total Coverage = 0.9

Class Coverage = 0.9

Class Coverage = 0.99

Greedy Expansion, Class Coverage ≥ 0.99
Ambiguity in Practice
Inference Via Sample Splitting
Inference Via Sample Splitting

Split data into two halves: $D_1$ and $D_2$.
Inference Via Sample Splitting

Split data into two halves: $D_1$ and $D_2$. Choose model using $D_1$. Very old idea. (First reference I could find is Tukey 1977 but it is probably much, much older).
Split data into two halves: $D_1$ and $D_2$.
Choose model using $D_1$.
Gives a subset of covariates $S$. Let $K_n = |S|$. 
Inference Via Sample Splitting

Split data into two halves: $D_1$ and $D_2$.
Choose model using $D_1$.
Gives a subset of covariates $S$. Let $K_n = |S|$.
Use $D_2$ for inference.
Inference Via Sample Splitting

Split data into two halves: $D_1$ and $D_2$.
Choose model using $D_1$.
Gives a subset of covariates $S$. Let $K_n = |S|$.
Use $D_2$ for inference.
Very old idea. (First reference I could find is Tukey 1977 but it is probably much, much older).
Inference Via Sample Splitting

Why do this?

Can infer the projection parameter: $\beta_S = \arg\min E(Y - \beta^T X_S)^2$

where $X_S = (X(j): j \in S)$.

Gives valid inferences after complicated (realistic) model selection.

Textbook model selection: choose $S$ using the lasso.

Real model selection: apply lasso, look at residuals, transform some variables, remove some outliers, re-do the lasso, add some interactions, unconscious bias, etc.

Splitting ensures that

$$\lim \inf_{n \to \infty} \inf_{P \in P} \inf_{w \in W} P(\beta \in C) \geq 1 - \alpha$$

where $W$ is the set of all selection procedures.
Inference Via Sample Splitting

Why do this?
Can infer the projection parameter:

$$\beta_S = \arg\min \mathbb{E}(Y - \beta^T X_S)^2$$

where $X_S = (X(j) : j \in S)$. 
Inference Via Sample Splitting

Why do this?
Can infer the projection parameter:

$$\beta_S = \arg\min \mathbb{E}(Y - \beta^T X_S)^2$$

where $X_S = (X(j) : j \in S)$.
Gives valid inferences after complicated (realistic) model selection.
Inference Via Sample Splitting

Why do this?

Can infer the projection parameter:

\[ \beta_S = \arg\min \mathbb{E}(Y - \beta^TX_S)^2 \]

where \( X_S = (X(j) : j \in S) \).

Gives valid inferences after complicated (realistic) model selection. Textbook model selection: choose \( S \) using the lasso.
Inference Via Sample Splitting

Why do this?
Can infer the projection parameter:

$$
\beta_S = \arg\min \mathbb{E}(Y - \beta^T X_S)^2
$$

where $X_S = (X(j) : j \in S)$.

Gives valid inferences after complicated (realistic) model selection.

Textbook model selection: choose $S$ using the lasso.

Real model selection: apply lasso, look at residuals, transform some variables, remove some outliers, re-do the lasso, add some interactions, unconscious bias, etc.
Inference Via Sample Splitting

Why do this?
Can infer the projection parameter:

$$\beta_S = \arg\min_{\beta} \mathbb{E}(Y - \beta^T X_S)^2$$

where $X_S = (X(j) : j \in S)$.

Gives valid inferences after complicated (realistic) model selection.

Textbook model selection: choose $S$ using the lasso.

Real model selection: apply lasso, look at residuals, transform some variables, remove some outliers, re-do the lasso, add some interactions, unconscious bias, etc.

Splitting ensures that

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{w \in \mathcal{W}} P(\beta \in C) \geq 1 - \alpha$$

where $\mathcal{W}$ is the set of all selection procedures.
Inferring $\beta_S$

Use either Normal approximation to $\hat{\beta}_S$ or the bootstrap. Can also infer mean LOCO $\gamma_S$ or median LOCO $\phi_S$. 

There is a new phenomenon: Prediction-Inference Tradeoff. Splitting can lead to less accurate predictions. But it gives more accurate (distribution free) inferences.
Inferring $\beta_S$

Use either Normal approximation to $\hat{\beta}_S$ or the bootstrap. Can also infer mean LOCO $\gamma_S$ or median LOCO $\phi_S$. We get

$$\lim\inf_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{w \in \mathcal{W}} P(\theta \in C_n) \geq 1 - \alpha$$

where $\mathcal{P}$ is a large set of distributions and $\mathcal{W}$ is all selectors. (For median LOCO, we can take all distributions).
Inferring $\beta_S$

Use either Normal approximation to $\hat{\beta}_S$ or the bootstrap. Can also infer mean LOCO $\gamma_S$ or median LOCO $\phi_S$. We get

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}} \inf_{w \in \mathcal{W}} P(\theta \in C_n) \geq 1 - \alpha$$

where $\mathcal{P}$ is a large set of distributions and $\mathcal{W}$ is all selectors. (For median LOCO, we can take all distributions).

There is a new phenomenon: Prediction-Inference Tradeoff. Splitting can lead to less accurate predictions. But it gives more accurate (distribution free) inferences.
The coverage error is

\[ \text{Coverage} - (1 - \alpha) \]

where \( \text{Coverage} = P(\theta \in C_n) \).
The coverage error is

\[
\text{Coverage} - (1 - \alpha)
\]

where Coverage = \( P(\theta \in C_n) \).

This can be decomposed into two pieces:

\[
\delta_1 = \text{error due to CLT}
\]

and

\[
\delta_2 = \text{error due to nonlinearity (\(\beta\) only)}
\]

Sparse fit: \( K_n < \infty \)

Non-sparse fit: \( K_n \to \infty \).
<table>
<thead>
<tr>
<th>Regime</th>
<th>Parameter</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse Fit</td>
<td>$\beta_S$</td>
<td>$O \left( \frac{1}{n} \right)^{1/6}$</td>
<td>$O \left( \frac{1}{n} \right)^{1/2}$</td>
</tr>
<tr>
<td>Sparse Fit</td>
<td>$\gamma_S$</td>
<td>$O \left( \frac{1}{n} \right)^{1/6}$</td>
<td>$0$</td>
</tr>
<tr>
<td>Sparse Fit</td>
<td>$\phi_S$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>Non-Sparse Fit</td>
<td>$\beta_S$</td>
<td>$O \left( \frac{K_n^{10}}{n} \right)^{1/6}$</td>
<td>$O(\Xi_n)$</td>
</tr>
<tr>
<td>Sparse Fit</td>
<td>$\gamma_S$</td>
<td>$O \left( \frac{\log K_n}{n} \right)^{1/6}$</td>
<td>$0$</td>
</tr>
<tr>
<td>Sparse Fit</td>
<td>$\phi_S$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

where

$$\Xi_n = \sqrt{\frac{K_n^{10} (\log K_n)^2}{n} \log \left( \frac{n}{K_n^6 (\log K_n)^2} \right)}.$$
Remarks

We conjecture that these rates are sharp. (Work in progress).
We conjecture that these rates are sharp. (Work in progress). Also, we conjecture that non-splitting (selector depends on all data) leads to sub-optimal rates (for inference). We can prove this in some cases.
Remarks

We conjecture that these rates are sharp. (Work in progress). Also, we conjecture that non-splitting (selector depends on all data) leads to sub-optimal rates (for inference). We can prove this in some cases.

Non-splitting is NOT uniform over $\mathcal{W}$. 
We conjecture that these rates are sharp. (Work in progress). Also, we conjecture that non-splitting (selector depends on all data) leads to sub-optimal rates (for inference). We can prove this in some cases.

Non-splitting is NOT uniform over $\mathcal{W}$. However, non-splitting gives more accurate predictors.
We conjecture that these rates are sharp. (Work in progress). Also, we conjecture that non-splitting (selector depends on all data) leads to sub-optimal rates (for inference). We can prove this in some cases.

Non-splitting is NOT uniform over $\mathcal{W}$.
However, non-splitting gives more accurate predictors.
Again, there is a prediction-inference tradeoff.
All methods that use the strong assumptions are fragile. Here we focus only on selective inference (inference based on conditioning on the selected linear model).
All methods that use the strong assumptions are fragile. Here we focus only on selective inference (inference based on conditioning on the selected linear model).

Pivot $T$ such that $T|\text{selection} \sim \text{Unif}(0, 1)$. Basically, is the cdf of $\hat{\beta}_j$ given that model $S$ was selected. This is a truncated Gaussian.
All methods that use the strong assumptions are fragile. Here we focus only on selective inference (inference based on conditioning on the selected linear model).

Pivot $T$ such that $T|\text{selection} \sim \text{Unif}(0, 1)$. Basically, is the cdf of $\hat{\beta}_j$ given that model $S$ was selected. This is a truncated Gaussian. Can construct tests and confidence intervals.
All methods that use the strong assumptions are fragile. Here we focus only on selective inference (inference based on conditioning on the selected linear model).

Pivot $T$ such that $T | \text{selection} \sim \text{Unif}(0, 1)$. Basically, is the cdf of $\hat{\beta}_j$ given that model $S$ was selected. This is a truncated Gaussian. Can construct tests and confidence intervals.

The resulting p-value is the ratio of two extreme Gaussian tail probabilities.
Tail Ratios

![Graph of tail ratios with two curves: one increasing and one decreasing.](image)
If $d$ fixed, Normality is not needed thanks to CLT.
Fragility

If $d$ fixed, Normality is not needed thanks to CLT.

If $d \log d/n \to \infty$, and $\epsilon$ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266): $T$ does not converge to $\text{Unif}(0, 1)$. 

Also, the non-normality is not detectable.
If $d$ fixed, Normality is not needed thanks to CLT.

If $d \log d/n \to \infty$, and $\epsilon$ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266): $T$ does not converge to $\text{Unif}(0, 1)$.

In fact, we create an example where

$$T \to 0$$

with probability at least $1/e$. 
If $d$ fixed, Normality is not needed thanks to CLT.
If $d \log d/n \to \infty$, and $\epsilon$ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266): 
$T$ does not converge to Unif$(0, 1)$.
In fact, we create an example where 

$$T \to 0$$

with probability at least $1/e$.
Also, the non-normality is not detectable.
Conclusion

Assumptions are dangerous. Procedures based on these assumptions are fragile. Strong assumptions are not needed. Conformal, sample splitting, others? Similar ideas for IDA (interactive data analysis) also called adaptive data analysis. Prediction-Inference tradeoff.
Assumptions are dangerous.
Conclusion

Assumptions are dangerous.
Procedures based on these assumptions are fragile.
Conclusion

Assumptions are dangerous.
Procedures based on these assumptions are fragile.
Strong assumptions are not needed.
Conclusion

Assumptions are dangerous.
Procedures based on these assumptions are fragile.
Strong assumptions are not needed.
Conformal, sample splitting, others?
Assumptions are dangerous.
Procedures based on these assumptions are fragile.
Strong assumptions are not needed.
Conformal, sample splitting, others?
Similar ideas for IDA (interactive data analysis) also called adaptive data analysis.
Conclusion

Assumptions are dangerous.
Procedures based on these assumptions are fragile.
Strong assumptions are not needed.
Conformal, sample splitting, others?
Similar ideas for IDA (interactive data analysis) also called adaptive data analysis.
Prediction-Inference tradeoff.
THE END