Machine Learning Overview
(Inference For High Dimensional Regression)

Larry Wasserman
Carnegie Mellon University

OUTLINE

OUTLINE

- What is ML (and how does it differ from econometrics)?

OUTLINE

- What is ML (and how does it differ from econometrics)?
- Lasso and random forests.

OUTLINE

- What is ML (and how does it differ from econometrics)?
- Lasso and random forests.
- Inference for high-dimensional regression.

-debiasing
-conditioning
-uniform
-sample splitting
-subsampling
-conformal

OUTLINE

- What is ML (and how does it differ from econometrics)?
- Lasso and random forests.
- Inference for high-dimensional regression.

-debiasing
-conditioning
-uniform
-sample splitting
-subsampling
-conformal

- A few comments on causal inference.

(Susan and Victor will give talks about this.)

WARNING

WARNING

My view is quite highly biased and controversial.

WARNING

My view is quite highly biased and controversial.

> *Investigators who use [regression] are not paying adequate attention to the connection - if any - between the models and the phenomena they are studying. ... By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian ...*

—David Freedman

WARNING

My view is quite highly biased and controversial.

> *Investigators who use [regression] are not paying adequate attention to the connection - if any - between the models and the phenomena they are studying. ... By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian ...*

—David Freedman

Panglossian: characterized by or given to extreme optimism, especially in the face of unrelieved hardship or adversity

WARNING

My view is quite highly biased and controversial.

> *Investigators who use [regression] are not paying adequate attention to the connection - if any - between the models and the phenomena they are studying. ... By the time the models are deployed, the scientific position is nearly hopeless. Reliance on models in such cases is Panglossian ...*

—David Freedman

Panglossian: characterized by or given to extreme optimism, especially in the face of unrelieved hardship or adversity

I am skeptical of assumptions, especially in high-dimensional settings. (But I seem to be an outlier.)

WHAT IS ML?

Field	Assumptions	Goals

WHAT IS ML?

Field	Assumptions	Goals
Machine Learning	light	high-dimensional prediction

WHAT IS ML?

Field	Assumptions	Goals
Machine Learning	light	high-dimensional prediction
Statistics	heavier	prediction and inference

WHAT IS ML?

Field	Assumptions	Goals
Machine Learning	light	high-dimensional prediction
Statistics	heavier	prediction and inference
Economics	very heavy	prediction, inference and causation

PREDICTION

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

$X_i \in \mathbb{R}^d$. (Possibly $d > n$).

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

$X_i \in \mathbb{R}^d$. (Possibly $d > n$).

$Y_i \in \mathbb{R} \quad \text{regression, or} \qquad Y_i \in \{0, 1\} \quad \text{classification.}$

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

$X_i \in \mathbb{R}^d$. (Possibly $d > n$).

$Y_i \in \mathbb{R} \quad \text{regression, or} \qquad Y_i \in \{0, 1\} \quad \text{classification.}$

$$\mu(x) = \mathbb{E}[Y|X = x]$$

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

$X_i \in \mathbb{R}^d$. (Possibly $d > n$).

$Y_i \in \mathbb{R} \quad \text{regression, or} \qquad Y_i \in \{0, 1\} \quad \text{classification.}$

$$\mu(x) = \mathbb{E}[Y|X = x]$$

ML goal: given new pair (X, Y), minimize prediction error

$$\mathbb{E}[(Y - \widehat{\mu}(X))^2].$$

Requires balancing bias and variance.

PREDICTION

Observe $Z_1, \ldots, Z_n \sim P$ where $Z_i = (X_i, Y_i)$

$X_i \in \mathbb{R}^d$. (Possibly $d > n$).

$Y_i \in \mathbb{R} \quad \text{regression, or} \qquad Y_i \in \{0,1\} \quad \text{classification.}$

$$\mu(x) = \mathbb{E}[Y|X = x]$$

ML goal: given new pair (X, Y), minimize prediction error

$$\mathbb{E}[(Y - \widehat{\mu}(X))^2].$$

Requires balancing bias and variance.

(In contrast, causal inference is a semiparametric problem and bias is worse than variance.)

Three Popular Prediction Methods For High Dimensional Problems

Three Popular Prediction Methods For High Dimensional Problems

- Lasso: high-dimensional linear regression

Three Popular Prediction Methods For High Dimensional Problems

- Lasso: high-dimensional linear regression
- Random Forests: best off-the-shelf nonparametric prediction method

Three Popular Prediction Methods For High Dimensional Problems

- Lasso: high-dimensional linear regression
- Random Forests: best off-the-shelf nonparametric prediction method
- Deep learning: huge breakthrough or snake oil?

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\mathrm{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

(3) Can prove things about it. (Much harder for forward stepwise regression.)

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

(3) Can prove things about it. (Much harder for forward stepwise regression.)

Questions:

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

(3) Can prove things about it. (Much harder for forward stepwise regression.)

Questions:

1. What is the meaning of β?

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

(3) Can prove things about it. (Much harder for forward stepwise regression.)

Questions:

1. What is the meaning of β?
2. How do we choose λ?

The Lasso for Linear Regression

Recall that the lasso estimator is

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1 \right)$$

where $||\beta||_1 = \sum_j |\beta_j|$.

Why does everyone use it?

(1) This is convex and can be solved quickly.

(2) The resulting $\widehat{\beta}$ is sparse (most $\widehat{\beta}(j)$'s are 0).

(3) Can prove things about it. (Much harder for forward stepwise regression.)

Questions:

1. What is the meaning of β?
2. How do we choose λ?
3. How do we do inference?

Random Forests

To describe random forests, we begin with trees.

Random Forests

To describe random forests, we begin with trees.

A regression tree is a nonparametric estimator $\widehat{\mu}$ where $\widehat{\mu}$ is a piecewise constant over rectangles. (The fit is done by recursive splitting).

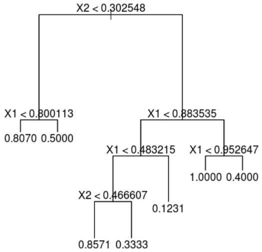
Random Forests

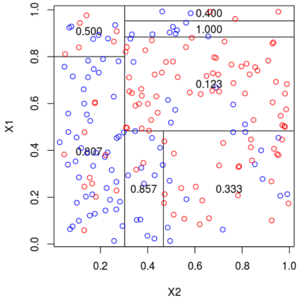
To describe random forests, we begin with trees.

A regression tree is a nonparametric estimator $\widehat{\mu}$ where $\widehat{\mu}$ is a piecewise constant over rectangles. (The fit is done by recursive splitting).

Same for classification (binary regression).

Tree





source: https://www.r-bloggers.com/regression-tree-using-ginis-index/

Forest

Draw subsamples $D_1, \ldots, D_N$. Fit trees $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ on the subsamples. (Usually one also draws random subsets of covariates.)

Forest

Draw subsamples $D_1, \ldots, D_N$. Fit trees $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ on the subsamples. (Usually one also draws random subsets of covariates.)

Random forest:

$$\widehat{\mu}(x) = \frac{1}{N} \sum_j \widehat{\mu}_j(x).$$

Forest

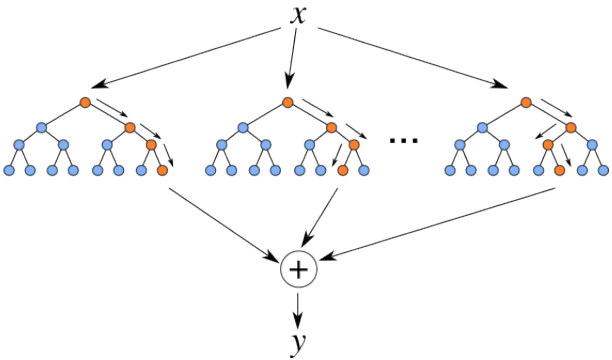
Draw subsamples $D_1, \ldots, D_N$. Fit trees $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ on the subsamples. (Usually one also draws random subsets of covariates.)

Random forest:

$$\widehat{\mu}(x) = \frac{1}{N} \sum_j \widehat{\mu}_j(x).$$

One of the best general purpose prediction methods.

Forest



source:
http://kazoo04.hatenablog.com/entry/2013/12/04/175402

A Property of the Lasso

Let

$$r(\beta) = E[(Y - \beta^T X)^2].$$

A Property of the Lasso

Let

$$r(\beta) = E[(Y - \beta^T X)^2].$$

Define the best, ℓ_1-sparse linear predictor β_* by

$$r(\beta_*) = \inf_{\beta \in B(L)} r(\beta)$$

where

$$B_L = \{\beta : \ ||\beta||_1 \leq L\}.$$

Thus, β_* is the population version of the lasso estimator.

A Property of the Lasso

Let

$$r(\beta) = E[(Y - \beta^T X)^2].$$

Define the best, ℓ_1-sparse linear predictor β_* by

$$r(\beta_*) = \inf_{\beta \in B(L)} r(\beta)$$

where

$$B_L = \{\beta : \ ||\beta||_1 \leq L\}.$$

Thus, β_* is the population version of the lasso estimator.

Let $\widehat{\beta}$ be the lasso estimator. With probabilty at least $1 - \delta$,

$$r(\widehat{\beta}) \leq r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n} \log\left(\frac{\sqrt{2}\, d}{\sqrt{\delta}}\right)}.$$

A Property of the Lasso

Let

$$r(\beta) = E[(Y - \beta^T X)^2].$$

Define the best, ℓ_1-sparse linear predictor β_* by

$$r(\beta_*) = \inf_{\beta \in B(L)} r(\beta)$$

where

$$B_L = \{\beta : \ ||\beta||_1 \leq L\}.$$

Thus, β_* is the population version of the lasso estimator.

Let $\widehat{\beta}$ be the lasso estimator. With probabilty at least $1 - \delta$,

$$r(\widehat{\beta}) \leq r(\beta_*) + \sqrt{\frac{16(L+1)^4 B^2}{n} \log\left(\frac{\sqrt{2}\, d}{\sqrt{\delta}}\right)}.$$

No assumptions! (Not even linearity).

INFERENCE?

The ‘True’ Parameter Versus the Projection Parameter

The ‘True’ Parameter Versus the Projection Parameter

- If we assume that

$$Y = \sum_j \beta_j X(j) + \epsilon$$

then there is a true β.

The ‘True’ Parameter Versus the Projection Parameter

- If we assume that

$$Y = \sum_j \beta_j X(j) + \epsilon$$

then there is a true β.

- Probably a bogus assumption.

The ‘True’ Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

The 'True' Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

There is a best linear predictor $\beta_*^T x$ that minimizes

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

The 'True' Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

There is a best linear predictor $\beta_*^T x$ that minimizes

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

No need to assume that $\mu(x)$ is linear.

The 'True' Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

There is a best linear predictor $\beta_*^T x$ that minimizes

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

No need to assume that $\mu(x)$ is linear.

We call β_* the *projection parameter*.

The 'True' Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

There is a best linear predictor $\beta_*^T x$ that minimizes

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

No need to assume that $\mu(x)$ is linear.

We call β_* the *projection parameter*.

For a subset $S \subset \{1, \ldots, d\}$, β_S is the projection parameter for S.

The 'True' Parameter Versus the Projection Parameter

- Let $\mu(x) = \mathbb{E}[Y|X = x]$ be arbitrary.

There is a best linear predictor $\beta_*^T x$ that minimizes

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

No need to assume that $\mu(x)$ is linear.

We call β_* the *projection parameter*.

For a subset $S \subset \{1, \ldots, d\}$, β_S is the projection parameter for S.

Note: β_S is a random parameter. (And it is not smooth.)

LOCO

LOCO (Leave Out COvariates)
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (arXiv:1604.04173)

LOCO

LOCO (Leave Out COvariates)
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (arXiv:1604.04173)

Example:

$$\gamma_j = \mathbb{E}\Bigg[|Y - \widehat{\mu}^T_{(-j)} X| - |Y - \widehat{\mu}^T X| \Bigg]$$

LOCO

LOCO (Leave Out COvariates)
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (arXiv:1604.04173)

Example:

$$\gamma_j = \mathbb{E}\Bigg[|Y - \widehat{\mu}_{(-j)}^T X| - |Y - \widehat{\mu}^T X|\Bigg]$$

Here, $\widehat{\mu}_{(-j)}$ is obtained by removing $X(j)$ and re-running the whole algorithm.

LOCO

LOCO (Leave Out COvariates)
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (arXiv:1604.04173)

Example:

$$\gamma_j = \mathbb{E}\left[|Y - \widehat{\mu}_{(-j)}^T X| - |Y - \widehat{\mu}^T X| \right]$$

Here, $\widehat{\mu}_{(-j)}$ is obtained by removing $X(j)$ and re-running the whole algorithm.

γ_j: the increase in prediction error due to not having $X(j)$.

LOCO

LOCO (Leave Out COvariates)
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (arXiv:1604.04173)

Example:

$$\gamma_j = \mathbb{E}\Bigg[|Y - \widehat{\mu}_{(-j)}^T X| - |Y - \widehat{\mu}^T X| \Bigg]$$

Here, $\widehat{\mu}_{(-j)}$ is obtained by removing $X(j)$ and re-running the whole algorithm.

γ_j: the increase in prediction error due to not having $X(j)$.

Does not depend on linearity or model correctness. Interpretable.

True versus Projection versus LOCO

True parameter approach:
Javanmard and Montanari (2014), Nickl and van de Geer (2013),
Belloni, Chernozhukov and Kato (2013), others ...

True versus Projection versus LOCO

True parameter approach:
Javanmard and Montanari (2014), Nickl and van de Geer (2013), Belloni, Chernozhukov and Kato (2013), others ...

Projection parameter approach:
Berk, Brown, Buja, Zhang, Zhao (2013), Lee, Sun, Taylor (2016), Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2016), ...

True versus Projection versus LOCO

True parameter approach:
Javanmard and Montanari (2014), Nickl and van de Geer (2013), Belloni, Chernozhukov and Kato (2013), others ...

Projection parameter approach:
Berk, Brown, Buja, Zhang, Zhao (2013), Lee, Sun, Taylor (2016), Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2016), ...

LOCO:
Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2016a,b), Hooker and Mentch (2016).

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes
Sample Splitting	projection	none	$\sqrt{\log k/n}$	Easy	Yes

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes
Sample Splitting	projection	none	$\sqrt{\log k/n}$	Easy	Yes
Sample Splitting	LOCO	none	$\sqrt{\log k/n}$	Easy	Yes

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes
Sample Splitting	projection	none	$\sqrt{\log k/n}$	Easy	Yes
Sample Splitting	LOCO	none	$\sqrt{\log k/n}$	Easy	Yes
Conformal	prediction	none	NA	Easy	Yes

Inference

Method	Parameter	Assumptions	Accuracy	Computation	Robust
Debiasing	'true' β	Very Strong	$1/\sqrt{n}$	Good	No
Conditional	projection	Strong	?	Good	No
Uniform	projection	weak	$\sqrt{k/n}$	NP hard	Yes
Sample Splitting	projection	none	$\sqrt{\log k/n}$	Easy	Yes
Sample Splitting	LOCO	none	$\sqrt{\log k/n}$	Easy	Yes
Conformal	prediction	none	NA	Easy	Yes

The forgotten methods: unsupervised dimension reduction (variable clustering, PCA, non-linear dimension reduction, etc).

Types of coverage

Types of coverage

Uniform $\quad \inf_{P \in \mathrm{ALL}} P^n(\theta \in C) \geq 1 - \alpha$

Types of coverage

Uniform $$\inf_{P \in \mathrm{ALL}} P^n(\theta \in C) \geq 1 - \alpha$$

Honest $$\liminf_{n \to \infty} \inf_{P \in \mathrm{BIG}} P^n(\theta \in C) \geq 1 - \alpha$$

Types of coverage

Uniform $\inf_{P \in \mathrm{ALL}} P^n(\theta \in C) \geq 1 - \alpha$

Honest $\liminf_{n \to \infty} \inf_{P \in \mathrm{BIG}} P^n(\theta \in C) \geq 1 - \alpha$

Parametric $\liminf_{n \to \infty} \inf_{P \in \mathrm{SMALL}} P^n(\theta \in C) \geq 1 - \alpha$

Types of coverage

Uniform $\inf_{P \in \mathrm{ALL}} P^n(\theta \in C) \geq 1 - \alpha$

Honest $\liminf_{n \to \infty} \inf_{P \in \mathrm{BIG}} P^n(\theta \in C) \geq 1 - \alpha$

Parametric $\liminf_{n \to \infty} \inf_{P \in \mathrm{SMALL}} P^n(\theta \in C) \geq 1 - \alpha$

Pointwise $P^n(\theta \in C) \to 1 - \alpha$

Types of coverage

Uniform $\inf_{P \in \mathrm{ALL}} P^n(\theta \in C) \geq 1 - \alpha$

Honest $\liminf_{n \to \infty} \inf_{P \in \mathrm{BIG}} P^n(\theta \in C) \geq 1 - \alpha$

Parametric $\liminf_{n \to \infty} \inf_{P \in \mathrm{SMALL}} P^n(\theta \in C) \geq 1 - \alpha$

Pointwise $P^n(\theta \in C) \to 1 - \alpha$

We really want: Robust and Honest:

$$\liminf_{n \to \infty} \inf_{w \in W_n} \inf_{P \in \mathrm{BIG}} P^n(\theta \in C) \geq 1 - \alpha$$

where $\mathcal{W}_n =$ all model selection rules.

Debiasing Methods

Javanmard and Montanari (2014)

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Set

$$\widehat{\beta} \longleftarrow \widehat{\beta} + \frac{1}{n} M X^T (Y - X\widehat{\beta})$$

where M is an estimate of Σ^{-1}.

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Set

$$\widehat{\beta} \longleftarrow \widehat{\beta} + \frac{1}{n} M X^T (Y - X\widehat{\beta})$$

where M is an estimate of Σ^{-1}.

Then

$$\sqrt{n}(\widehat{\beta} - \beta) = \text{Normal} \ + \ \text{small}$$

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Set

$$\widehat{\beta} \longleftarrow \widehat{\beta} + \frac{1}{n} M X^T (Y - X\widehat{\beta})$$

where M is an estimate of Σ^{-1}.

Then

$$\sqrt{n}(\widehat{\beta} - \beta) = \text{Normal} \ + \ \text{small}$$

Then we can construct confidence intervals. (Bonferroni over all parameters.)

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Set

$$\widehat{\beta} \longleftarrow \widehat{\beta} + \frac{1}{n} M X^T (Y - X\widehat{\beta})$$

where M is an estimate of Σ^{-1}.

Then

$$\sqrt{n}(\widehat{\beta} - \beta) = \mathrm{Normal} \ + \ \mathrm{small}$$

Then we can construct confidence intervals. (Bonferroni over all parameters.)

Very clean.

Debiasing Methods

Javanmard and Montanari (2014)

Lasso to get $\widehat{\beta}$.

Set

$$\widehat{\beta} \longleftarrow \widehat{\beta} + \frac{1}{n} M X^T (Y - X\widehat{\beta})$$

where M is an estimate of Σ^{-1}.

Then

$$\sqrt{n}(\widehat{\beta} - \beta) = \mathrm{Normal} \ + \ \mathrm{small}$$

Then we can construct confidence intervals. (Bonferroni over all parameters.)

Very clean.

But, it requires: linear model correct, incoherence, sparsity, constant variance, very carefully chosen tuning parameter.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Focus on $\beta_S(j)$ where β_S is the projection parameter.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Focus on $\beta_S(j)$ where β_S is the projection parameter.

Carefully chosen event E_n.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Focus on $\beta_S(j)$ where β_S is the projection parameter.

Carefully chosen event E_n.

By sufficiency:

$$\sqrt{n}(\widehat{\beta}(j) - \beta(j)) \;\Big|\; E_n$$

has a distribution (truncated Normal) indexed by one parameter.
Test and invert.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Focus on $\beta_S(j)$ where β_S is the projection parameter.

Carefully chosen event E_n.

By sufficiency:

$$\sqrt{n}(\widehat{\beta}(j) - \beta(j)) \; \Big| \; E_n$$

has a distribution (truncated Normal) indexed by one parameter.
Test and invert.

Advantage: no linearity. No incoherence.

Conditional Methods

Lee, Sun and Taylor (2014).
Target is the projection parameter.

Assume: normality, known, constant variance. Fixed X.

Select model $S \subset \{1, \ldots, d\}$ by lasso.

Focus on $\beta_S(j)$ where β_S is the projection parameter.

Carefully chosen event E_n.

By sufficiency:

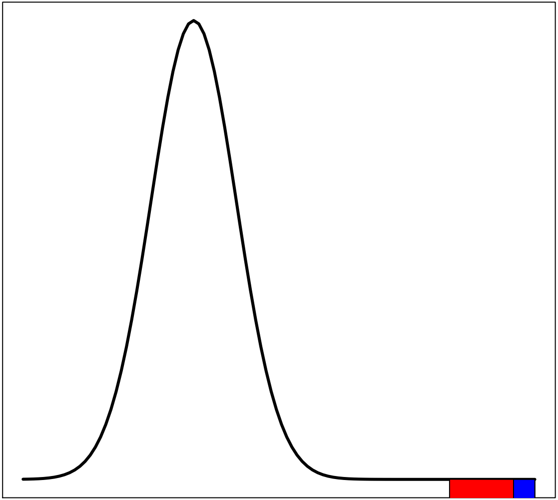
$$\sqrt{n}(\widehat{\beta}(j) - \beta(j)) \,\Big|\, E_n$$

has a distribution (truncated Normal) indexed by one parameter.
Test and invert.

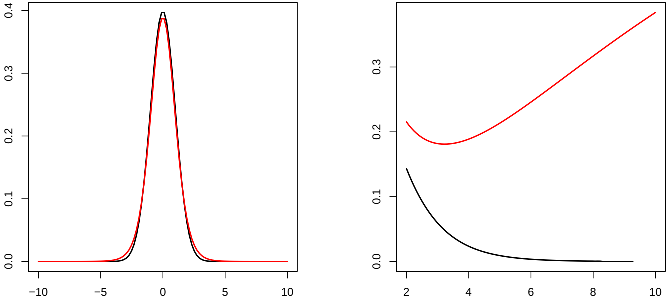
Advantage: no linearity. No incoherence.

Disadvantages: depends on the tails of the Normal. (fragile)

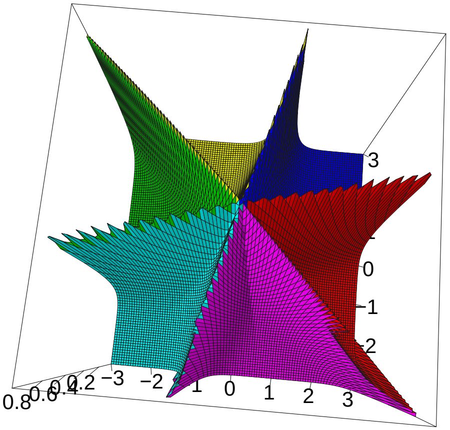
Tail Ratios



Tail Ratios



The Pivot



Fragility

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 ϵ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266):

Fragility

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

If $d \log d/n \to \infty$, and ϵ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266):

T does not converge to $\mathrm{Unif}(0, 1)$.

Fragility

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

If $d \log d/n \to \infty$, and ϵ not Normal, then by Theorem 12 of Tibshirani, Rinaldo, Tibshirani and Wasserman (arXiv:1506.06266):

T does not converge to $\mathrm{Unif}(0,1)$.

In fact, we create an example where

$$T \to 0$$

with probability at least $1/e$.

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Let $\mathcal{S}$ be all possible models that can be selected.

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Let $\mathcal{S}$ be all possible models that can be selected.

Let

$$F_n(t) = P\left(\sup_{S \in \mathcal{S}} \sqrt{n}||\widehat{\beta}_S - \beta_S||_\infty \le t\right).$$

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Let $\mathcal{S}$ be all possible models that can be selected.

Let

$$F_n(t) = P\left(\sup_{S\in\mathcal{S}} \sqrt{n}||\widehat{\beta}_S - \beta_S||_\infty \leq t\right).$$

Then $\widehat{\beta}_S(j) \pm F_n^{-1}(1-\alpha)$ is a valid confidence interval for any S and any $j \in S$.

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Let $\mathcal{S}$ be all possible models that can be selected.

Let

$$F_n(t) = P\left(\sup_{S \in \mathcal{S}} \sqrt{n}||\widehat{\beta}_S - \beta_S||_\infty \leq t\right).$$

Then $\widehat{\beta}_S(j) \pm F_n^{-1}(1-\alpha)$ is a valid confidence interval for any S and any $j \in S$.

Advantages: no linear model. No incoherence asumptions. Honest coverage.

Uniform Methods

Berk, Brown, Buja, Zhang, Zhao (2013),

Let $\mathcal{S}$ be all possible models that can be selected.

Let

$$F_n(t) = P\left(\sup_{S\in\mathcal{S}} \sqrt{n}||\widehat{\beta}_S - \beta_S||_\infty \leq t\right).$$

Then $\widehat{\beta}_S(j) \pm F_n^{-1}(1-\alpha)$ is a valid confidence interval for any S and any $j \in S$.

Advantages: no linear model. No incoherence asumptions. Honest coverage.

Disadvantages: cannot estimate F_n unless we make extra assumptions.

Sample Splitting

Hartigan (1969), Moran (1973), Barnard (1974), Cox (1975), Mosteller and Tukey (1977, p 37), Picard and Berk (1990), Miller (1990, p13) and Faraway (1995), G'Sell, Lei, Rinaldo, Tibshirani, Wasserman (2016).

Sample Splitting

Hartigan (1969), Moran (1973), Barnard (1974), Cox (1975), Mosteller and Tukey (1977, p 37), Picard and Berk (1990), Miller (1990, p13) and Faraway (1995), G'Sell, Lei, Rinaldo, Tibshirani, Wasserman (2016).

Barnard:
" ... the simple idea of splitting a sample in two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics ..."

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Choose model S from $\mathcal{D}_1$.

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Choose model S from $\mathcal{D}_1$.

Use Normal approximation or bootstrap on $\mathcal{D}_1$.

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Choose model S from $\mathcal{D}_1$.

Use Normal approximation or bootstrap on $\mathcal{D}_1$.

Then

$$\liminf_n \inf_{w \in \mathcal{W}} \inf_{P \in \mathcal{P}_n} P^n(\beta_S \in C) \geq 1 - \alpha$$

where $\mathcal{W}$ is all possible selection rules.

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Choose model S from $\mathcal{D}_1$.

Use Normal approximation or bootstrap on $\mathcal{D}_1$.

Then

$$\liminf_n \inf_{w \in \mathcal{W}} \inf_{P \in \mathcal{P}_n} P^n(\beta_S \in C) \geq 1 - \alpha$$

where $\mathcal{W}$ is all possible selection rules.

Advantage: No assumptions! Robust to selection rule.

Sample Splitting

Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Choose model S from $\mathcal{D}_1$.

Use Normal approximation or bootstrap on $\mathcal{D}_1$.

Then

$$\liminf_n \inf_{w \in \mathcal{W}} \inf_{P \in \mathcal{P}_n} P^n(\beta_S \in C) \geq 1 - \alpha$$

where $\mathcal{W}$ is all possible selection rules.

Advantage: No assumptions! Robust to selection rule.

Disadvantage: We lose some prediction accuracy.

Sample Splitting + LOCO

G'Sell, Lei, Rinaldo, Tibshirani and Wasserma (2016).
Define variable importance directly:

$$\phi_S(j) = \mathrm{median}\left(|Y - \widehat{\mu}_{(-j)}(X)| - |Y - \widehat{\mu}(X)| \;\middle|\; \mathcal{D}_1 \right)$$

Sample Splitting + LOCO

G'Sell, Lei, Rinaldo, Tibshirani and Wasserma (2016).
Define variable importance directly:

$$\phi_S(j) = \mathrm{median}\left(|Y - \widehat{\mu}_{(-j)}(X)| - |Y - \widehat{\mu}(X)| \;\middle|\; \mathcal{D}_1 \right)$$

Using order statistics from $\mathcal{D}_2$ we get a confidence interval C such that

$$\inf_{w} \inf_{\mathrm{all}\ P} P^n(\phi_S(j) \in C) \geq 1 - \alpha.$$

Sample Splitting + LOCO

G'Sell, Lei, Rinaldo, Tibshirani and Wasserma (2016).
Define variable importance directly:

$$\phi_S(j) = \text{median}\left(|Y - \widehat{\mu}_{(-j)}(X)| - |Y - \widehat{\mu}(X)| \;\middle|\; \mathcal{D}_1 \right)$$

Using order statistics from $\mathcal{D}_2$ we get a confidence interval C such that

$$\inf_{w} \inf_{\text{all } P} P^n(\phi_S(j) \in C) \geq 1 - \alpha.$$

Truly distribution free.

A Subsampling Approach

Mentch and Hooker (2016).

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

Use test statistic: $\widehat{D}(x) = \widehat{\mu}(x) - \widehat{\mu}_{(-j)}(x)$.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

Use test statistic: $\widehat{D}(x) = \widehat{\mu}(x) - \widehat{\mu}_{(-j)}(x)$.

Note: we could also use this for linear models.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

Use test statistic: $\widehat{D}(x) = \widehat{\mu}(x) - \widehat{\mu}_{(-j)}(x)$.

Note: we could also use this for linear models.

Advantages: nonparametric, assumption-free. No splitting needed.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

Use test statistic: $\widehat{D}(x) = \widehat{\mu}(x) - \widehat{\mu}_{(-j)}(x)$.

Note: we could also use this for linear models.

Advantages: nonparametric, assumption-free. No splitting needed.

Disadvantages: Does not return a simple model. Non-uniform CLT.

A Subsampling Approach

Mentch and Hooker (2016).

Let $\widehat{\mu}$ be a random forest.

Recall that a U-statistic has the form

$$\frac{1}{\binom{n}{r}} \sum_{\beta} h(Z_{\beta_1}, \ldots, Z_{\beta_r}).$$

However, for a random forest, we have an incomplete, infinite order U-statistic.

They show that $\widehat{\mu}(x) \approx \mathrm{Normal}$. Then they do a LOCO test:

$$H_0 : \mu(x) = \mu_{(-j)}(x).$$

Use test statistic: $\widehat{D}(x) = \widehat{\mu}(x) - \widehat{\mu}_{(-j)}(x)$.

Note: we could also use this for linear models.

Advantages: nonparametric, assumption-free. No splitting needed.

Disadvantages: Does not return a simple model. Non-uniform CLT.

Wager and Athey (2015) have a different approach for random forests for treatment effects.

Conformalization (arXiv:1604.04173)

Suppose we want a model-free confidence set for a future Y.

Conformalization (arXiv:1604.04173)

Suppose we want a model-free confidence set for a future Y.

Linear working model (assumed to be wrong).

Conformalization (arXiv:1604.04173)

Suppose we want a model-free confidence set for a future Y.

Linear working model (assumed to be wrong).

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

Conformalization (arXiv:1604.04173)

Suppose we want a model-free confidence set for a future Y.

Linear working model (assumed to be wrong).

Conformal inference was invented by Vovk et al (1990's).
Construct $C(x)$ such that

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \qquad \text{for all } P.$$

Basic idea

Observe

$$Y_1, \ldots, Y_n$$

Predict new Y_{n+1}.

Basic idea

Observe

$$Y_1, \ldots, Y_n$$

Predict new Y_{n+1}.

1. Fix y. We will test: $H_0 : Y_{n+1} = y$.

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 $Y_{n+1} = y$.

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:

$$|Y_i - \overline{Y}_y|.$$

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:
   $$|Y_i - \overline{Y}_y|.$$
4. Test $H_0 : Y_{n+1} = y$.

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:
   $$|Y_i - \overline{Y}_y|.$$
4. Test $H_0 : Y_{n+1} = y$.
   The p-value is $p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i \geq R_{n+1})$.

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:
   $$|Y_i - \overline{Y}_y|.$$
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).

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:
   $$|Y_i - \overline{Y}_y|.$$
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:
   $$C_n(y) = \{y : \ p(y) \geq \alpha\}.$$

Validity

For any P and any n,

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

Validity

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.

Linear Regression (with model selection)

- Data $\Longrightarrow$ model selection $\Longrightarrow \widehat{\beta}$

Linear Regression (with model selection)

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

Linear Regression (with model selection)

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

Linear Regression (with model selection)

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

Linear Regression (with model selection)

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

Linear Regression (with model selection)

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

Linear Regression (with model selection)

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

Conformalization

Methods for high-dimensional conformal inference are developed in arXiv:1604.04173.

Conformalization

Methods for high-dimensional conformal inference are developed in arXiv:1604.04173.

Includes: new theory, simulation studies.

Conformalization

Methods for high-dimensional conformal inference are developed in arXiv:1604.04173.

Includes: new theory, simulation studies.

R package conformalInference
https://github.com/ryantibs/conformal

CAUSAL INFERENCE

We have treatment variable $W_i \in \{0, 1\}$. Counterfactuals $(Y(0), Y(1))$.

CAUSAL INFERENCE

We have treatment variable $W_i \in \{0, 1\}$. Counterfactuals $(Y(0), Y(1))$.

$$Y = \begin{cases} Y(0) & \text{if } W = 0 \qquad\qquad (Y(1) \text{ is unobserved}) \\ Y(1) & \text{if } W = 1 \qquad\qquad (Y(0) \text{ is unobserved}). \end{cases}$$

CAUSAL INFERENCE

We have treatment variable $W_i \in \{0,1\}$. Counterfactuals $(Y(0), Y(1))$.

$$Y = \begin{cases} Y(0) & \text{if } W = 0 \qquad (Y(1) \text{ is unobserved}) \\ Y(1) & \text{if } W = 1 \qquad (Y(0) \text{ is unobserved}). \end{cases}$$

$$\mu(x, w) = \mathbb{E}[Y|X = x, W = w]$$

ML vs Statistics vs Economics

In ML we want low prediction error

$$\mathbb{E}[(Y - \widehat{\mu}(X))^2].$$

Balance bias and variance.

ML vs Statistics vs Economics

In ML we want low prediction error

$$\mathbb{E}[(Y - \widehat{\mu}(X))^2].$$

Balance bias and variance.

In economics, we may want to estimate the causal effect

$$\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

Not identifiable unless we assume no unmeasured confounding:

$$W \perp\!\!\!\perp (Y(1), Y(0)) \;\Big|\; X.$$

In that case

$$\theta = \int [\mu(x,1) - \mu(x,0)] dP(x)$$

The Causal Effect

$$\theta = \int [\mu(x,1) - \mu(x,0)] dP(x)$$

The Causal Effect

$$\theta = \int [\mu(x, 1) - \mu(x, 0)] dP(x)$$

In principle:

$$\widehat{\theta} = \frac{1}{n} \sum_i [\widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)]$$

but:

1. $\mu(x, w)$ may be high-dimensional.
2. This is a semiparametric problem not a prediction problem so we don't want to balance bias and variance.
3. We also want a confidence interval.
4. We want to assess the sensitivity to the "no confounding" assumption.

(See Susan and Victor's talks.)

Sensitivity Analysis

No unmeasured confounding:

$$W \perp\!\!\!\perp (Y(1), Y(0)) \;\Big|\; X.$$

Sensitivity to this assumption.

Sensitivity Analysis

No unmeasured confounding:

$$W \perp\!\!\!\perp (Y(1), Y(0)) \;\Big|\; X.$$

Sensitivity to this assumption.

Large literature:
Robins (1999), Rosenbaum (2002), Richardosn et al (2014), Imbens (2003), Brumback et al (2004), Blackwell (2013), ...

Sensitivity Analysis

No unmeasured confounding:

$$W \perp\!\!\!\perp (Y(1), Y(0)) \;\Big|\; X.$$

Sensitivity to this assumption.

Large literature:
Robins (1999), Rosenbaum (2002), Richardosn et al (2014), Imbens (2003), Brumback et al (2004), Blackwell (2013), ...

Let

$$q(1, x) = \mathbb{E}[Y(1)|W = 1, X = x] - \mathbb{E}[Y(1)|W = 0, X = x]$$

$$q(0, x) = \mathbb{E}[Y(0)|W = 0, X = x] - \mathbb{E}[Y(0)|W = 1, X = x].$$

These are 0 if there is no unmeasured confounding.

Sensitivity Analysis

Assume $q(1, x) = q(0, x) = a$.

Sensitivity Analysis

Assume $q(1, x) = q(0, x) = a$.

Let

$$\widetilde{Y}_i = Y_i - a\,\pi(1 - W_i|X_i)$$

where

$$\pi(w|x) = P(W = w|X = x).$$

Sensitivity Analysis

Assume $q(1, x) = q(0, x) = a$.

Let

$$\widetilde{Y}_i = Y_i - a\,\pi(1 - W_i | X_i)$$

where

$$\pi(w|x) = P(W = w | X = x).$$

Replace Y_i's with $\widetilde{Y}_i$'s. Repeat for every value of a.

Sensitivity Analysis

Assume $q(1,x) = q(0,x) = a$.

Let

$$\widetilde{Y}_i = Y_i - a\,\pi(1 - W_i|X_i)$$

where

$$\pi(w|x) = P(W = w|X = x).$$

Replace Y_i's with $\widetilde{Y}_i$'s. Repeat for every value of a.

But in practice, we need to estimate $\pi(w|x)$ (another high-dimensional regression).

Sensitivity Analysis

Assume $q(1,x) = q(0,x) = a$.

Let

$$\widetilde{Y}_i = Y_i - a\,\pi(1 - W_i|X_i)$$

where

$$\pi(w|x) = P(W = w|X = x).$$

Replace Y_i's with $\widetilde{Y}_i$'s. Repeat for every value of a.

But in practice, we need to estimate $\pi(w|x)$ (another high-dimensional regression).

We should do a sensitivity analysis for that too.

Sensitivity Analysis

Assume $q(1, x) = q(0, x) = a$.

Let

$$\widetilde{Y}_i = Y_i - a\,\pi(1 - W_i|X_i)$$

where

$$\pi(w|x) = P(W = w|X = x).$$

Replace Y_i's with $\widetilde{Y}_i$'s. Repeat for every value of a.

But in practice, we need to estimate $\pi(w|x)$ (another high-dimensional regression).

We should do a sensitivity analysis for that too.

Double sensitivity analysis?

CONCLUSION

CONCLUSION

There are now many methods for inference with high-dimensional models.

CONCLUSION

There are now many methods for inference with high-dimensional models.

Sample splitting is simple and gives valid, robust confidence intervals.

CONCLUSION

There are now many methods for inference with high-dimensional models.

Sample splitting is simple and gives valid, robust confidence intervals.

Forget about β; there are better parameters to estimate.

CONCLUSION

There are now many methods for inference with high-dimensional models.

Sample splitting is simple and gives valid, robust confidence intervals.

Forget about β; there are better parameters to estimate.

Conformal methods for distribution free predictive inference.

CONCLUSION

There are now many methods for inference with high-dimensional models.

Sample splitting is simple and gives valid, robust confidence intervals.

Forget about β; there are better parameters to estimate.

Conformal methods for distribution free predictive inference.

For causal inference, need to develop method to assess sensitivity to unobserved confounding in the high-dimensional setting.

CONCLUSION

There are now many methods for inference with high-dimensional models.

Sample splitting is simple and gives valid, robust confidence intervals.

Forget about β; there are better parameters to estimate.

Conformal methods for distribution free predictive inference.

For causal inference, need to develop method to assess sensitivity to unobserved confounding in the high-dimensional setting.

THE END