# The jackknife+: distribution-free prediction

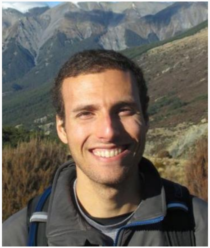
Rina Foygel Barber
(with Emmanuel Candès, Aaditya Ramdas, Ryan Tibshirani)

http://www.stat.uchicago.edu/~rina/

# Collaborators

Emmanuel Candès

Aaditya Ramdas



Ryan Tibshirani

- Thanks to American Institute of Math (AIM)

# The prediction problem

Setting:

- Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ $\rightsquigarrow$ fit model $\widehat{\mu}(X_i) \approx Y_i$
- Test point $(X_{n+1}, Y_{n+1})$ from same distribution
- If $\widehat{\mu}$ overfits to training data,

$$|Y_{n+1} - \widehat{\mu}(X_{n+1})| \gg \frac{1}{n} \sum_{i=1}^{n} |Y_i - \widehat{\mu}(X_i)|$$

# The prediction problem

Goal: build prediction band $C$ as a function of the training data, such that $C(X_{n+1})$ is likely to contain $Y_{n+1}$

- Want to be <u>distribution-free</u> — coverage holds w/o assumptions on distrib. of $(X, Y)$
- Want to be <u>efficient</u> — minimize width of interval $C(X_{n+1})$

# Defining distribution-free coverage

**Definition:**

A method satisfies distribution-free coverage at level $1-\alpha$ if

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

w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$, for any $P$.

# Defining distribution-free coverage

**Definition:**

A method satisfies distribution-free coverage at level $1 - \alpha$ if

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

w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \stackrel{\text{iid}}{\sim} P$, for any $P$.

The coverage rate is averaged over training data & over test points

# Preliminaries

1. Quantiles of a sample:

   $\mathrm{Q}_{1-\alpha}(x_1, \ldots, x_n) = \lceil (1-\alpha)(n+1) \rceil$-th smallest value of $x_1, \ldots, x_n$

   Abusing notation...

   $\mathrm{Q}_{\alpha}(x_1, \ldots, x_n) = \lfloor \alpha(n+1) \rfloor$-th smallest value of $x_1, \ldots, x_n$

# Preliminaries

1. Quantiles of a sample:

   $Q_{1-\alpha}(x_1, \ldots, x_n) = \lceil (1-\alpha)(n+1) \rceil$-th smallest value of $x_1, \ldots, x_n$

   Abusing notation...

   $Q_{\alpha}(x_1, \ldots, x_n) = \lfloor \alpha(n+1) \rfloor$-th smallest value of $x_1, \ldots, x_n$

2. Fitted regression function

   $$\widehat{\mu} = \underbrace{\mathcal{A}\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)}_{\text{invariant to ordering of training points}}$$

   For any subset $S \subset \{1, \ldots, n\}$,

   $$\widehat{\mu}_{-S} = \mathcal{A}\Big((X_i, Y_i) : i \in \{1, \ldots, n\} \backslash S\Big)$$

# Some options...

Naive method:

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\Big(|Y_i - \widehat{\mu}(X_i)|\Big)$$

— Computational cost: one regression $\widehat{\mu}$

— Poor coverage in practice (overfitting)

# Some options...

Holdout method:

Choose holdout set $S_{\text{hold}} \subset \{1, \ldots, n\}$

$$C(X_{n+1}) = \widehat{\mu}_{-S_{\text{hold}}}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\Big(|Y_i - \widehat{\mu}_{-S_{\text{hold}}}(X_i)| \ : \ i \in S_{\text{hold}}\Big)$$

— Computational cost: one regression $\widehat{\mu}_{-S_{\text{hold}}}$

— Wider intervals (reduced sample size $n - |S_{\text{hold}}| < n$)

[1] Papadopoulos 2008, Vovk 2012, Lei et al. 2018

# Some options...

Holdout method:

Choose holdout set $S_{\text{hold}} \subset \{1, \ldots, n\}$

$$C(X_{n+1}) = \widehat{\mu}_{-S_{\text{hold}}}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\Big( |Y_i - \widehat{\mu}_{-S_{\text{hold}}}(X_i)| \ : \ i \in S_{\text{hold}} \Big)$$

— Computational cost: one regression $\widehat{\mu}_{-S_{\text{hold}}}$

— Wider intervals (reduced sample size $n - |S_{\text{hold}}| < n$)

— Distribution-free coverage due to exchangeability[1]
of $|Y_i - \widehat{\mu}_{-S_{\text{hold}}}(X_i)|$, $i \in S_{\text{hold}} \cup \{n+1\}$

[1] Papadopoulos 2008, Vovk 2012, Lei et al. 2018

# Some options...

Cross-validation:

Split data into $S_1 \cup \cdots \cup S_K$

For each $i \in S_k$, $R_i^{\mathrm{CV}} = |Y_i - \widehat{\mu}_{-S_k}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\Big(R_i^{\mathrm{CV}}\Big)$$

— Computational cost: $K + 1$ regressions

# Some options...

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)
Residuals $R_i^{\mathsf{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathsf{Q}_{1-\alpha}\left(R_i^{\mathsf{LOO}}\right)$$

[2] Steinberger & Leeb 2018

# Some options...

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)
Residuals $R_i^{\mathrm{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(R_i^{\mathrm{LOO}}\right)$$

— Predictive coverage holds under assumptions[2]

Asymptotic stability: $\widehat{\mu}(X_{n+1}) \approx \widehat{\mu}_{-i}(X_{n+1})$

[2] Steinberger & Leeb 2018

# Some options...

Cross-conformal prediction:[3]

Split data into $S_1 \cup \cdots \cup S_K$,

& find all $y$ overlapping $\geq n - (1-\alpha)(n+1)$

of the intervals $\widehat{\mu}_{-S_k}(X_{n+1}) \pm R_i^{\text{CV}}$

— Computational cost: $K+1$ regressions

— Distribution-free theory

Coverage $\geq 1 - 2\alpha - 2K/n$

[3] Vovk 2015, Vovk et al 2018

# Our methods

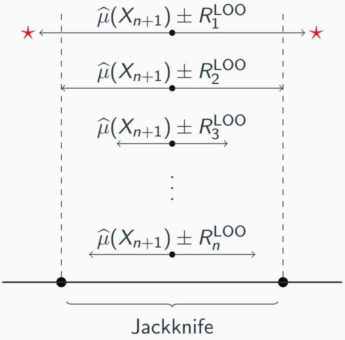
Jackknife+:

$$C(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left(\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}\right),\ \mathrm{Q}_{1-\alpha}\left(\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}\right)\right]$$

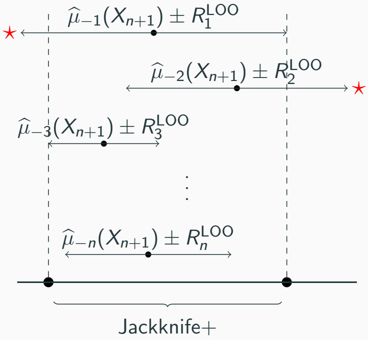
Compare to jackknife:

$$C(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left(\widehat{\mu}(X_{n+1}) - R_i^{\mathrm{LOO}}\right),\ \mathrm{Q}_{1-\alpha}\left(\widehat{\mu}(X_{n+1}) + R_i^{\mathrm{LOO}}\right)\right]$$

# Our methods



$\widehat{\mu}(X_{n+1}) \pm R_1^{\mathrm{LOO}}$
$\widehat{\mu}(X_{n+1}) \pm R_2^{\mathrm{LOO}}$
$\widehat{\mu}(X_{n+1}) \pm R_3^{\mathrm{LOO}}$
$\vdots$
$\widehat{\mu}(X_{n+1}) \pm R_n^{\mathrm{LOO}}$
Jackknife

$\widehat{\mu}_{-1}(X_{n+1}) \pm R_1^{\mathrm{LOO}}$
$\widehat{\mu}_{-2}(X_{n+1}) \pm R_2^{\mathrm{LOO}}$
$\widehat{\mu}_{-3}(X_{n+1}) \pm R_3^{\mathrm{LOO}}$
$\vdots$
$\widehat{\mu}_{-n}(X_{n+1}) \pm R_n^{\mathrm{LOO}}$
Jackknife+

# Our methods

Extension to CV+:

$$C(X_{n+1}) = \left[ \mathsf{Q}_{\alpha}\left( \widehat{\mu}_{-S_k}(X_{n+1}) - R_i^{\mathrm{CV}} \right),\ \mathsf{Q}_{1-\alpha}\left( \widehat{\mu}_{-S_k}(X_{n+1}) + R_i^{\mathrm{CV}} \right) \right]$$

CV+ interval $\supseteq$ Vovk's cross-conformal prediction set

# Distribution-free coverage

**Theorem:** For any distrib. $P$ and any $\mathcal{A}$, jackknife+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - 2\alpha.$$

# Distribution-free coverage

**Theorem:** For any distrib. $P$ and any $\mathcal{A}$, jackknife+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - 2\alpha.$$

**Theorem:** For any distrib. $P$ and any $\mathcal{A}$, $K$-fold CV+ satisfies

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq \begin{cases} 1 - 2\alpha - 1/K & \text{(new)} \\ 1 - 2\alpha - 2K/n & \text{(Vovk et al)} \end{cases}$$

$$\rightsquigarrow \quad \geq 1 - 2\alpha - \sqrt{2/n}.$$

# Proof sketch (jackknife+)

Define:

- Regression $\widetilde{\mu}_{-\{i,j\}}$ for each $i, j \in \{1, \ldots, n+1\}$
- Residuals $R_{ij} = |Y_i - \widetilde{\mu}_{-\{i,j\}}(X_i)|$
- Comparison matrix for $R_{ij}$ vs $R_{ji}$:

$$A_{ij} = \mathbf{1}\{R_{ij} > R_{ji}\} \quad \text{(and } A_{ii} = 0\text{)}$$

# Proof sketch (jackknife+)

Define:

- Regression $\widetilde{\mu}_{-\{i,j\}}$ for each $i, j \in \{1, \ldots, n+1\}$
- Residuals $R_{ij} = |Y_i - \widetilde{\mu}_{-\{i,j\}}(X_i)|$
- Comparison matrix for $R_{ij}$ vs $R_{ji}$:

$$A_{ij} = \mathbf{1}\{R_{ij} > R_{ji}\} \quad \text{(and } A_{ii} = 0\text{)}$$

Verify: if $Y_{n+1} \notin C(X_{n+1})$ then

$$\sum_{i=1}^{n} \mathbf{1}\Big\{ |Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1})| > |Y_i - \widehat{\mu}_{-i}(X_i)| \Big\} \geq (1-\alpha)(n+1)$$

# Proof sketch (jackknife+)

Define:

- Regression $\widetilde{\mu}_{-\{i,j\}}$ for each $i, j \in \{1, \ldots, n+1\}$
- Residuals $R_{ij} = |Y_i - \widetilde{\mu}_{-\{i,j\}}(X_i)|$
- Comparison matrix for $R_{ij}$ vs $R_{ji}$:

$$A_{ij} = \mathbf{1}\{R_{ij} > R_{ji}\} \quad (\text{and } A_{ii} = 0)$$

Verify: if $Y_{n+1} \notin C(X_{n+1})$ then

$$\sum_{i=1}^{n} \mathbf{1}\left\{ \overbrace{|Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1})|}^{R_{n+1,i}} > \overbrace{|Y_i - \widehat{\mu}_{-i}(X_i)|}^{R_{i,n+1}} \right\} \geq (1-\alpha)(n+1)$$

# Proof sketch (jackknife+)

Define:

- Regression $\widetilde{\mu}_{-\{i,j\}}$ for each $i, j \in \{1, \ldots, n+1\}$
- Residuals $R_{ij} = |Y_i - \widetilde{\mu}_{-\{i,j\}}(X_i)|$
- Comparison matrix for $R_{ij}$ vs $R_{ji}$:

$$A_{ij} = \mathbf{1}\{R_{ij} > R_{ji}\} \quad \text{(and } A_{ii} = 0\text{)}$$

Verify: if $Y_{n+1} \notin C(X_{n+1})$ then

$$\sum_{i=1}^{n} \underbrace{\mathbf{1}\Big\{ \overbrace{|Y_{n+1} - \widehat{\mu}_{-i}(X_{n+1})|}^{R_{n+1,i}} > \overbrace{|Y_i - \widehat{\mu}_{-i}(X_i)|}^{R_{i,n+1}} \Big\}}_{A_{n+1,i}} \geq (1-\alpha)(n+1)$$

# Proof sketch (jackknife+)

Summary so far:

$$Y_{n+1} \notin C(X_{n+1}) \quad \Rightarrow \quad \sum_i A_{n+1,i} \geq (1-\alpha)(n+1)$$

# Proof sketch (jackknife+)

Summary so far:

$$Y_{n+1} \notin C(X_{n+1}) \quad \Rightarrow \quad \sum_i A_{n+1,i} \geq (1-\alpha)(n+1)$$

Data points are exchangeable, so

$$\mathbb{P}\left\{\sum_i A_{n+1,i} \geq (1-\alpha)(n+1)\right\} = \frac{\mathbb{E}\left[\#\text{ rows } j \text{ with } \sum_i A_{ji} \geq (1-\alpha)(n+1)\right]}{n+1}$$

# Proof sketch (jackknife+)

Summary so far:

$$Y_{n+1} \notin C(X_{n+1}) \quad \Rightarrow \quad \sum_i A_{n+1,i} \geq (1-\alpha)(n+1)$$

Data points are exchangeable, so

$$\mathbb{P}\left\{\sum_i A_{n+1,i} \geq (1-\alpha)(n+1)\right\} = \frac{\mathbb{E}\left[\#\text{ rows } j \text{ with } \sum_i A_{ji} \geq (1-\alpha)(n+1)\right]}{n+1}$$

Deterministically, at most $2\alpha(n+1)$ such rows $\quad \Rightarrow \quad \mathbb{P}\{\ldots\} \leq 2\alpha$

# Proof sketch (jackknife+)

Worst case for # rows $j$ with $\sum_i A_{ji} \geq (1-\alpha)(n+1)$:

| | | |
|---|---|---|
| 50% 1's | all 1's | $2\alpha(n+1)$ rows |
| all 0's | all 0's | $(1-2\alpha)(n+1)$ rows |

# Jackknife-minmax

$$\left[\min_i \widehat{\mu}_{-i}(X_{n+1}) - \mathrm{Q}_{1-\alpha}\left(R_i^{\mathrm{LOO}}\right),\ \max_i \widehat{\mu}_{-i}(X_{n+1}) + \mathrm{Q}_{1-\alpha}\left(R_i^{\mathrm{LOO}}\right)\right]$$

**Theorem:** For any distrib. $P$ and any $\mathcal{A}$, jackknife-minmax satisfies

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

# Results so far

| Method | Assumption-free theory | Typical empirical coverage |
| --- | --- | --- |
| Naive | No guarantee | $< 1 - \alpha$ |
| Jackknife | No guarantee | $\approx 1 - \alpha$ |
| Jackknife+ | $1 - 2\alpha$ coverage | $\approx 1 - \alpha$ |
| Jackknife-minmax | $1 - \alpha$ coverage | $> 1 - \alpha$ |

# Matching lower bound

**Theorem:** There exists a data distribution $P$ and a regression algorithm $\mathcal{A}$ s.t.

1. For the naive method, coverage $= 0$
2. For jackknife, coverage $= 0$

Furthermore if $\alpha \leq \frac{1}{2}$, there exist $P$ and $\mathcal{A}$ s.t.

3. For jackknife+, coverage $\leq 1 - 2\alpha + o(1)$

# Stability

Out-of-sample stability:

$$\mathbb{P}\left\{\left|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})\right| \leq \epsilon\right\} \geq 1 - \nu$$

In-sample stability:

$$\mathbb{P}\left\{\left|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)\right| \leq \epsilon\right\} \geq 1 - \nu$$

Stability $\Rightarrow$ generalization bounds[4], predictive coverage[5]

[4] Bousquet & Elisseeff 2002
[5] Steinberger & Leeb 2018

# Stability

Out-of-sample stability:

$$\mathbb{P}\left\{\left|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})\right| \leq \epsilon\right\} \geq 1 - \nu$$

In-sample stability:

$$\mathbb{P}\left\{\left|\widehat{\mu}(X_i) - \widehat{\mu}_{-i}(X_i)\right| \leq \epsilon\right\} \geq 1 - \nu$$

Stability $\Rightarrow$ generalization bounds[4], predictive coverage[5]

- Example: $K$-nearest-neighbors satisfies out-of-sample stability with $\epsilon = 0$ and $\nu = K/n$

[4]Bousquet & Elisseeff 2002

[5]Steinberger & Leeb 2018

# Stability

**Theorem:** With out-of-sample stability, for jackknife,

$$\mathbb{P}\left\{\mathrm{dist}(Y_{n+1}, C(X_{n+1})) \leq \epsilon\right\} \geq 1 - \alpha - 2\sqrt{\nu}$$

For jackknife+,

$$\mathbb{P}\left\{\mathrm{dist}(Y_{n+1}, C(X_{n+1})) \leq 2\epsilon\right\} \geq 1 - \alpha - 4\sqrt{\nu}$$

If we also assume in-sample stability, then for the naive method,

$$\mathbb{P}\left\{\mathrm{dist}(Y_{n+1}, C(X_{n+1})) \leq 2\epsilon\right\} \geq 1 - \alpha - 4\sqrt{\nu}$$

# Summary of theory

| Method | Assumption-free theory | Out-of-sample stability | In-sample and out-of-sample stab. |
|---|---|---|---|
| Naive | No guarantee | No guarantee | $\approx 1-\alpha$ |
| Jackknife | No guarantee | $\approx 1-\alpha$ | $\approx 1-\alpha$ |
| Jackknife+ | $1-2\alpha$ | $\approx 1-\alpha$ | $\approx 1-\alpha$ |
| Jackknife-mm | $1-\alpha$ | $1-\alpha$ | $1-\alpha$ |

# Simulation

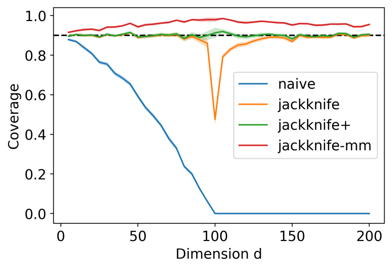
- $n = 100$, $d = 5, 10, \ldots, 200$
- $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
- $Y_i = X_i^\top \beta + \mathcal{N}(0, 1)$
- Regression method $\mathcal{A}$:

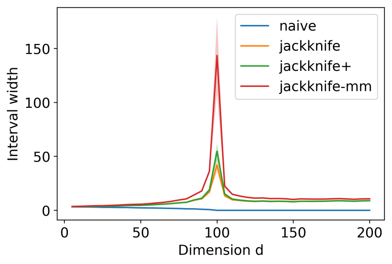
  Least squares with min $\ell_2$ norm (ridge with penalty 0)

  Stable[6] if $d \ll n$ or $d \gg n$, unstable if $d \approx n$

[6] Hastie et al 2019, *Ridgeless Least Squares Interpolation.*

# Simulation


Coverage
Dimension d
naive
jackknife
jackknife+
jackknife-mm
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200


Interval width
Dimension d
naive
jackknife
jackknife+
jackknife-mm
150
100
50
0
0
50
100
150
200

# Real data

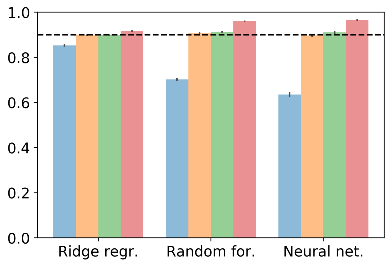
| Data set | # samples | # features |
|---|---|---|
| Communities & crime | 1994 | 99 |
| BlogFeedback | 52397 | 280 |
| Medical Expenditures Panel | 33005 | 107 |

- Training data size $n = 200$
- Algorithms: ridge regression, random forests, neural nets

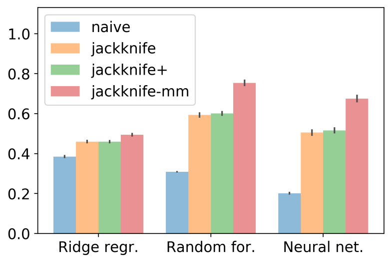
# Real data

Communities & crime data:

Coverage



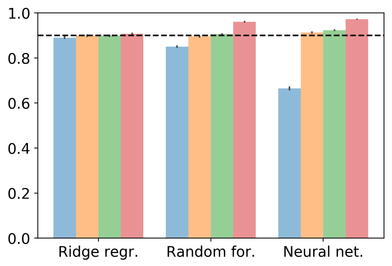

Interval width



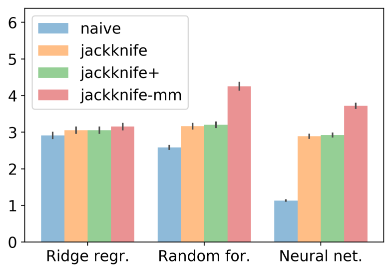
# Real data

BlogFeedback data:

Coverage



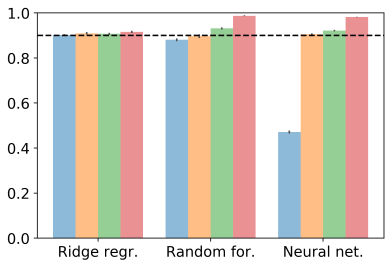

Interval width



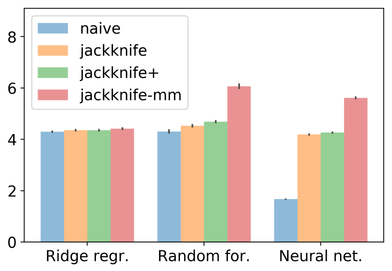
# Real data

Medical Expenditure Panel data:

Coverage




Interval width



# Summary & open questions

- Jackknife can fail to cover if the regression is unstable
- Jackknife+ accounts for instability in $\widehat{\mu}$
  to gain assumption-free $1-2\alpha$ coverage
- If $\widehat{\mu}$ is stable, jackknife+ $\approx$ jackknife (both have $1-\alpha$ cov.)

- What algorithms/data sets cause jackknife to fail in practice?
- How to use jackknife+ or CV+ for model selection / tuning?