Uncertainties in Predictive Inference: Conformal Inference and Cross-Validation

Jing Lei
Department of Statistics and Data Science, Carnegie Mellon University

Conference on Predictive Inference and Its Applications
Iowa State University, Ames, IA

May 8, 2018

Research partially supported by NSF grants DMS-1407771, DMS-1553884

Regression and Prediction

Data: $(X_i, Y_i)_{i=1}^n$ i.i.d from joint distribution with

$$Y = \mu(X) + \varepsilon$$

where

$$\mathbb{E}(\varepsilon \mid X) = 0.$$

Goal

1. learn about μ (estimation).
2. predict Y for future observations of X.

Predictive inference

- We would like to quantify the uncertainty of Y for each X observed in the future or in the sample.
  1. Noise uncertainty: even if we knew μ perfectly, we never observe ε.
  2. Sampling uncertainty: empirical distribution as approximation to underlying population.
  3. Modeling uncertainly: popular assumptions, such as Gaussianity of ε, linearity/smoothness of μ, sparsity, etc, may not be exactly correct.

Examples of assumptions

- Classical nonparametric regression
  - μ is smooth (e.g., Hölder class)
  - X has density bounded away from 0
  - $(\varepsilon \mid X) \sim N(0, \sigma^2)$ or similar
- High dimensional regression
  - $\mu(x) = \beta^T x$ and β is sparse
  - the design matrix is nice (incoherence, RIP, etc)
  - $(\varepsilon \mid X) \sim N(0, \sigma^2)$ or similar
- Neural network: μ can be written as compositions of (structured) multiple index models.
- Inferences based on these assumptions may not be robust.

Outline

- Conformal inference: reliable prediction band under no structural assumptions (joint work with L. Wasserman, R. J. Tibshirani, M. G'Sell, A. Rinaldo)
- Cross-validation with confidence: make better use of validated loss in sampling-splitting.

A naive prediction band

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- Estimate $\hat{\mu}$ (OLS, local polynomial, lasso, NN, etc)
- $R_i = |Y_i - \hat{\mu}(X_i)|$, or any other loss function.
- Prediction band:
  $\hat{\mu}(X_{n+1}) \pm$ upper α-quantile of $\{R_i : 1 \leq i \leq n\}$.
- OK only if $\hat{\mu}$ is very accurate, which requires standard assumptions, as well as good choices of tuning parameters.
- Overfitting: this prediction band tends to be too narrow, because the fitted residuals are smaller than the true values.

Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.

Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.

Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.
- Let $R_i^{(y)} = |Y_i - \hat{\mu}^{(y)}(X_i)|$, $1 \leq i \leq n+1$.

Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.
- Let $R_i^{(y)} = |Y_i - \hat{\mu}^{(y)}(X_i)|$, $1 \le i \le n+1$.
- Quality score: $\pi_n(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}(R_i^{(y)} \le R_{n+1}^{(y)})$

Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.
- Let $R_i^{(y)} = |Y_i - \hat{\mu}^{(y)}(X_i)|$, $1 \le i \le n+1$.
- Quality score: $\pi_n(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}(R_i^{(y)} \le R_{n+1}^{(y)})$
- Output $\hat{C}(X_{n+1}) = \{y \in \mathbb{R} : \pi_n(y) \le 1 - \alpha\}$.

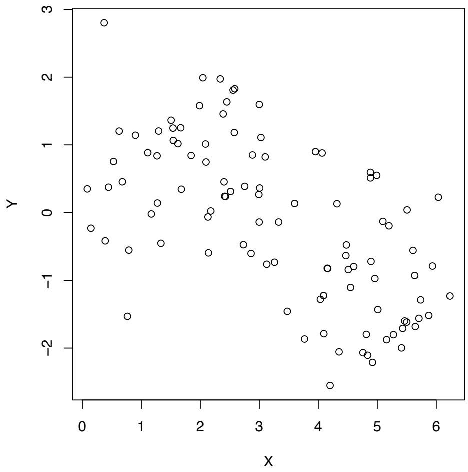
Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.
- Let $R_i^{(y)} = |Y_i - \hat{\mu}^{(y)}(X_i)|$, $1 \leq i \leq n+1$.
- Quality score: $\pi_n(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}(R_i^{(y)} \leq R_{n+1}^{(y)})$
- Output $\hat{C}(X_{n+1}) = \{y \in \mathbb{R} : \pi_n(y) \leq 1 - \alpha\}$.
- The fitting of $\hat{\mu}^{(y)}$ involves (X_{n+1}, y), and hence $\hat{C}$ is immune to overfitting.

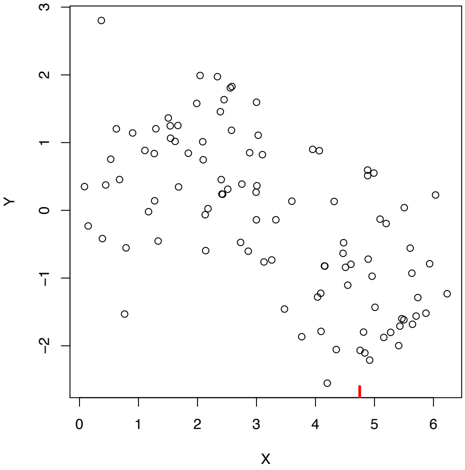
Conformal Prediction

- Data: $(X_i, Y_i)_{i=1}^n$; Goal: predict Y_{n+1} for a future X_{n+1}.
- For each $y \in \mathbb{R}$, let $\hat{\mu}^{(y)}$ be the fitted regression function using the augmented data set $(X_i, Y_i)_{i=1}^{n+1}$ with $Y_{n+1} = y$.
- Let $R_i^{(y)} = |Y_i - \hat{\mu}^{(y)}(X_i)|$, $1 \le i \le n+1$.
- Quality score: $\pi_n(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}(R_i^{(y)} \le R_{n+1}^{(y)})$
- Output $\hat{C}(X_{n+1}) = \{y \in \mathbb{R} : \pi_n(y) \le 1 - \alpha\}$.
- The fitting of $\hat{\mu}^{(y)}$ involves (X_{n+1}, y), and hence $\hat{C}$ is immune to overfitting.
- Theorem: $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \ge 1 - \alpha$, if $(X_i, Y_i)_{i=1}^{n+1}$ is iid.

Example: conformal prediction interval using smoothing splines

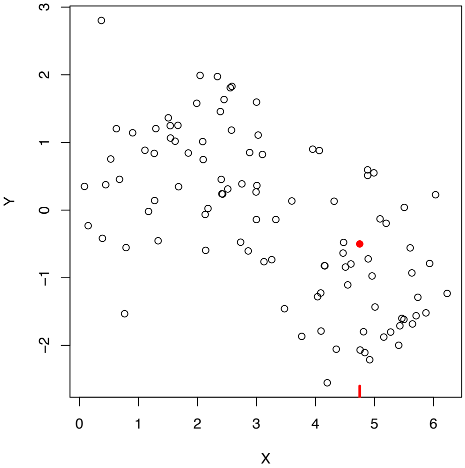


Example: conformal prediction interval using smoothing splines



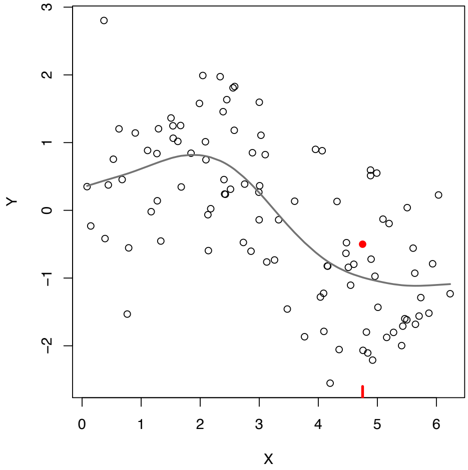
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



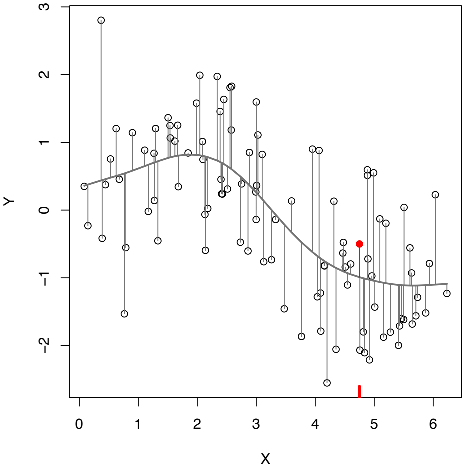
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



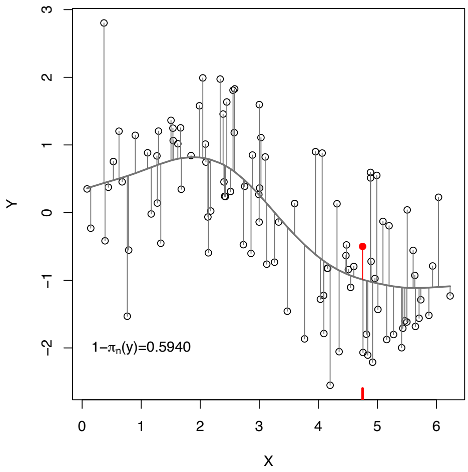
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



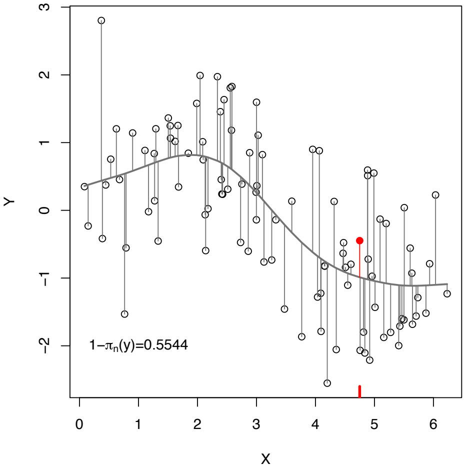
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



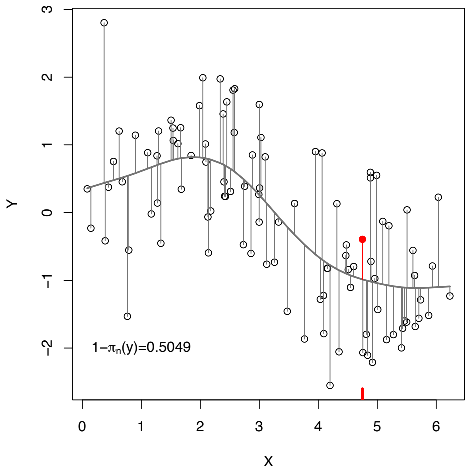
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



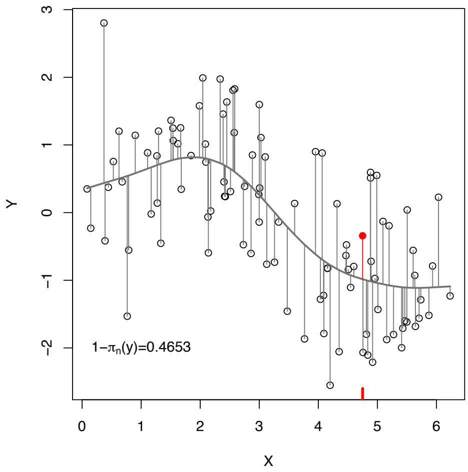
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



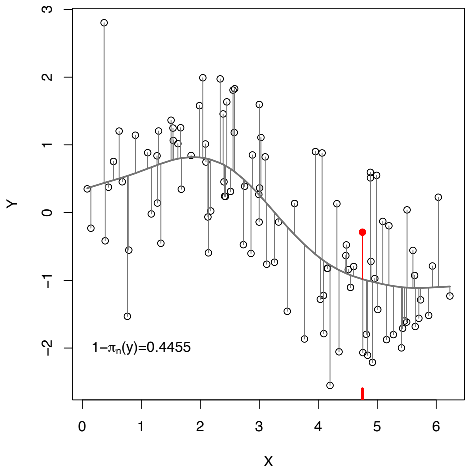
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



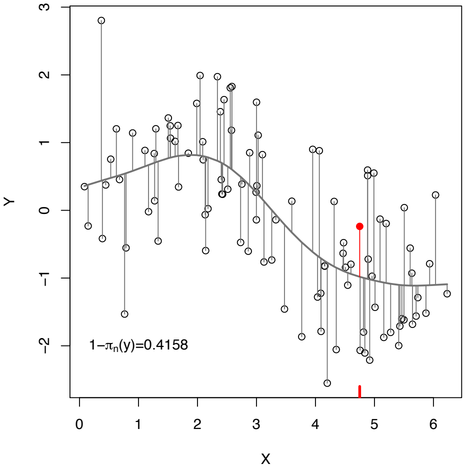
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



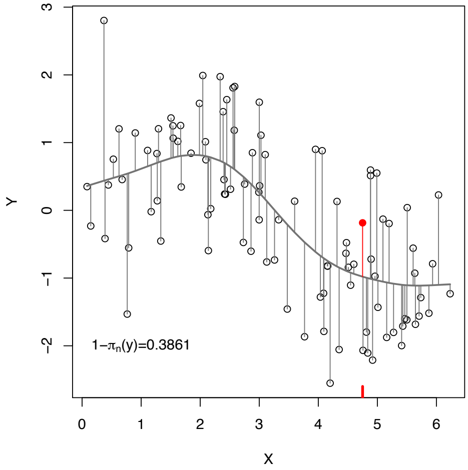
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



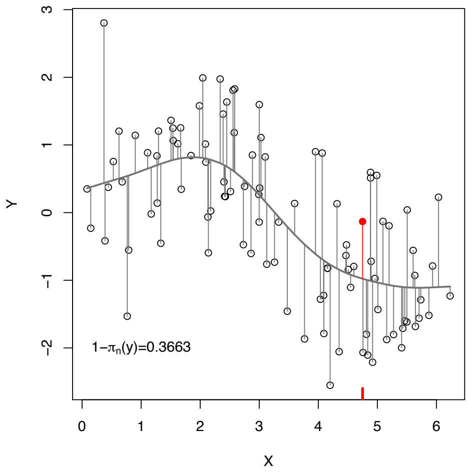
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



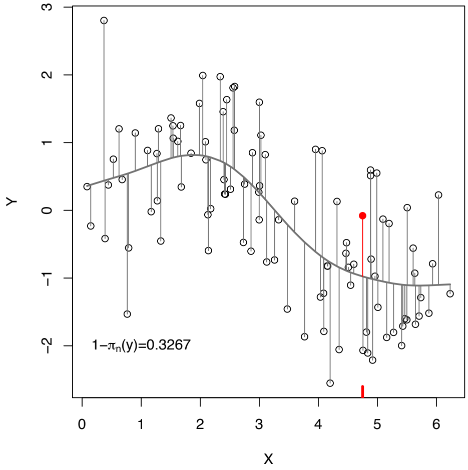
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



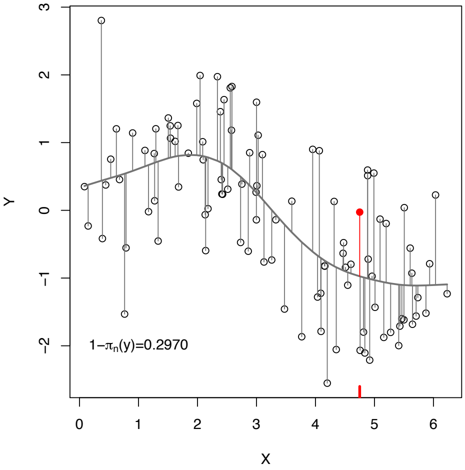
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



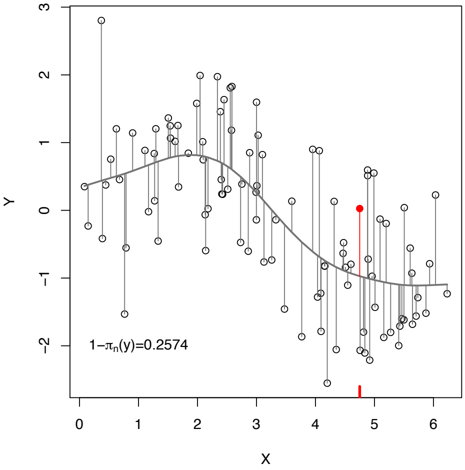
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



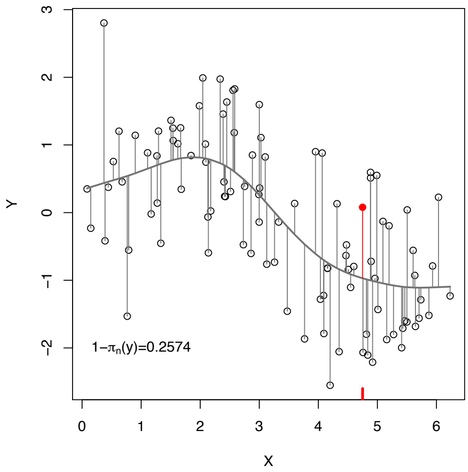
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



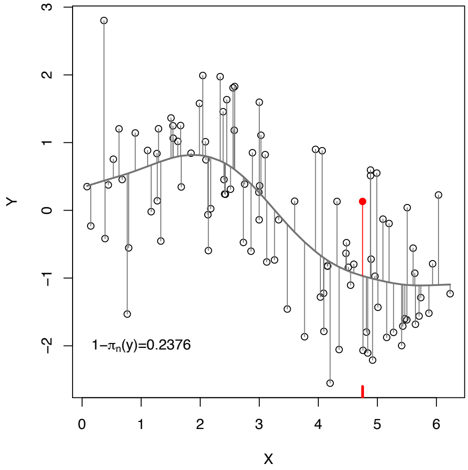
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



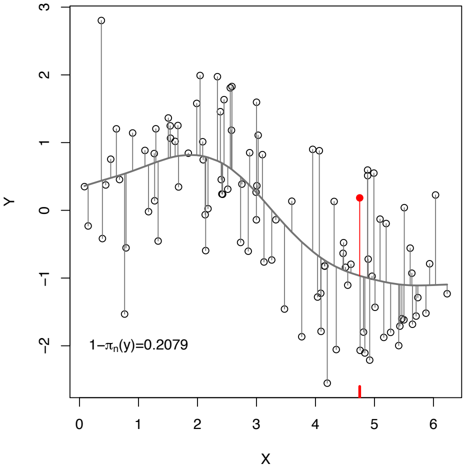
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



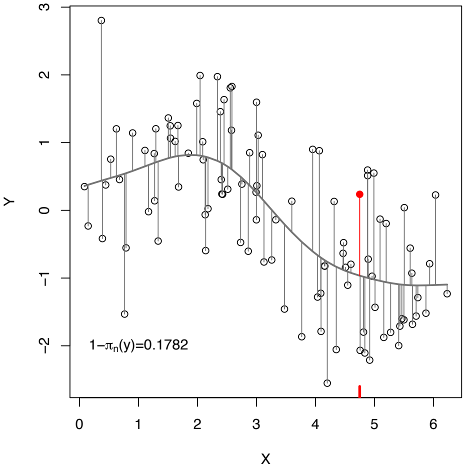
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



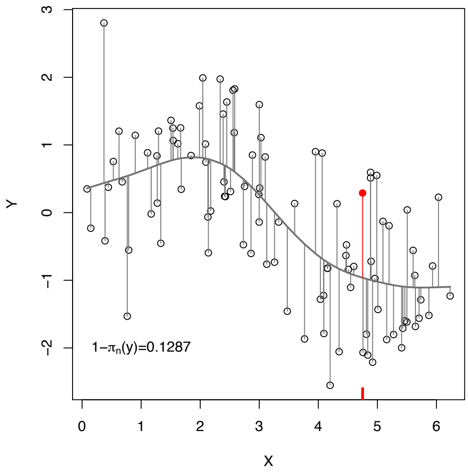
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



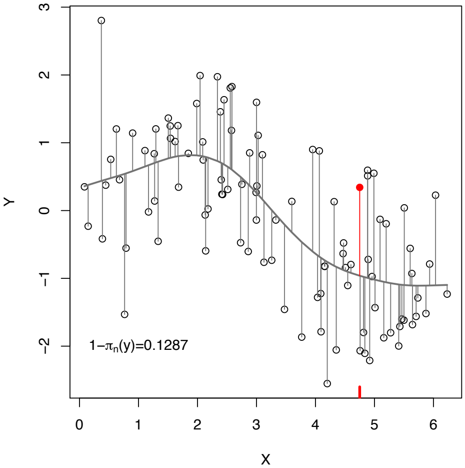
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



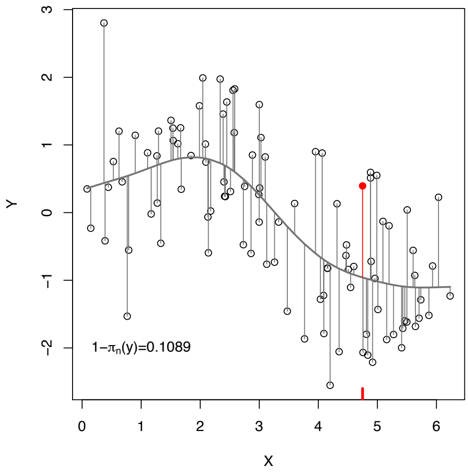
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



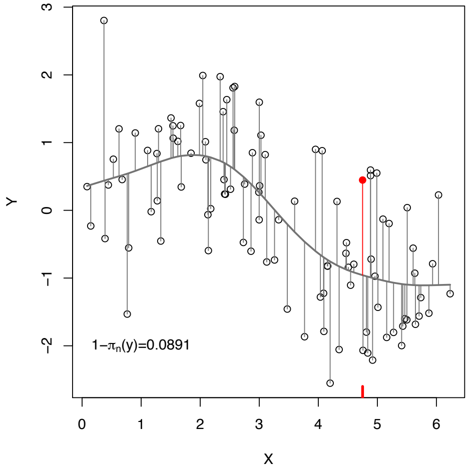
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



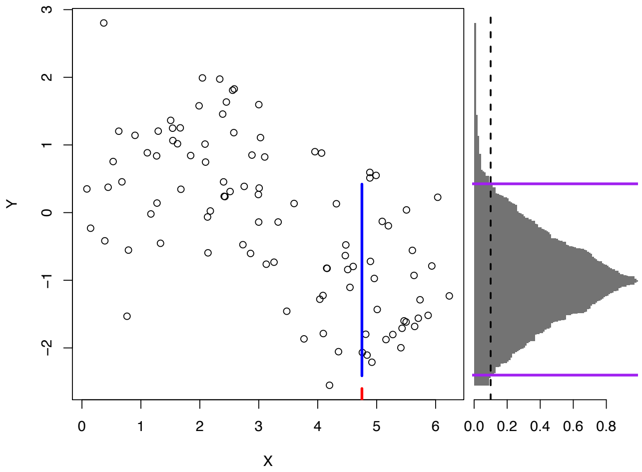
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

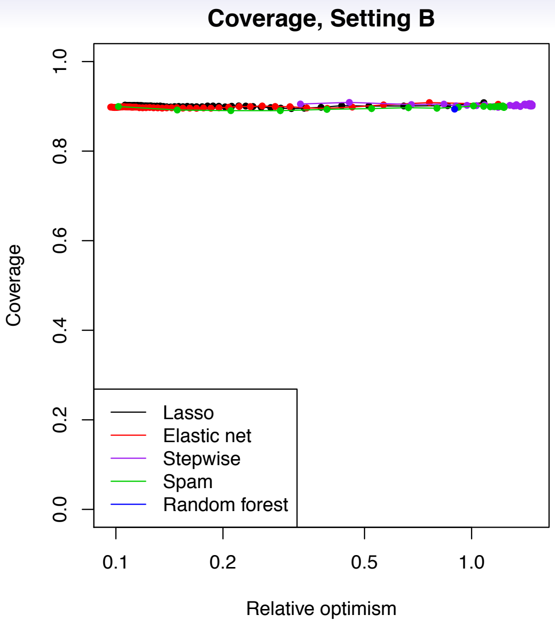
Example: conformal prediction interval using smoothing splines

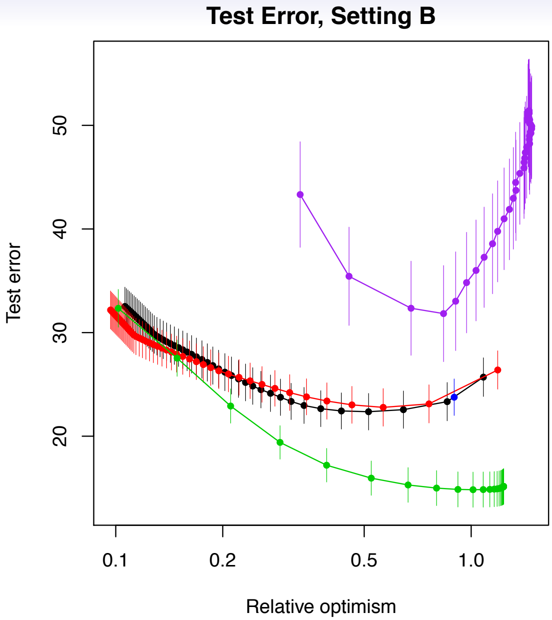


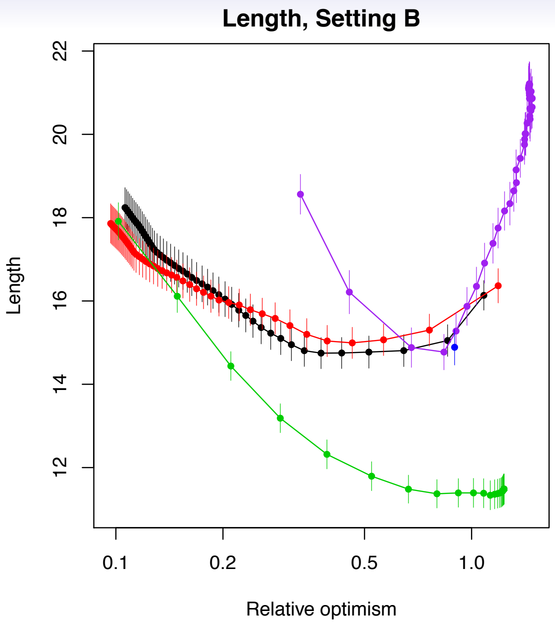
Invert p-values to get conformal interval

A high-dimensional example

- $n = 200, p = 2000$
- $\mathbb{E}(Y|X)$ is mixed additive B-splines on 5 variables.
- $X \sim N(0, I_{2000})$.
- $(\varepsilon \mid X = x) \sim t_2$


Coverage, Setting B
Coverage
1.0
0.8
0.6
0.4
0.2
0.0
Lasso
Elastic net
Stepwise
Spam
Random forest
0.1
0.2
0.5
1.0
Relative optimism


Test Error, Setting B
Test error
Relative optimism
50
40
30
20
0.1
0.2
0.5
1.0


Length, Setting B
Length
Relative optimism
22
20
18
16
14
12
0.1
0.2
0.5
1.0

Remarks

- The coverage is always $1-\alpha$ (anti-conservative) regardless of fitting method and value of tuning parameter.
- Good $\hat{\mu}$ gives short prediction intervals.
- The coverage guarantee is marginal, over the $(n+1)$-tuple $(X_i, Y_i)_{i=1}^{n+1}$.
- Can be combined with almost any point estimator $\hat{\mu}$.

A brief history of conformal prediction

- Developed, since 1996, by V. Vovk and collaborators as a generic tool for online sequential prediction.
- Lei, Robins, & Wasserman (2013): tolerance region.
- Lei & Wasserman (2014): nonparametric regression.
- Lei (2014): binary classification.
- Lei, Rinaldo, & Wasserman (2015): functional clustering.
- Sadinle, Lei, & Wasserman (2015): multi-class classification.
- Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2016): high dimensional regression, variable importance, further insights, R package "`conformalInference`".
- Lei (2017): Fast computation for the Lasso.
- Chernozhukov et al (2018): time series.

Variable importance

- Assume $X \in \mathbb{R}^d$, where d can be large; $\hat{\mu}$ is a fitting algorithm.
- For $j = 1, \ldots, d$, let $\hat{\mu}_{-j}$ be fitted without the jth coordinate of X.
- The jth variable is important if $|Y - \hat{\mu}_{-j}(X)|$ is larger than $|Y - \hat{\mu}(X)|$.
- Need to watch out for overfitting when using $|Y_i - \hat{\mu}_{-j}(X_i)| - |Y_i - \hat{\mu}(X_i)|$.

Variable importance

- Assume $X \in \mathbb{R}^d$, where d can be large; $\hat{\mu}$ is a fitting algorithm.
- For $j = 1, \ldots, d$, let $\hat{\mu}_{-j}$ be fitted without the jth coordinate of X.
- The jth variable is important if $|Y - \hat{\mu}_{-j}(X)|$ is larger than $|Y - \hat{\mu}(X)|$.
- Need to watch out for overfitting when using $|Y_i - \hat{\mu}_{-j}(X_i)| - |Y_i - \hat{\mu}(X_i)|$.
- Idea: make a conformal prediction interval for

$$D_{ij} = |Y_i' - \hat{\mu}_{-j}(X_i)| - |Y_i' - \hat{\mu}(X_i)|$$

where Y_i' is a fresh draw from $(Y|X = X_i)$.

Variable importance

- Let $\tilde{C}(X_i)$ be a valid prediction interval for Y_i' and define

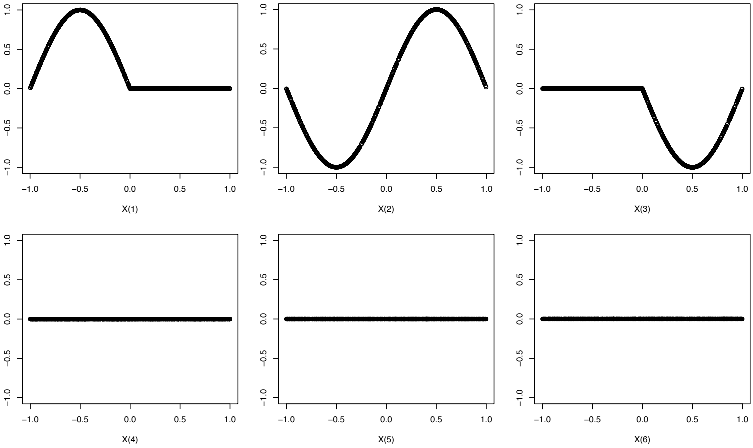
$$V_{ij} = \{|y - \hat{\mu}_{-j}(X_i)| - |y - \hat{\mu}(X_i)| : y \in \tilde{C}(X_i)\}$$

- Fact: $Y_i' \in \tilde{C}(X_i) \Rightarrow D_{ij} \in V_{ij}$, and $\mathbb{P}(D_{ij} \in V_{ij},\ \forall\, j) \geq 1 - \alpha$.
- Can construct conformal prediction band $\tilde{C}(X)$ such that

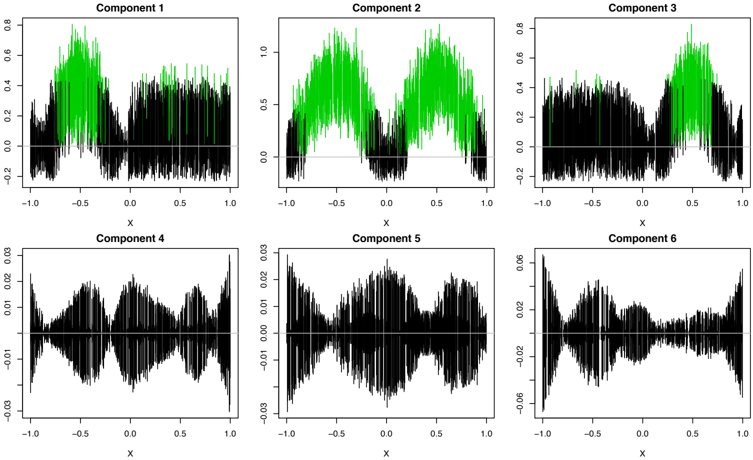
$$\mathbb{P}\left[n^{-1} \sum_{i=1}^{n} \mathbf{1}(D_{ij} \in V_{ij},\ \forall\, j) \geq 1 - \alpha - \varepsilon\right] \geq 1 - 2e^{-cn\varepsilon^2}$$

Example: Additive Model

$$Y = \sum_{j=1}^{6} f_j(X(j)) + N(0,1)$$



How do V_{ij}'s look like?



The jth variable is likely to be important if some of $\{D_{ij} : 1 \leq i \leq n\}$ are above 0.

A higher dimensional example

- $n = 200, p = 100$
- $Y = X^T\beta + \varepsilon$
- $\varepsilon \sim N(0,1)$, independent of X
- $\beta = (2,2,2,0,...,0)^T$
- Design matrix

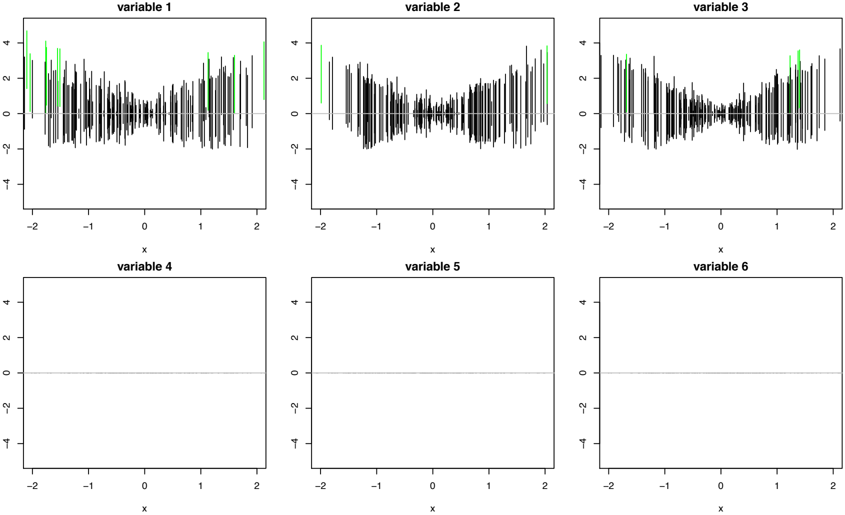
  Case 1: $\mathbb{E}(XX^T) = I$ (all standard assumptions hold)

  Case 2: $\mathrm{corr}(X(j), X(j')) = 0.7$ if $j \neq j'$ (strong correlation)
- Fitting methods

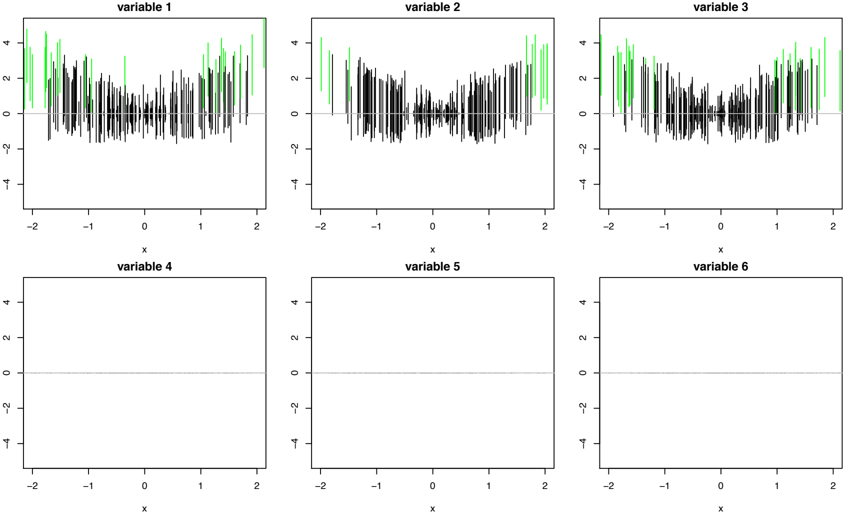
  (a) Lasso with $\lambda = 0.3$

  (b) Forward Stepwise with 3 steps

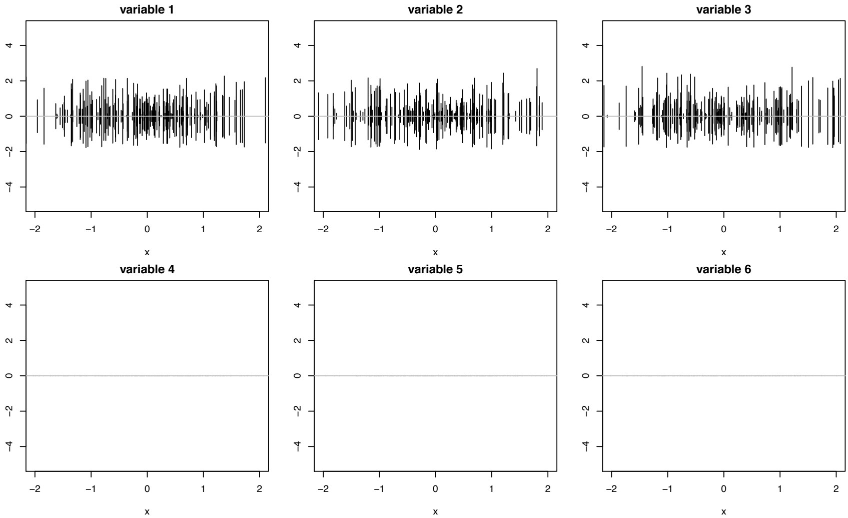
Uncorrelated case, Lasso



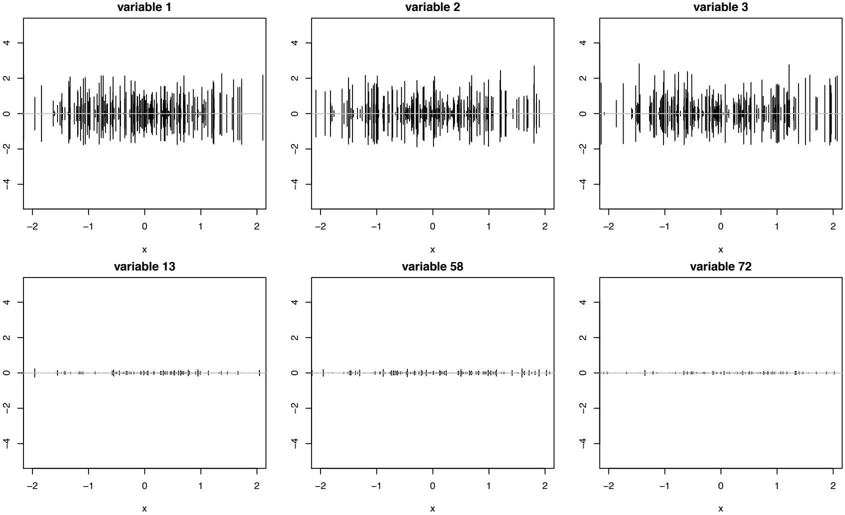
Uncorrelated case, Forward Stepwise



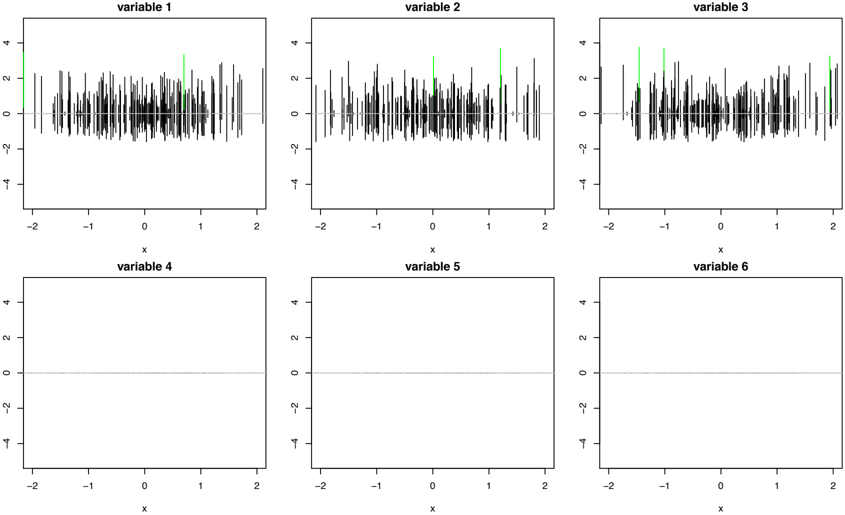
Correlated case, Lasso



Correlated case, Lasso



Correlated case, Forward Stepwise



Construction of $\tilde{C}(X)$

In-sample split conformal:

1. Split data into $\mathscr{D}_1$ and $\mathscr{D}_2$
2. For $k = 1, 2$
   2.1 Let $\hat{\mu}_k$ be fitted using $\mathscr{D}_k$, $k = 1, 2$.
   2.2 Let $\hat{F}_k$ be the empirical CDF of $\{|Y_i - \hat{\mu}_{3-k}(X_i)| : (X_i, Y_i) \in \mathscr{D}_k\}$.
   2.3 For each $X_i \in \mathscr{D}_k$,

$$\tilde{C}(X_i) = [\hat{\mu}_{3-k}(X_i) \pm \hat{F}_k^{-1}(1 - \alpha)]$$

Requires only two fits and two order statistics of cross-validated residuals.

Other topics

- Fast computation: avoid re-fitting $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y.
- Higher order correction: conformal prediction band with adaptive width.
- Theory: when $\hat{\mu}$ is a good estimator, then the conformal band is nearly optimal (requires standard assumptions, mainly relies on stability of $\hat{\mu}$).

From conformalization to cross-validation

- The construction of $\tilde{C}(X)$ reminds us of cross-validation, with just one difference:

  CV looks at the empirical mean of the validated loss, while $\tilde{C}(X)$ looks at the empirical quantiles.

- Idea: there is more information in the validated loss than just the empirical mean.

Cross-validation with confidence

	Parameter est.	Model selection
Point est.	MLE, M-est., ...	Cross-validation
Interval est.	Confidence interval	CVC

In the regression setting

- Data: $D = \{(X_i, Y_i) : 1 \le i \le n\}$, i.i.d from joint distribution P on $\mathbb{R}^p \times \mathbb{R}^1$
- $Y = \mu(X) + \varepsilon$, with $E(\varepsilon \mid X) = 0$
- Loss function: $\ell(\cdot, \cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$
- Goal: find $\hat{\mu} \approx \mu$ so that

$$Q(\hat{\mu}) \equiv \mathbb{E}\left[\ell(\hat{\mu}(X), Y) \mid \hat{\mu}\right]$$

is small.

Model selection

- Candidate set: $\mathscr{M} = \{1, ..., M\}$. Each $m \in \mathscr{M}$ corresponds to a candidate model.
- Given m and data D, there is an estimate $\hat{\mu}(D, m)$ of μ.
- Model selection: find the best m such that it minimizes $Q(\hat{\mu})$ over all $m \in \mathscr{M}$ with high probability.

Cross-validation

- Sample split: Let I_{tr} and I_{te} be a partition of $\{1,...,n\}$.
- Fitting: $\hat{\mu}_m = \hat{\mu}(D_{\text{tr}}, m)$, where $D_{\text{tr}} = \{(X_i, Y_i) : i \in I_{\text{tr}}\}$.
- Validation: $\hat{Q}(\hat{\mu}_m) = n_{\text{te}}^{-1} \sum_{i \in I_{\text{te}}} \ell(\hat{\mu}_m(X_i), Y_i)$.
- CV model selection: $\hat{m}_{\text{cv}} = \arg\min_{m \in \mathscr{M}} \hat{Q}(\hat{\mu}_m)$.
- V-fold cross-validation:
  1. For $V \geq 2$, split the data into V folds.
  2. Rotate over each fold as I_{tr} to obtain $\hat{Q}^{(v)}(\hat{\mu}_m^{(v)})$
  3. $\hat{m} = \arg\min V^{-1} \sum_{v=1}^{V} \hat{Q}^{(v)}(\hat{\mu}_m^{(v)})$
  4. Popular choices of V: 10 and 5.
  5. $V = n$: leave-one-out cross-validation

A simple negative example

- Model: $Y = \mu + \varepsilon$, where $\varepsilon \sim N(0,1)$.
- $\mathscr{M} = \{1,2\}$. $m = 1$: $\mu = 0$; $m = 2$: $\mu \in \mathbb{R}$.
- Truth: $\mu = 0$
- Consider a single split: $\hat{\mu}_1 \equiv 0$, $\hat{\mu}_2 = \bar{\varepsilon}_{\mathrm{tr}}$.
- $\hat{m}_{\mathrm{cv}} = 1 \;\; \Leftrightarrow \;\; 0 < \hat{Q}(\hat{\mu}_2) - \hat{Q}(\hat{\mu}_1) = \bar{\varepsilon}_{\mathrm{tr}}^2 - 2\bar{\varepsilon}_{\mathrm{tr}}\bar{\varepsilon}_{\mathrm{te}}$.
- If $n_{\mathrm{tr}}/n_{\mathrm{te}} \asymp 1$, then $\sqrt{n}\bar{\varepsilon}_{\mathrm{tr}}$ and $\sqrt{n}\bar{\varepsilon}_{\mathrm{te}}$ are independent normal random variables with constant variances. So $\mathbb{P}(\hat{m}_{\mathrm{cv}} = 1)$ is bounded away from 1.

A simple negative example

- Model: $Y = \mu + \varepsilon$, where $\varepsilon \sim N(0,1)$.
- $\mathscr{M} = \{1,2\}$. $m = 1$: $\mu = 0$; $m = 2$: $\mu \in \mathbb{R}$.
- Truth: $\mu = 0$
- Consider a single split: $\hat{\mu}_1 \equiv 0$, $\hat{\mu}_2 = \bar{\varepsilon}_{\text{tr}}$.
- $\hat{m}_{\text{cv}} = 1 \quad \Leftrightarrow \quad 0 < \hat{Q}(\hat{\mu}_2) - \hat{Q}(\hat{\mu}_1) = \bar{\varepsilon}_{\text{tr}}^2 - 2\bar{\varepsilon}_{\text{tr}}\bar{\varepsilon}_{\text{te}}$.
- If $n_{\text{tr}}/n_{\text{te}} \asymp 1$, then $\sqrt{n}\bar{\varepsilon}_{\text{tr}}$ and $\sqrt{n}\bar{\varepsilon}_{\text{te}}$ are independent normal random variables with constant variances. So $\mathbb{P}(\hat{m}_{\text{cv}} = 1)$ is bounded away from 1.
- (Shao 93, Zhang 93, Yang 07) $\hat{m}_{\text{cv}}$ is inconsistent unless $n_{\text{tr}} = o(n)$.

A simple negative example

- Model: $Y = \mu + \varepsilon$, where $\varepsilon \sim N(0,1)$.
- $\mathscr{M} = \{1,2\}$. $m = 1$: $\mu = 0$; $m = 2$: $\mu \in \mathbb{R}$.
- Truth: $\mu = 0$
- Consider a single split: $\hat{\mu}_1 \equiv 0$, $\hat{\mu}_2 = \bar{\varepsilon}_{\mathrm{tr}}$.
- $\hat{m}_{\mathrm{cv}} = 1 \quad \Leftrightarrow \quad 0 < \hat{Q}(\hat{\mu}_2) - \hat{Q}(\hat{\mu}_1) = \bar{\varepsilon}_{\mathrm{tr}}^2 - 2\bar{\varepsilon}_{\mathrm{tr}}\bar{\varepsilon}_{\mathrm{te}}$.
- If $n_{\mathrm{tr}}/n_{\mathrm{te}} \asymp 1$, then $\sqrt{n}\bar{\varepsilon}_{\mathrm{tr}}$ and $\sqrt{n}\bar{\varepsilon}_{\mathrm{te}}$ are independent normal random variables with constant variances. So $\mathbb{P}(\hat{m}_{\mathrm{cv}} = 1)$ is bounded away from 1.
- (Shao 93, Zhang 93, Yang 07) $\hat{m}_{\mathrm{cv}}$ is inconsistent unless $n_{\mathrm{tr}} = o(n)$.
- V-fold does not help!

Cross-Validation with Confidence

- Now suppose we have a set of candidate models $\mathscr{M} = \{1, ..., M\}$.
- Split the data into D_{tr} and D_{te}, and use D_{tr} to obtain $\hat{\mu}_m$ for each m.
- Recall that the model quality is $Q(\hat{\mu}) = \mathbb{E}[\ell(\hat{\mu}(X), Y) \mid \hat{\mu}]$.
- For each m, test hypothesis (conditioning on $\hat{\mu}_1, ..., \hat{\mu}_M$)

$$H_{0,m} : \min_{j \neq m} Q(\hat{\mu}_j) \geq Q(\hat{\mu}_m) .$$

- Let $\hat{p}_m$ be a valid p-value.
- $\mathscr{A}_{\text{cvc}} = \{m : \hat{p}_m > \alpha\}$ is our confidence set for the best fitted model: $\mathbb{P}(m^* \in \mathscr{A}_{\text{cvc}}) \geq 1 - \alpha$, where $m^* = \arg\min_m Q(\hat{\mu}_m)$.

Calculating $\hat{p}_m$

- Recall $H_{0,m} : \min_{j \neq m} Q(\hat{\mu}_j) \geq Q(\hat{\mu}_m)$.
- Consider $n_{\text{te}} \times (M-1)$ matrix (I_{te} is the index set of D_{te})

$$\left[\xi_{m,j}^{(i)}\right]_{i \in I_{\text{te}},\, j \neq m}, \quad \text{where} \quad \xi_{m,j}^{(i)} = \ell(\hat{\mu}_m(X_i), Y_i) - \ell(\hat{\mu}_j(X_i), Y_i)$$

- Multivariate mean testing. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0,\ \forall\, j \neq m$.

Calculating $\hat{p}_m$

- $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \; \forall\, j \neq m.$
- Let $\hat{\mu}_{m,j}$ and $\hat{\sigma}_{m,j}$ be the sample mean and standard deviation of $(\xi_{m,j}^{(i)} : i \in I_{\mathrm{te}})$.
- Naturally, one would reject $H_{0,m}$ for large values of

$$\max_{j \neq m} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}\,.$$

- Approximate the null distribution using high dimensional Gaussian comparison [Chernozhukov et al '12].

Studentized Gaussian Multiplier Bootstrap

1. $T_m = \max_{j \neq m} \sqrt{n_{\text{te}}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}$
2. Let B be the bootstrap sample size. For $b = 1, \ldots, B$,
   - 2.1 Generate iid standard Gaussian ζ_i, $i \in I_{\text{te}}$.
   - 2.2 $T_b^* = \max_{j \neq m} \frac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \frac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i$
3. $\hat{p}_m = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_b^* > T_m)$. correlation.

Properties of CVC

- $\mathscr{A}_{\text{cvc}} = \{m : \hat{p}_m > \alpha\}$.
- Let $\hat{m}_{\text{cv}} = \arg\min_m \hat{Q}(\hat{\mu}_m)$.

Proposition

If $\alpha < 0.5$, then $\mathbb{P}(\hat{m}_{\text{cv}} \in \mathscr{A}_{\text{cvc}}) \to 1$ as $B \to \infty$.

- Can view $\hat{m}_{\text{cv}}$ as the "center" of the confidence set.

Coverage of $\mathscr{A}_{\text{cvc}}$

- Recall $\xi_{m,j} = \ell(\hat{\mu}_m(X), Y) - \ell(\hat{\mu}_j(X), Y)$.
- Let $\mu_{m,j} = \mathbb{E}[\xi_{m,j} \mid \hat{\mu}_m, \hat{\mu}_j]$, $\sigma^2_{m,j} = \text{Var}[\xi_{m,j} \mid \hat{\mu}_m, \hat{\mu}_j]$.

Theorem

Assume that $(\xi_{m,j} - \mu_{m,j})/(A_n \sigma_{m,j})$ has sub-exponential tail for all $m \neq j$ and some $A_n \geq 1$ such that for some $c > 0$

$$A_n^6 \log^7(M \vee n) = O(n^{1-c}).$$

1. If $\max_{j \neq m} \left(\frac{\mu_{m,j}}{\sigma_{m,j}}\right)_+ = o\left(\sqrt{\frac{1}{n \log(M \vee n)}}\right)$, then $\mathbb{P}(m \in \mathscr{A}_{\text{cvc}}) \geq 1 - \alpha + o(1)$.
2. If $\max_{j \neq m} \left(\frac{\mu_{m,j}}{\sigma_{m,j}}\right)_+ \geq C A_n \sqrt{\frac{\log(M \vee n)}{n}}$ for some constant C, and $\alpha \geq n^{-1}$, then $\mathbb{P}(m \in \mathscr{A}_{\text{cvc}}) = o(1)$.

Proof of coverage

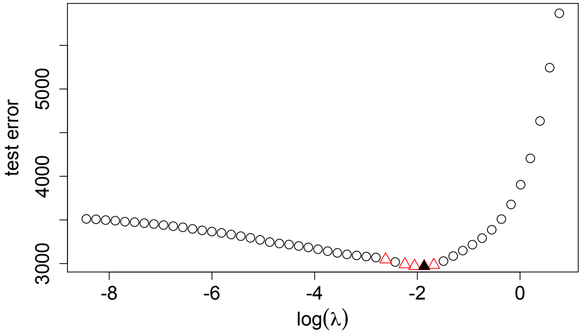
- Let $Z(\Sigma) = \max N(0, \Sigma)$, and $z(1-\alpha, \Sigma)$ its $1-\alpha$ quantile.
- Let $\hat{\Gamma}$ and Γ be sample and population correlation matrices of $(\xi_{m,j}^{(i)})_{i \in I_{\text{te}}, j \neq m}$. When $B \to \infty$,

$$\mathbb{P}(\hat{p}_m \leq \alpha) = \mathbb{P}\left[\max_j \sqrt{n_{\text{te}}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \geq z(1-\alpha, \hat{\Gamma})\right]$$

- Tools (2, 3 are due to Chernozhukov et al.)
  1. Concentration: $\sqrt{n_{\text{te}}} \frac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \leq \sqrt{n_{\text{te}}} \frac{\hat{\mu}_{m,j} - \mu_{m,j}}{\sigma_{m,j}} + o(1/\sqrt{\log M})$
  2. Gaussian comparison: $\max_j \sqrt{n_{\text{te}}} \frac{\hat{\mu}_{m,j} - \mu_{m,j}}{\sigma_{m,j}} \overset{d}{\approx} Z(\Gamma) \overset{d}{\approx} Z(\hat{\Gamma})$
  3. Anti-concentration: $Z(\hat{\Gamma})$ and $Z(\Gamma)$ have densities $\lesssim \sqrt{\log M}$

Example: the diabetes data (Efron et al 04)

- $n = 442$, with 10 covariates: age, sex, bmi, blood pressure, etc.
- Response is diabetes progression after one year.
- Including all quadratic terms, $p = 64$.
- 5-fold CVC with $\alpha = 0.05$, using Lasso with 50 values of λ.



Triangle: models in $\mathscr{A}_{\mathrm{cvc}}$, solid triangle: $\hat{m}_{\mathrm{cv}}$.

The most parsimonious model in $\mathscr{A}_{\text{cvc}}$

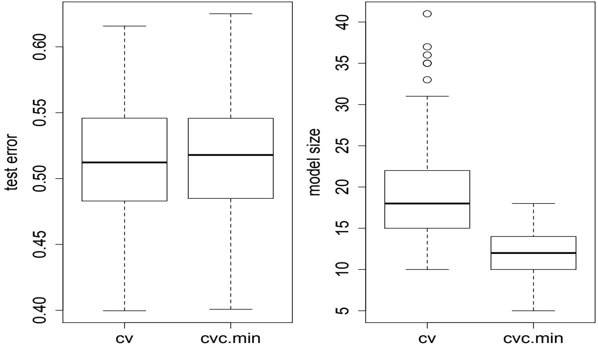
- Let J_m be the subset of variables selected using model m

$$\hat{m}_{\text{cvc.min}} = \arg \min_{m \in \mathscr{A}_{\text{cvc}}} |J_m| .$$

- $\hat{m}_{\text{cvc.min}}$ is the simplest model that gives a similar predictive risk as $\hat{m}_{\text{cv}}$.
- Consistent in low-dimensional linear models with conventional V-fold implement.

The diabetes data revisited

- Split $n = 442$ into 300 (estimation) and 142 (risk approximation).
- 5-fold CVC applied on the 300 sample points, with a final re-fit.
- The final estimate is evaluated using the 142 hold-out sample.
- Repeat 100 times, using Lasso with 50 values of λ.



Summary

- Conformal prediction uses symmetry and out-of-sample fitting to add protection against model misspecification.
- CVC uses hypothesis tests to produce confidence sets for model selection
- Both methods are applicable to many learning algorithms, even black-box type algorithms.

Thanks!

Questions?

"Distribution Free Predictive Inference for Regression"

arXiv:1604.04173 with Wasserman, Tibshirani, G'Sell, Rinaldo

"Cross-Validation with Confidence", arxiv.org/1703.07904

http://www.stat.cmu.edu/~jinglei/talk.shtml