Uncertainties in Predictive Inference: Out-of-Sample Fitting and Cross-Validation

Jing Lei

Department of Statistics, Carnegie Mellon University

Amazon Research Seminar, Palo Alto, Feb 19 2018

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.
2. predict Y for future observations of X.

Popular assumptions for $\hat{\mu}$ in statistics

- 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
- We call these standard assumptions.
- These assumptions lead to practical procedures with good insights, e.g. kernel, local polynomial, Lasso, OMP, etc.

In machine learning

Assumptions about μ are more general and implicit. For example

- μ can be approximated using functions in an RKHS.
- μ can be represented by a neural network with a particular structure.
- Other choices we make when fitting our model: loss function, batch size, number of iterations, etc...

These choices reflect our belief (assumptions) about the underlying function μ and the joint distribution of (X, ε).

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: our assumptions may not be exactly correct. For example, Gaussianity of ε, linearity/smoothness of μ.

Outline

- Conformal inference: reliable prediction band under no structural assumptions.
- Cross-validation with confidence: choosing tuning parameters with better accuracy-interpretability trade-off.

Conformal inference

- What it is
  1. a general framework for predictive inference;
  2. can be combined with (almost) any existing or new regression estimator.
- What it does
  1. converts a point estimate $\hat{\mu}$ to a prediction band
  2. maintains good properties of the original estimator if standard assumptions hold
  3. always guarantees finite sample coverage, with no assumptions other than iid.
- Key idea: when prediction is of interest, we include the potential future data point in our fitting procedure (*conformalization*).

The starting point: sample quantile

- If $Y_1, ..., Y_n \overset{iid}{\sim} P$.
- Let $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$ be the order statistics.
- Let $\alpha \in (0, 1)$ be a constant.
- Then

$$\mathbb{P}\left[Y_{n+1} \leq Y_{(\lceil (n+1)(1-\alpha) \rceil)}\right] \geq 1 - \alpha.$$

- Reason: the rank of Y_{n+1} is uniform on $\{1, ..., n+1\}$.
- Roughly speaking, a $(1-\alpha)$ prediction set for Y_{n+1} is $(-\infty, \hat{F}_n^{-1}(1-\alpha)]$.

How to apply it to regression?

- 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.
- Naïve 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: 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 $R_i^{(y)} = Y_i - \hat{\mu}^{(y)}(X_i)$, $1 \leq i \leq n+1$.
- Quality score: $\pi_n(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}(|R_i^{(y)}| \leq |R_{n+1}^{(y)}|)$

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 $R_i^{(y)} = Y_i - \hat{\mu}^{(y)}(X_i)$, $1 \le i \le n+1$.
- 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.

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.

Conformal Prediction

Idea: The procedure essentially tests the null hypothesis that (X_{n+1}, y) is an independent sample from the same distribution.

Proof: By iid assumption and symmetry, $(R_i^{(Y_{n+1})})_{i=1}^{n+1}$ are exchangeable. Thus $\pi_n(Y_{n+1})$ is a valid p-value.

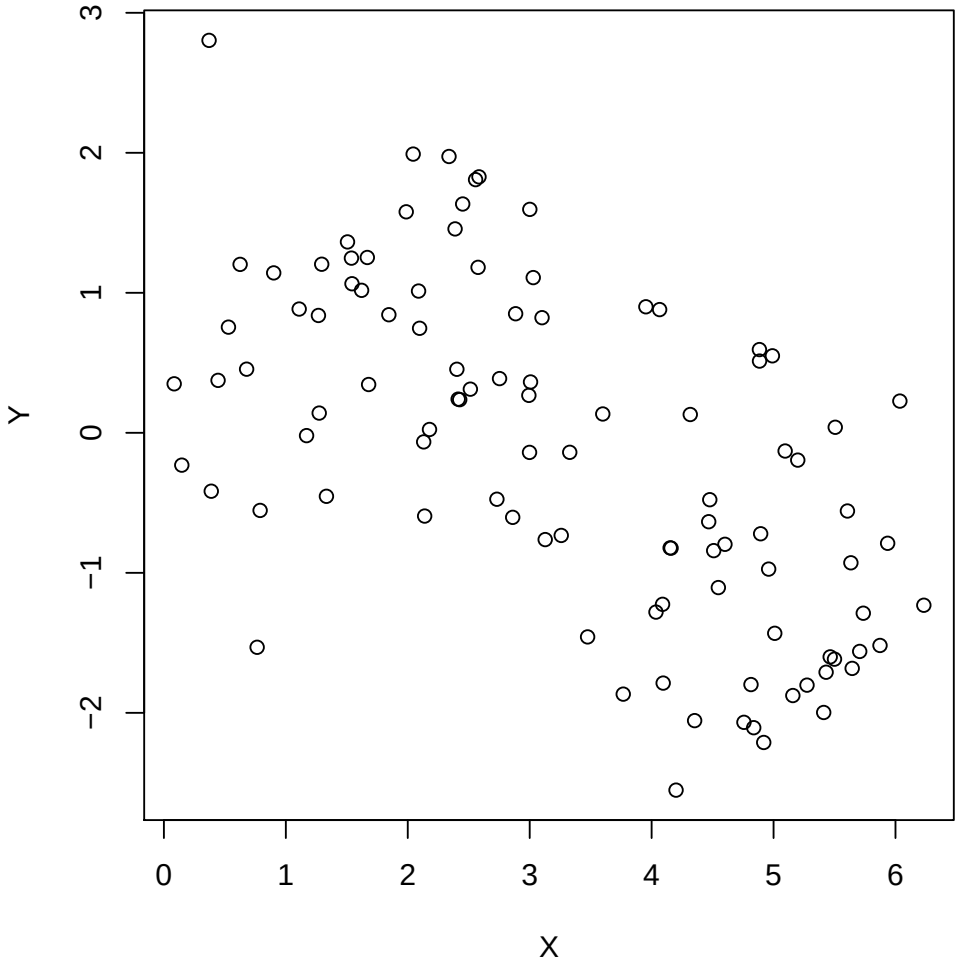
Remark: Can replace $R_i^{(y)}$ by

$$\sigma_i^{(y)} := f(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_{n+1}; Z_i)$$

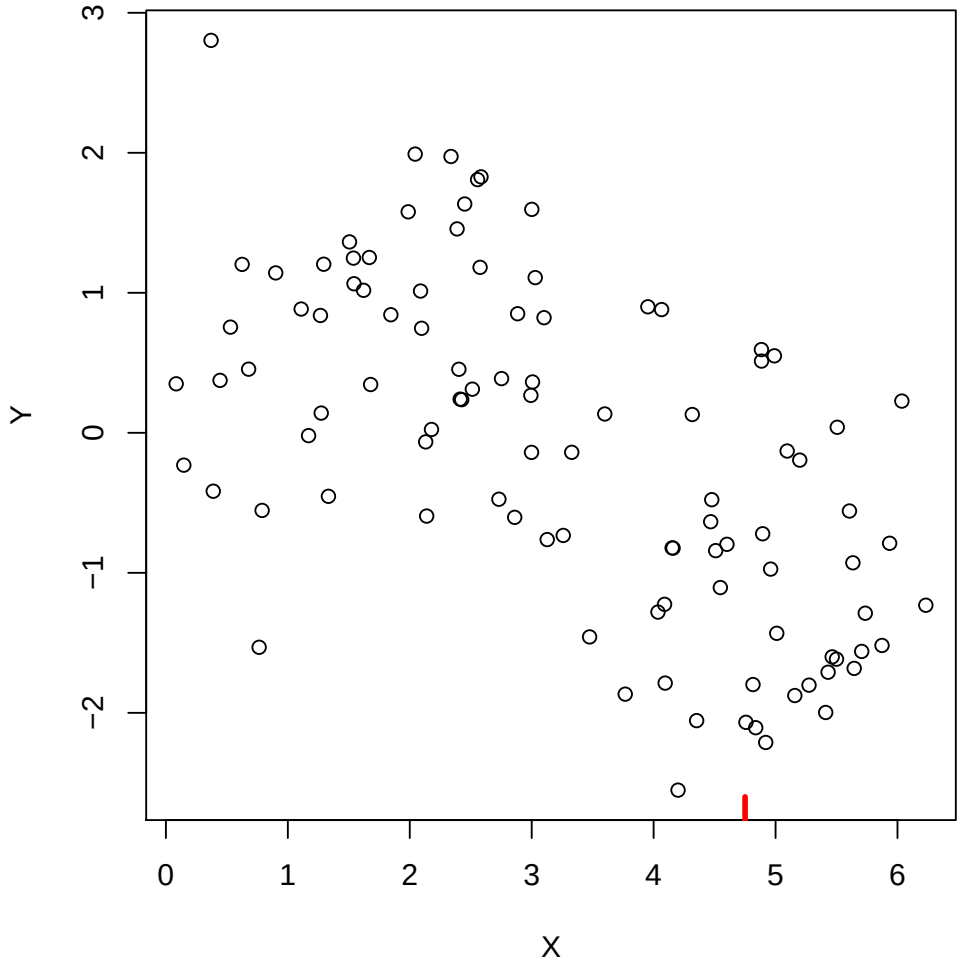
with $Z_i = (X_i, Y_i)$, $Y_{n+1} = y$, for any f that is symmetric in the first n arguments.

f is called the conformity score function.

Example: conformal prediction interval using smoothing splines

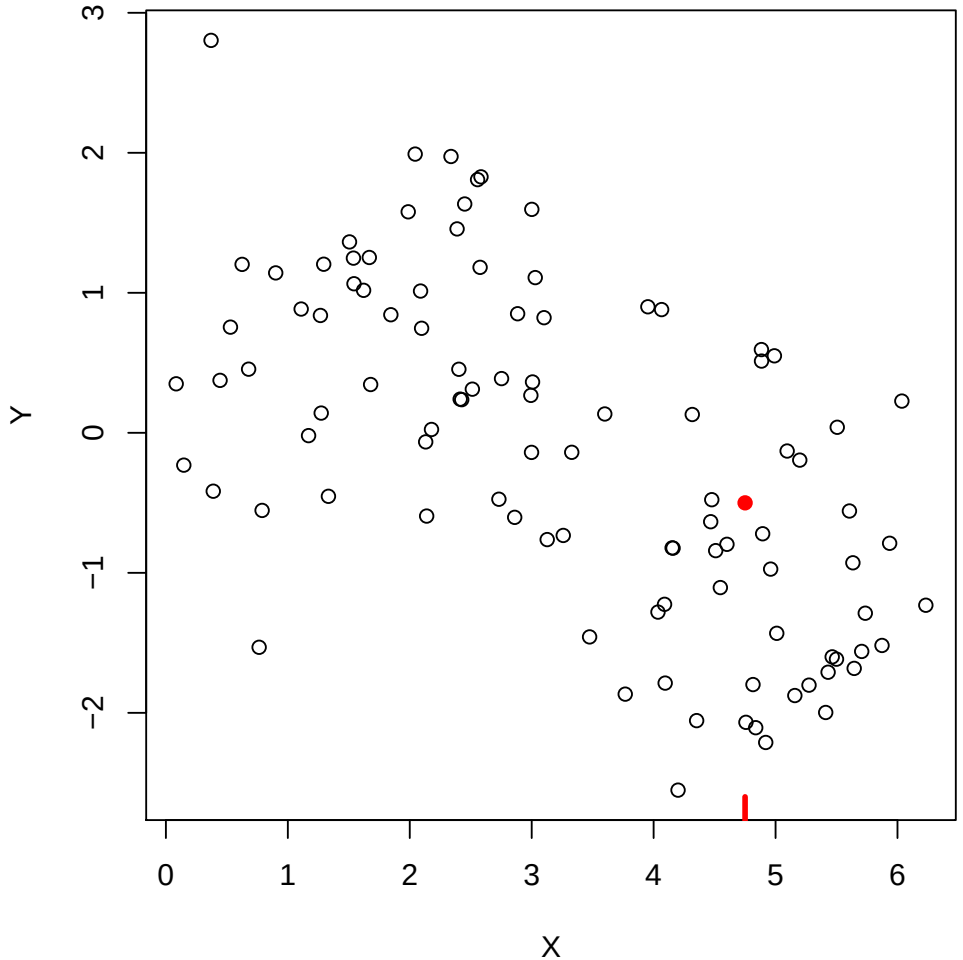


Example: conformal prediction interval using smoothing splines



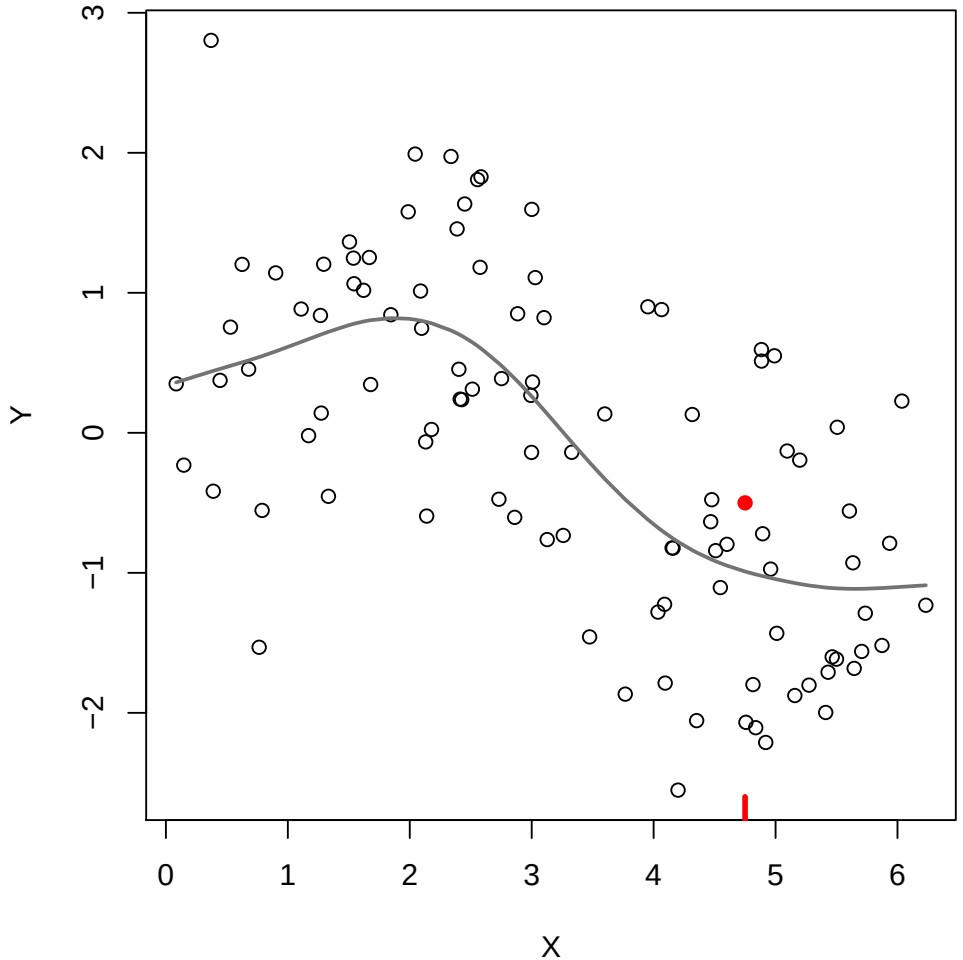
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



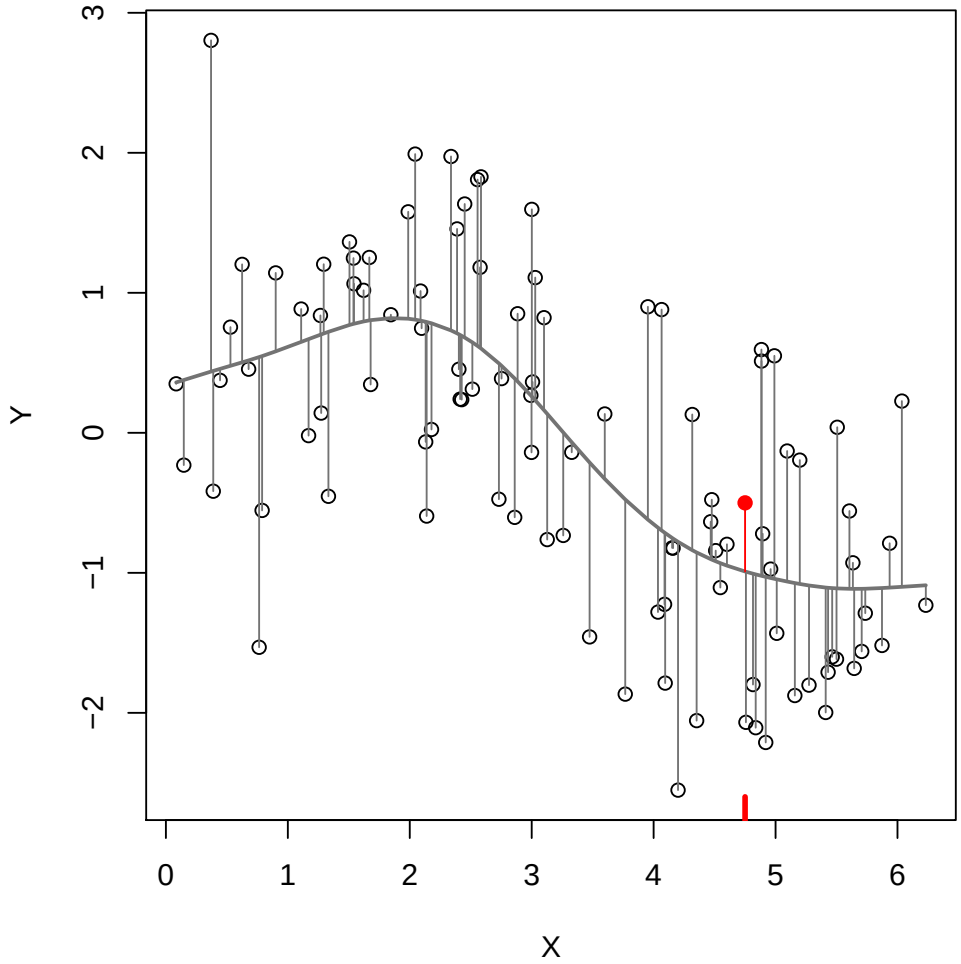
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



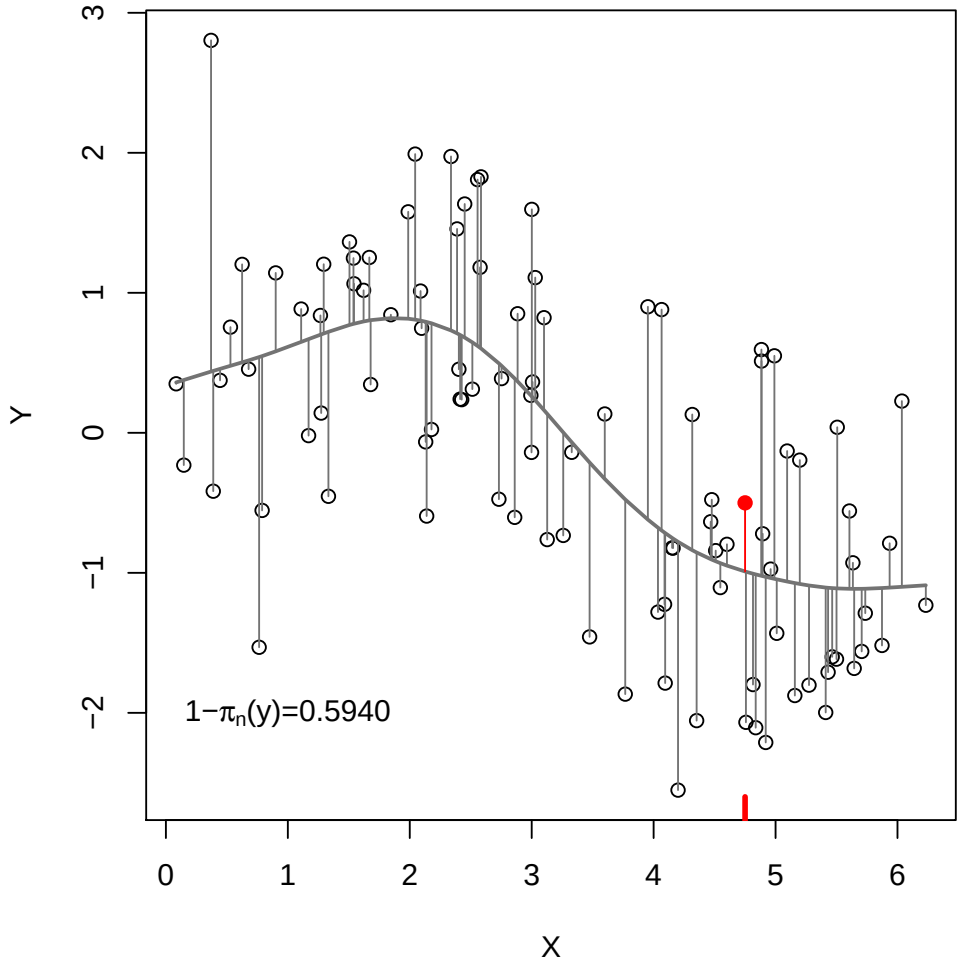
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



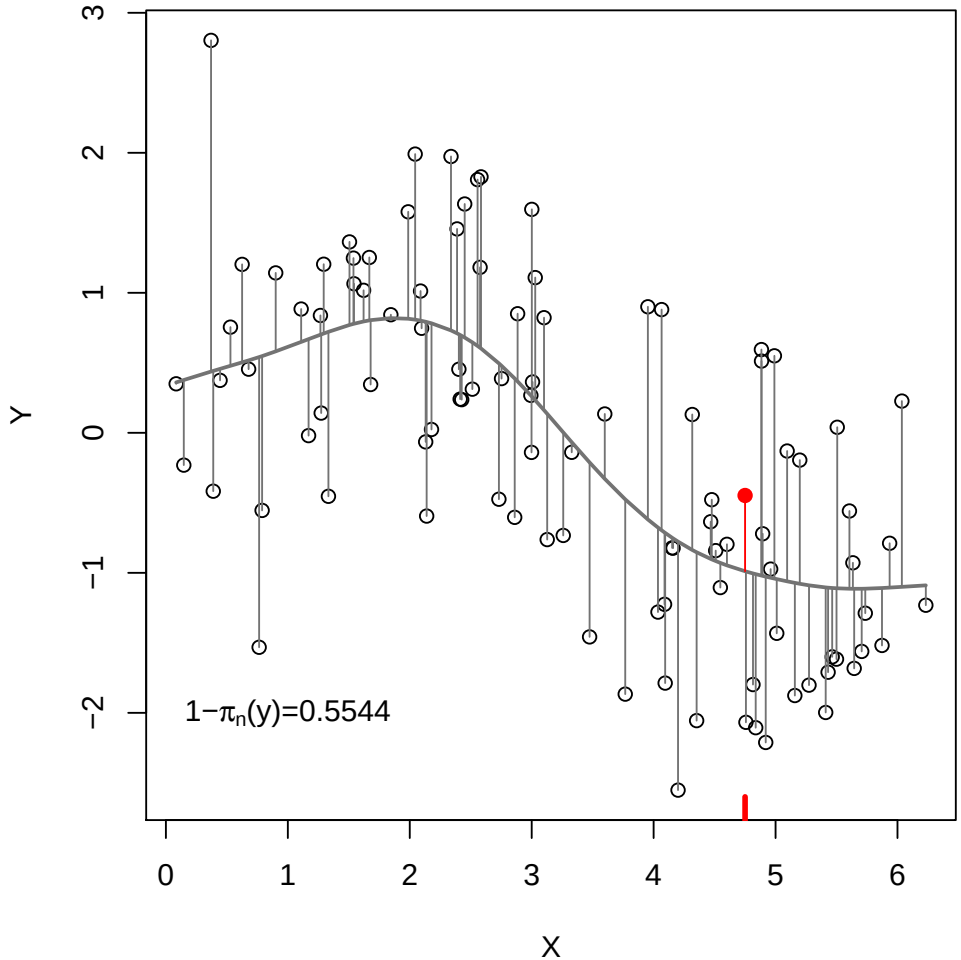
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



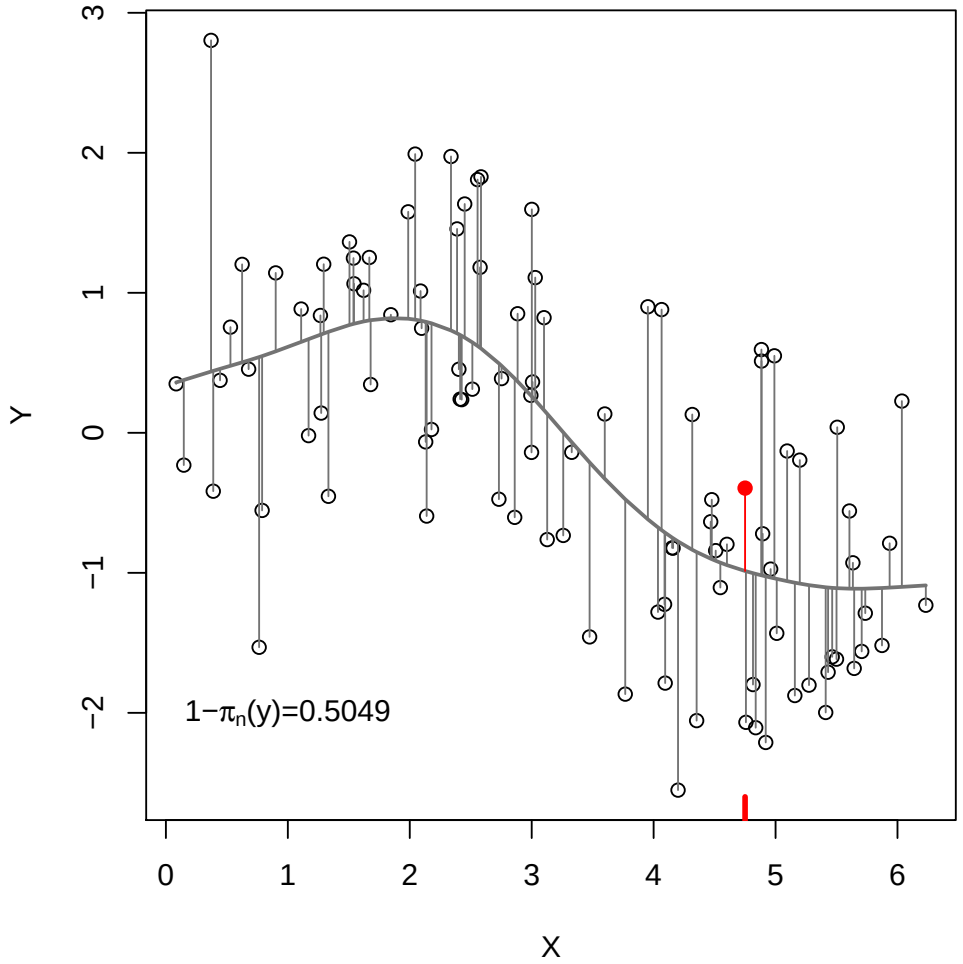
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



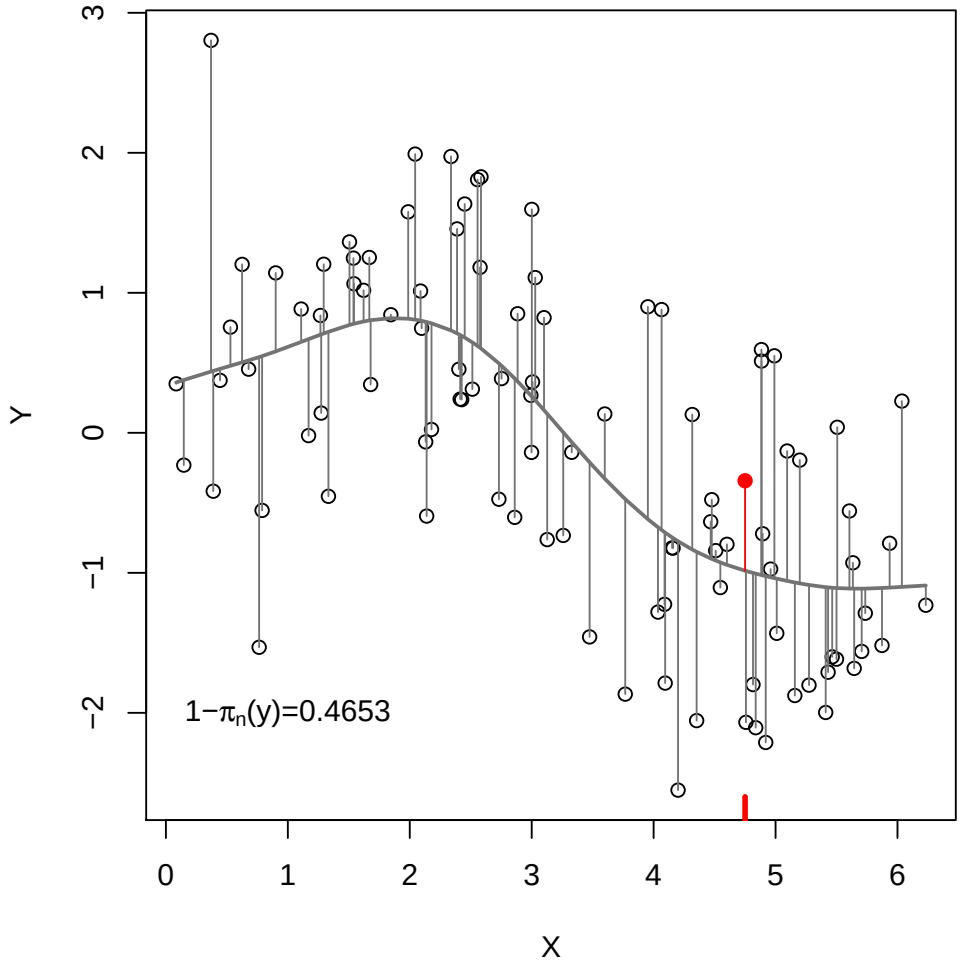
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



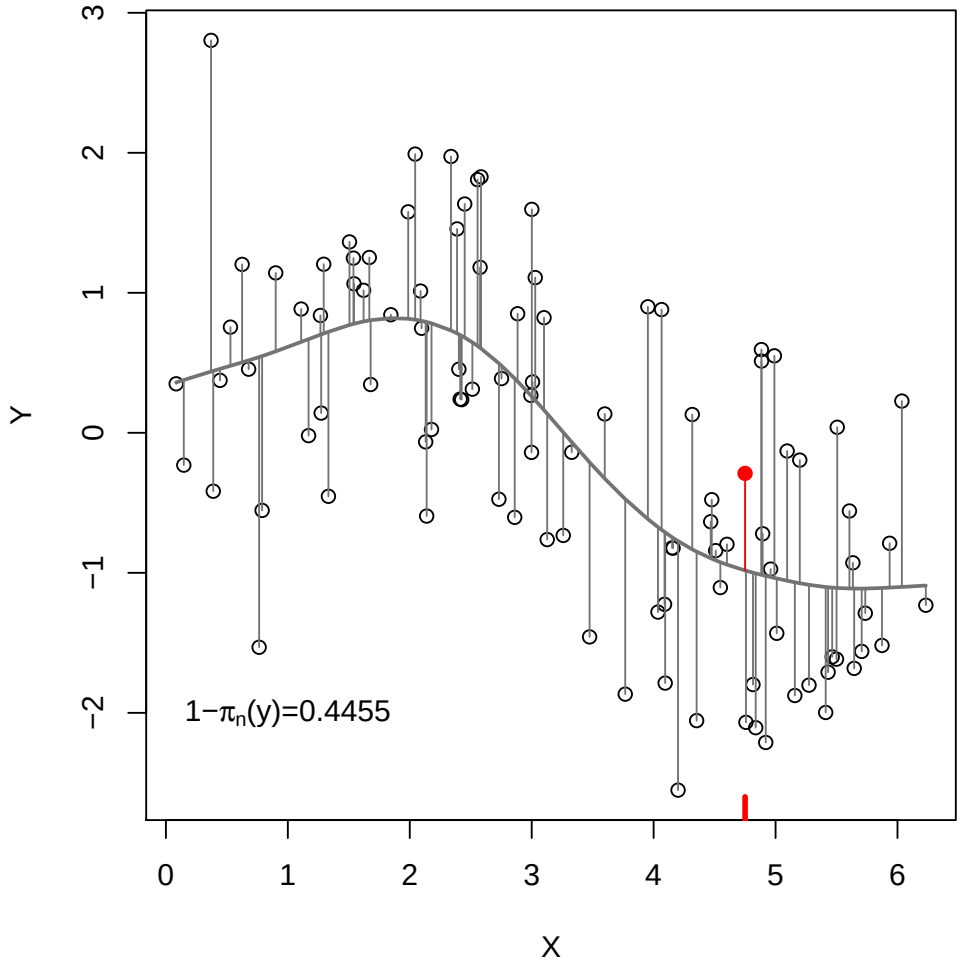
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



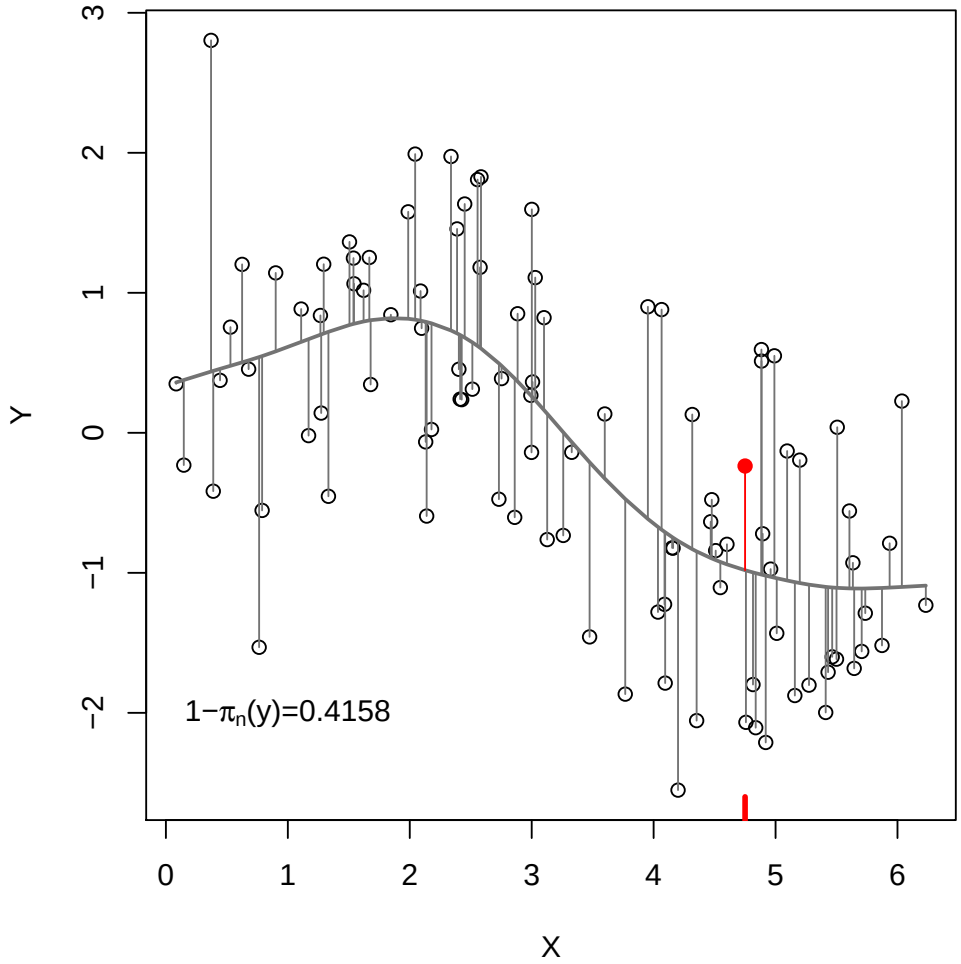
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



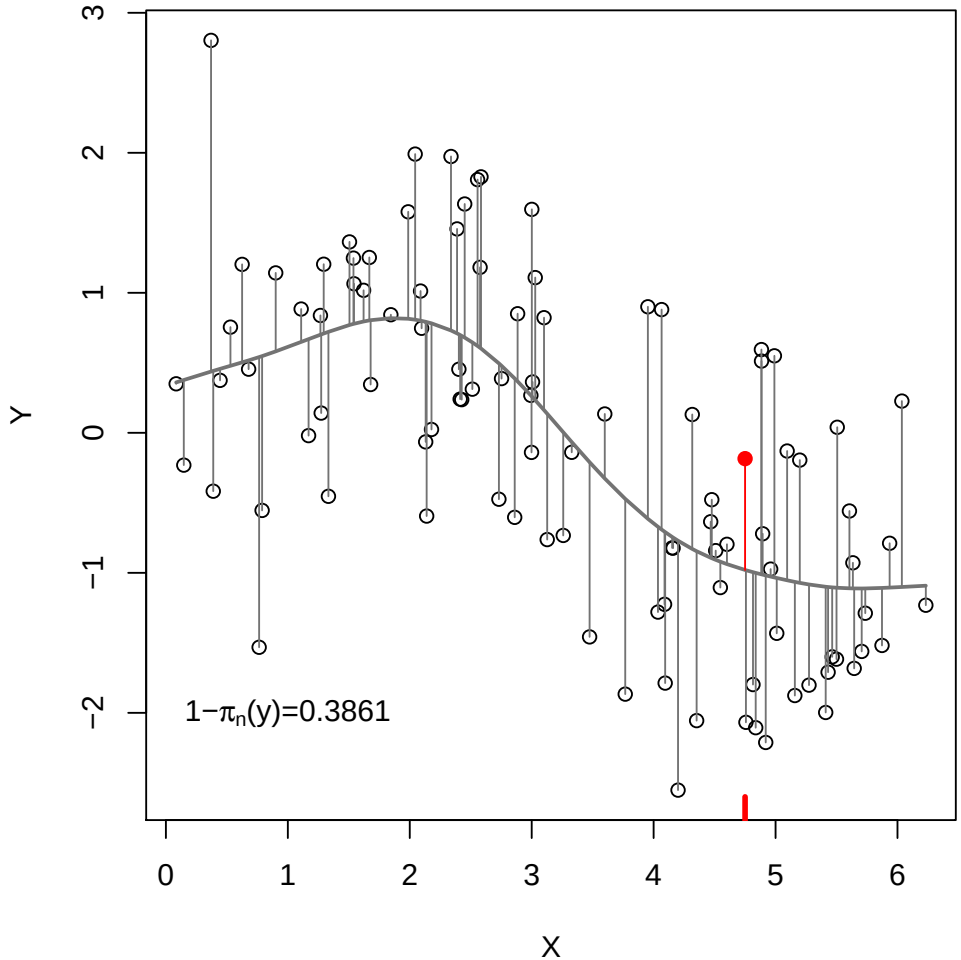
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



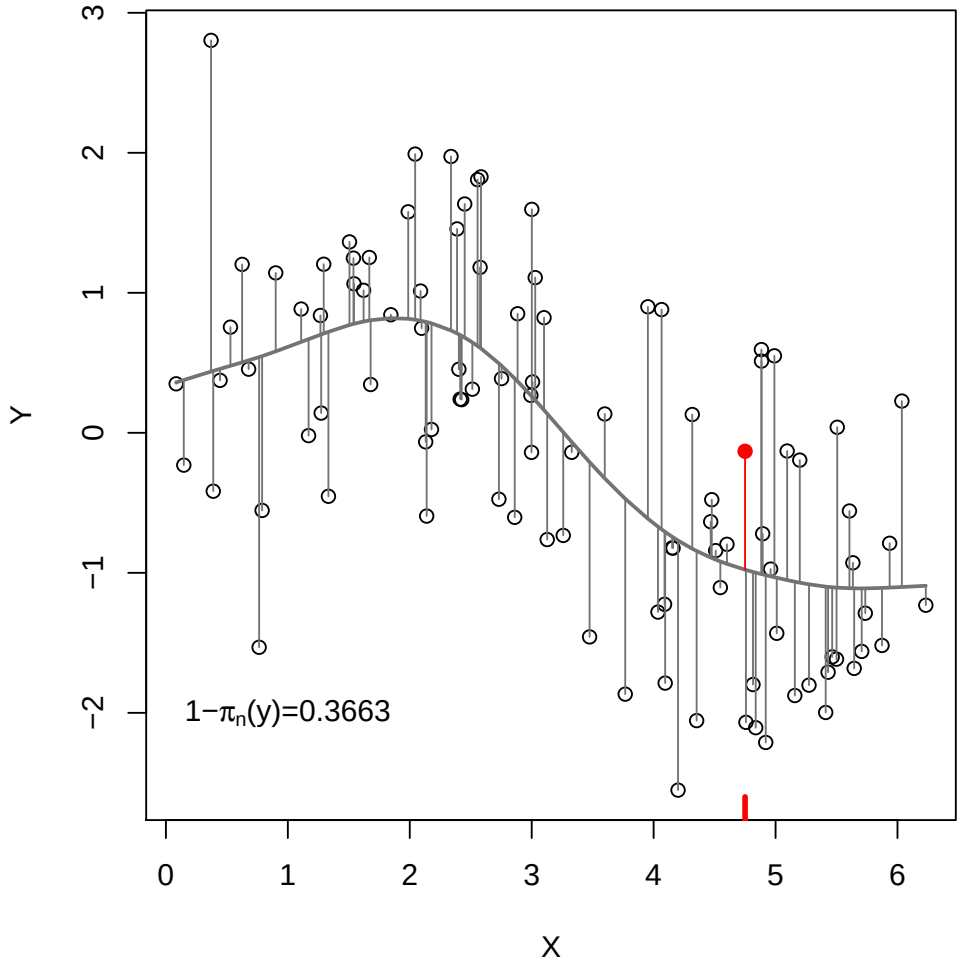
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



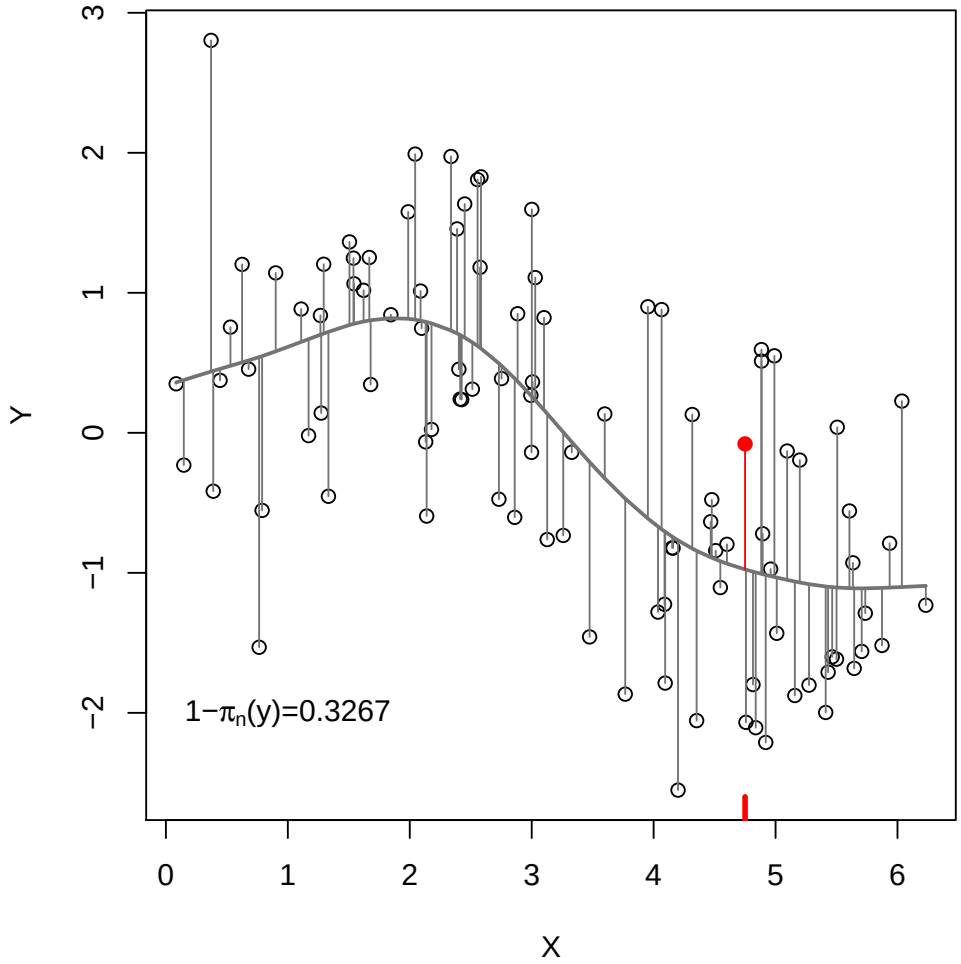
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



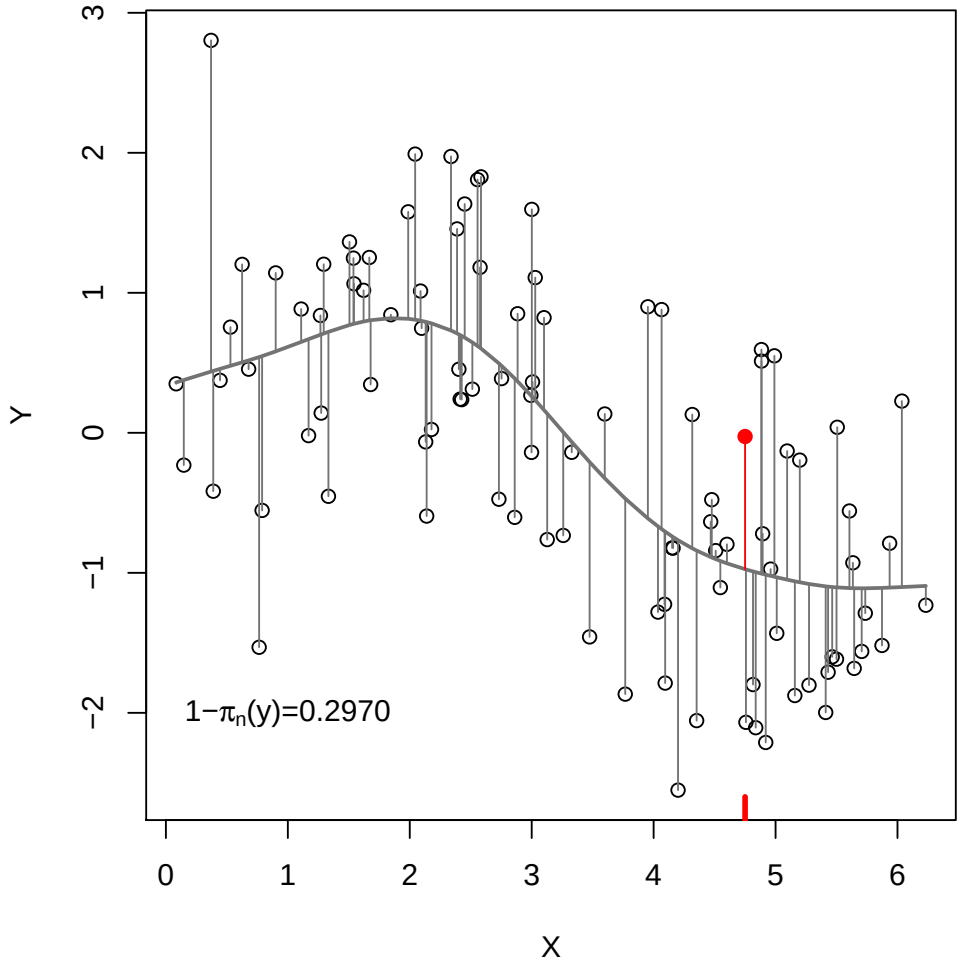
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



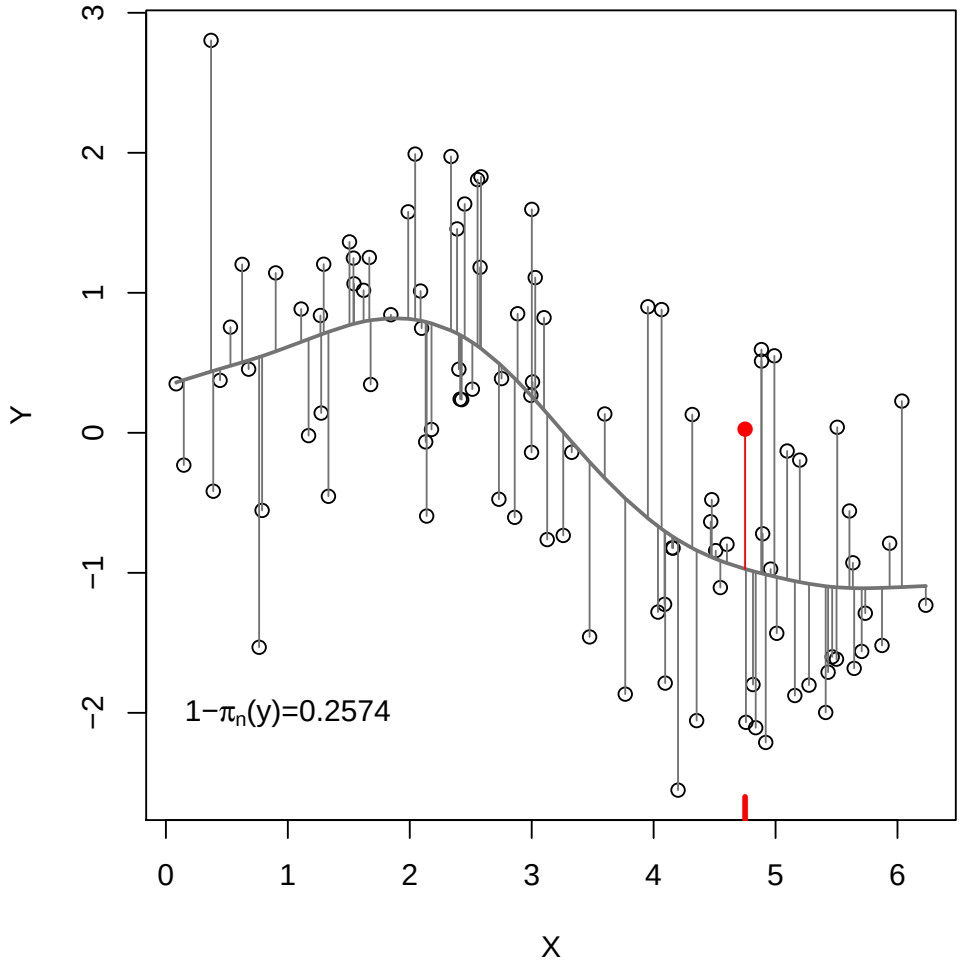
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



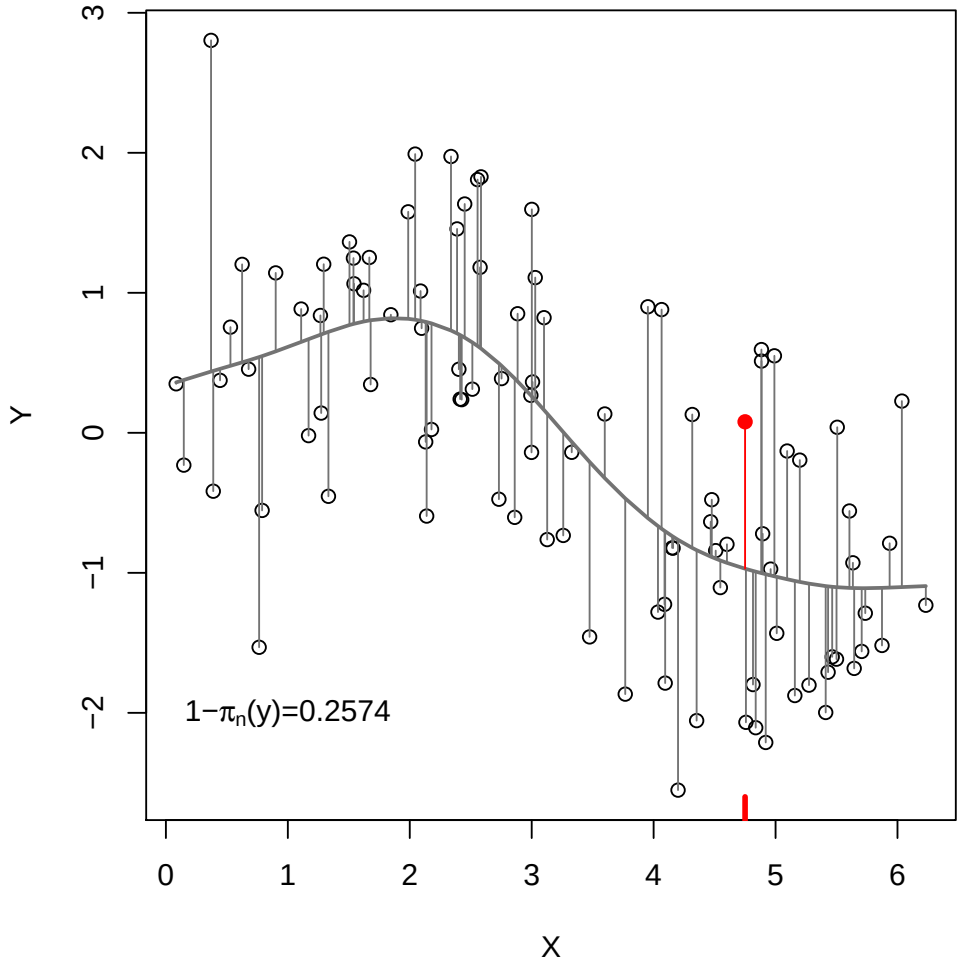
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



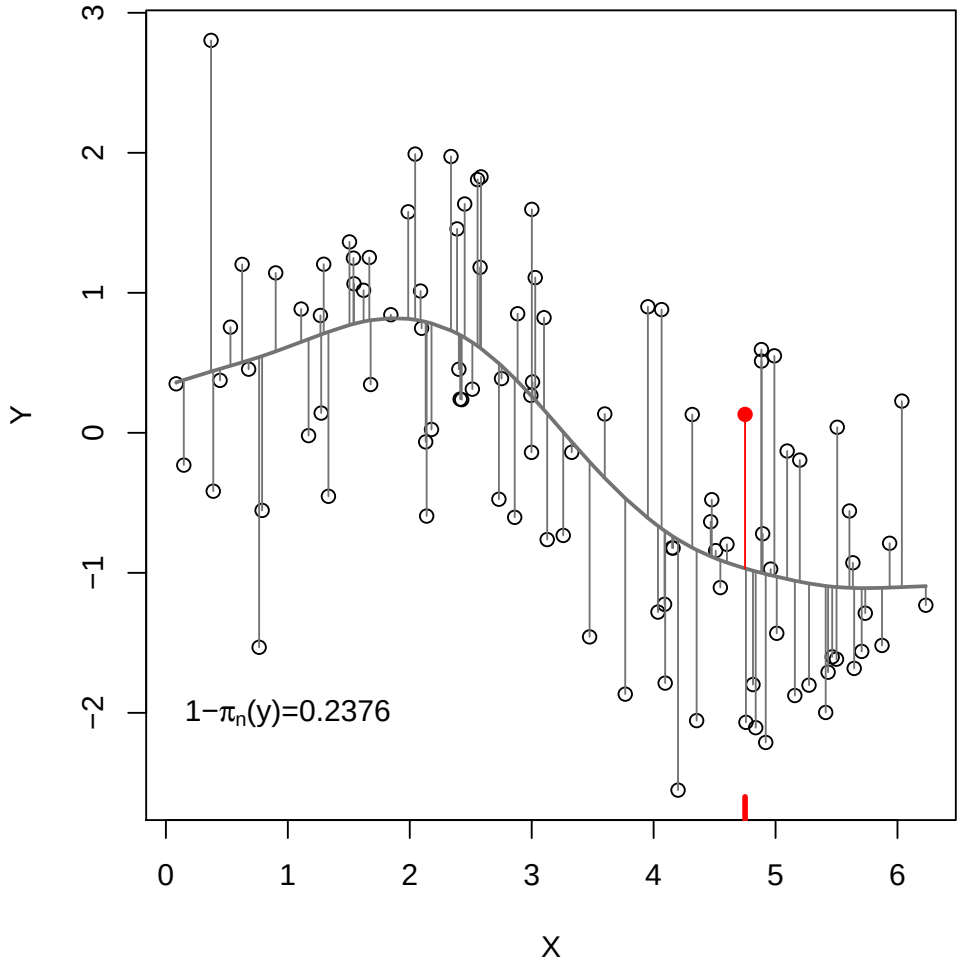
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



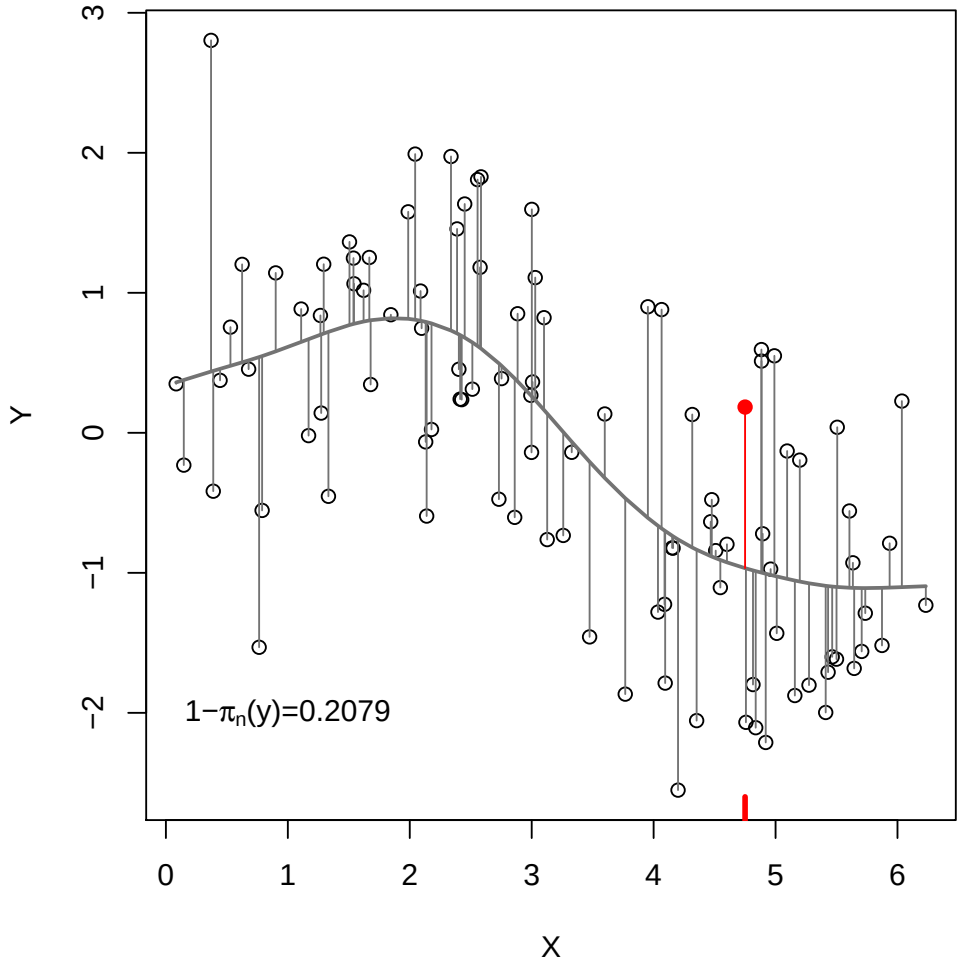
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



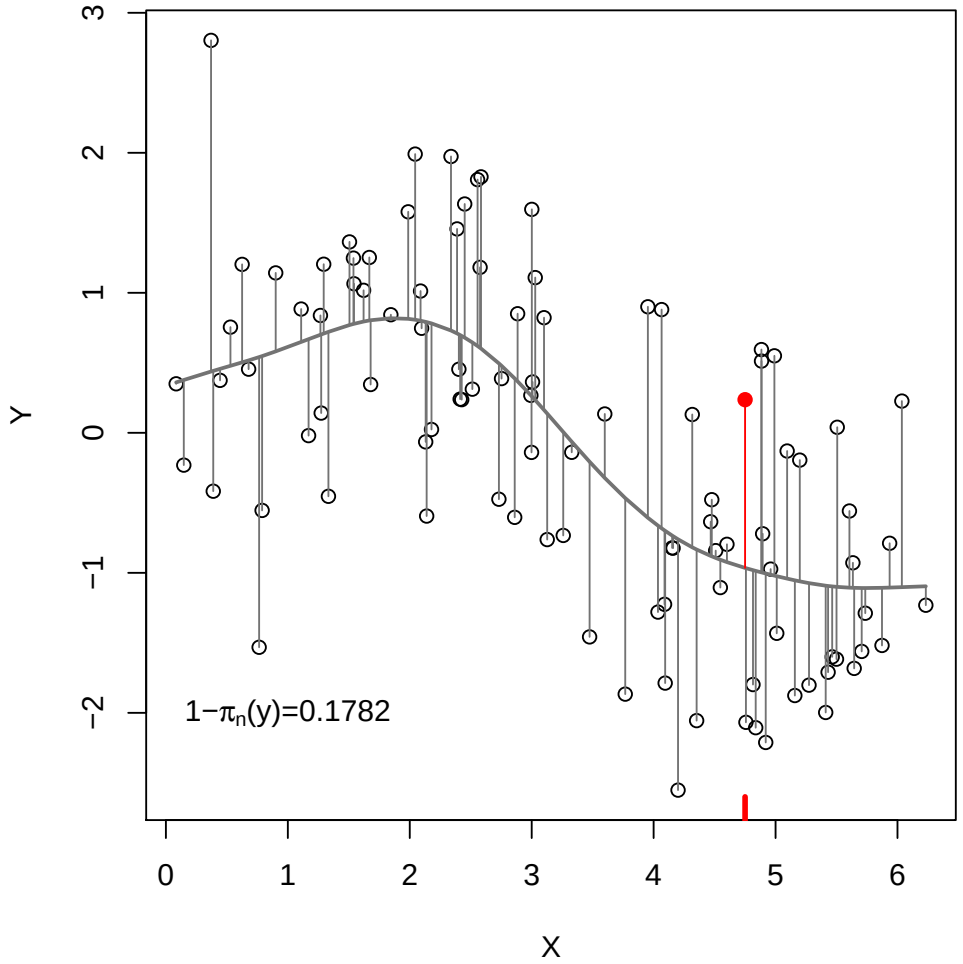
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



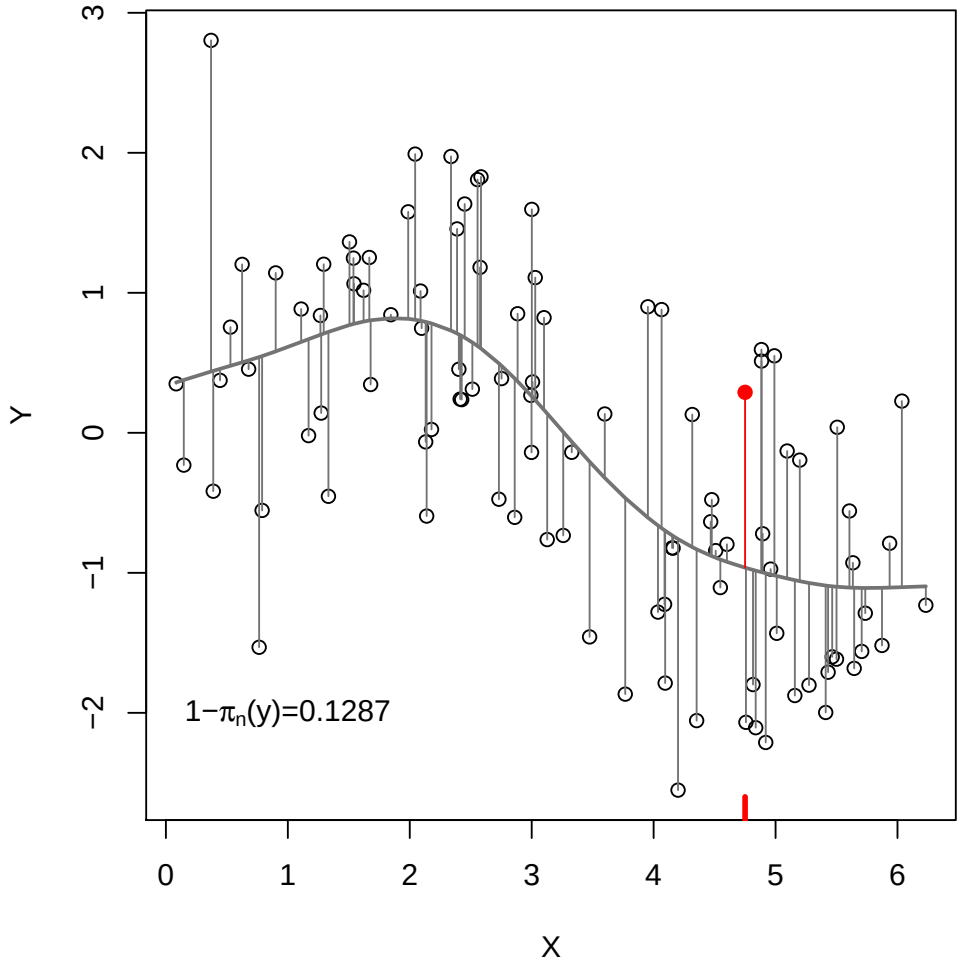
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



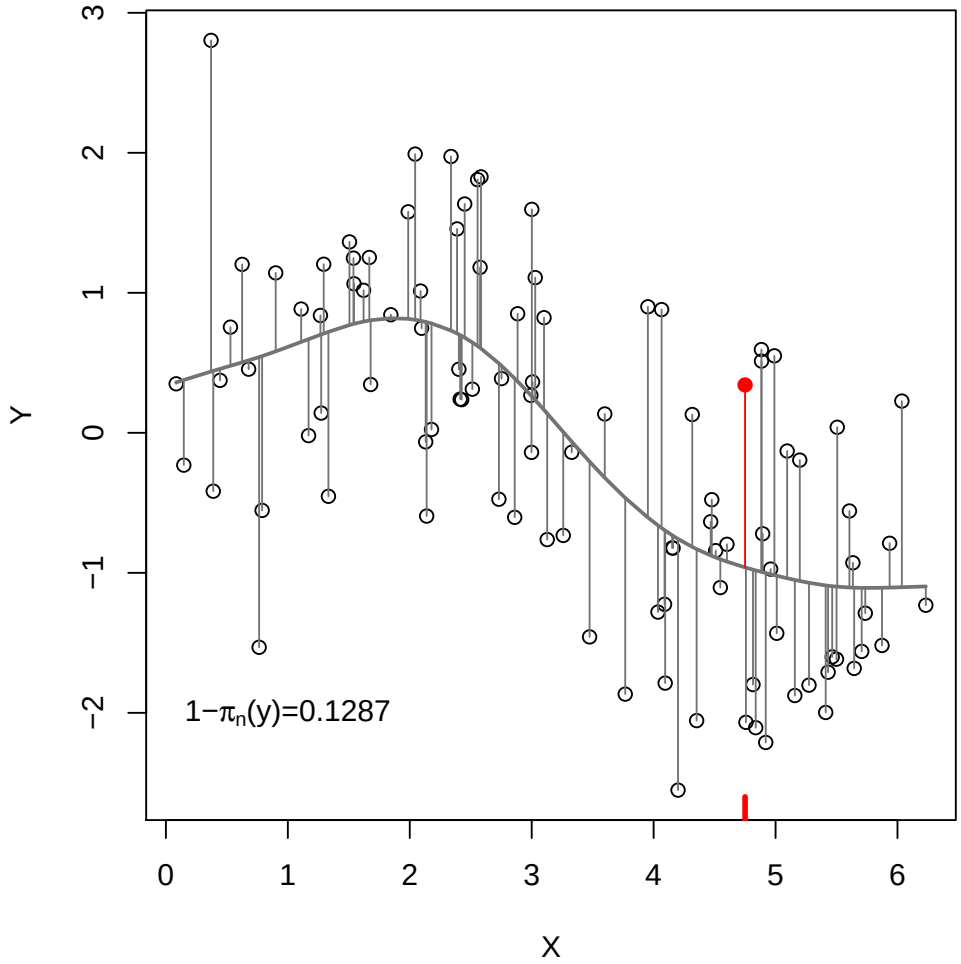
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



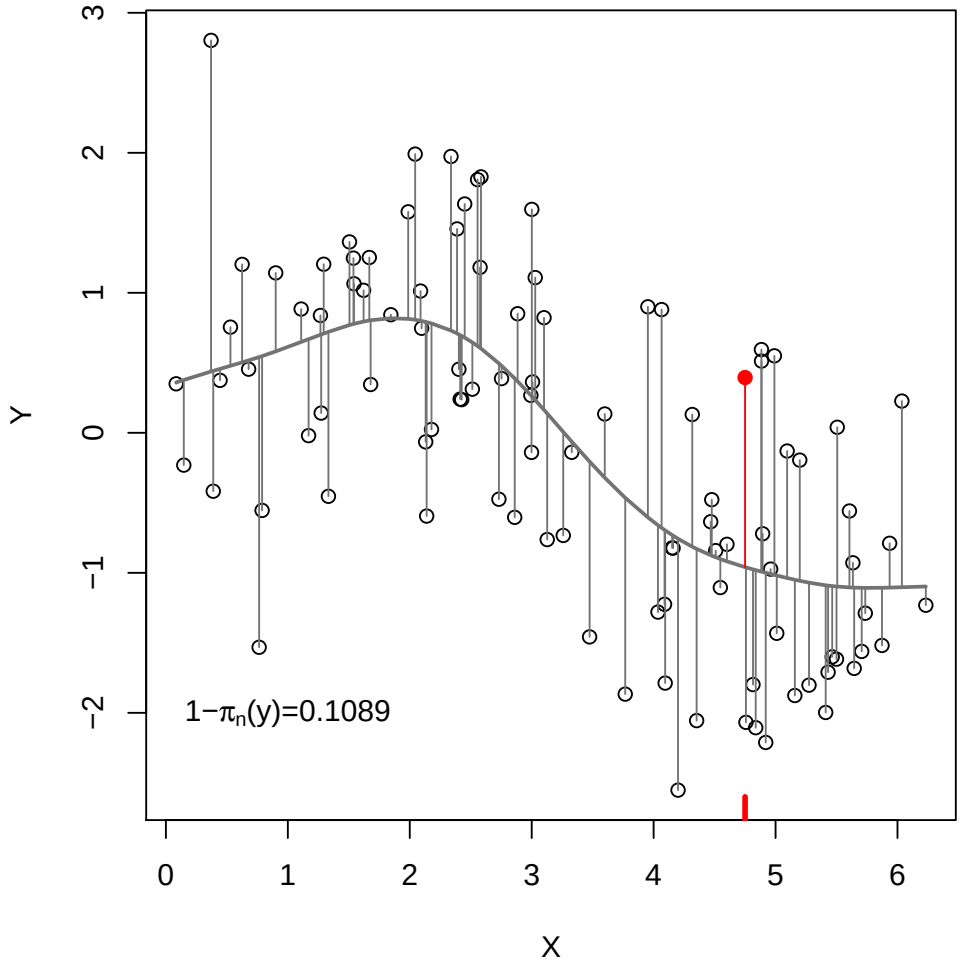
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



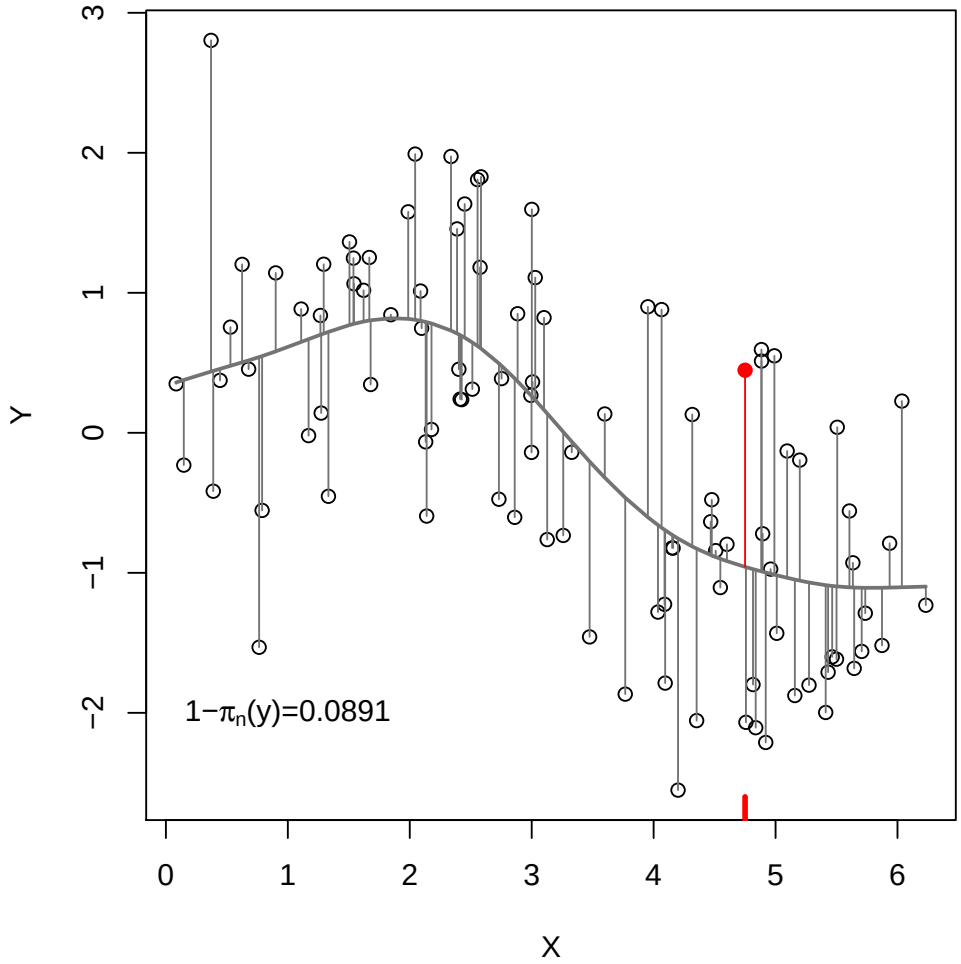
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



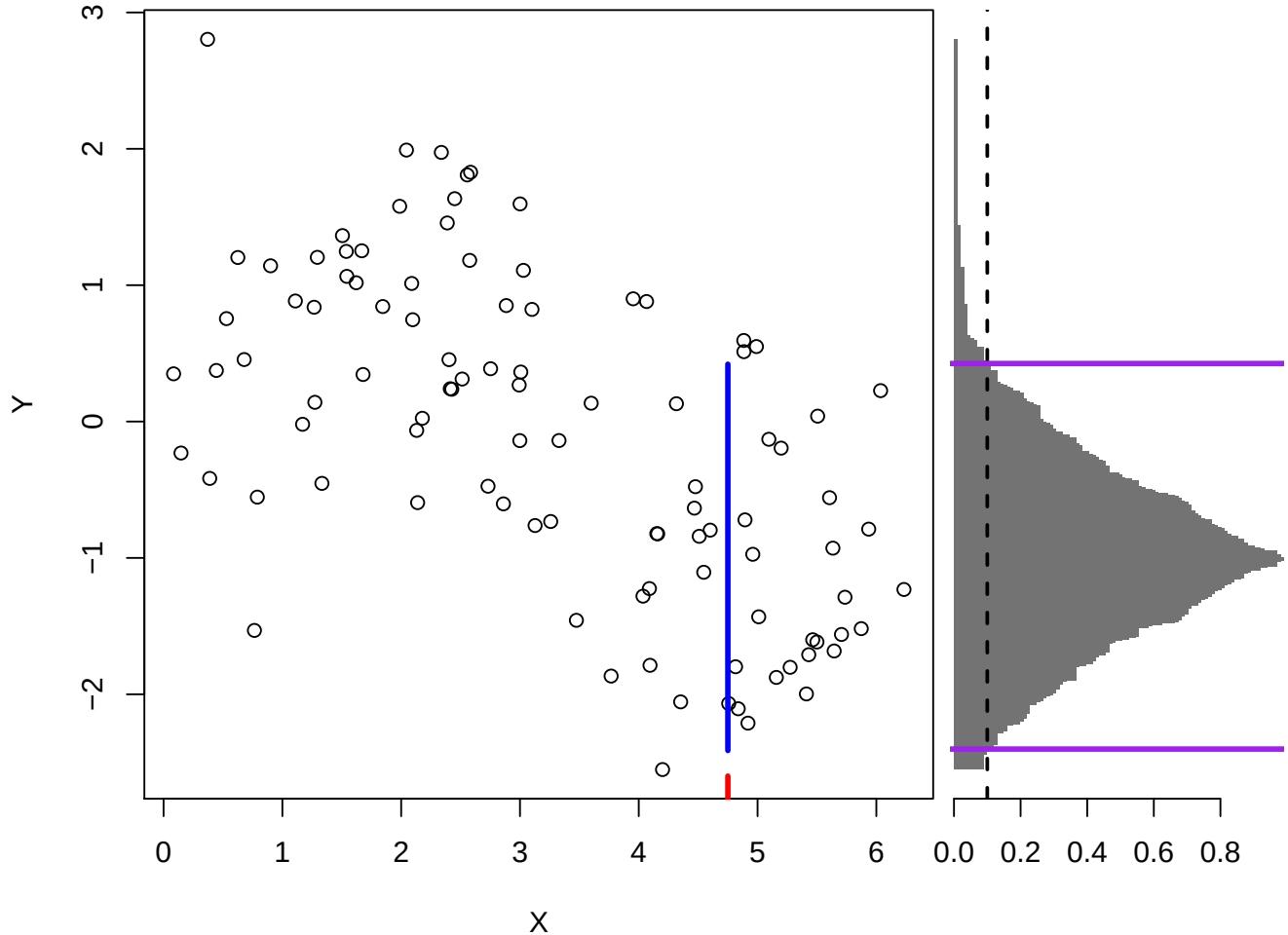
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



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

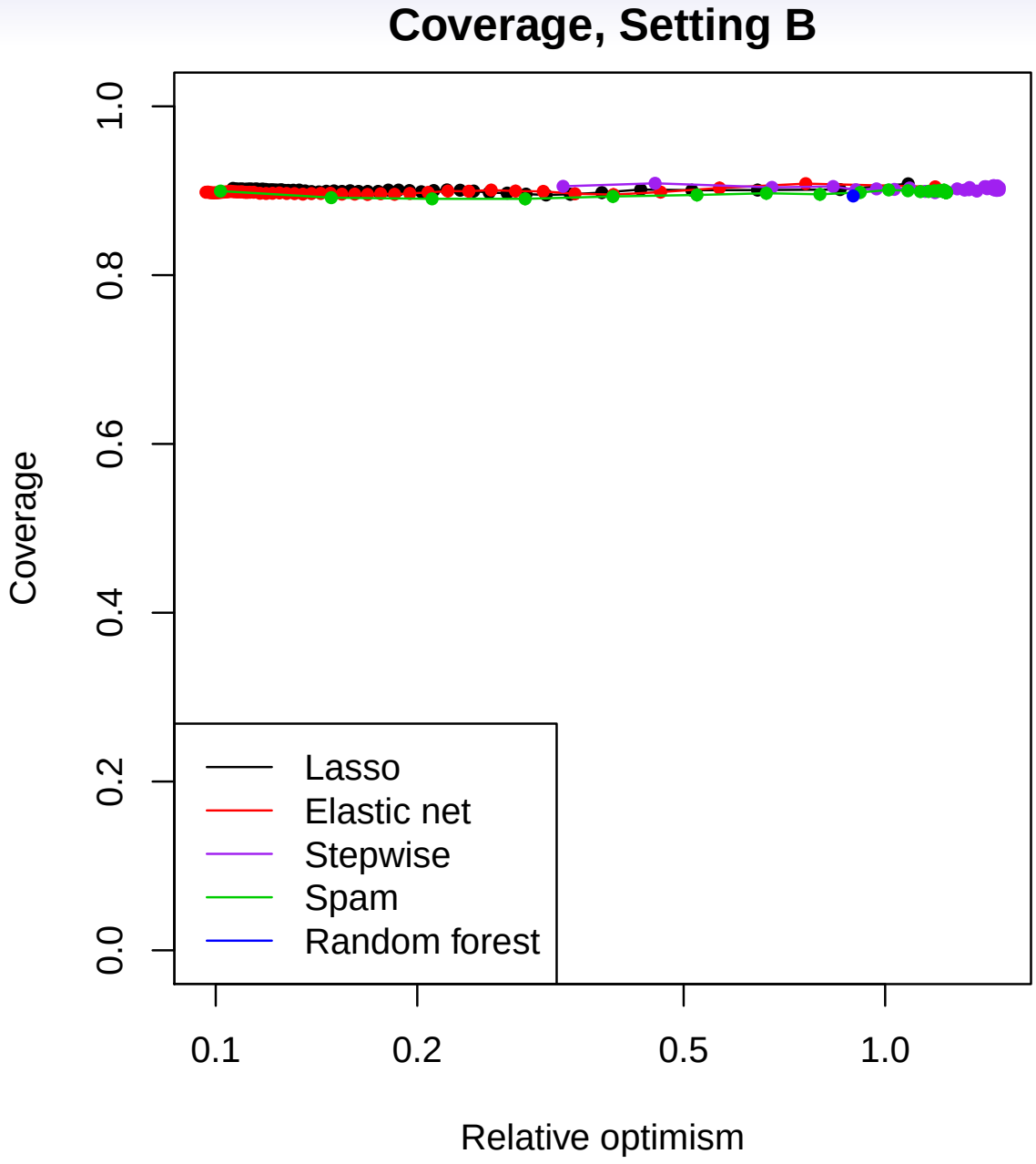
Example: conformal prediction interval using smoothing splines

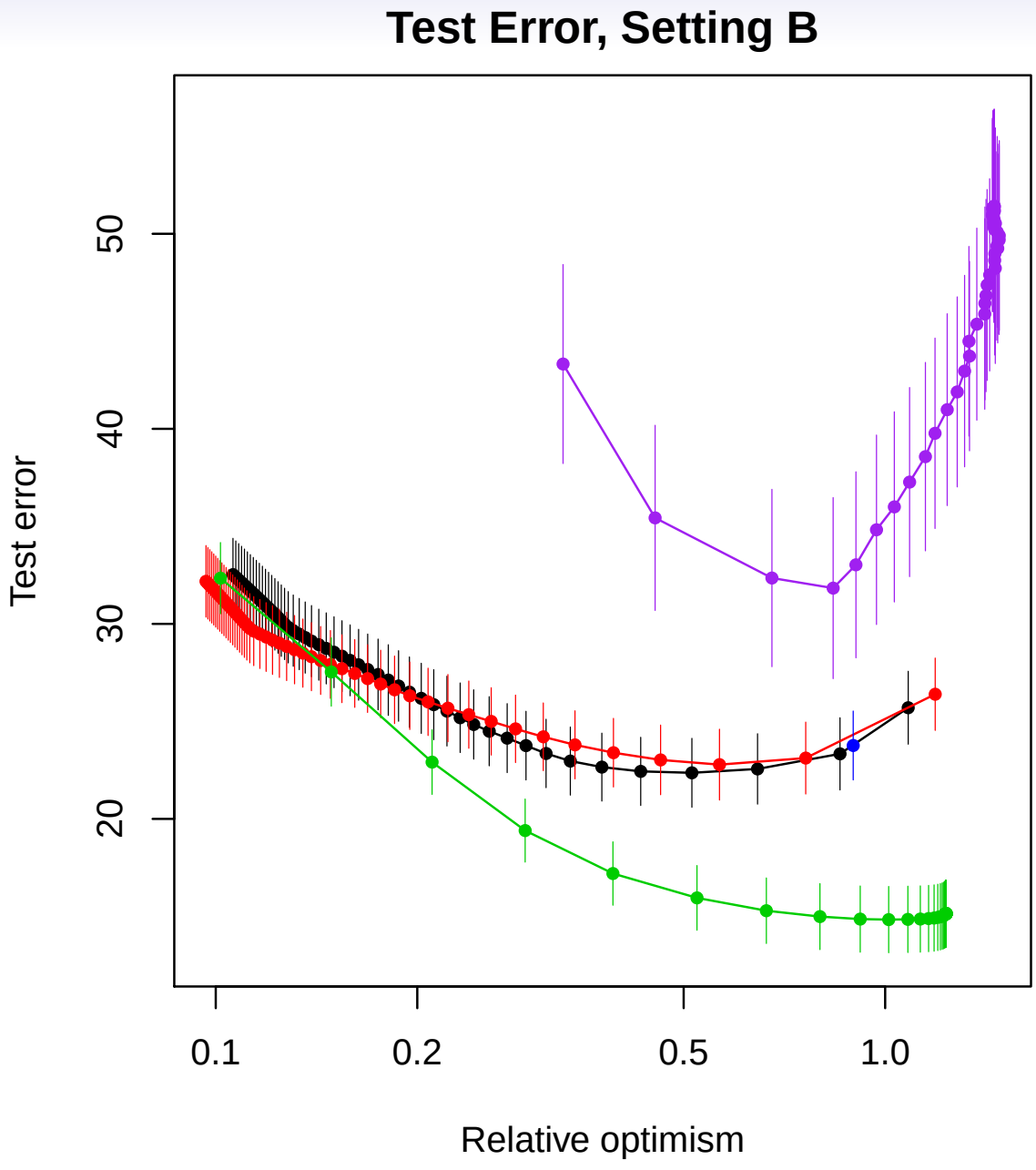


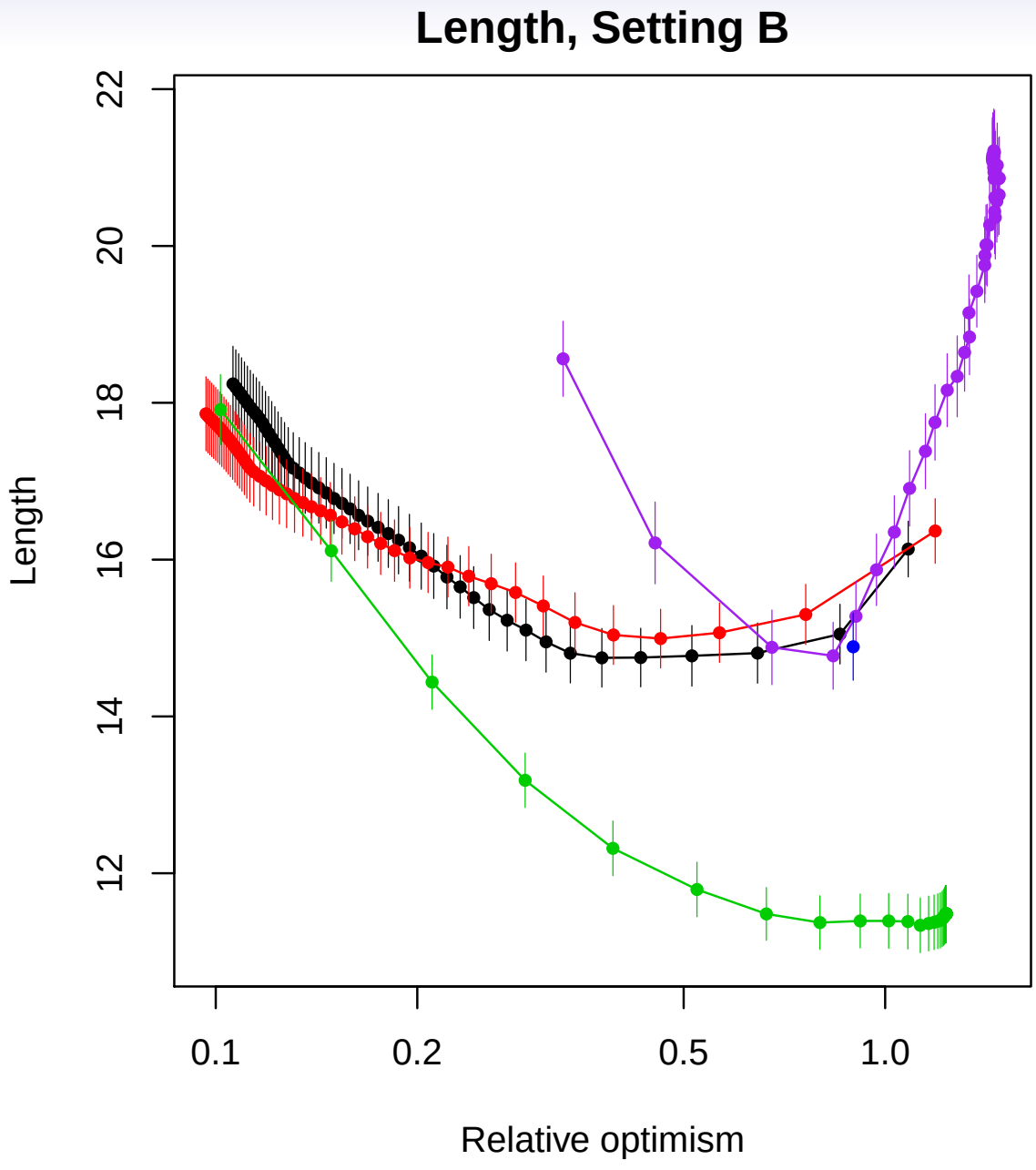
Invert p-values to get conformal interval

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


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


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

Observations

- The coverage is always $1 - \alpha$, regardless of fitting method and value of tuning parameter.
- Good methods and good tuning parameters give short prediction intervals.

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): 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.

Extensions

- Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y?
- Variable selection/importance?

Extensions

- Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y?
- Variable selection/importance?

Fast computation by sample splitting

- Original conformal prediction requires re-fitting $\hat{\mu}$ with new data point (X_{n+1}, y) for all values of X_{n+1} and y.
- Fast approximation available for kernel smoothing methods (Lei, Robins, & Wasserman 13; Lei & Wasserman 14).
- Fast exact conformalization available for Lasso (Lei, 2017).
- A general solution by detaching the fitting and ranking steps (Lei, Rinaldo, & Wasserman 15).

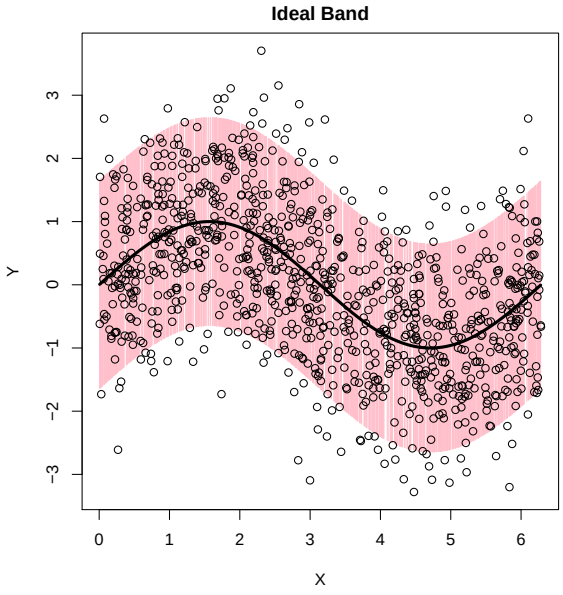
Split Conformal (Lei, Rinaldo, Wasserman 15)

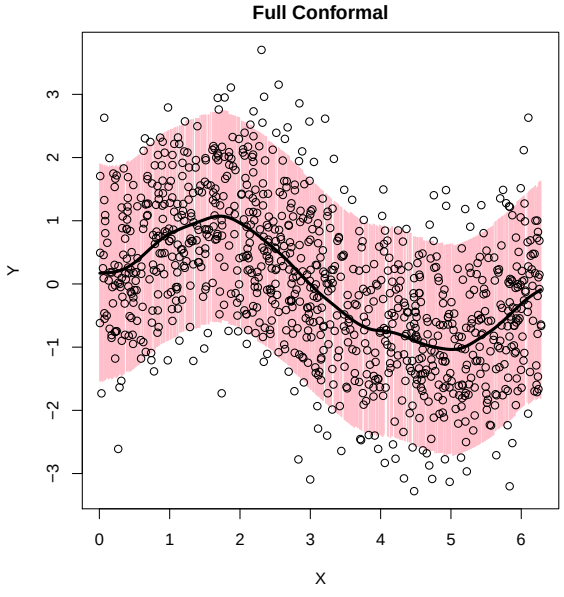
- Randomly split the data into two subsets, say, D_1 and D_2.
- Fit $\hat{\mu}$ on D_1.
- Let $\hat{F}$ be the empirical CDF of $\{|Y_i - \hat{\mu}(X_i)| : (X_i, Y_i) \in D_2\}$.
- Output $\tilde{C}(X_{n+1})$

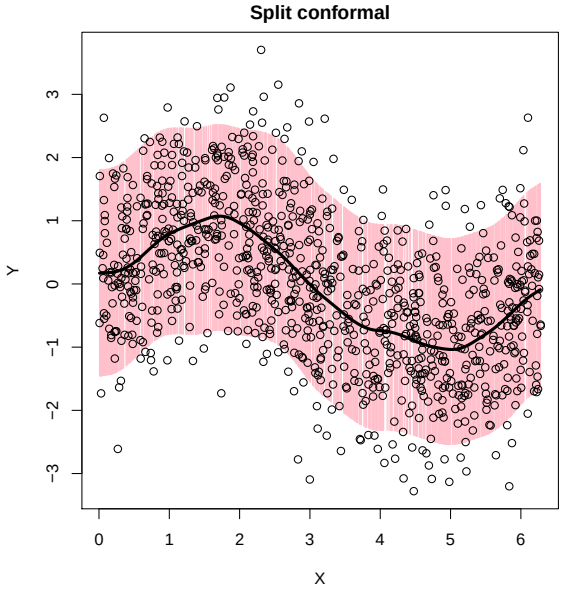
$$\tilde{C}(X_{n+1}) = [\hat{\mu}(X_{n+1}) \pm \hat{F}^{-1}(1-\alpha)]$$

- Can compute $\hat{C}(X_{n+1})$ for all values of X_{n+1} with a single fitted $\hat{\mu}$.
- Theorem: $\mathbb{P}(Y_{n+1} \in \tilde{C}(X_{n+1})) \geq 1 - \alpha$.

$Y = \sin(X) + N(0,1)$, $\hat{\mu}$: *smooth.spline*, *df*$= 12$







Split conformal offers in-sample validity

- Conformal prediction works for a future observation X_{n+1} not yet in the training sample.
- Can we get valid prediction at points $(X_i : 1 \leq i \leq n)$ in the sample?
- Theorem: $\mathbb{P}(Y_i' \in \tilde{C}(X_i)) \geq 1 - \alpha$, for all $X_i \in D_2$, where Y_i' is an independent copy of $(Y|X = X_i)$, and

$$\mathbb{P}\left[(2/n)\sum_{i \in D_2} \mathbf{1}(Y_i \in \tilde{C}(X_i)) \geq 1 - \alpha - \varepsilon\right] \geq c_1 \exp(-c_2 n \varepsilon^2).$$

- Switch D_1 and D_2 to cover points in D_1.

Extensions

- ✓ Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y.
- Variable selection/importance?

Extensions

- ✓ Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y.
- Variable selection/importance?

Variable importance

- Assume $X \in \mathbb{R}^d$, where d can be large.
- 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.
- 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 $|Y_i - \hat{\mu}_{-j}(X_i)| - |Y_i - \hat{\mu}(X_i)|$.
- Can use conformal prediction to obtain a valid prediction interval for

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

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

Variable importance

- Recall that we want a prediction interval for

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

  where Y_i' is a fresh draw from $(Y|X = X_i)$.
- Let $\tilde{C}(X_i)$ be a valid prediction interval for Y_i and define

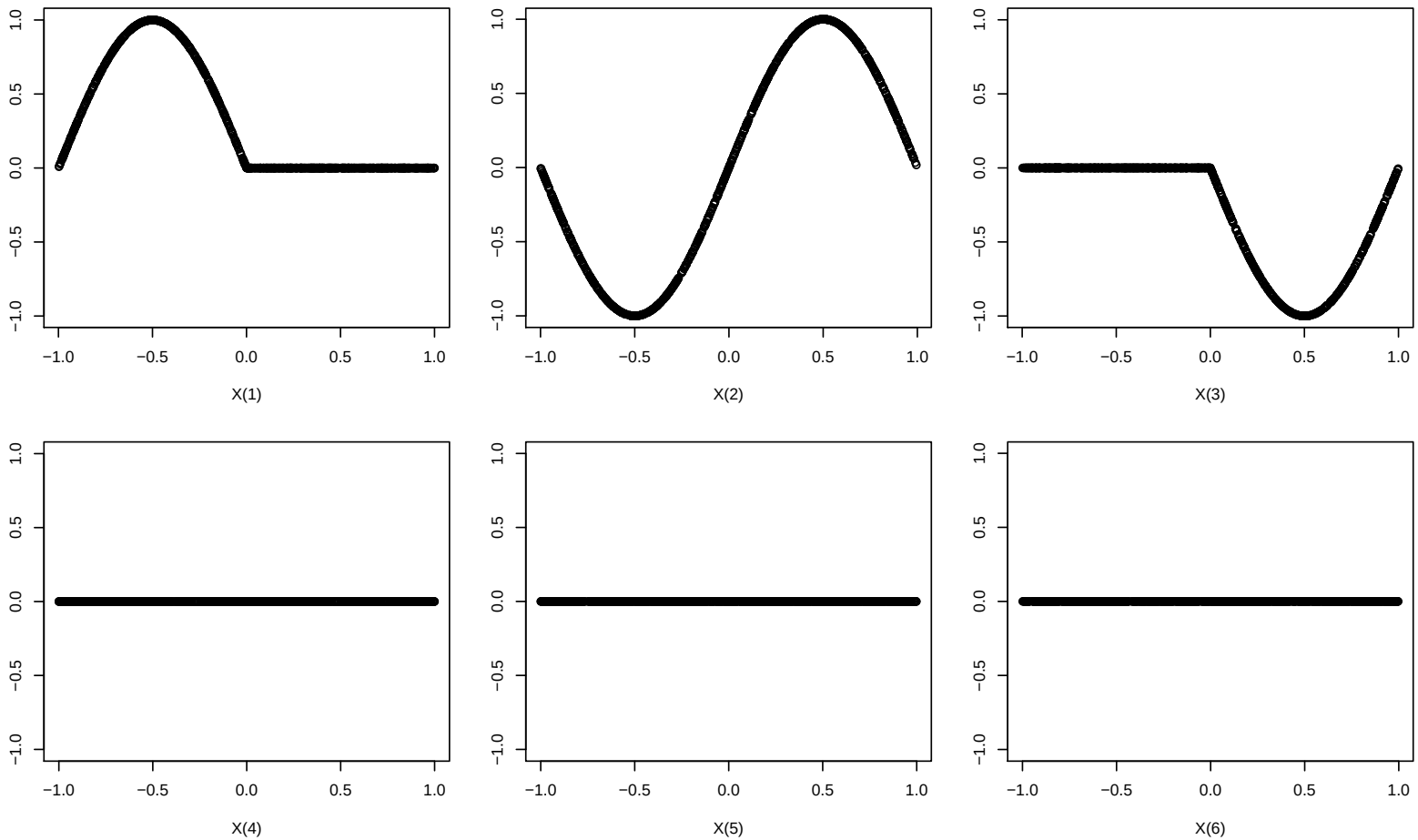
$$D_{ij} = \{|y - \hat{\mu}_{-j}(X_i)| - |y - \hat{\mu}(X_i)| : y \in \tilde{C}(X_i)\}$$

- Fact: $Y_i' \in \tilde{C}(X_i) \Rightarrow V_{ij} \in D_{ij}$, and $\mathbb{P}(V_{ij} \in D_{ij},\ \forall\, j) \geq 1 - \alpha$.
- Corollary: If $\tilde{C}(X_i)$ is obtained from split conformal, then

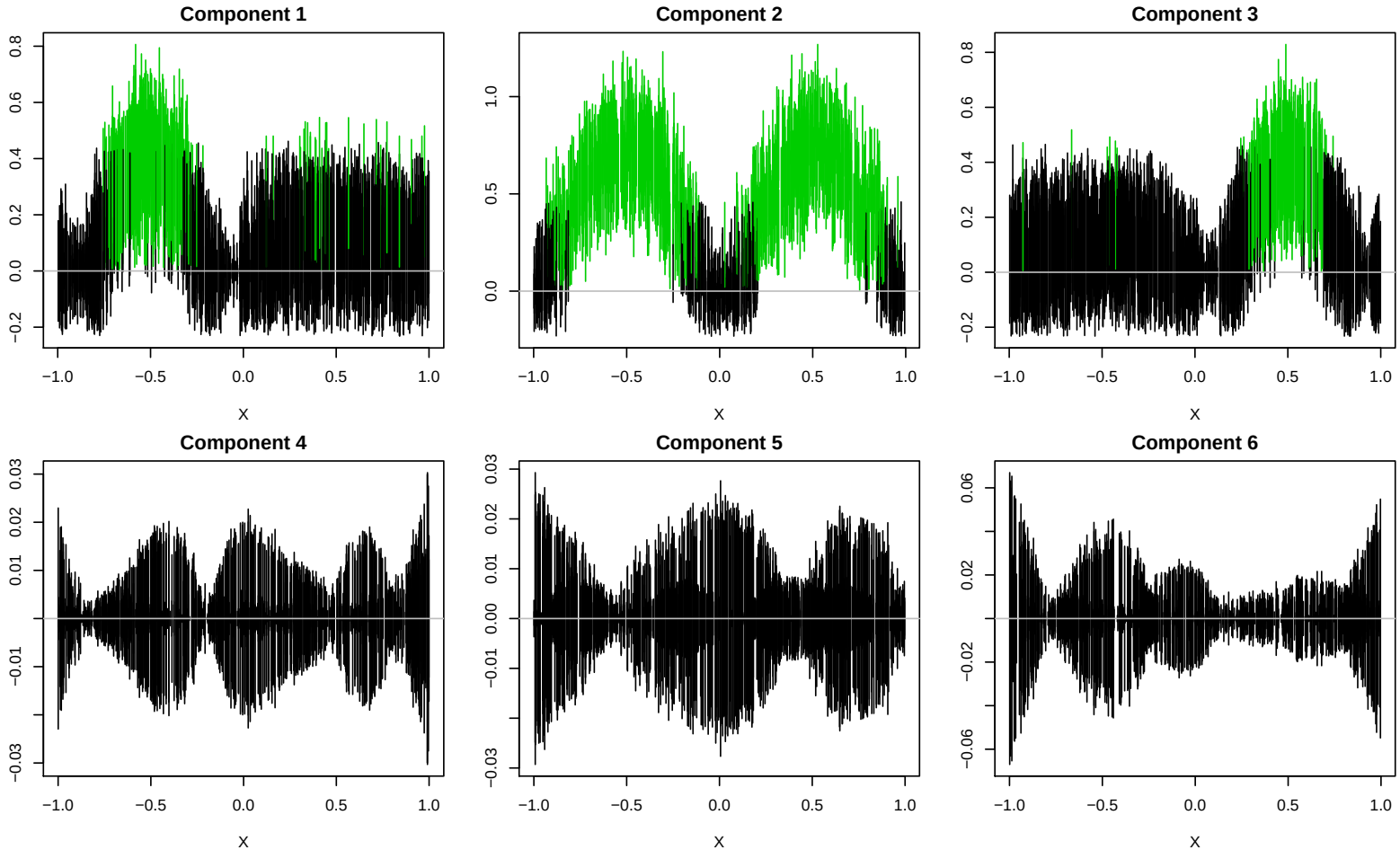
$$\mathbb{P}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}(V_{ij} \in D_{ij},\ \forall\, j) \geq 1 - \alpha - \varepsilon\right] \geq 1 - 2e^{-cn\varepsilon^2}$$

Example: Additive Model

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



How do D_{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

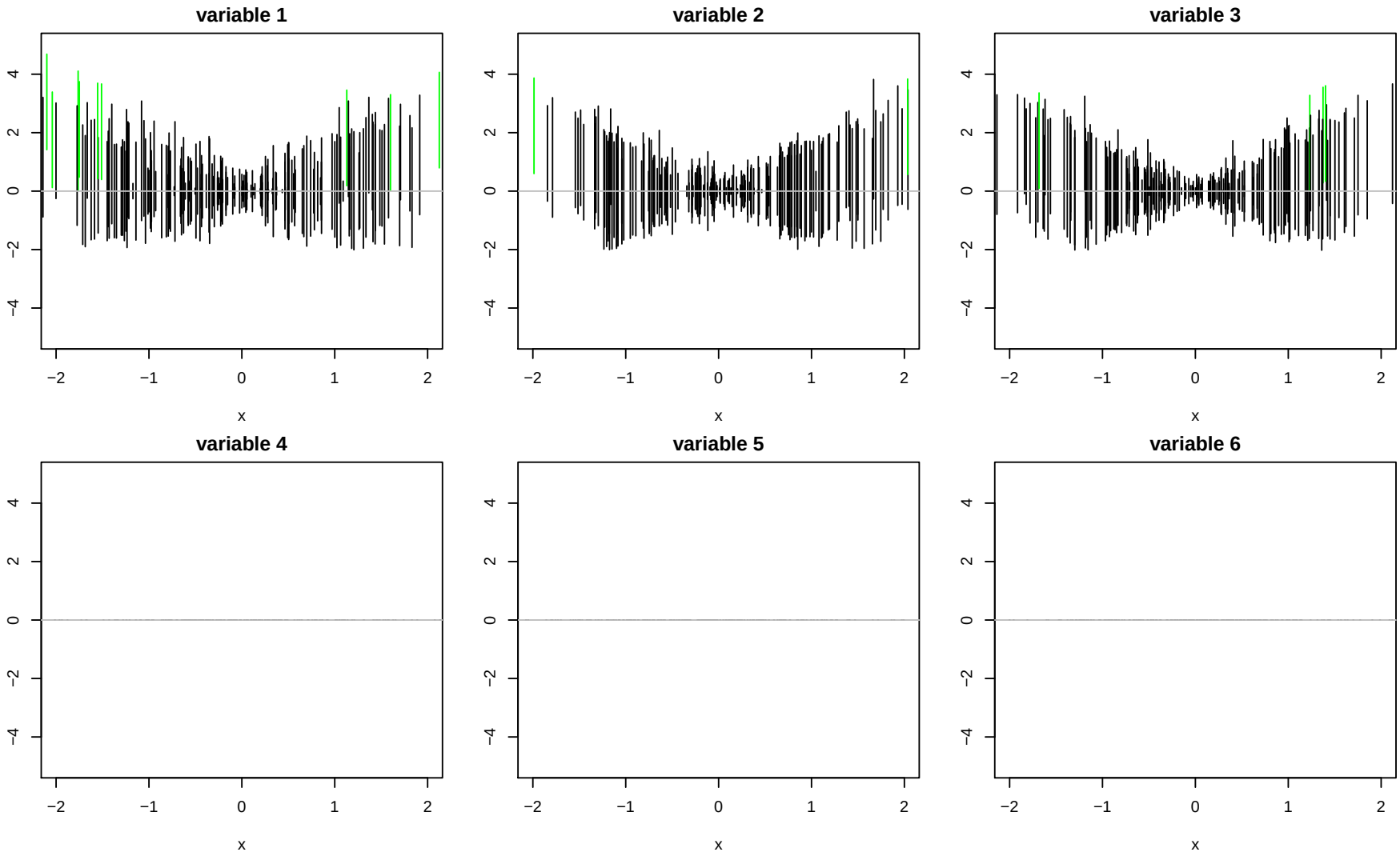
- $n = 200, p = 100$
- $Y = X^T\beta + \varepsilon$
- $\varepsilon \sim N(0,1)$, independent of X
- $\beta = (2,2,2,0,...,0)^T$
- Design matrix

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

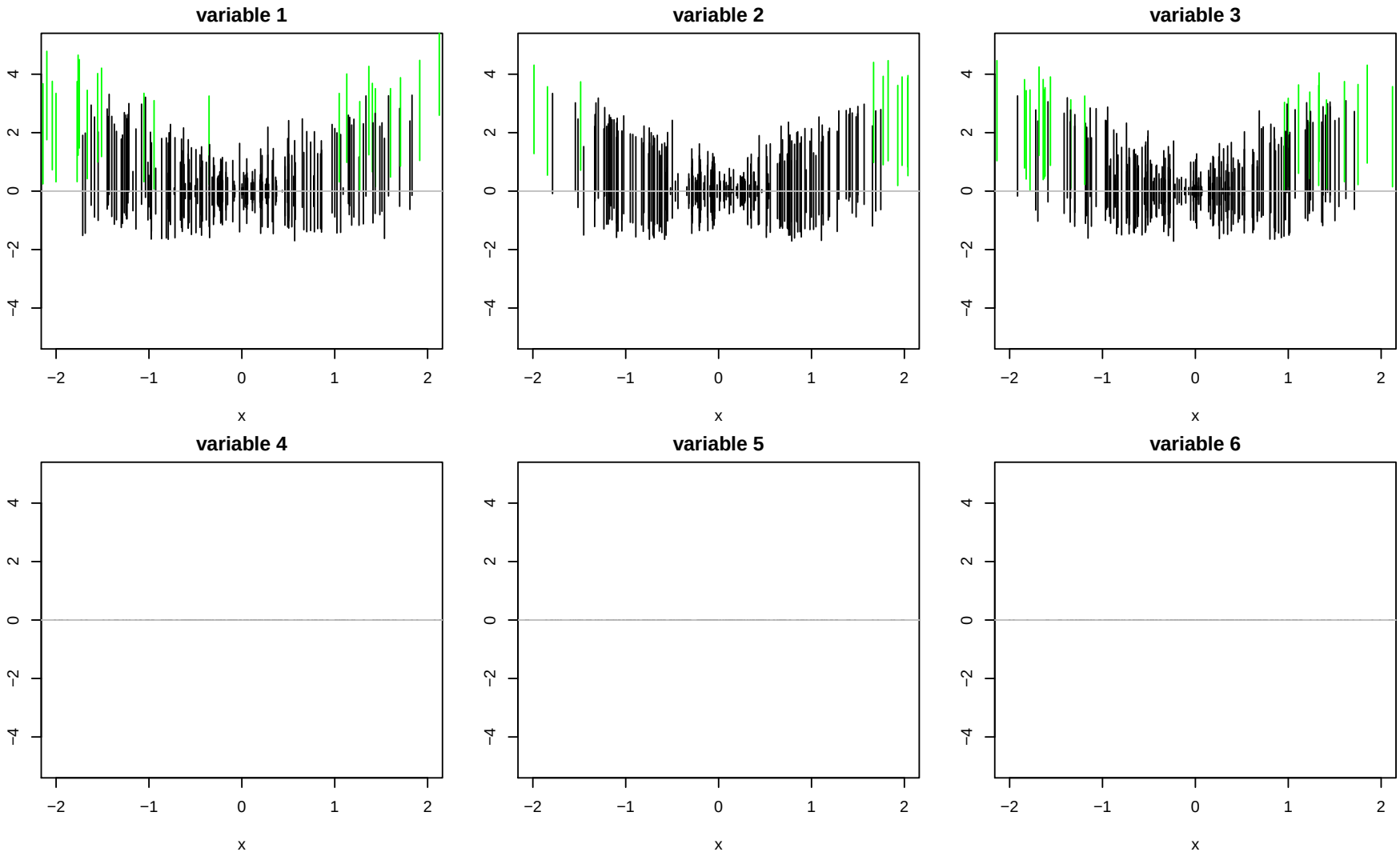
  Case 2: $\text{corr}(X(j), X(j')) = 0.7$ if $j \neq j'$ (strong correlation)

- Fitting methods
  - *(a)* Lasso with $\lambda = 0.3$
  - *(b)* Forward Stepwise with 3 steps

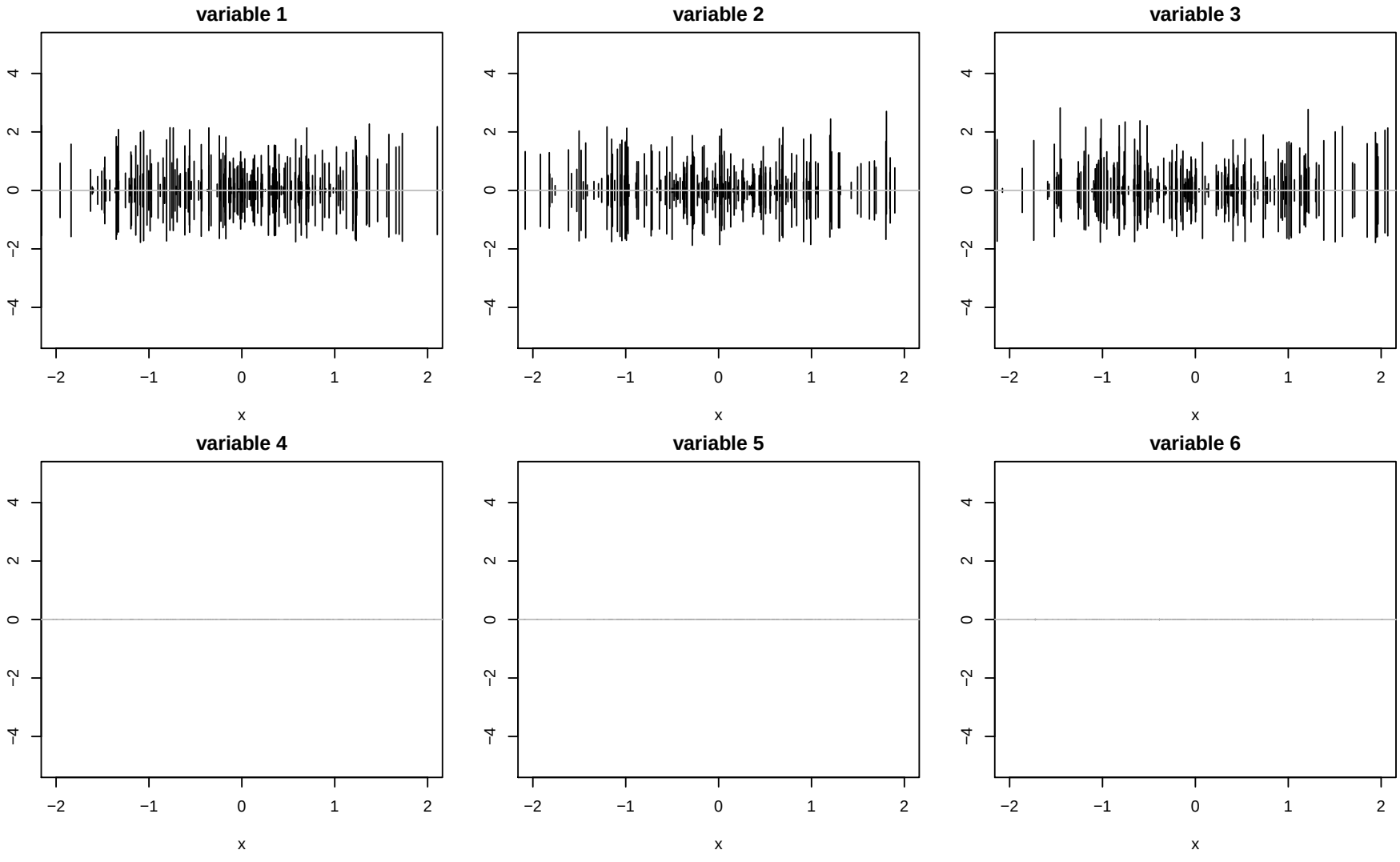
Uncorrelated case, Lasso



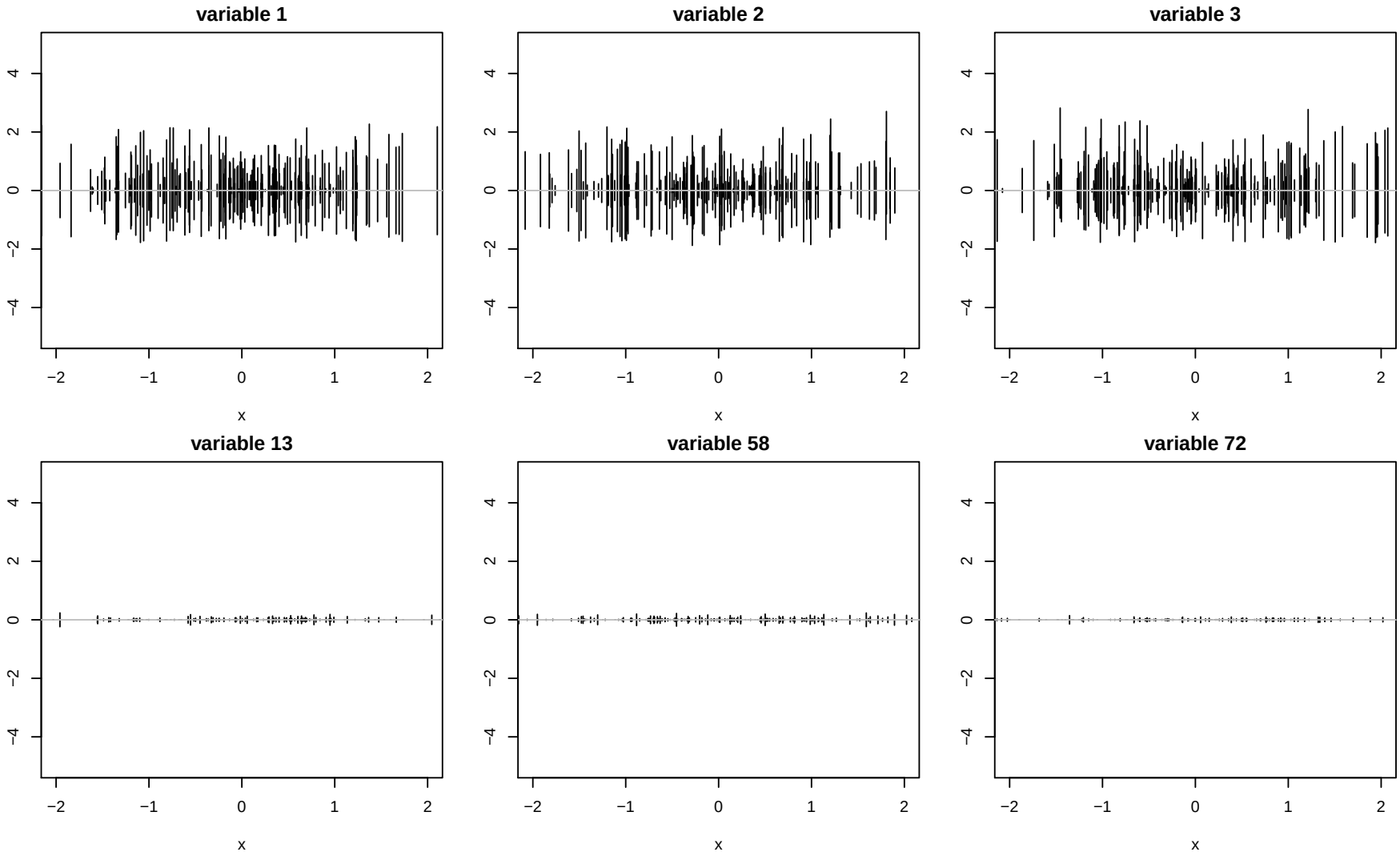
Uncorrelated case, Forward Stepwise



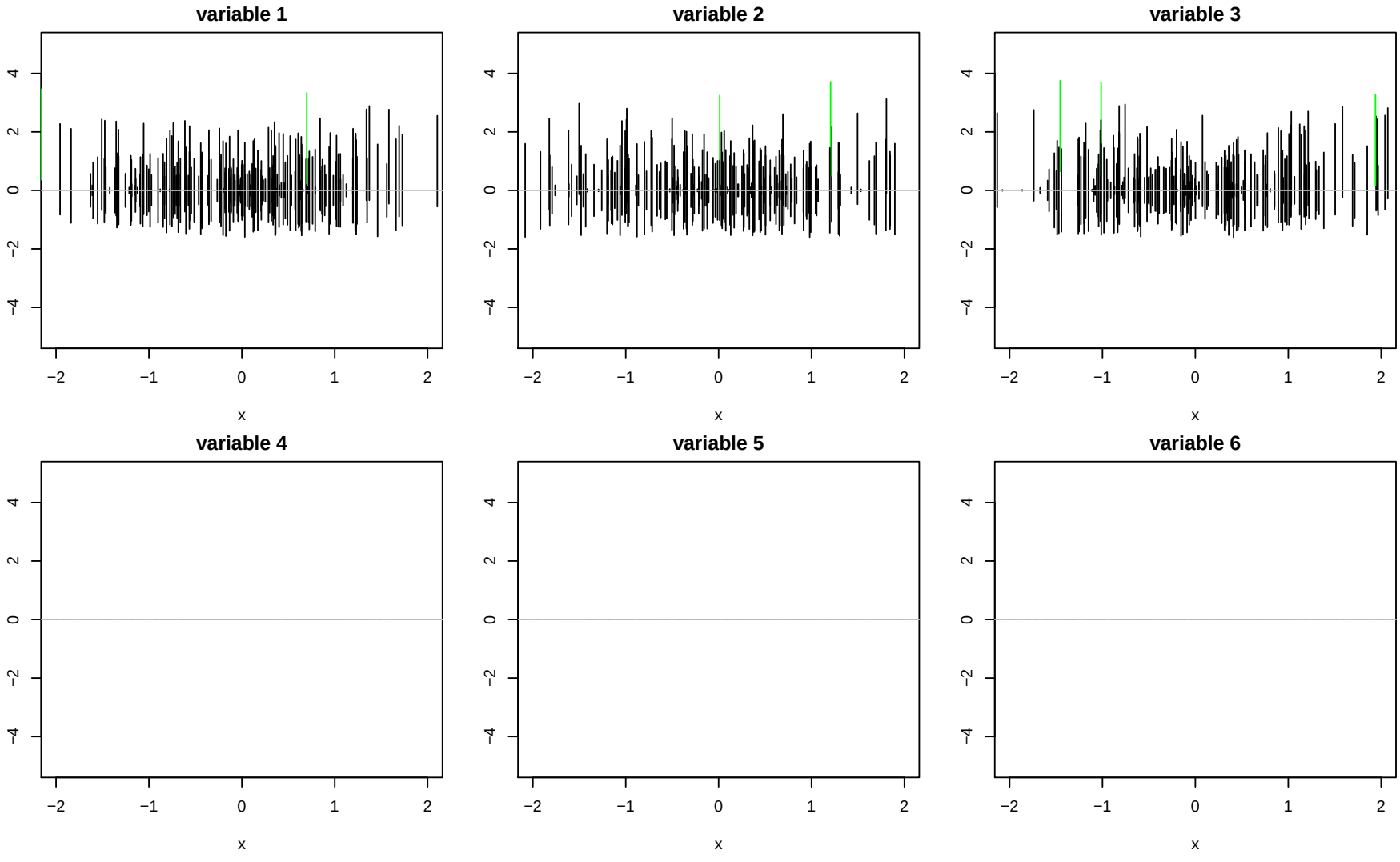
Correlated case, Lasso



Correlated case, Lasso



Correlated case, Forward Stepwise



Extensions

- ✓ Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y.
- ✓*(?)* Variable selection/importance?

Extensions

- ✓ Fast computation: can we avoid having to re-fit $\hat{\mu}$ with extra data point (X_{n+1}, y) for all values of X_{n+1} and all y.
- ✓*(?)* Variable selection/importance?
- Higher order correction: can we produce prediction band with adaptive width?
- Theory: when $\hat{\mu}$ is a good estimator, then the conformal band is nearly optimal (requires standard assumptions).

Cross-validation with confidence

From conformalization to cross-validation

- Another look at the variable importance method:
  1. Is the prediction worse without variable j?
  2. Split the sample, fit both with and without X_j using half data.
  3. Compare the risk on the other half.
- This looks very much like cross-validation/sample-splitting, with just one difference:

  CV looks at the empirical mean of the validated loss, but conformal looks at the empirical quantiles.
- Idea: there is probably more information in the validated loss than just the empirical mean.

Overview

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

In the regression setting

- Data: $D = \{(X_i, Y_i) : 1 \leq i \leq n\}$, 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.

In the regression setting

- Data: $D = \{(X_i, Y_i) : 1 \leq i \leq n\}$, 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.
- The framework can be extended to unsupervised learning problems, including network data.

Model selection

- Candidate set: $\mathscr{M} = \{1, \ldots, M\}$. Each $m \in \mathscr{M}$ corresponds to a candidate model.
  1. m can represent a competing theory about P (e.g., μ is linear, μ is quadratic, variable j is irrelevant, etc).
  2. m can represent a particular value of a tuning parameter of a certain algorithm to calculate $\hat{\mu}$ (e.g., λ in the lasso, choice of loss function, structure of NN)
- 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 $\{1,\ldots,n\}$.
- 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: $\hat{\mu}_1 \equiv 0$, $\hat{\mu}_2 = \bar{\varepsilon}_{\text{tr}}$.
- $\hat{m}_{\text{cv}} = 1 \quad \Leftrightarrow \quad 0 < \hat{Q}(\hat{\mu}_2) - \hat{Q}(\hat{\mu}_1) = \bar{\varepsilon}_{\text{tr}}^2 - 2\bar{\varepsilon}_{\text{tr}}\bar{\varepsilon}_{\text{te}}$.
- 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.

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: $\hat{\mu}_1 \equiv 0$, $\hat{\mu}_2 = \bar{\varepsilon}_{\text{tr}}$.
- $\hat{m}_{\text{cv}} = 1 \quad \Leftrightarrow \quad 0 < \hat{Q}(\hat{\mu}_2) - \hat{Q}(\hat{\mu}_1) = \bar{\varepsilon}_{\text{tr}}^2 - 2\bar{\varepsilon}_{\text{tr}}\bar{\varepsilon}_{\text{te}}$.
- 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 $n_{\text{tr}} = o(n)$.
- V-fold does not help!

A fix for the simple example: hypothesis testing

- The fundamental question: When we see $\hat{Q}(\hat{\mu}_2) < \hat{Q}(\hat{\mu}_1)$, do we feel confident to say $Q(\hat{\mu}_2) < Q(\hat{\mu}_1)$?

A fix for the simple example: hypothesis testing

- The fundamental question: When we see $\hat{Q}(\hat{\mu}_2) < \hat{Q}(\hat{\mu}_1)$, do we feel confident to say $Q(\hat{\mu}_2) < Q(\hat{\mu}_1)$?
- A standard solution uses hypothesis testing

$$H_0 : Q(\hat{\mu}_1) \leq Q(\hat{\mu}_2)$$

conditioning on $\hat{\mu}_1$, $\hat{\mu}_2$.

A fix for the simple example: hypothesis testing

- The fundamental question: When we see $\hat{Q}(\hat{\mu}_2) < \hat{Q}(\hat{\mu}_1)$, do we feel confident to say $Q(\hat{\mu}_2) < Q(\hat{\mu}_1)$?
- A standard solution uses hypothesis testing

$$H_0 : Q(\hat{\mu}_1) \leq Q(\hat{\mu}_2)$$

  conditioning on $\hat{\mu}_1$, $\hat{\mu}_2$.
- Can do this using a paired sample t-test, say with type I error level α.

CVC for the simple example

- Recall that $H_0 : Q(\hat{\mu}_1) \leq Q(\hat{\mu}_2)$.
- When H_0 is not rejected, does it mean we shall just pick $m = 1$?

CVC for the simple example

- Recall that $H_0 : Q(\hat{\mu}_1) \leq Q(\hat{\mu}_2)$.
- When H_0 is not rejected, does it mean we shall just pick $m = 1$?
- No. Because if we consider $H_0' : Q(\hat{\mu}_2) \leq Q(\hat{\mu}_1)$. H_0' will not be rejected either (probability of rejecting H_0' is bounded away from 0.)
- Most likely, we do not reject H_0 or H_0'.

CVC for the simple example

- Recall that $H_0 : Q(\hat{\mu}_1) \leq Q(\hat{\mu}_2)$.
- When H_0 is not rejected, does it mean we shall just pick $m = 1$?
- No. Because if we consider $H_0' : Q(\hat{\mu}_2) \leq Q(\hat{\mu}_1)$. H_0' will not be rejected either (probability of rejecting H_0' is bounded away from 0.)
- Most likely, we do not reject H_0 or H_0'.
- We accept both fitted models $\hat{\mu}_1$ and $\hat{\mu}_2$, as they are very similar and the difference cannot be noticed from the data.

Existing work

- Hansen et al (2011, Econometrica): sequential testing, only for low dimensional problems.
- Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.

Existing work

- Hansen et al (2011, Econometrica): sequential testing, only for low dimensional problems.
- Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.
- Our approach: one step, with provable coverage and power under mild assumptions in high dimensions.

Existing work

- Hansen et al (2011, Econometrica): sequential testing, only for low dimensional problems.
- Ferrari and Yang (2014): F-tests, need a good variable screening procedure in high dimensions.
- Our approach: one step, with provable coverage and power under mild assumptions in high dimensions.
- Key technique: high-dimensional Gaussian comparison of sample means (Chernozhukov et al).

CVC in general

- Now suppose we have a set of candidate models $\mathscr{M} = \{1, \ldots, M\}$.
- 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 $Q(\hat{\mu}) = \mathbb{E}\left[\ell(\hat{\mu}(X), Y) \mid \hat{\mu}\right]$.

CVC in general

- Now suppose we have a set of candidate models $\mathscr{M} = \{1, \ldots, M\}$.
- 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 $Q(\hat{\mu}) = \mathbb{E}\left[\ell(\hat{\mu}(X), Y) \mid \hat{\mu}\right]$.
- For each m, test hypothesis (conditioning on $\hat{\mu}_1, \ldots, \hat{\mu}_M$)

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

- Let $\hat{p}_m$ be a valid p-value.

CVC in general

- Now suppose we have a set of candidate models $\mathscr{M} = \{1, \ldots, M\}$.
- 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 $Q(\hat{\mu}) = \mathbb{E}\left[\ell(\hat{\mu}(X), Y) \mid \hat{\mu}\right]$.
- For each m, test hypothesis (conditioning on $\hat{\mu}_1, \ldots, \hat{\mu}_M$)

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

- 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 that D_{tr} is the training data and D_{te} is the testing data.
- The test and p-values are conditional on D_{tr}.
- Data: $n_{\text{te}} \times (M-1)$ 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 $\hat{p}_m$

- Recall that D_{tr} is the training data and D_{te} is the testing data.
- The test and p-values are conditional on D_{tr}.
- Data: $n_{\text{te}} \times (M-1)$ 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. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \ \forall\, j \neq m.$
- Challenges

Calculating $\hat{p}_m$

- Recall that D_{tr} is the training data and D_{te} is the testing data.
- The test and p-values are conditional on D_{tr}.
- Data: $n_{\text{te}} \times (M-1)$ 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. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \ \forall\, j \neq m.$
- Challenges
  1. High dimensionality: M can be large.

Calculating $\hat{p}_m$

- Recall that D_{tr} is the training data and D_{te} is the testing data.
- The test and p-values are conditional on D_{tr}.
- Data: $n_{\text{te}} \times (M-1)$ 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. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \ \forall\, j \neq m.$
- Challenges
  1. High dimensionality: M can be large.
  2. Potentially high correlation between $\xi_{m,j}$ and $\xi_{m,j'}$.

Calculating $\hat{p}_m$

- Recall that D_{tr} is the training data and D_{te} is the testing data.
- The test and p-values are conditional on D_{tr}.
- Data: $n_{\text{te}} \times (M-1)$ 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. $H_{0,m} : \mathbb{E}(\xi_{m,j}) \leq 0, \ \forall\, j \neq m.$
- Challenges
  1. High dimensionality: M can be large.
  2. Potentially high correlation between $\xi_{m,j}$ and $\xi_{m,j'}$.
  3. Vastly different scaling: $\text{Var}(\xi_{m,j})$ can be $O(1)$ or $O(n^{-1})$.

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 $(\xi_{m,j}^{(i)} : i \in I_{\text{te}})$.
- 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.

Studentized Gaussian Multiplier Bootstrap

1. $T_m = \max_{j \neq m} \sqrt{n_{\text{te}}} \dfrac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}$
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 $T_b^* = \max_{j \neq m} \dfrac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \dfrac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i$
3. $\hat{p}_m = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_b^* > T_m)$.

Studentized Gaussian Multiplier Bootstrap

1. $T_m = \max_{j \neq m} \sqrt{n_{\text{te}}} \dfrac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}$
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 $T_b^* = \max_{j \neq m} \dfrac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \dfrac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i$
3. $\hat{p}_m = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_b^* > T_m)$.

- The studentization takes care of the scaling difference.

Studentized Gaussian Multiplier Bootstrap

1. $T_m = \max_{j \neq m} \sqrt{n_{\text{te}}} \dfrac{\hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}}$
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 $T_b^* = \max_{j \neq m} \dfrac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \dfrac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i$
3. $\hat{p}_m = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_b^* > T_m)$.

- The studentization takes care of the scaling difference.
- The bootstrap Gaussian comparison takes care of the dimensionality and correlation.

Properties of CVC

- $\mathscr{A}_{\text{cvc}} = \{m : \hat{p}_m > \alpha\}$.
- Let $\hat{m}_{\text{cv}} = \arg\min_m \hat{Q}(\hat{\mu}_m)$. By construction $T_{\hat{m}_{\text{cv}}} \leq 0$.

Proposition

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

- Proof: $\left[\frac{1}{\sqrt{n_{\text{te}}}} \sum_{i \in I_{\text{te}}} \frac{\xi_{m,j}^{(i)} - \hat{\mu}_{m,j}}{\hat{\sigma}_{m,j}} \zeta_i\right]_{j \neq m}$ is a zero-mean Gaussian random vector. So the upper α quantile of its maximum must be positive.
- Can view $\hat{m}_{\text{cv}}$ as the "center" of the confidence set.

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

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

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 $(\xi^{(i)}_{m,j})_{i \in I_{\text{te}}, j \neq m}$. 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 $\lesssim \sqrt{\log M}$

V-fold CVC

- Split data into V folds.

V-fold CVC

- Split data into V folds.
- Let v_i be the fold that contains data point i.

V-fold CVC

- Split data into V folds.
- Let v_i be the fold that contains data point i.
- Let $\hat{\mu}_{m,v}$ be the estimate using model m and all data but fold v.

V-fold CVC

- Split data into V folds.
- Let v_i be the fold that contains data point i.
- Let $\hat{\mu}_{m,v}$ be the estimate using model m and all data but fold v.
- $\xi_{m,j}^{(i)} = \ell(\hat{\mu}_{m,v_i}(X_i), Y_i) - \ell(\hat{\mu}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.

V-fold CVC

- Split data into V folds.
- Let v_i be the fold that contains data point i.
- Let $\hat{\mu}_{m,v}$ be the estimate using model m and all data but fold v.
- $\xi_{m,j}^{(i)} = \ell(\hat{\mu}_{m,v_i}(X_i), Y_i) - \ell(\hat{\mu}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.
- Treat folds as independent samples with group mean effects.

V-fold CVC

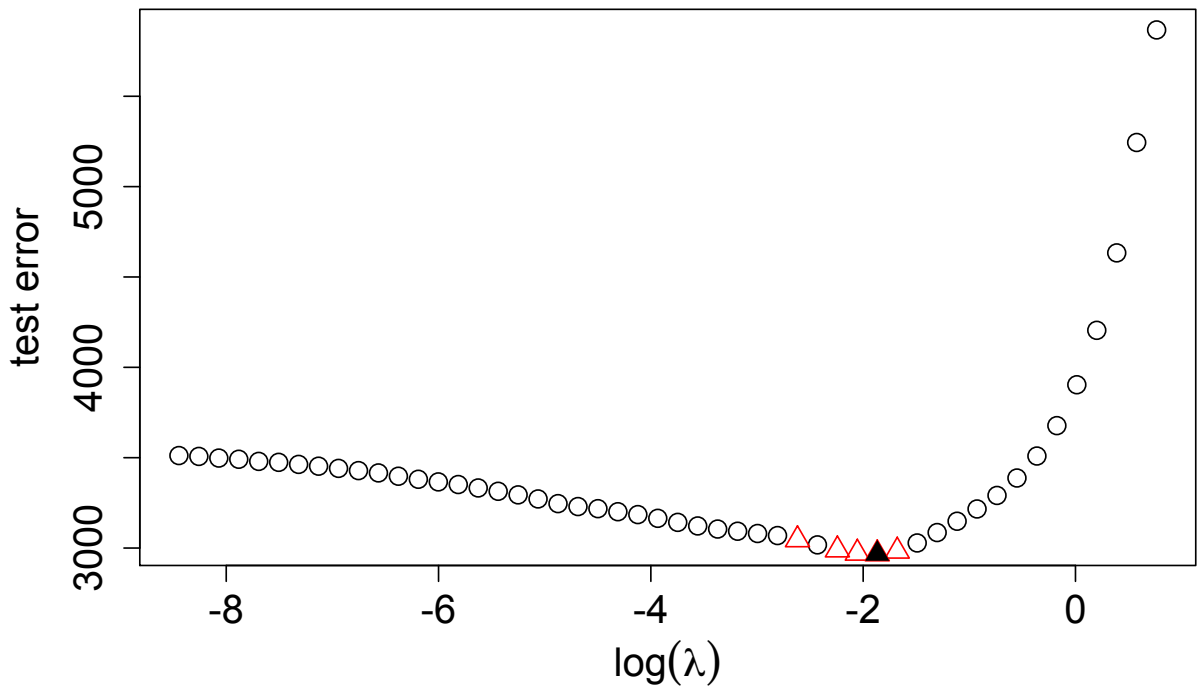
- Split data into V folds.
- Let v_i be the fold that contains data point i.
- Let $\hat{\mu}_{m,v}$ be the estimate using model m and all data but fold v.
- $\xi_{m,j}^{(i)} = \ell(\hat{\mu}_{m,v_i}(X_i), Y_i) - \ell(\hat{\mu}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.
- Treat folds as independent samples with group mean effects.
- Calculate T_m and T_b^* correspondingly using the $n \times (M-1)$ cross-validated error difference matrix $(\xi_{m,j}^{(i)})_{1 \leq i \leq n, j \neq m}$.

V-fold CVC

- Split data into V folds.
- Let v_i be the fold that contains data point i.
- Let $\hat{\mu}_{m,v}$ be the estimate using model m and all data but fold v.
- $\xi_{m,j}^{(i)} = \ell(\hat{\mu}_{m,v_i}(X_i), Y_i) - \ell(\hat{\mu}_{m,v_i}(X_i, Y_i))$, for all $1 \leq i \leq n$.
- Treat folds as independent samples with group mean effects.
- Calculate T_m and T_b^* correspondingly using the $n \times (M-1)$ cross-validated error difference matrix $(\xi_{m,j}^{(i)})_{1 \leq i \leq n, j \neq m}$.
- Rigorous justification is hard due to dependence between folds. But empirically much better.

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 $\mathscr{A}_{\text{cvc}}$, solid triangle: $\hat{m}_{\text{cv}}$.

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

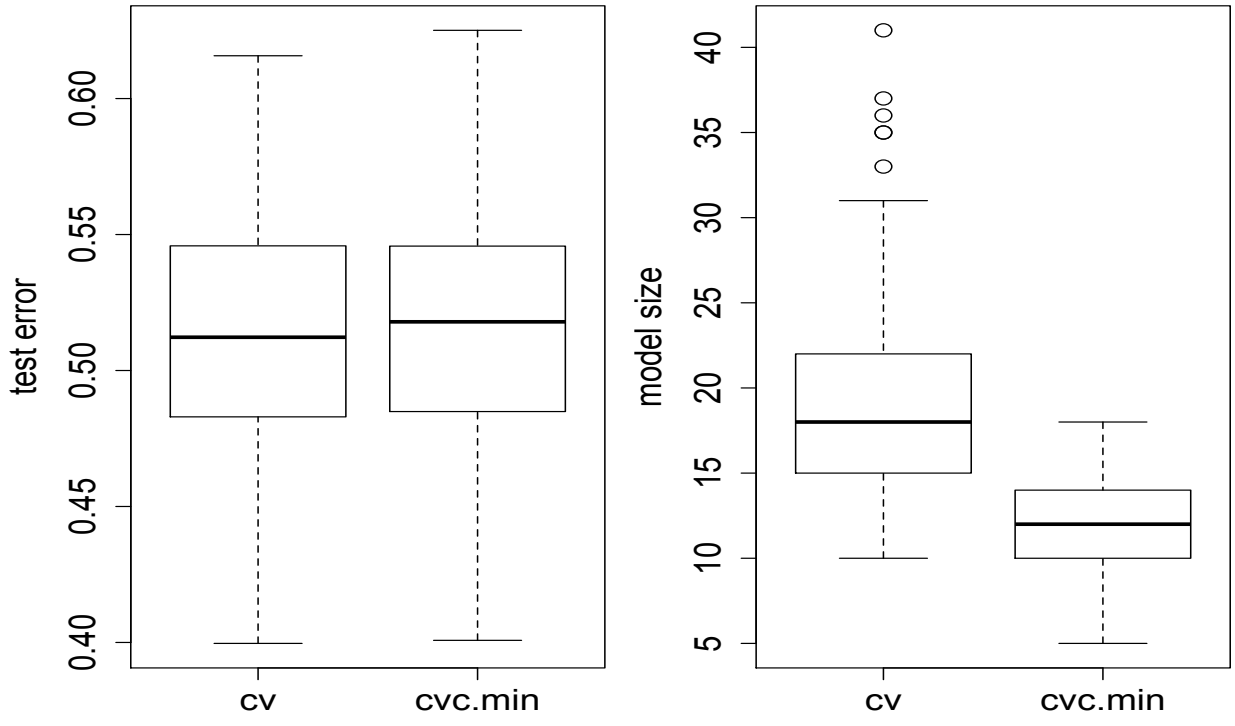
- Let J_m be the subset of variables selected using model m

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

- $\hat{m}_{\text{cvc.min}}$ is the simplest model that gives a similar predictive risk as $\hat{m}_{\text{cv}}$.

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

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

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

Theoretical analysis: basic setup

- Assume iid data from model $Y = \mu(X) + \varepsilon$, where the density of ε is symmetric, decreasing on $[0, \infty)$.
- Let $\hat{\mu}_n(\cdot)$ be any point estimator from a sample of size n.
- Super oracle band: $C_s^*(x) = [\mu(x) \pm q_\alpha]$, where q_α is the upper α quantile of $|\varepsilon|$.
- Oracle band: $C_o^*(x) = [\hat{\mu}_n(x) \pm q_{n,\alpha}]$, where $q_{n,\alpha}$ is the upper α quantile of $|Y - \hat{\mu}_n(X)|$.

Approximating the oracle

Let v_n be the width of the split conformal band obtained from $\hat{\mu}_n$.

Theorem

If $\hat{\mu}_n$ satisfies the *sampling stability*

$$\mathbb{P}(\|\hat{\mu}_n - \mu_0\|_\infty \geq \eta_n) \leq \rho_n$$

for some function μ_0, and $\eta_n \vee \rho_n = o(1)$, then

$$v_n - 2q_{n,\alpha} = o_P(1).$$

Remark

- Similar result is available for full conformal bands.
- μ_0 can be different from μ (e.g., undersmoothing).

Approximate the super oracle

Theorem

If the density function of $|\varepsilon|$ has continuous derivative that is uniformly bounded by a constant M, then

$$|q_\alpha - q_{n,\alpha}| \lesssim M\mathbb{E}(\hat{\mu}_n(X) - \hat{\mu}(X))^2.$$

where the expectation is taken over both $\hat{\mu}_n$ and a freshly drawn X.

As a consequence, the two oracle bands are close to each other if $\hat{\mu}_n(x) \approx \mu(x)$.

Approximate the super oracle (cont'd)

Theorem

Assuming additionally that $\mathbb{E}(\hat{\mu}_n(X) - \mu(X))^2 = o(1)$, then

$$\mathrm{Leb}(C_{\mathrm{split}}(X) \triangle C_s^*(X)) = o_P(1)\,.$$

Simulations: coverage of $\mathscr{A}_{\text{CVC}}$

- $Y = X^T\beta + \varepsilon$, $X \sim N(0,\Sigma)$, $\varepsilon \sim N(0,1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1,1,1,0,\ldots,0)^T$ (simple), or
  $\beta = (1,1,1,0.7,0.5,0.3,0,\ldots,0)^T$ (mixed).
- 5-fold CVC with $\alpha = 0.05$ using Lasso with 50 values of λ

setting of (Σ,β)	coverage	$\lvert\mathscr{A}_{\text{CVC}}\rvert$	cv is opt.
identity, simple	.92 (.03)	5.1 (.19)	.27 (.04)
identity, mixed	.95 (.02)	5.1 (.18)	.37 (.05)
correlated, simple	.96 (.02)	7.5 (.18)	.18 (.04)
correlated, mixed	.93 (.03)	7.4 (.23)	.19 (.04)

Simulations: coverage of $\mathscr{A}_{\text{CVC}}$

- $Y = X^T\beta + \varepsilon$, $X \sim N(0,\Sigma)$, $\varepsilon \sim N(0,1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1,1,1,0,\ldots,0)^T$ (simple), or
  $\beta = (1,1,1,0.7,0.5,0.3,0,\ldots,0)^T$ (mixed).
- 5-fold CVC with $\alpha = 0.05$ using forward stepwise

setting of (Σ,β)	coverage	$\lvert\mathscr{A}_{\text{CVC}}\rvert$	cv is opt.
identity, simple	1 (0)	3.7 (.29)	.87 (.03)
identity, mixed	.95 (.02)	5.2 (.33)	.58 (.05)
correlated, simple	.97 (.02)	4.1 (.31)	.80 (.04)
correlated, mixed	.93 (.03)	6.3 (.36)	.44 (.05)

How to use $\mathscr{A}_{\text{CVC}}$?

- We are often interested in picking one model, not a subset of models.
- $\mathscr{A}_{\text{cvc}}$ provides some flexibility of picking among a subset of highly competitive models.

How to use $\mathscr{A}_{\text{CVC}}$?

- We are often interested in picking one model, not a subset of models.
- $\mathscr{A}_{\text{CVC}}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathscr{A}_{\text{CVC}}$ may contain a model that includes a particularly interesting variable.

How to use $\mathscr{A}_{\text{CVC}}$?

- We are often interested in picking one model, not a subset of models.
- $\mathscr{A}_{\text{CVC}}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathscr{A}_{\text{CVC}}$ may contain a model that includes a particularly interesting variable.
  2. $\mathscr{A}_{\text{CVC}}$ can be used to answers questions like "Is fitting procedure A better than procedure B?"

How to use $\mathscr{A}_{\text{CVC}}$?

- We are often interested in picking one model, not a subset of models.
- $\mathscr{A}_{\text{CVC}}$ provides some flexibility of picking among a subset of highly competitive models.
  1. $\mathscr{A}_{\text{CVC}}$ may contain a model that includes a particularly interesting variable.
  2. $\mathscr{A}_{\text{CVC}}$ can be used to answers questions like "Is fitting procedure A better than procedure B?"
  3. We can also simply choose the most parsimonious model in $\mathscr{A}_{\text{CVC}}$.

A classical setting

- $Y = X^T\beta + \varepsilon$, $X \in \mathbb{R}^p$, $\mathrm{Var}(X) = \Sigma$ has full rank.
- ε has mean zero and variance $\sigma^2 < \infty$.
- Assume that (p, Σ, σ^2) are fixed and $n \to \infty$.
- $\mathscr{M}$ contains the true model m^*, and at least one overfitting model.
- $n_{\mathrm{tr}}/n_{\mathrm{te}} \asymp 1$.
- Using squared loss, the true model and all overfitting models give $\sqrt{n}$-consistent estimates.
- Early results (Shao 93, Zhang 93, Yang 07) show that $\mathbb{P}(\hat{m}_{\mathrm{cv}} \neq m^*)$ is bounded away from 0.

Consistency of $\hat{m}_{\text{cvc.min}}$

Theorem

Assume that X and ε are independent and sub-Gaussian, and $\mathscr{A}_{\text{cvc}}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n\to\infty} \mathbb{P}(\hat{m}_{\text{cvc.min}} = m^*) = 1\,.$$

Consistency of $\hat{m}_{\text{cvc.min}}$

Theorem

Assume that X and ε are independent and sub-Gaussian, and $\mathscr{A}_{\text{cvc}}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n\to\infty} \mathbb{P}(\hat{m}_{\text{cvc.min}} = m^*) = 1\,.$$

- Sub-Gaussianity of X and ε implies that $(Y - X^T\beta)^2$ is sub-exponential.

Consistency of $\hat{m}_{\text{cvc.min}}$

Theorem

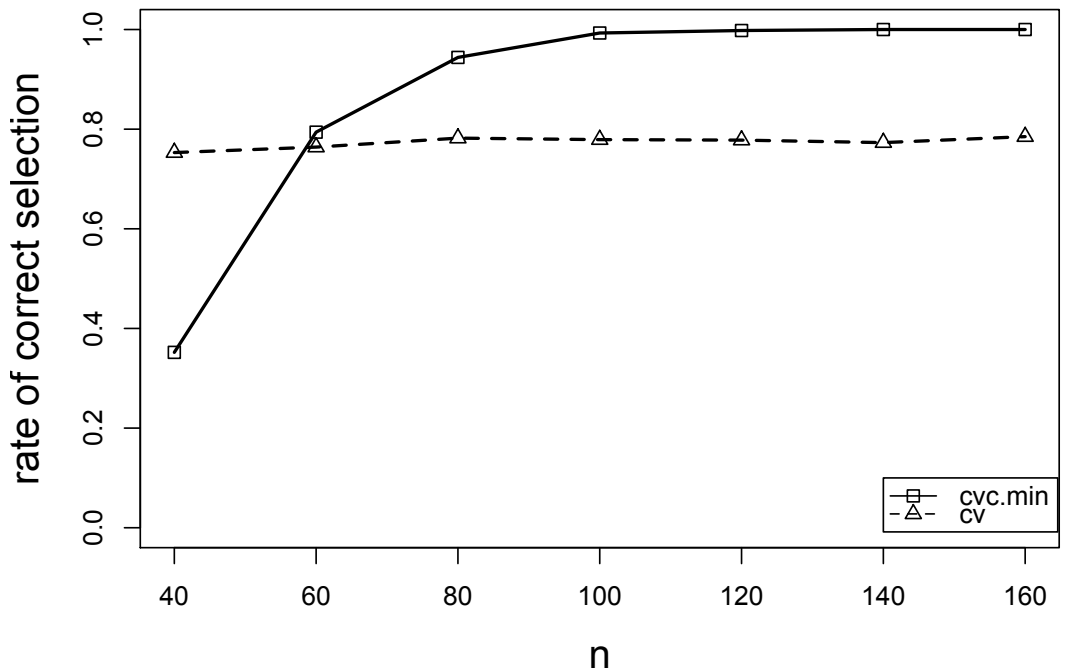
Assume that X and ε are independent and sub-Gaussian, and $\mathscr{A}_{\text{cvc}}$ is the output of CVC with $\alpha = o(1)$ and $\alpha \geq n^{-1}$, then

$$\lim_{n\to\infty} \mathbb{P}(\hat{m}_{\text{cvc.min}} = m^*) = 1\,.$$

- Sub-Gaussianity of X and ε implies that $(Y - X^T\beta)^2$ is sub-exponential.
- Can allow p to grow slowly as n using union bound.

Example in low-dim. variable selection

- Synthetic data with $p = 5$, $n = 40$, as in [Shao 93] .
- $Y = X^T\beta + \varepsilon$, $\beta = (2,9,0,4,8)^T$, $\varepsilon \sim N(0,1)$.
- Generated additional rows for $n = 60, 80, 100, 120, 140, 160$.
- Candidates: $(1,4,5), (1,2,4,5), (1,3,4,5), (1,2,3,4,5)$
- Repeated 1000 times, using OLS with 5-fold CVC.



Simulations: variable selection with $\hat{m}_{\text{cvc.min}}$

- $Y = X^T\beta + \varepsilon$, $X \sim N(0,\Sigma)$, $\varepsilon \sim N(0,1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1,1,1,0,...,0)^T$ (simple)
- 5-fold CVC with $\alpha = 0.05$ using forward stepwise

setting of (Σ, β)	oracle	$\hat{m}_{\text{cvc.min}}$	$\hat{m}_{\text{cv}}$
identity, simple	1	1	.87
correlated, simple	1	.97	.80

Proportion of correct model selection over 100 independent data sets.

Oracle method: the number of steps that gives smallest prediction risk.

Simulations: variable selection with $\hat{m}_{\text{cvc.min}}$

- $Y = X^T\beta + \varepsilon$, $X \sim N(0,\Sigma)$, $\varepsilon \sim N(0,1)$, $n = 200$, $p = 200$
- $\Sigma = I_{200}$ (identity), or $\Sigma_{jk} = 0.5 + 0.5\delta_{jk}$ (correlated).
- $\beta = (1,1,1,0,...,0)^T$ (simple)
- 5-fold CVC with $\alpha = 0.05$ using Lasso + Least Square

setting of (Σ, β)	oracle	$\hat{m}_{\text{cvc.min}}$	$\hat{m}_{\text{cv}}$
identity, simple	1	1	.88
correlated, simple	.87	.85	.71

Proportion of correct model selection over 100 independent data sets.

Oracle method: the λ value that gives smallest prediction risk.