A Framework for Assumption-Free Predictive Regression Inference

Jing Lei

Department of Statistics, Carnegie Mellon University

U. Pitt. Statistics Seminar, Feb 17 2017

Based on joint work with Larry Wasserman, Ryan Tibshirani,
Alessandro Rinaldo, Max G'Sell

Overview

- Predictive regression analysis under weak assumptions.
- Conformal inference: efficient and reliable prediction using nonparametric techniques.
  - Key idea: out-of-sample fitting.
  - Fast computation.
  - Variable importance measure.
  - Theory and further extensions.

Regression

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 $\hat{\mu}$

- 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.

Assumption-free inference

- Most of these standard assumptions are not checkable and hard to satisfy in practice.
- False assumptions may give misleading inference (Buja et al 2014).
- Inference based on these assumptions can be fragile (Tibshirani et al 2015).
- But we can still use these estimators for reliable prediction, even without standard assumptions.

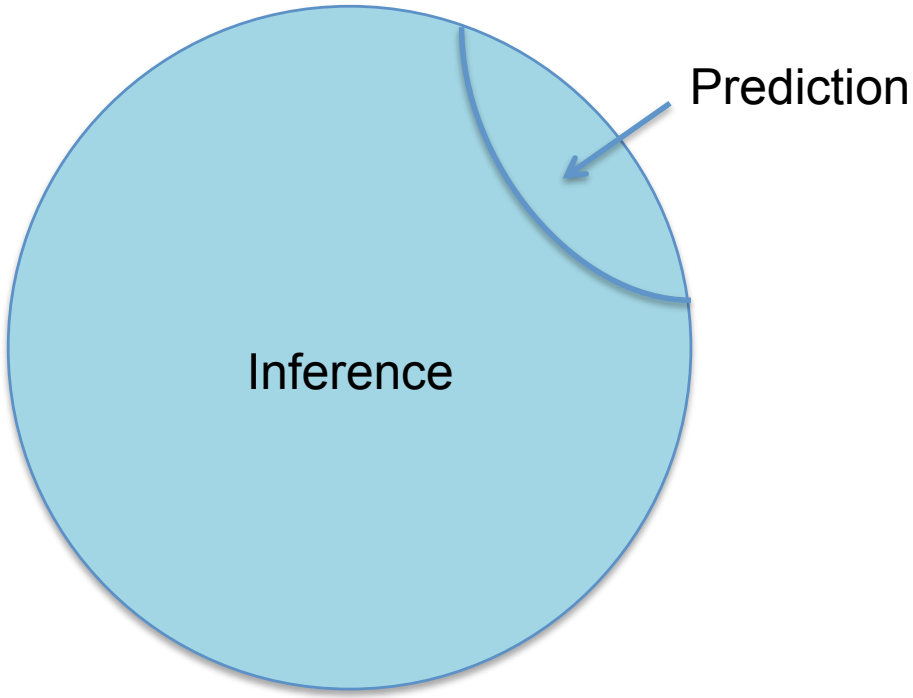
Predictive inference

- We would like to quantify the uncertainty of Y for each X observed in the future or in the sample.
- We also would like to provide insights on variable importance: which coordinates of X have significant predictive power for Y?

Inference and Prediction

or "Statistics and Machine Learning"

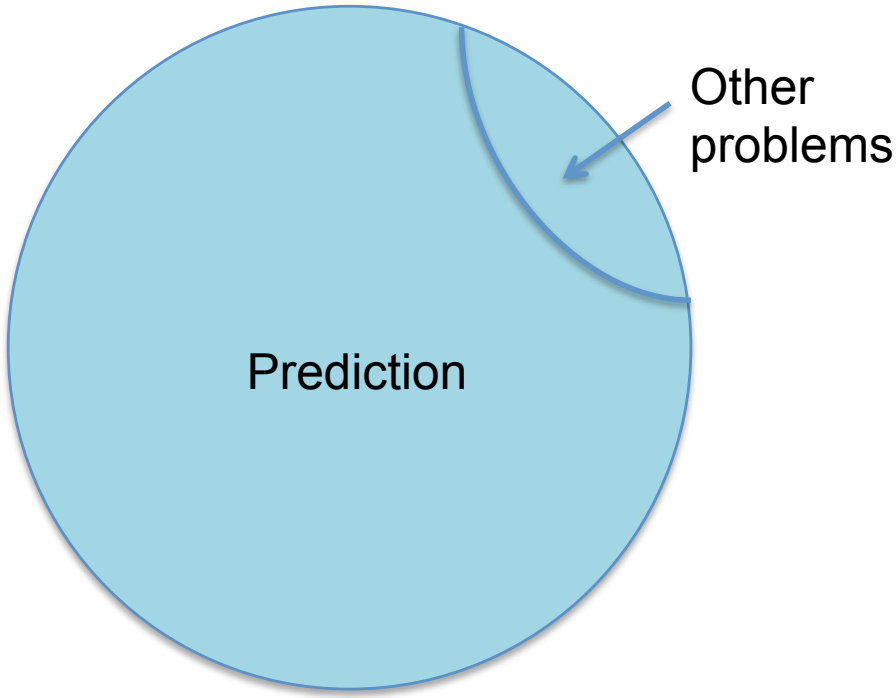
Statistician's perspective



Inference and Prediction

or "Statistics and Machine Learning"

Machine learner's perspective



Inference and Prediction

or "Statistics and Machine Learning"

Another perspective

Example: Leo Breiman, "Statistical Modeling: The Two Cultures", *Statist. Sci.* **16**(3) 2001.

Prediction vs Parameter Estimation

- Prediction usually requires less assumptions than estimation.
- In low dimensional linear regression, accurate prediction is possible even with strong colinearity.
- In high dimensional linear regression, Lasso estimators have near optimal predictive performance under essentially no assumption (Greenshtein & Ritov 04).
- In functional linear regression, prediction risk is much smaller than the estimation error (Cai & Hall 06).

Conformal inference

- What it is
  1. a general framework for predictive regression inference;
  2. can be combined with any existing or new regression estimator.
- What it does
  1. converts any 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, etc)
- $R_i = Y_i - \hat{\mu}(X_i)$
- 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.
- 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 $R_i^{(y)} = Y_i - \hat{\mu}^{(y)}(X_i)$, $1 \leq i \leq n+1$.
- 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 $R_i^{(y)} = Y_i - \hat{\mu}^{(y)}(X_i)$, $1 \leq i \leq n+1$.
- 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 $\hat{C}(X_{n+1}) = \{y \in \mathbb{R} : \pi_n(y) \leq 1 - \alpha\}$.

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.

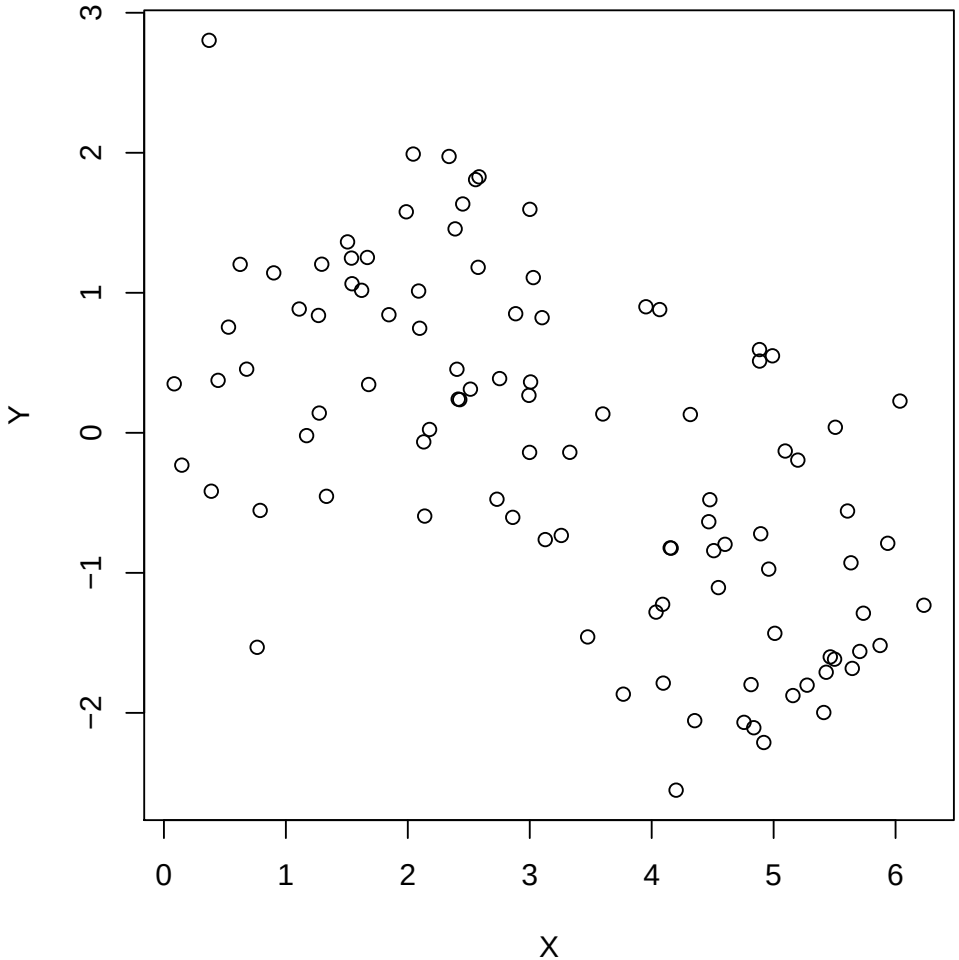
Remark: Can replace $R_i^{(y)}$ by

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

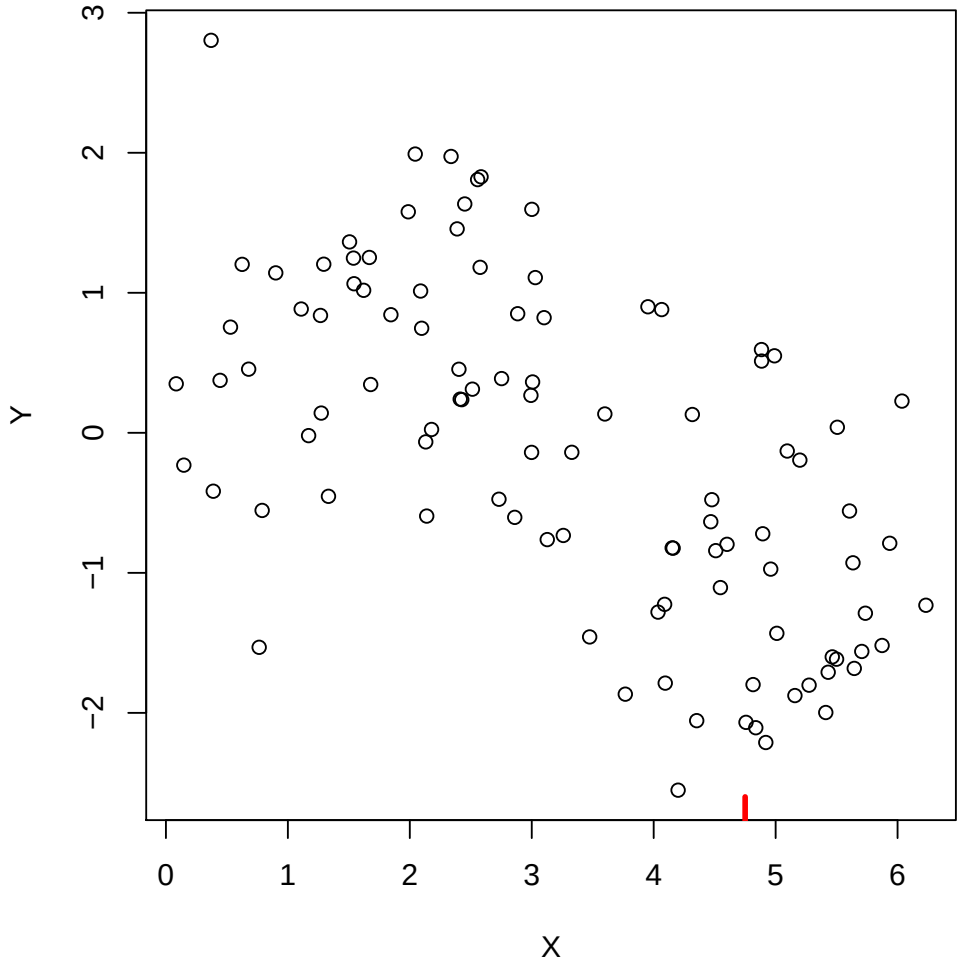
with $Z_i = (X_i, Y_i)$, $Y_{n+1} = y$, for any f that is symmetric in the first n arguments.

f is called the conformity score function.

Example: conformal prediction interval using smoothing splines

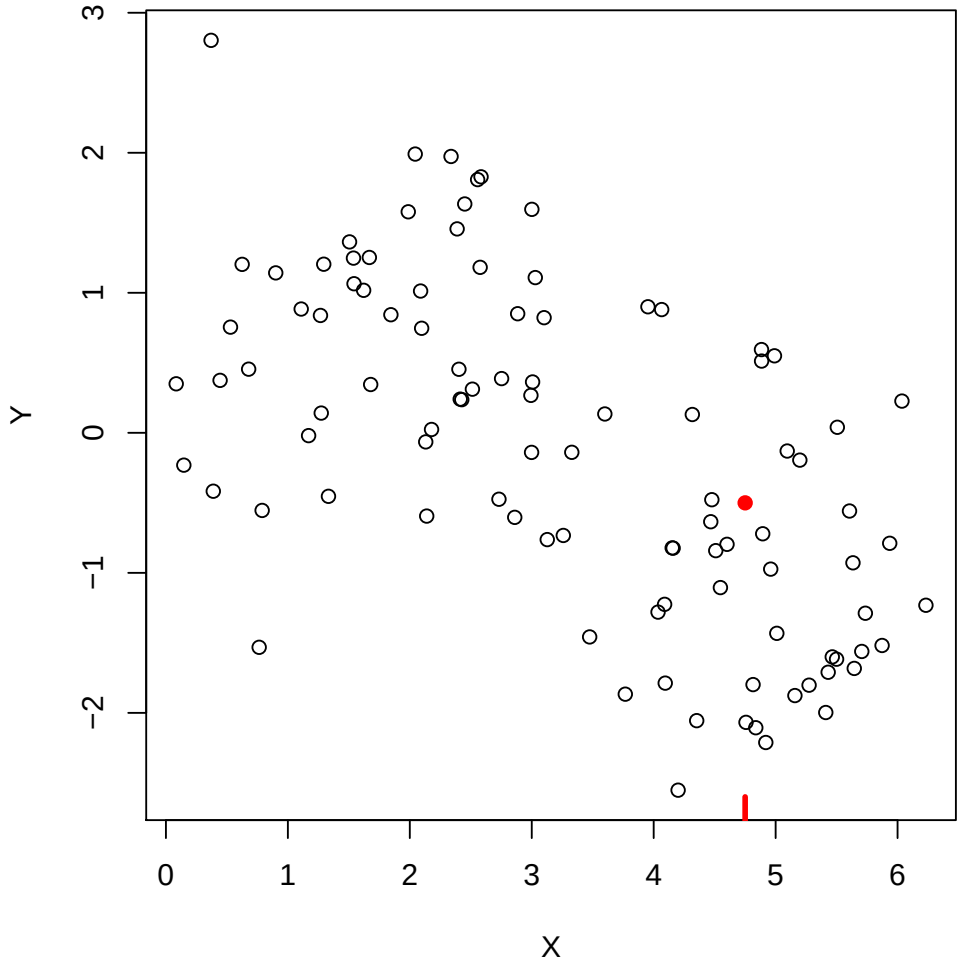


Example: conformal prediction interval using smoothing splines



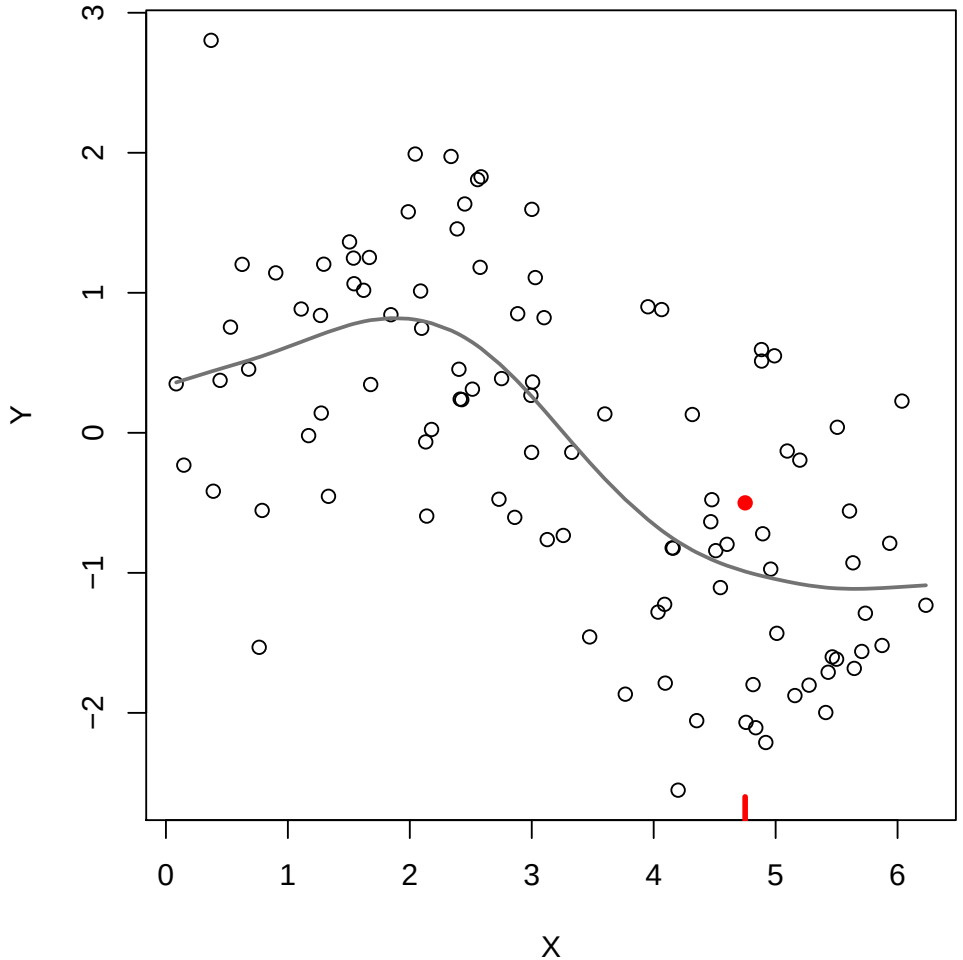
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



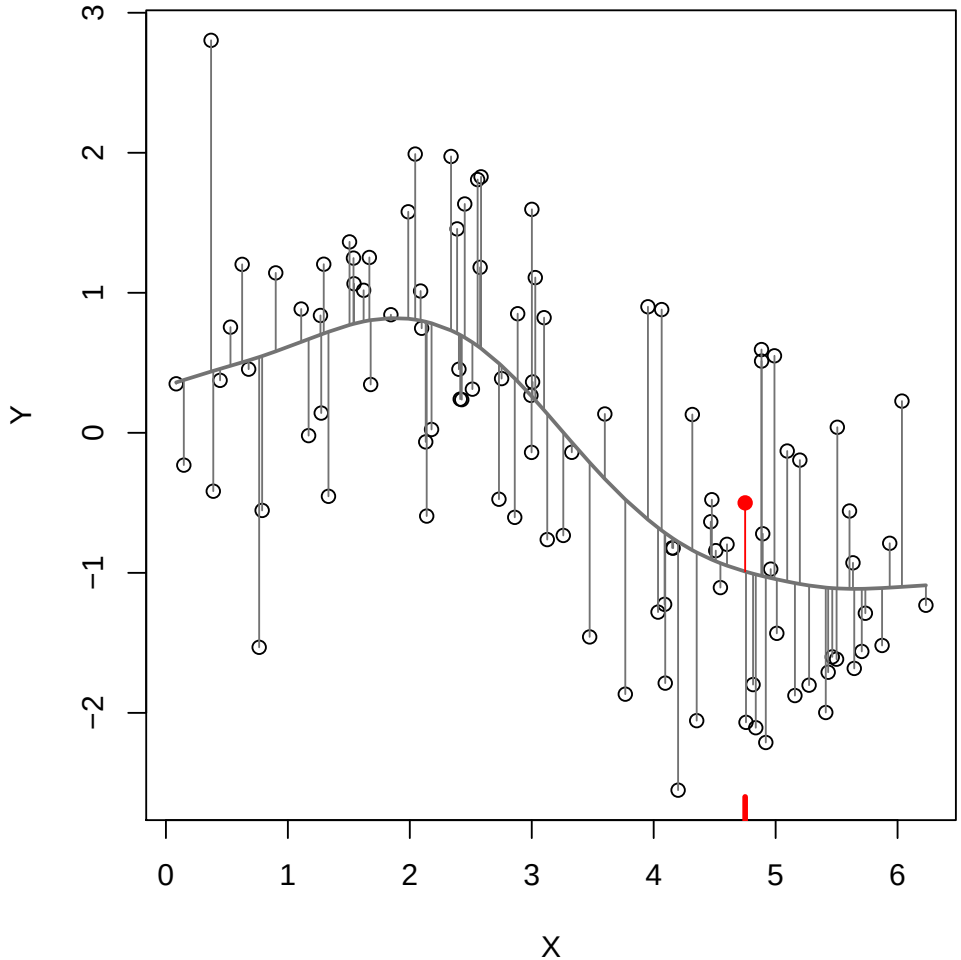
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



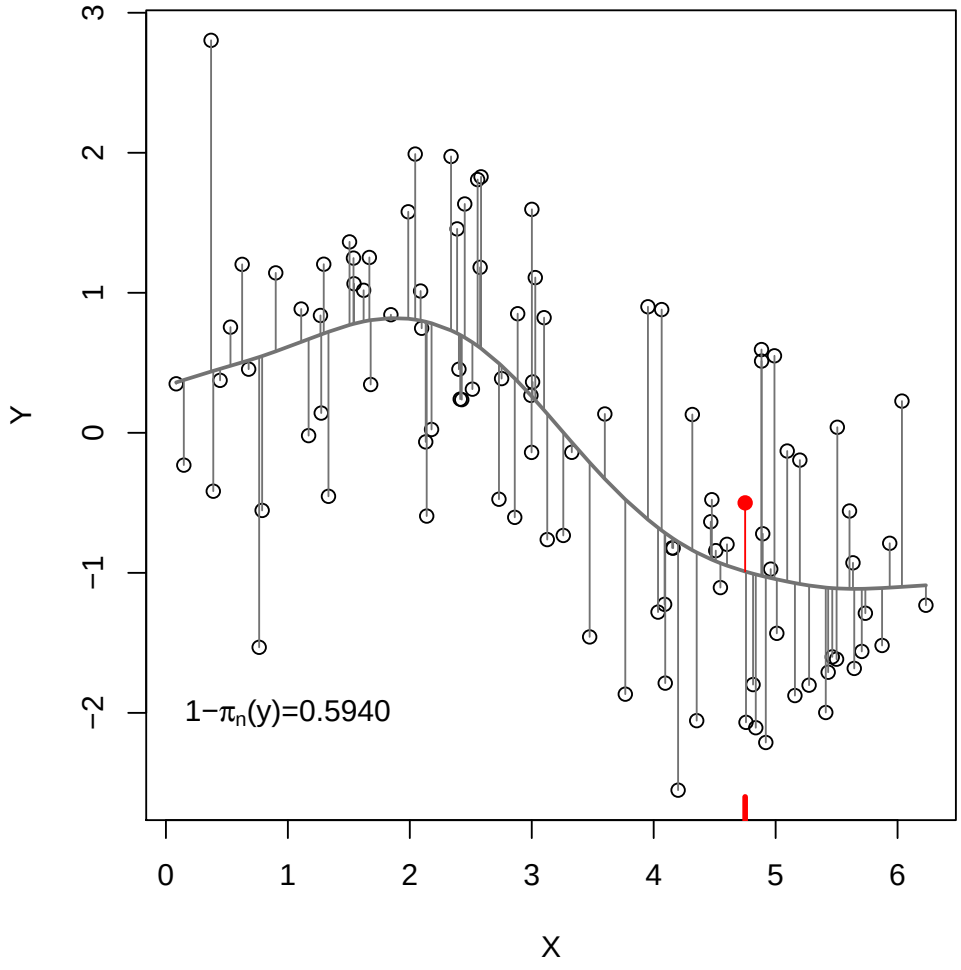
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



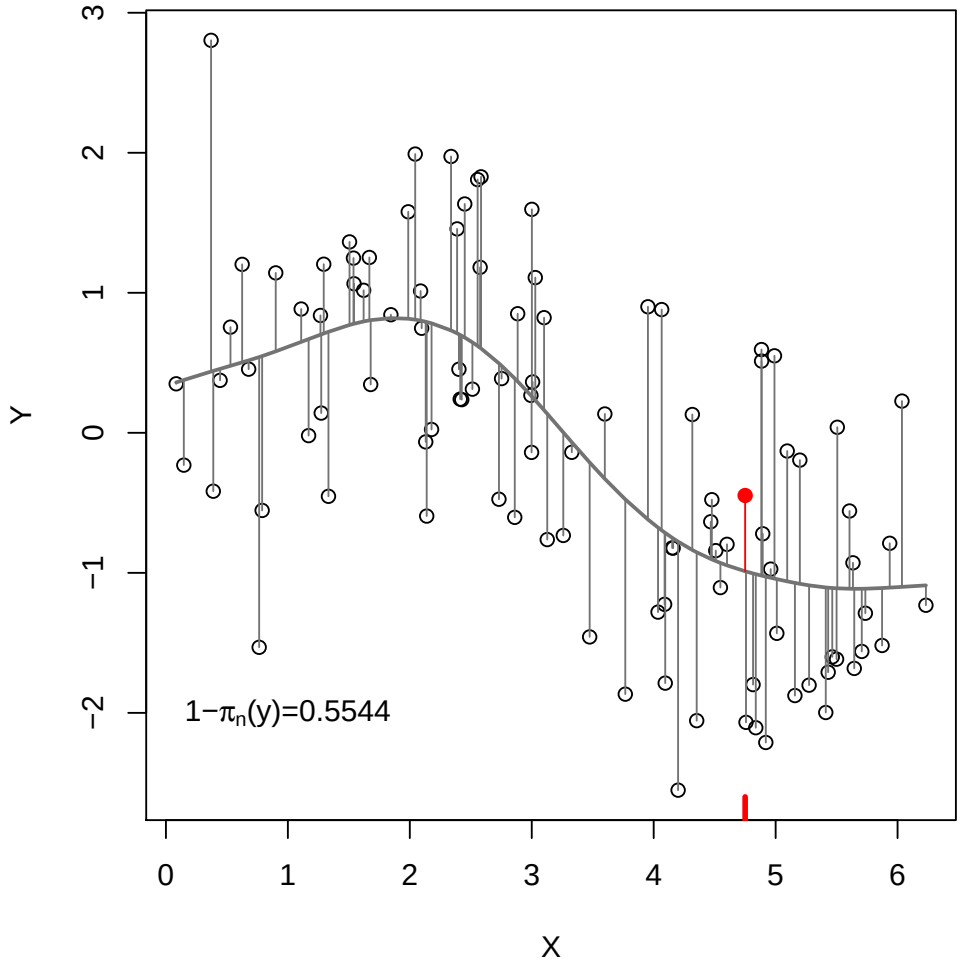
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



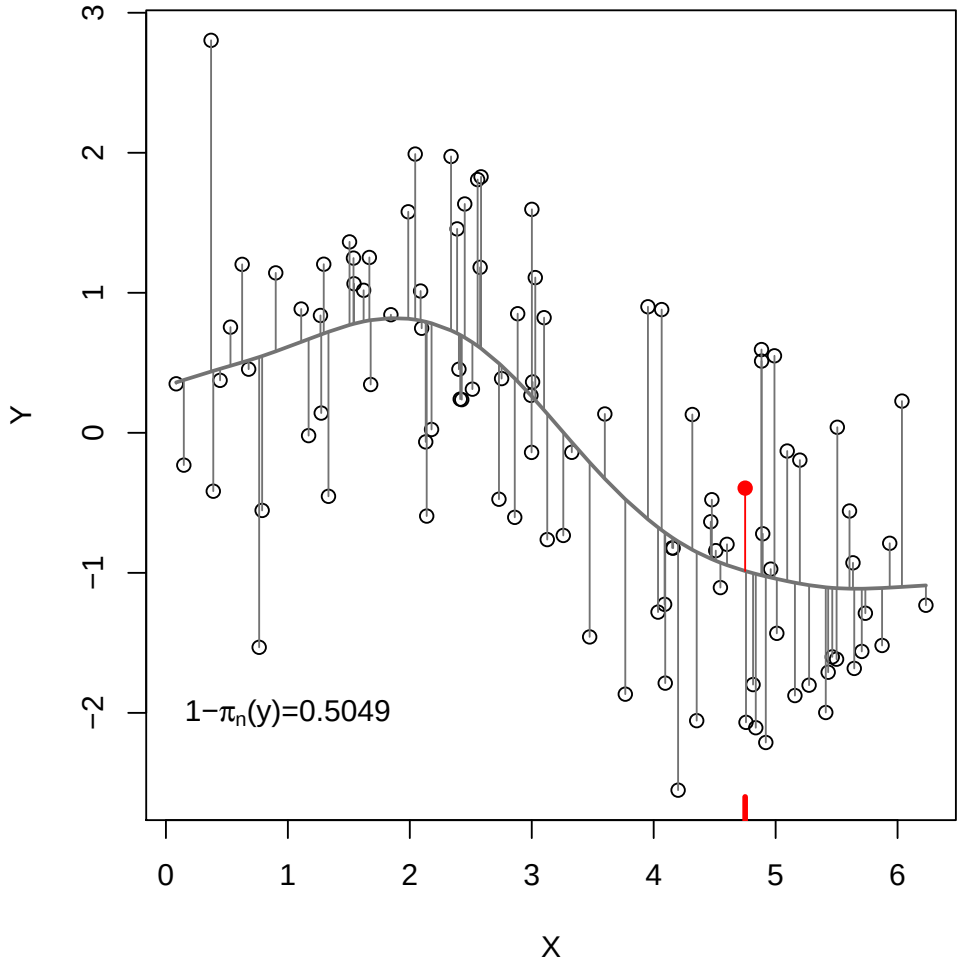
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



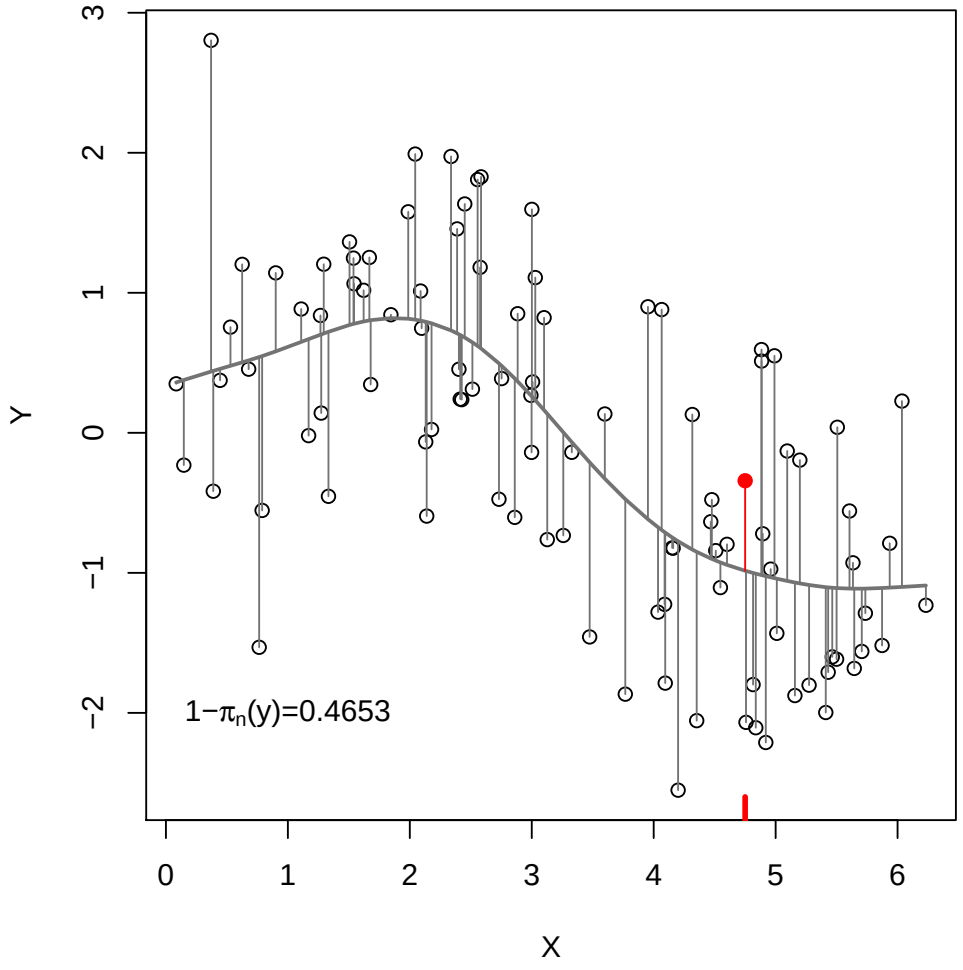
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



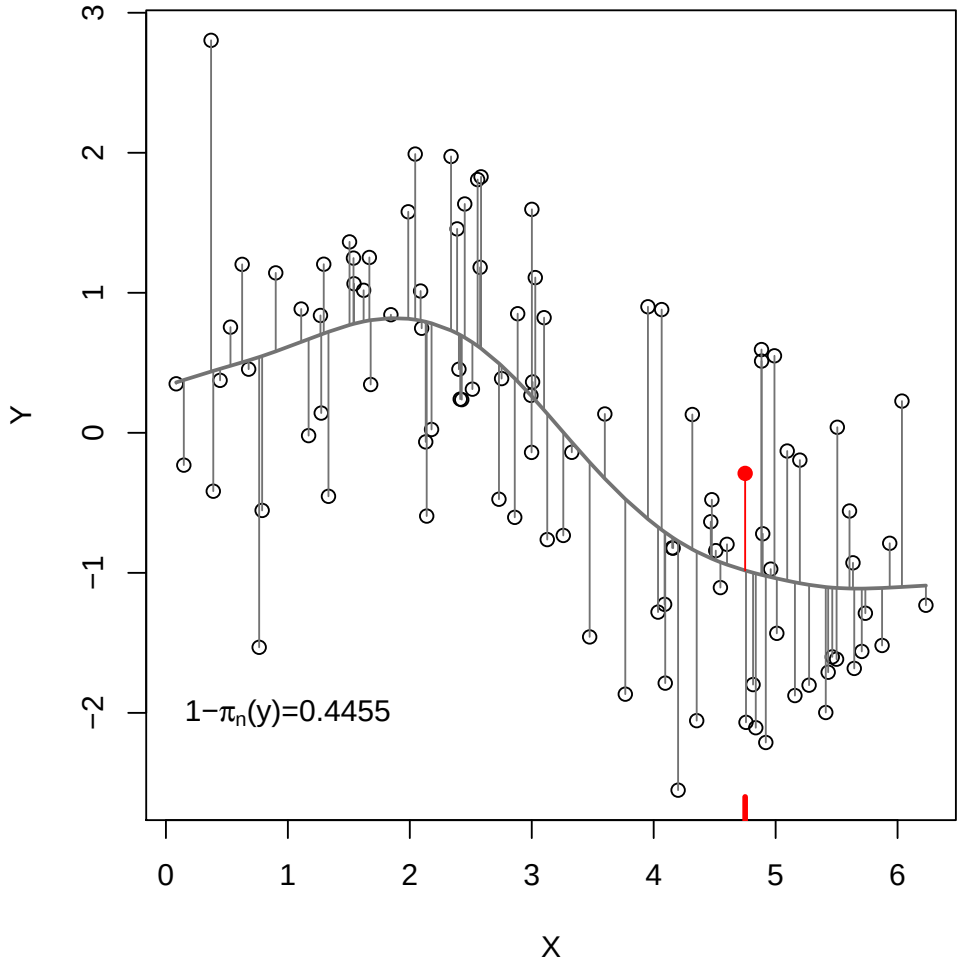
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



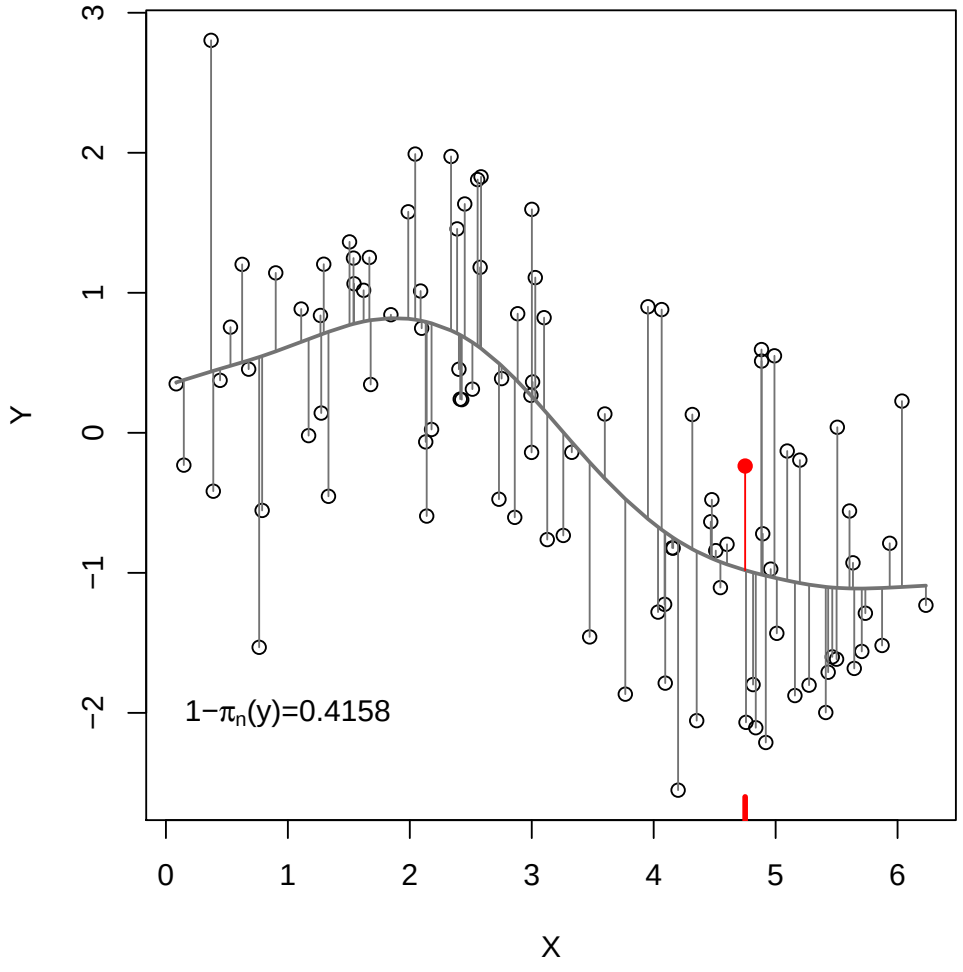
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



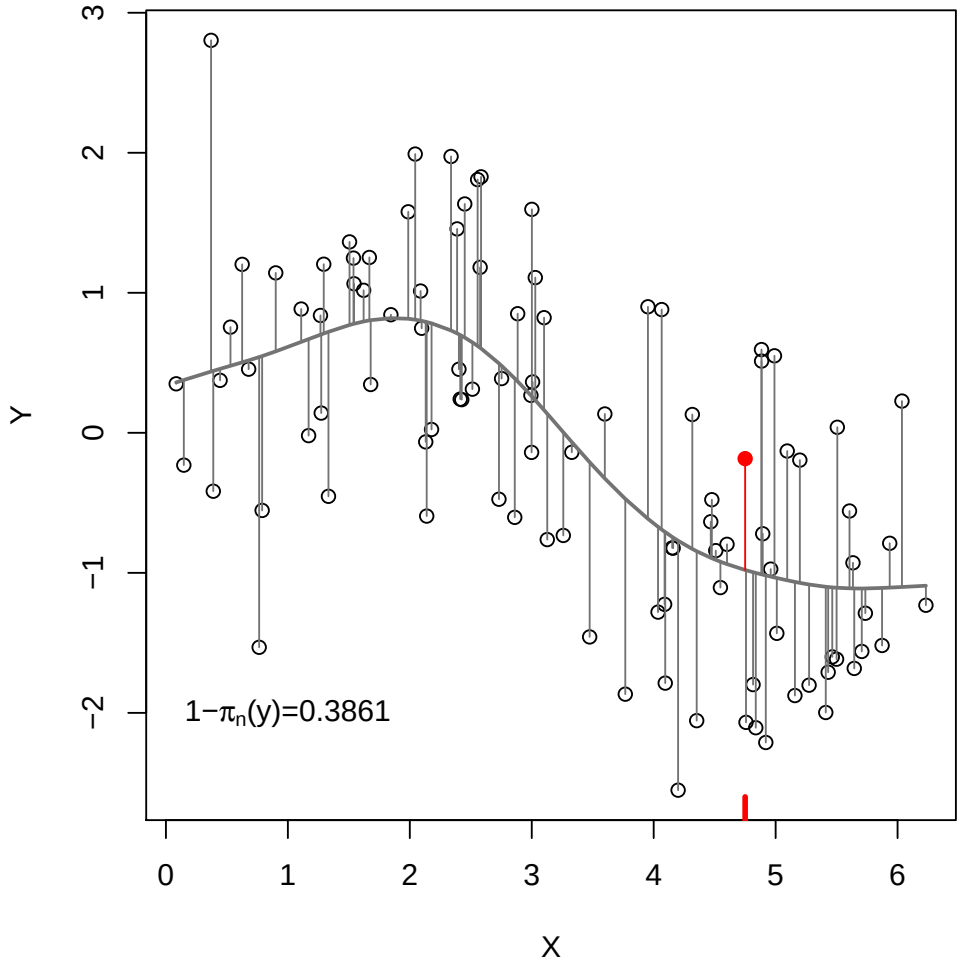
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



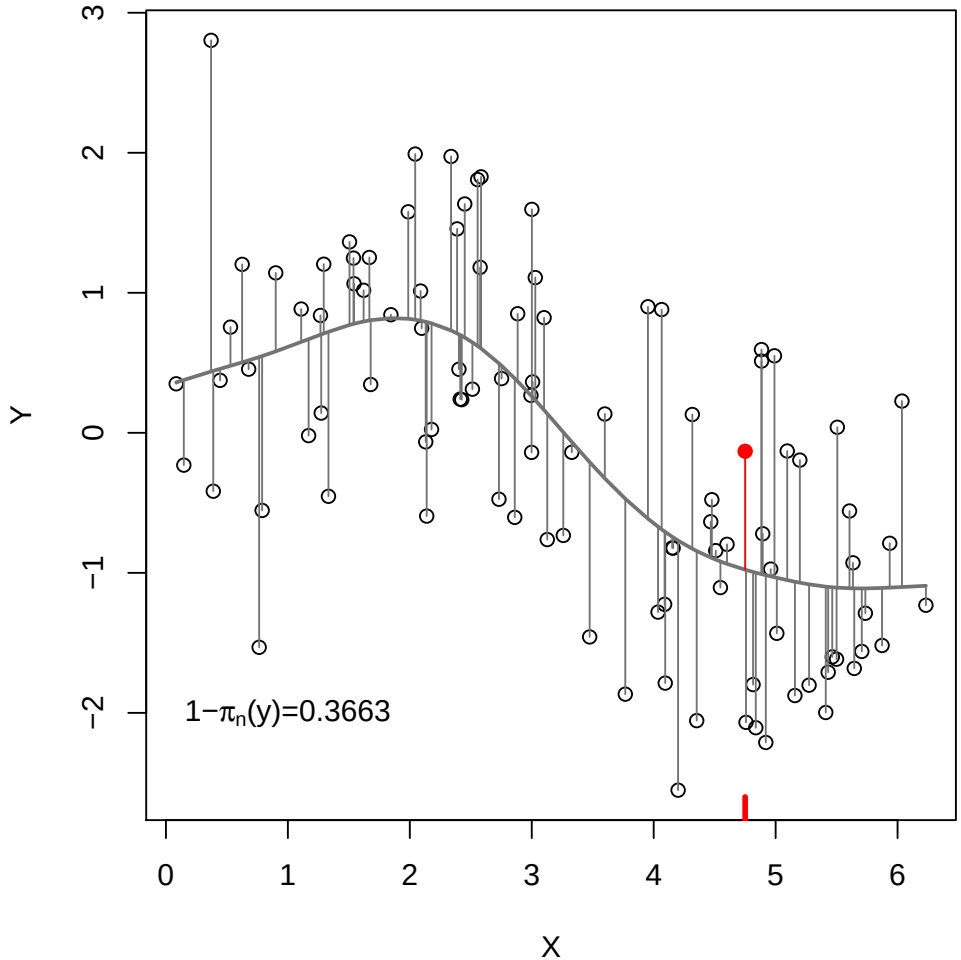
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



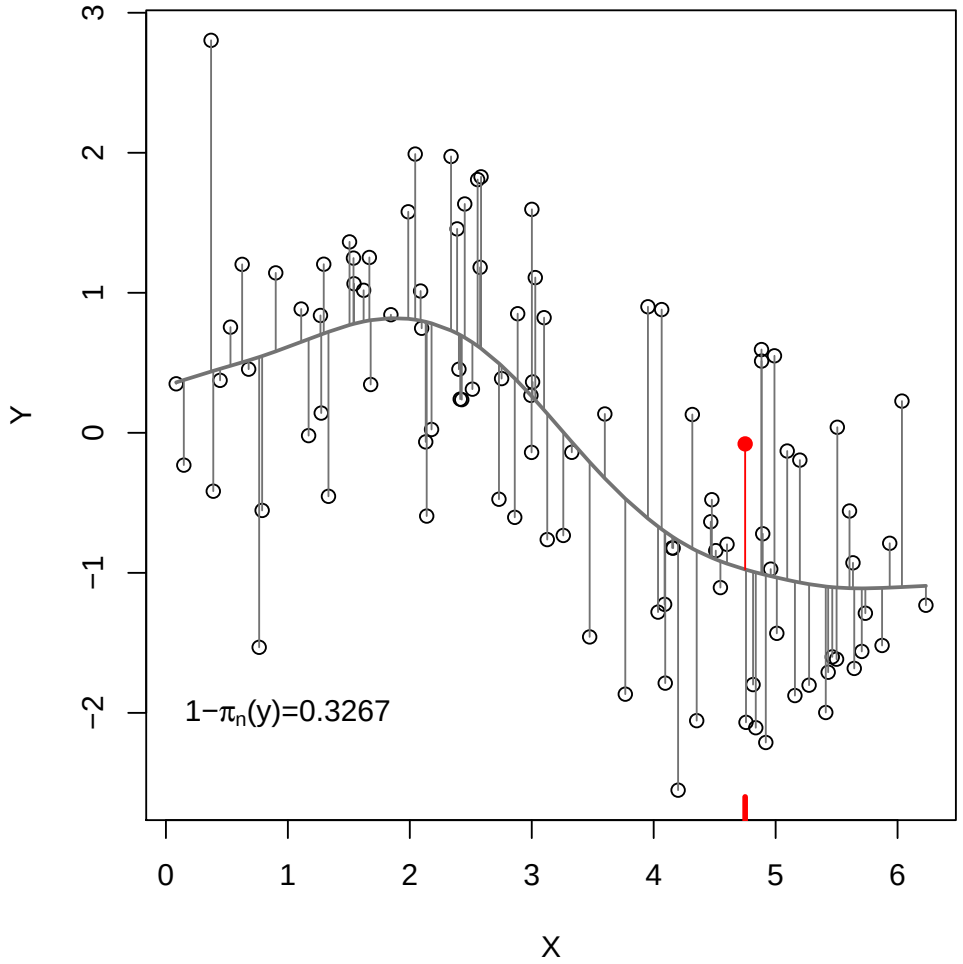
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



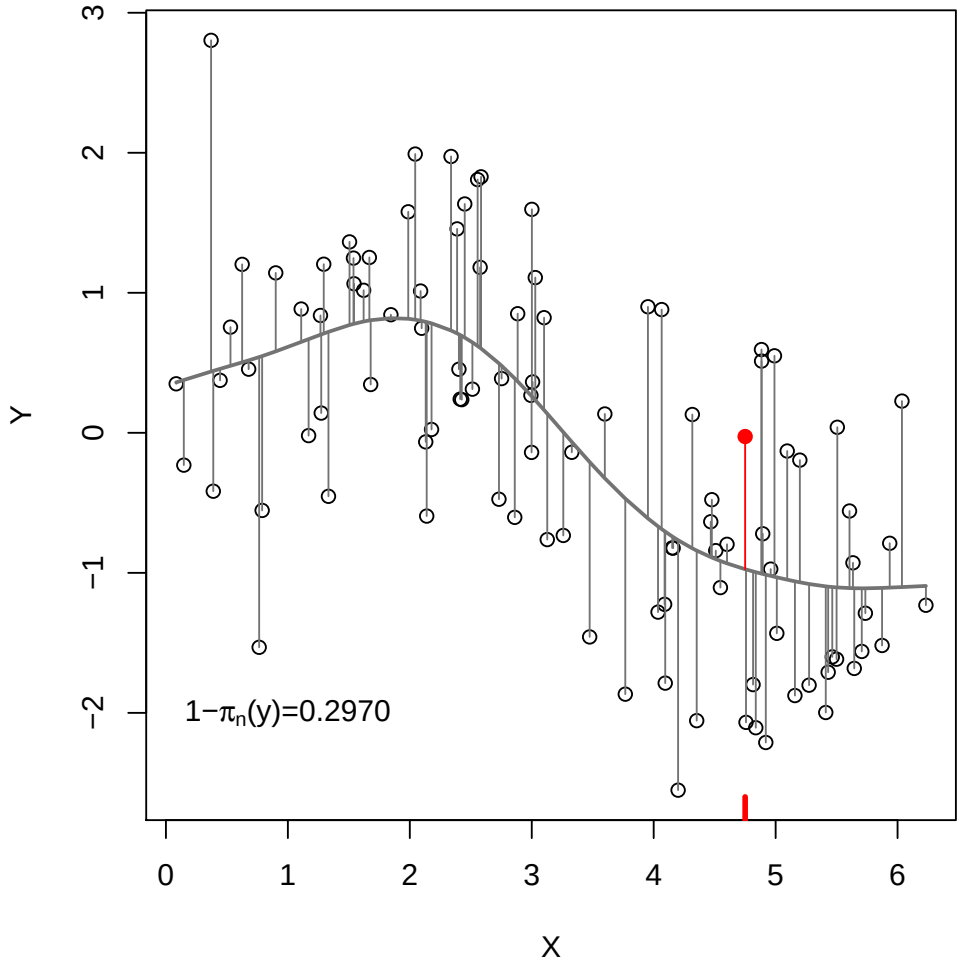
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



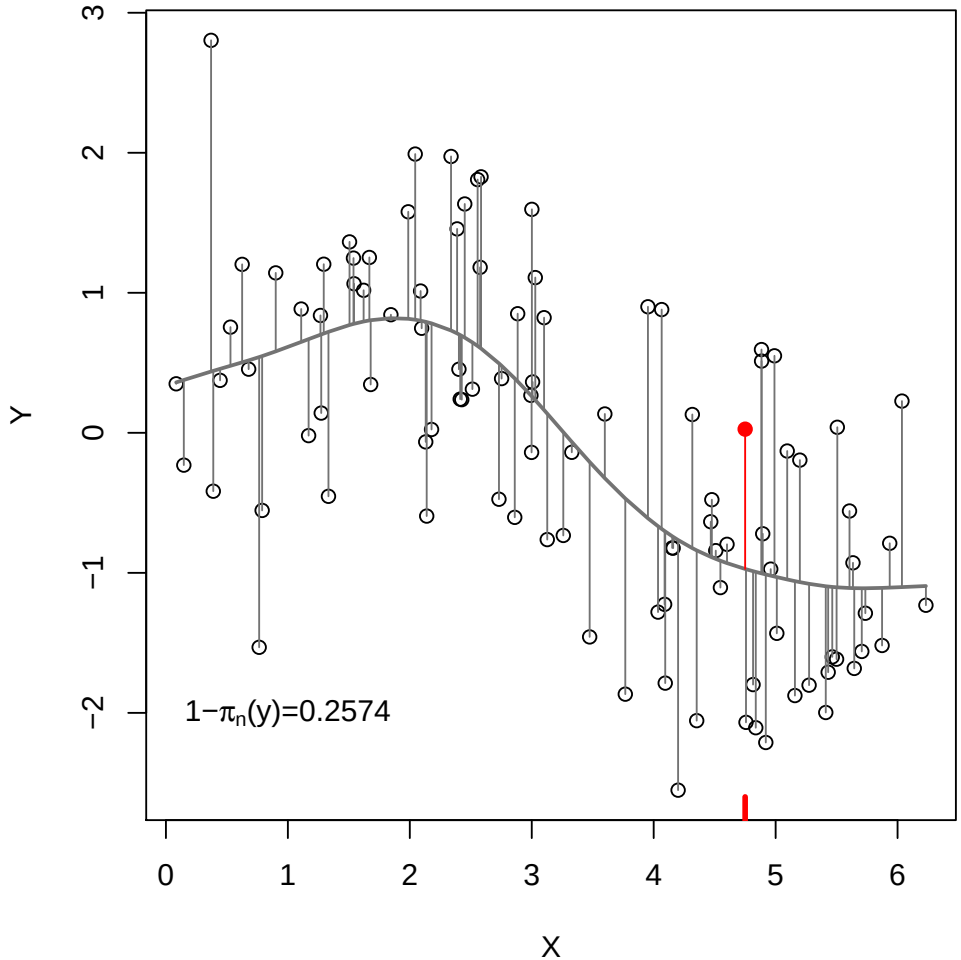
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



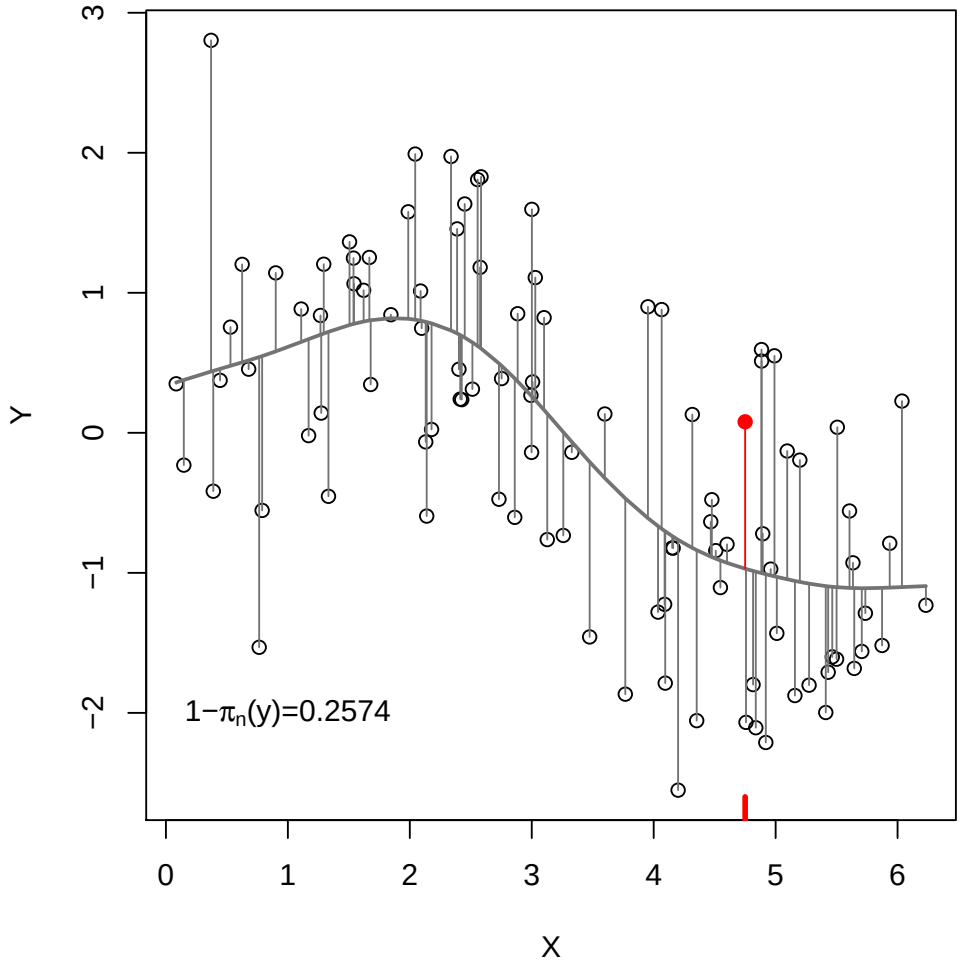
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



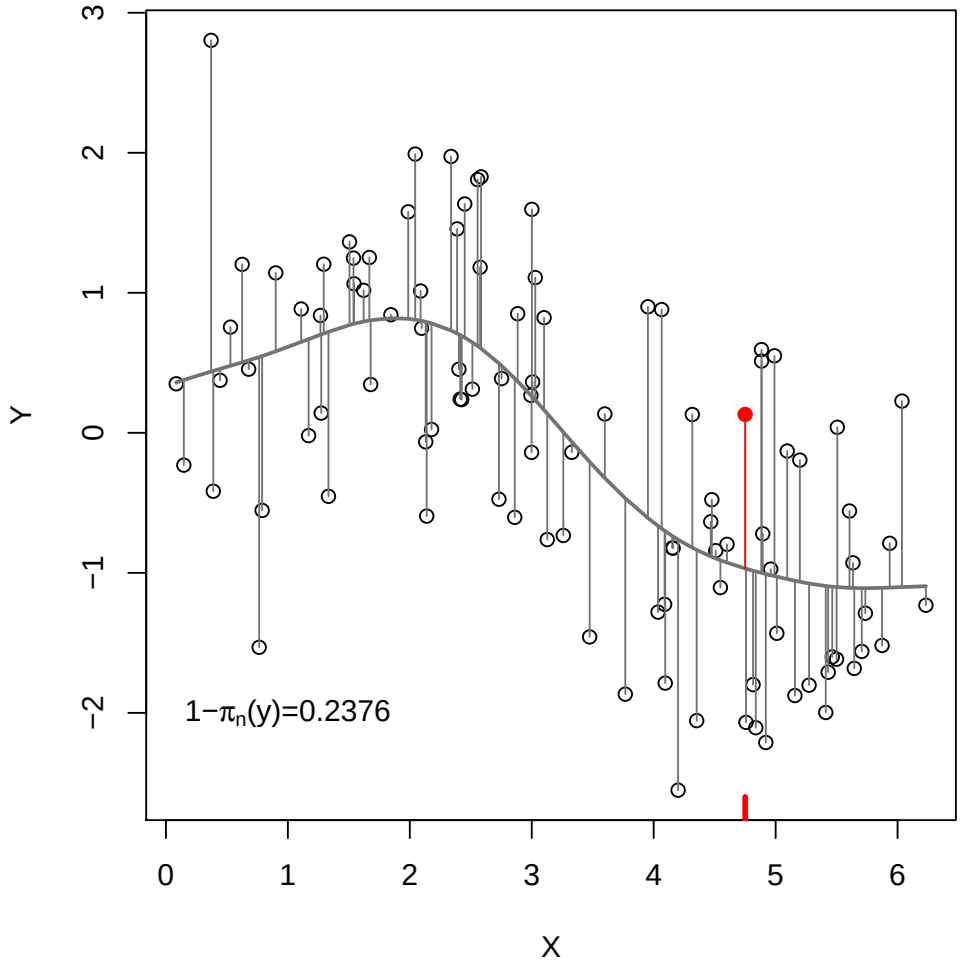
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



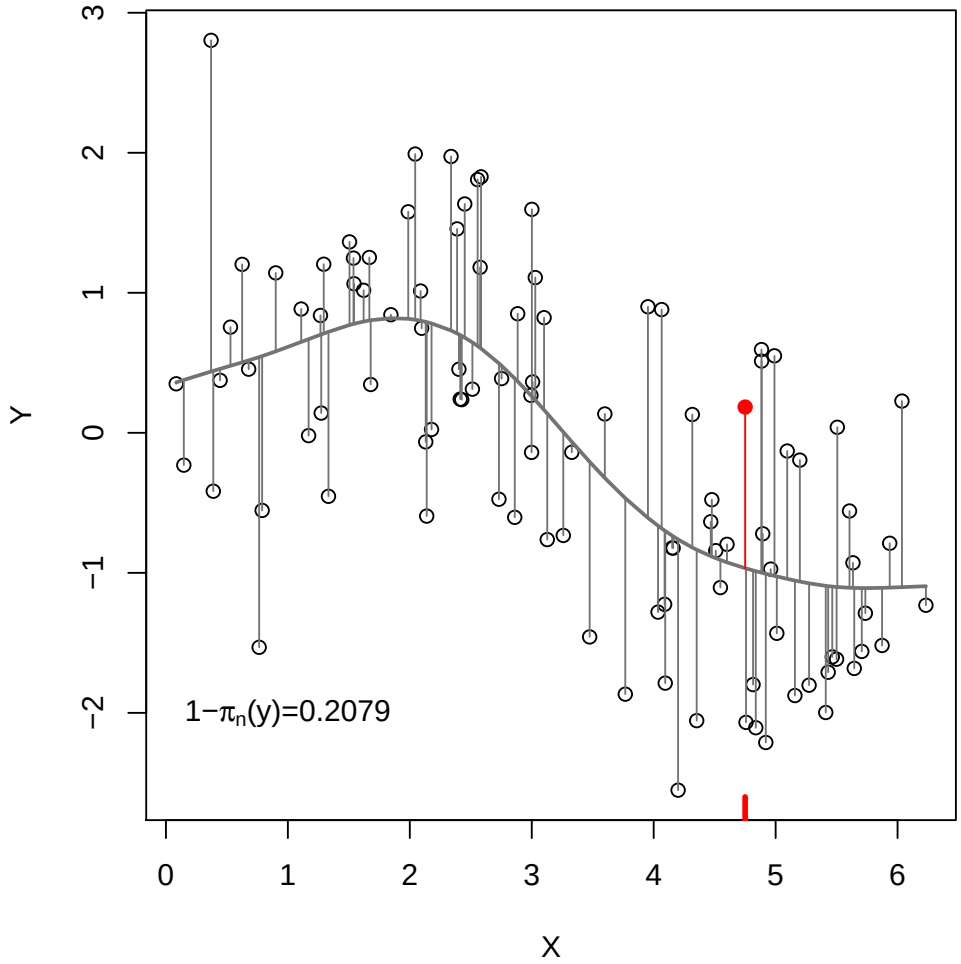
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



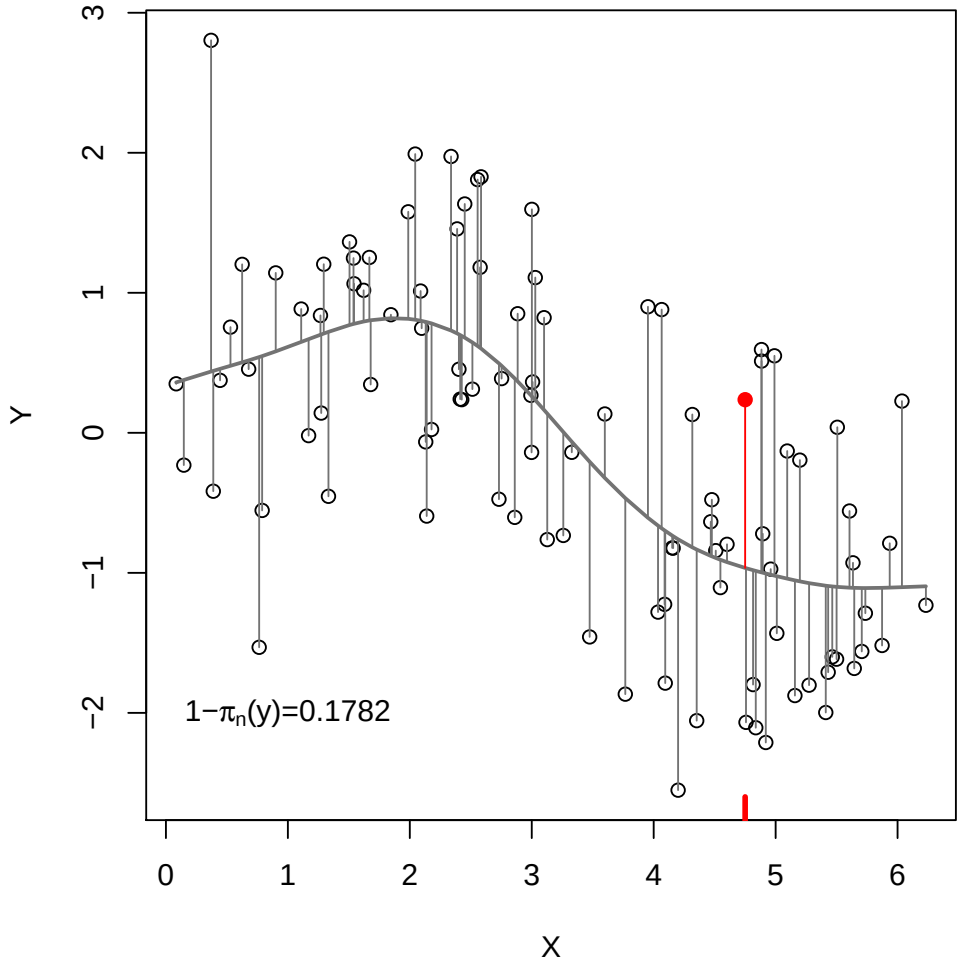
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



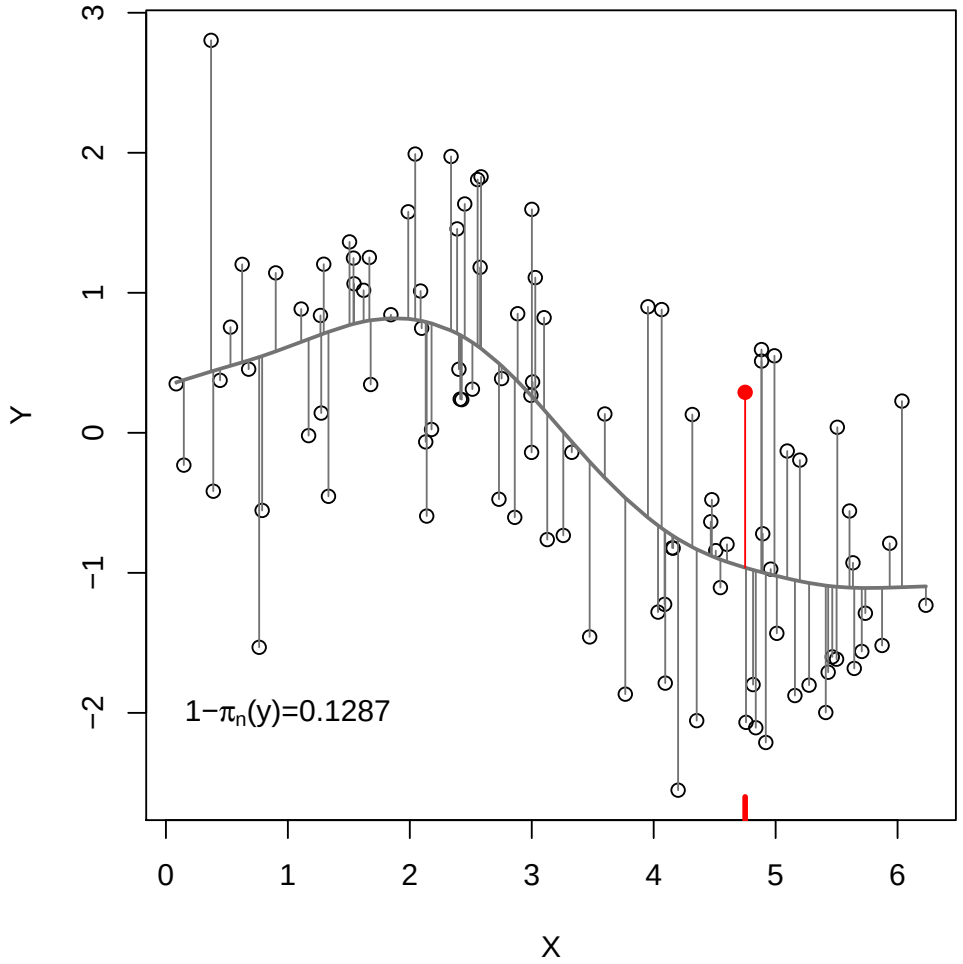
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



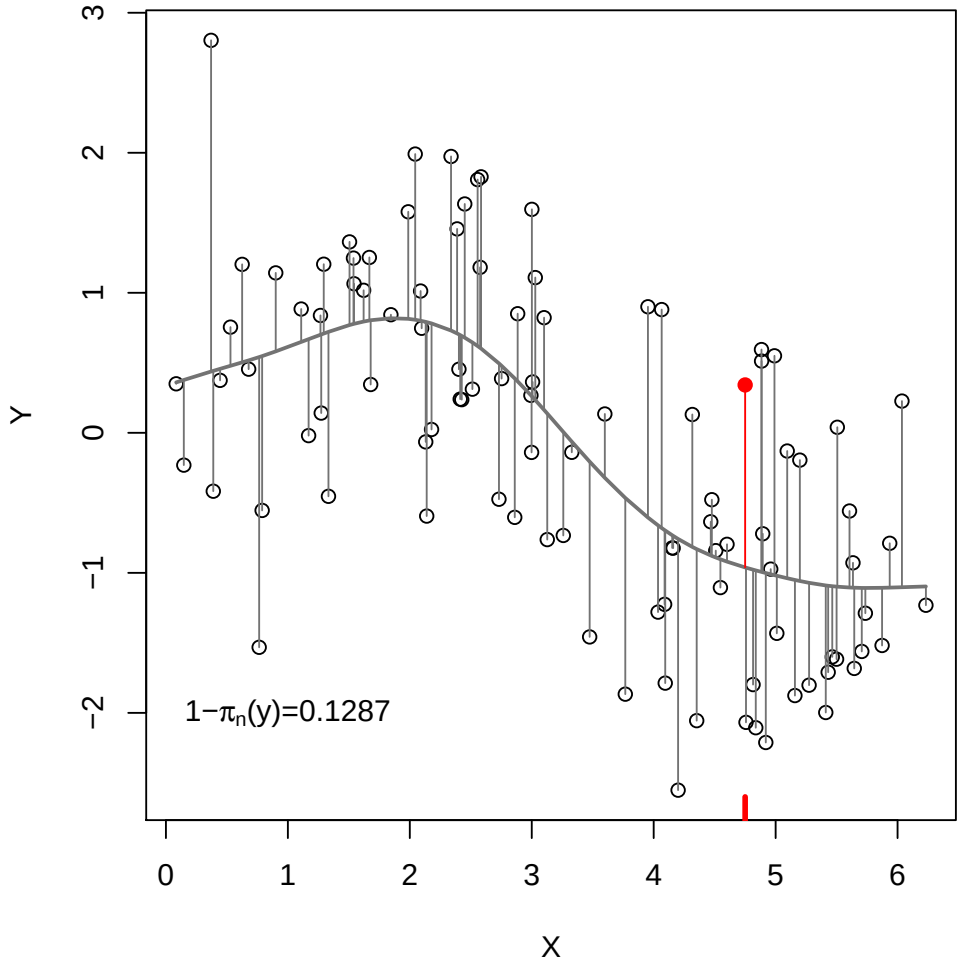
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



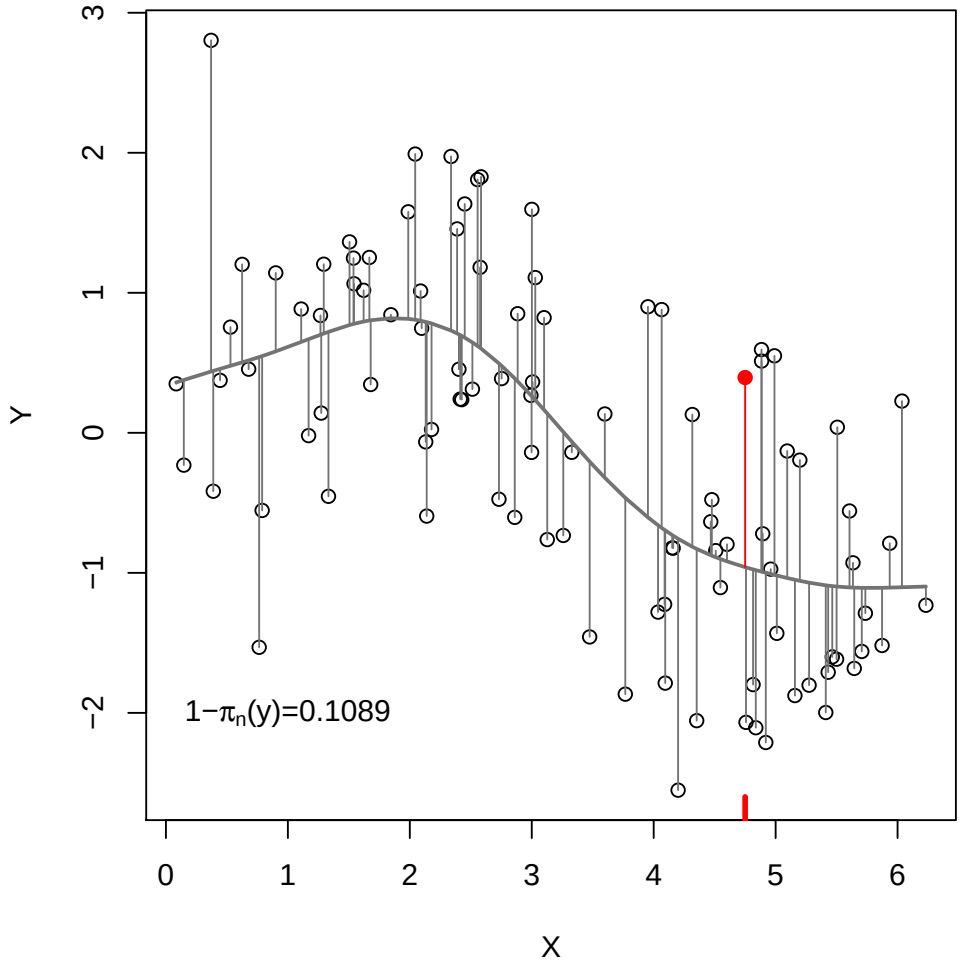
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



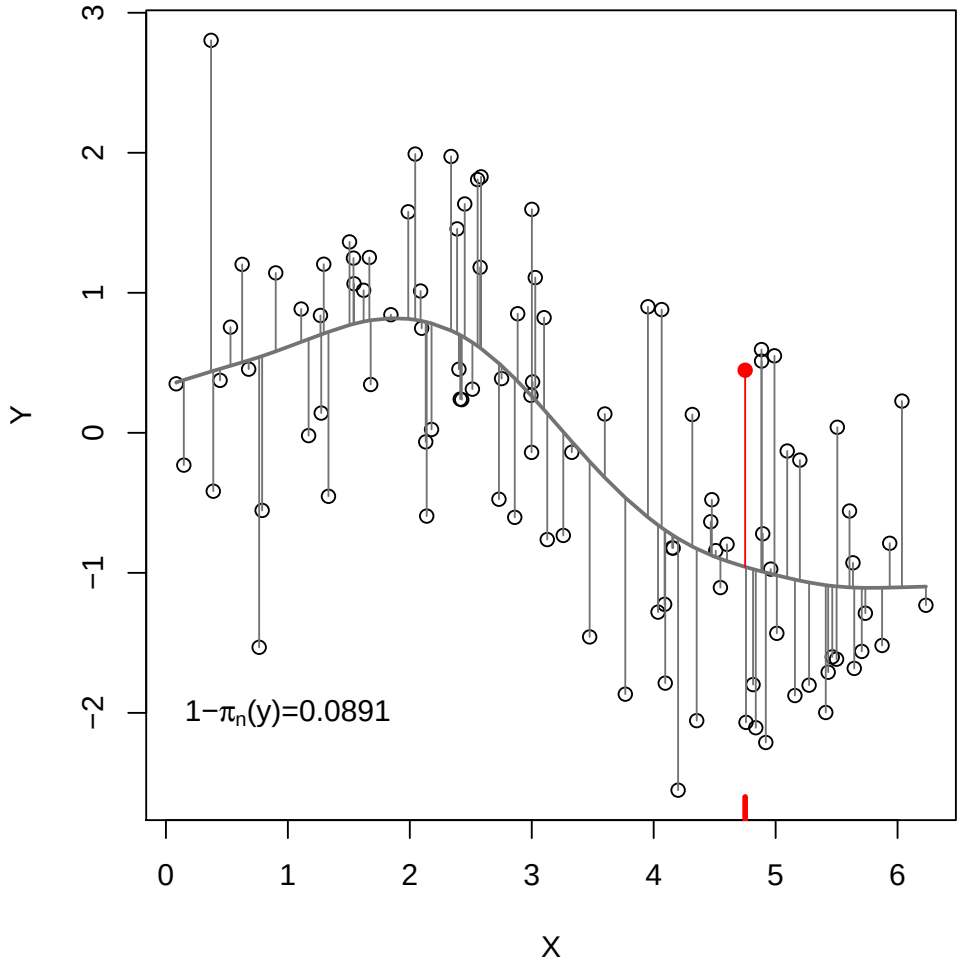
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



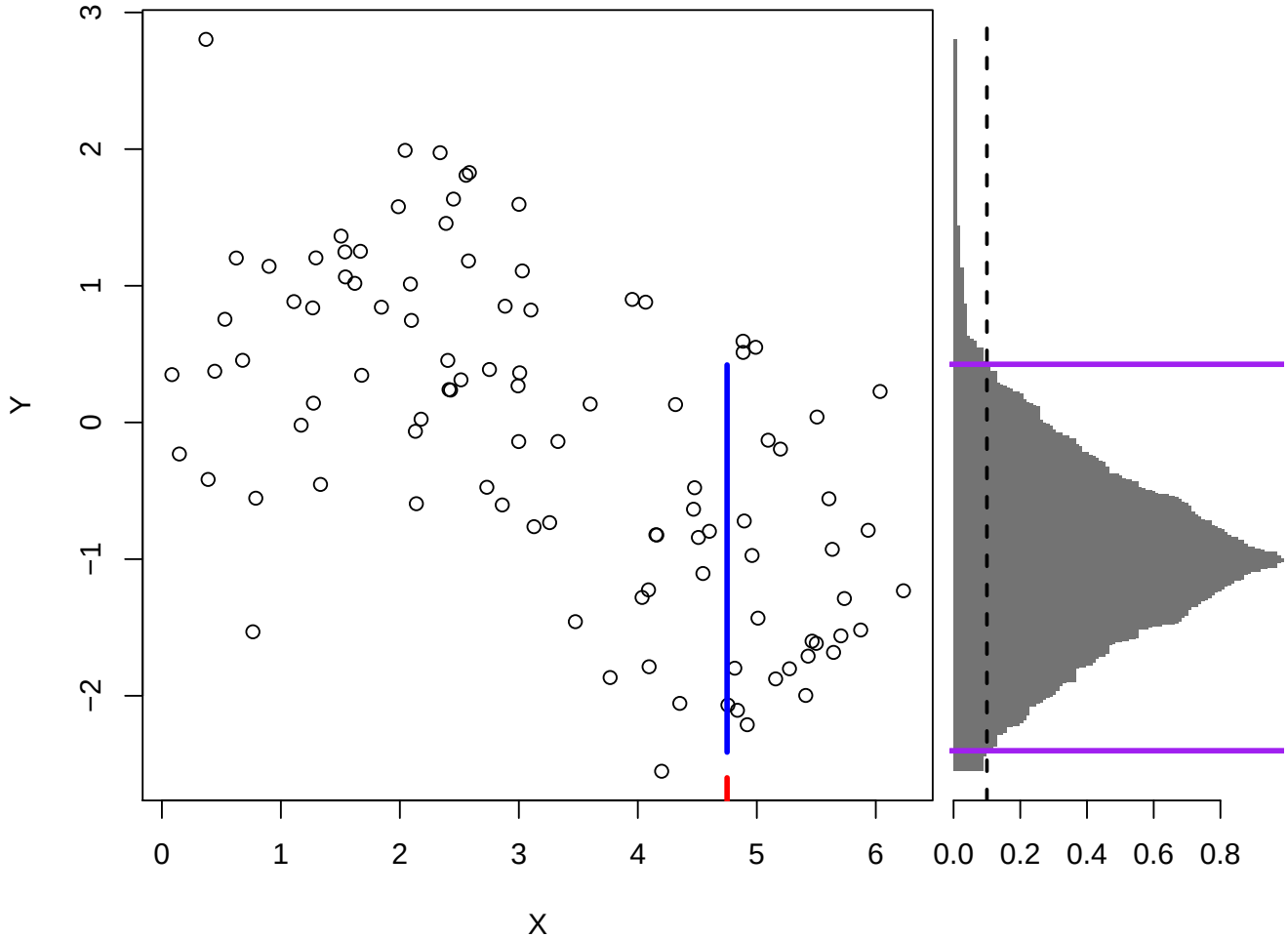
Suppose we want a prediction interval at $X_{n+1} = 4.75$, $\alpha = 0.1$

Example: conformal prediction interval using smoothing splines



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

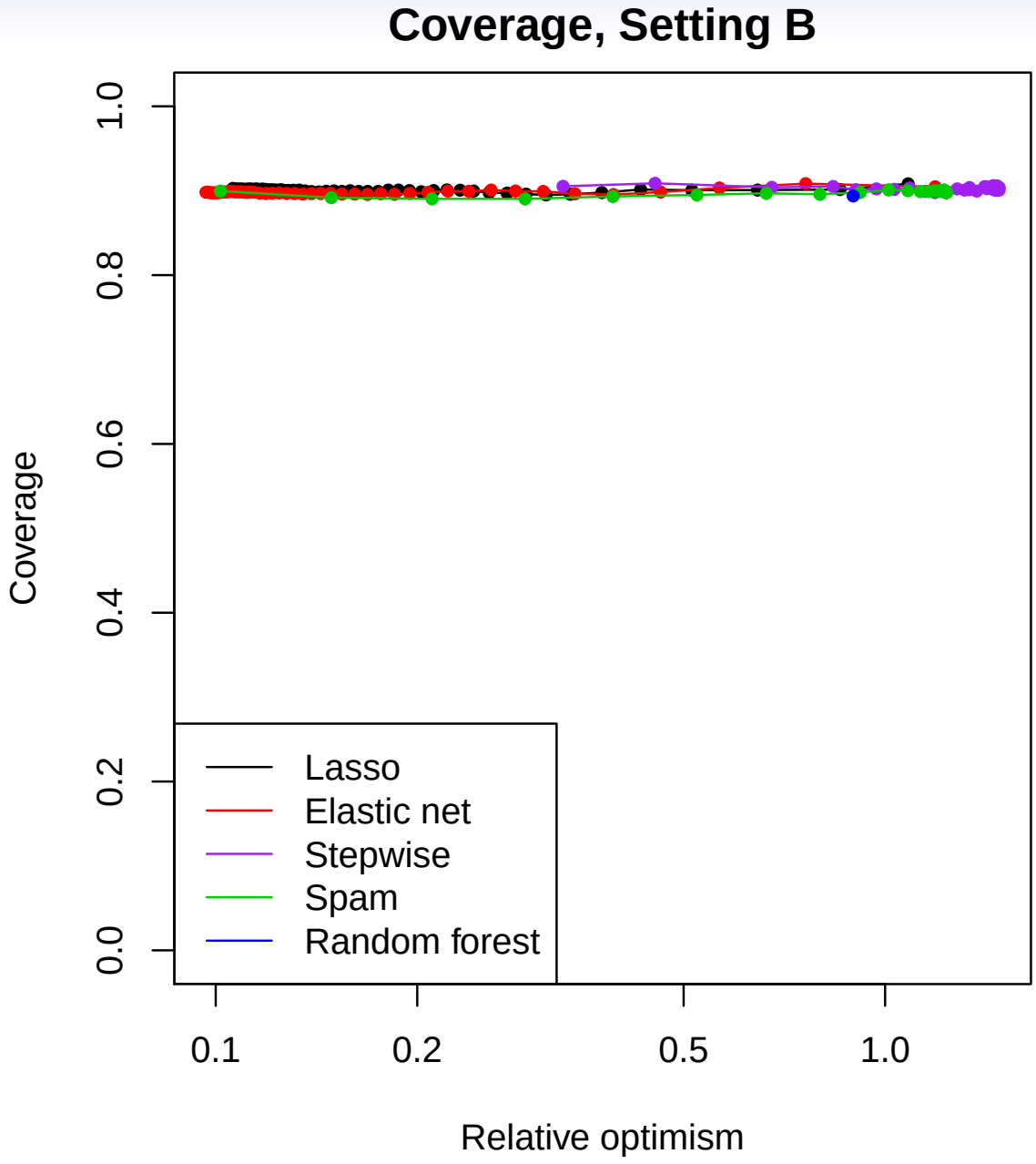
Example: conformal prediction interval using smoothing splines

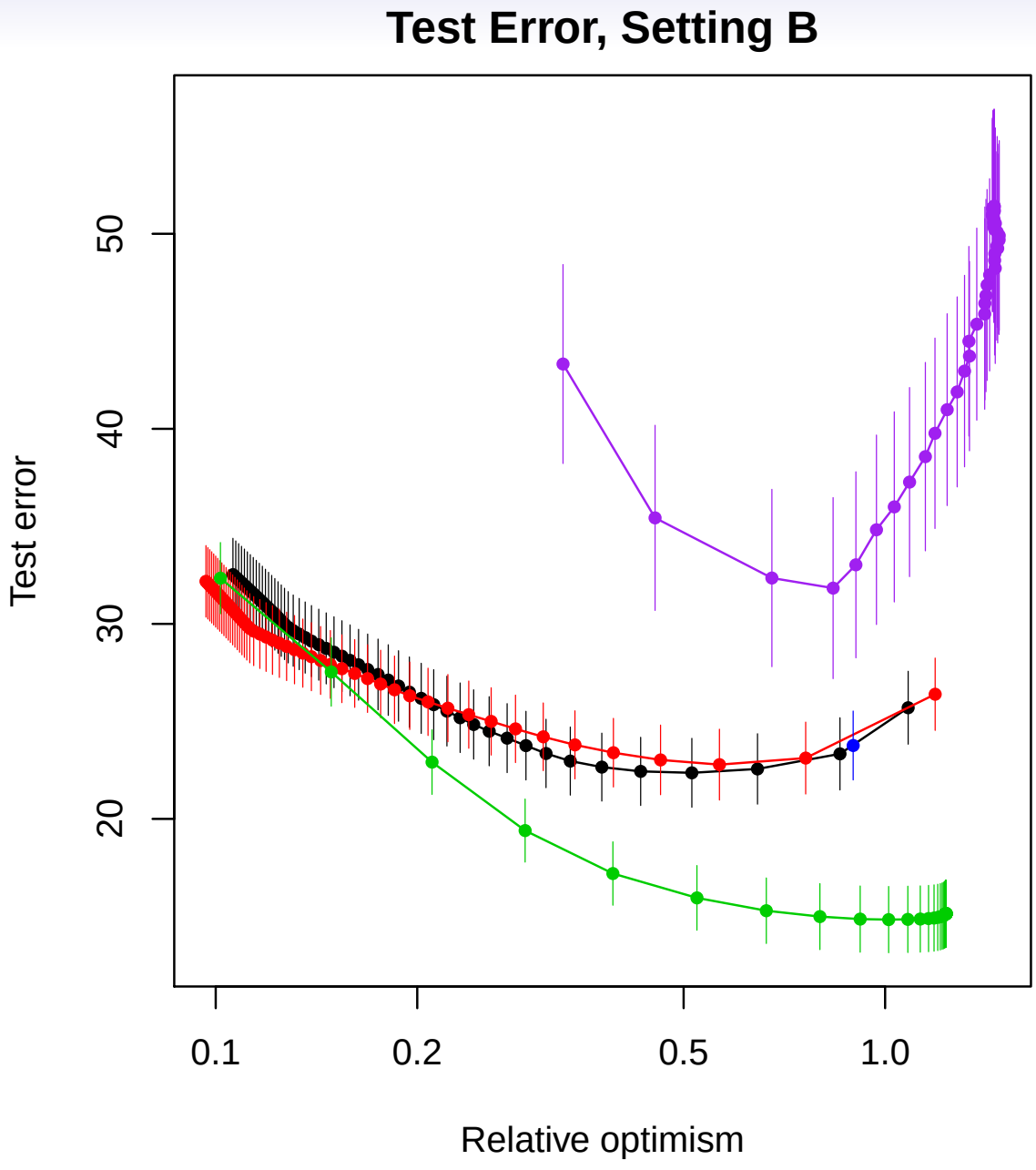


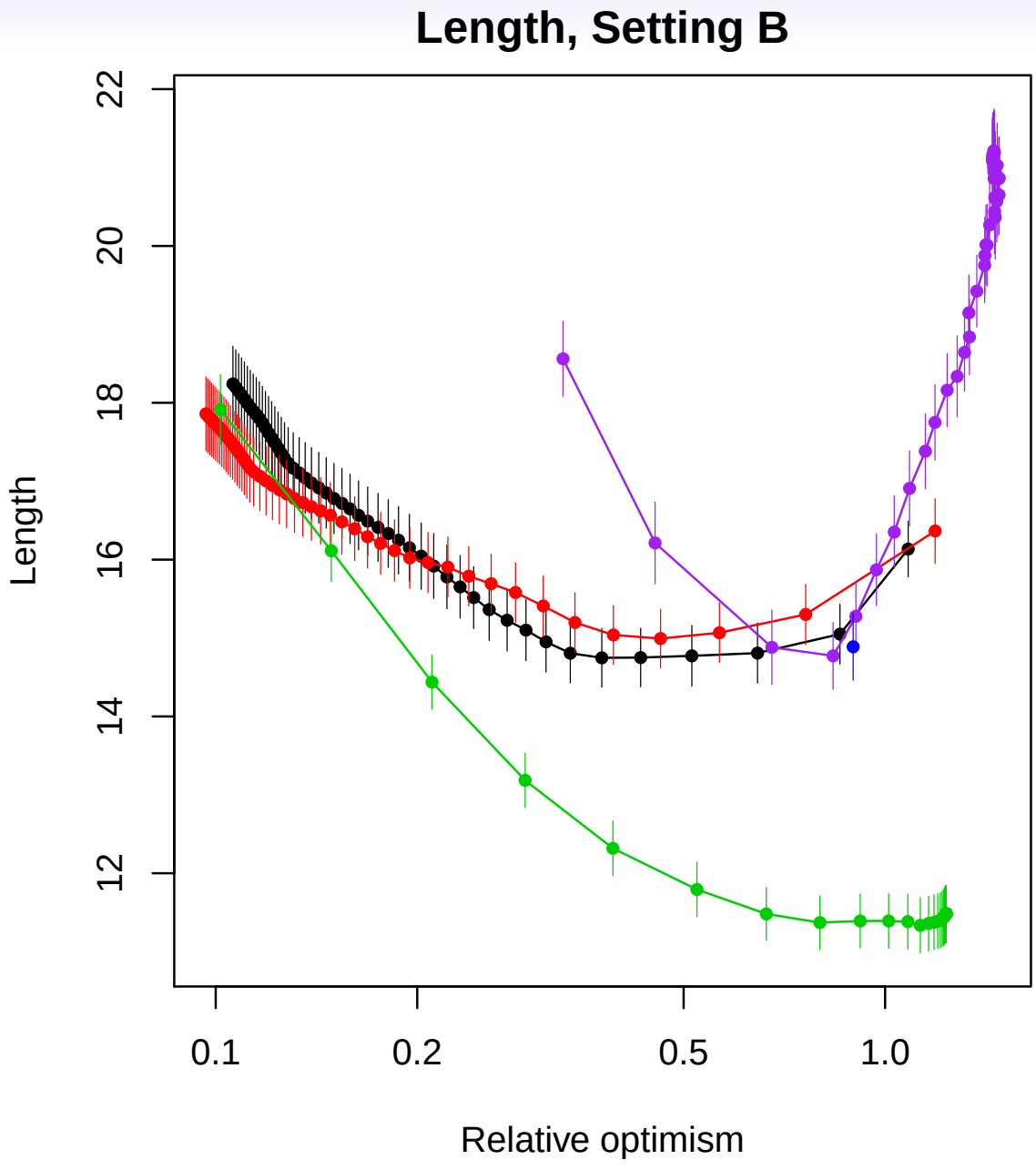
Invert p-values to get conformal interval

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

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.

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`".

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).
- A general solution has been developed in [Lei, Rinaldo, & Wasserman 15], by detaching the fitting and ranking steps.

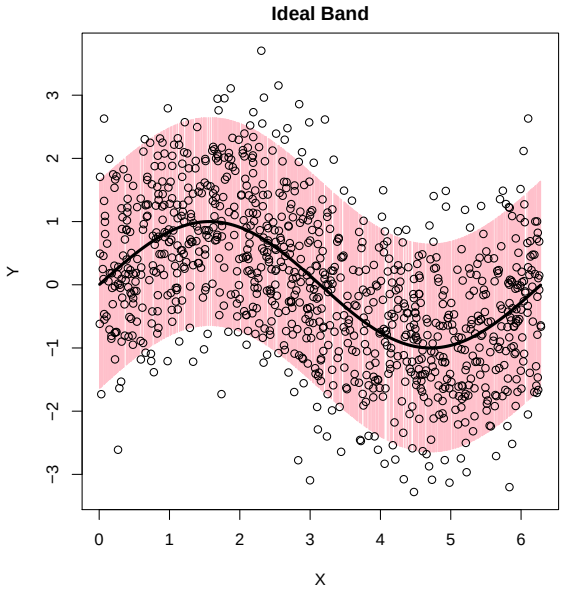
Split Conformal (Lei, Rinaldo, Wasserman 15)

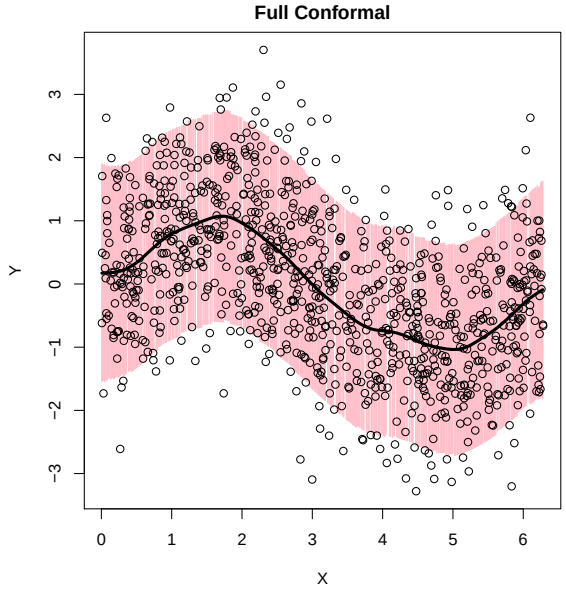
- Randomly split the data into two subsets, say, D_1 and D_2.
- Fit $\hat{\mu}$ on D_1.
- Let $\hat{F}$ be the empirical CDF of $\{|Y_i - \hat{\mu}(X_i)| : (X_i, Y_i) \in D_2\}$.
- Output $\tilde{C}(X_{n+1})$

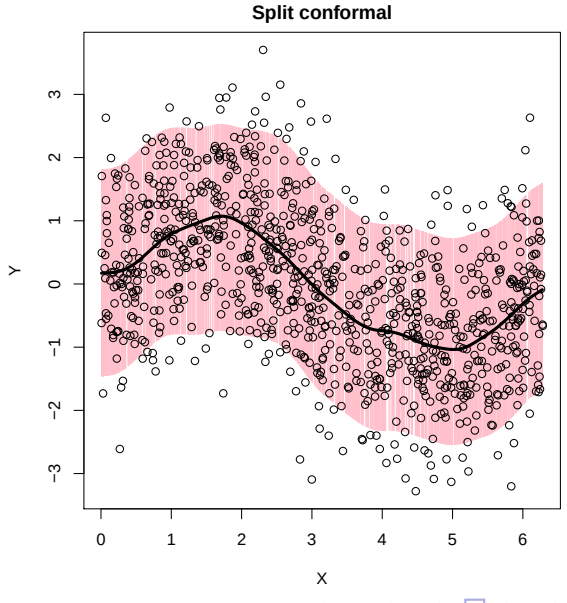
$$\tilde{C}(X_{n+1}) = [\hat{\mu}(X_{n+1}) \pm \hat{F}^{-1}(1-\alpha)]$$

- 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

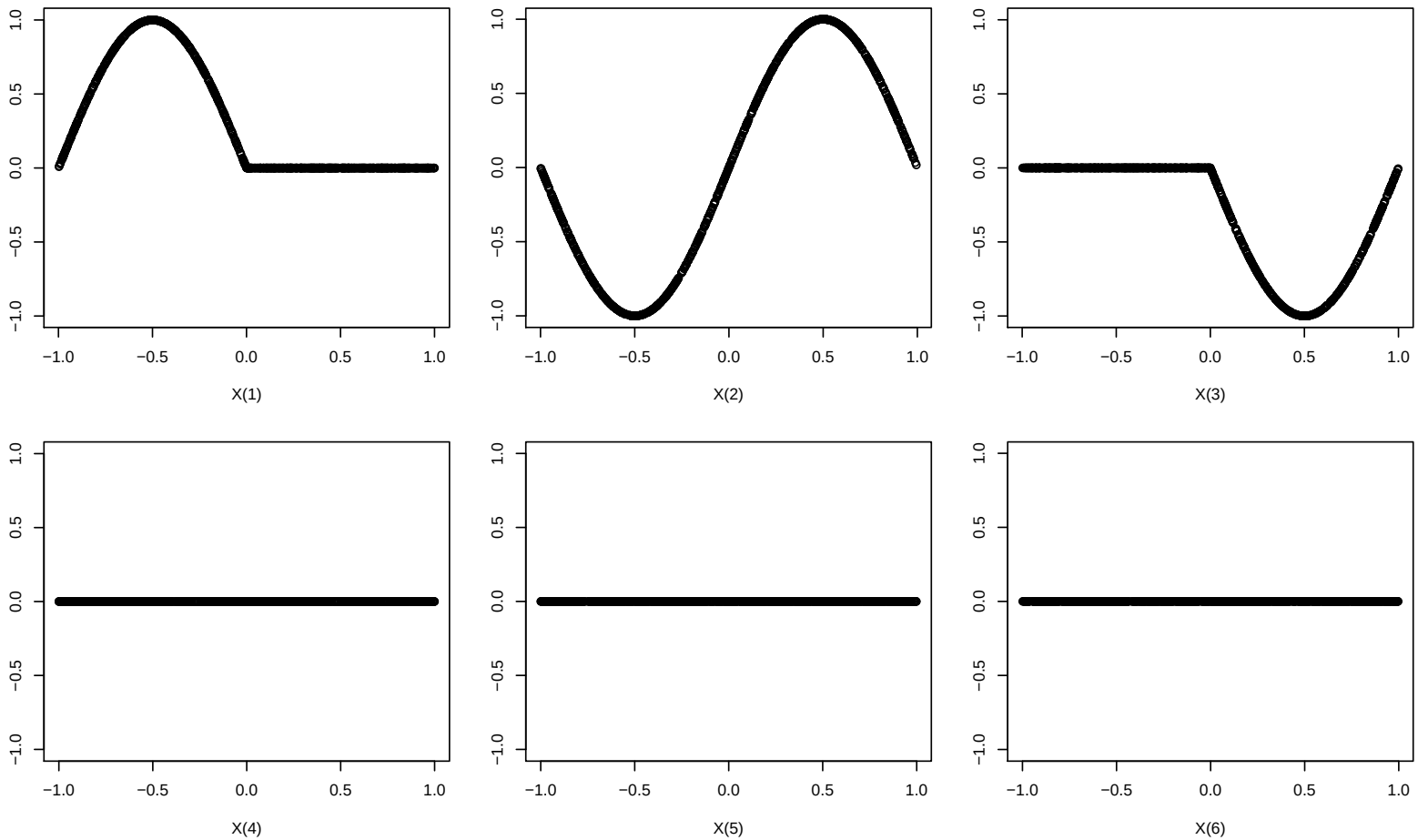
$$D_{ij} = \{|y - \hat{\mu}_{-j}(X_i)| - |y - \hat{\mu}(X_i)| : y \in \tilde{C}(X_i)\}$$

- 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

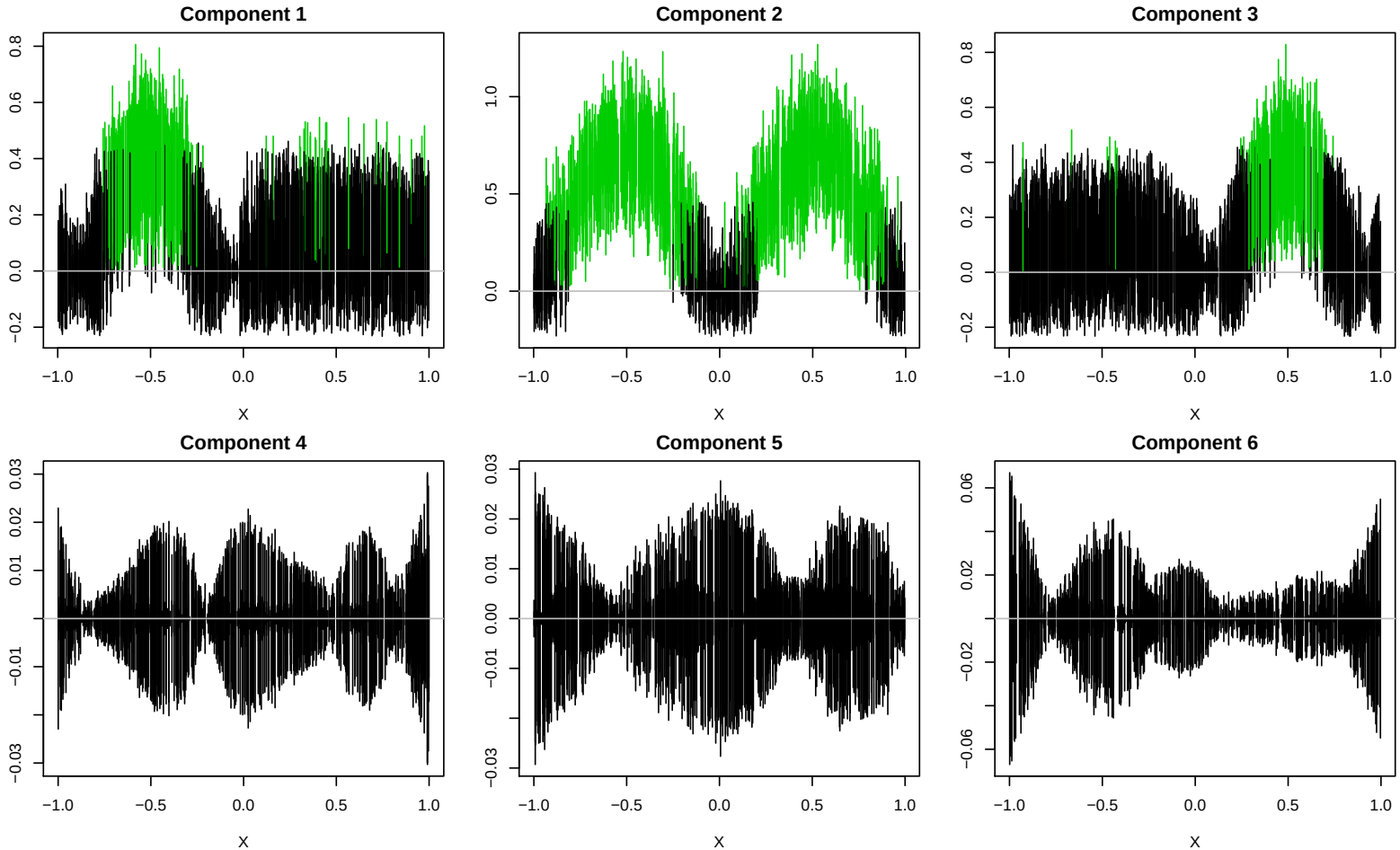
$$\mathbb{P}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}(V_{ij} \in D_{ij},\ \forall\, j) \geq 1 - \alpha - \varepsilon\right] \geq 1 - 2e^{-cn\varepsilon^2}$$

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.

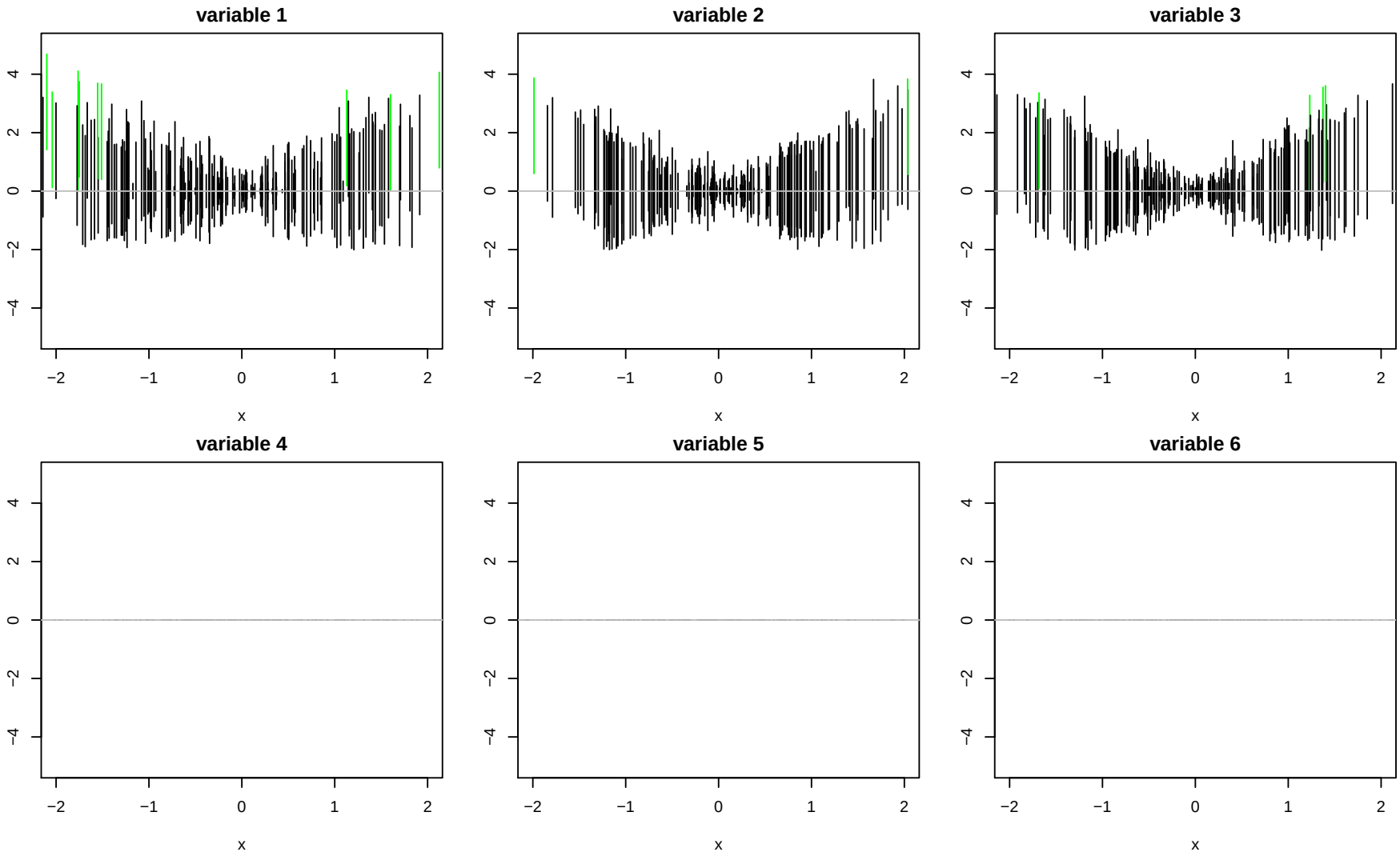
A higher dimensional example

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

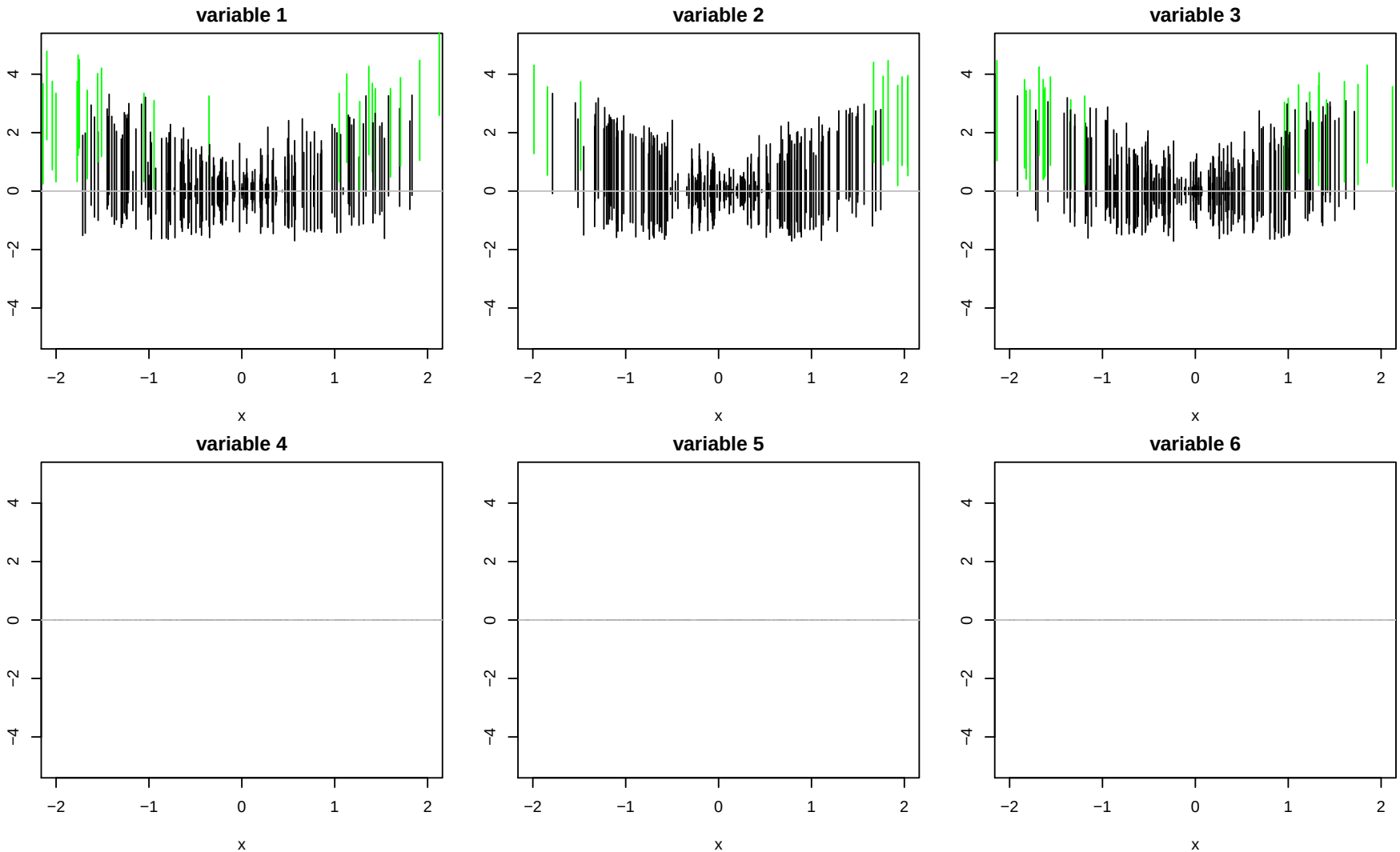
  Case 1: $\mathbb{E}(XX^T) = I$ (all standard assumptions hold)

  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

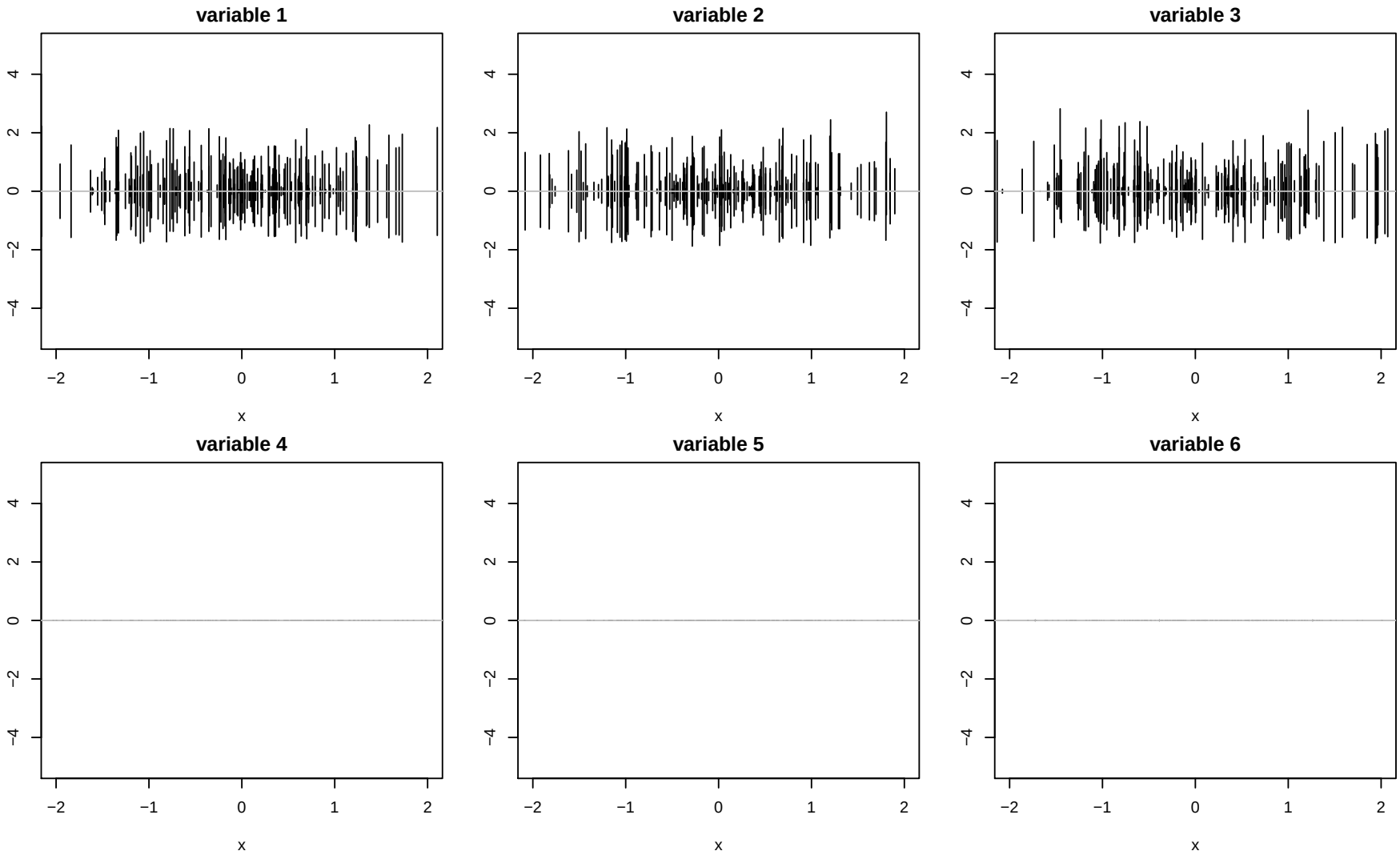
Uncorrelated case, Lasso



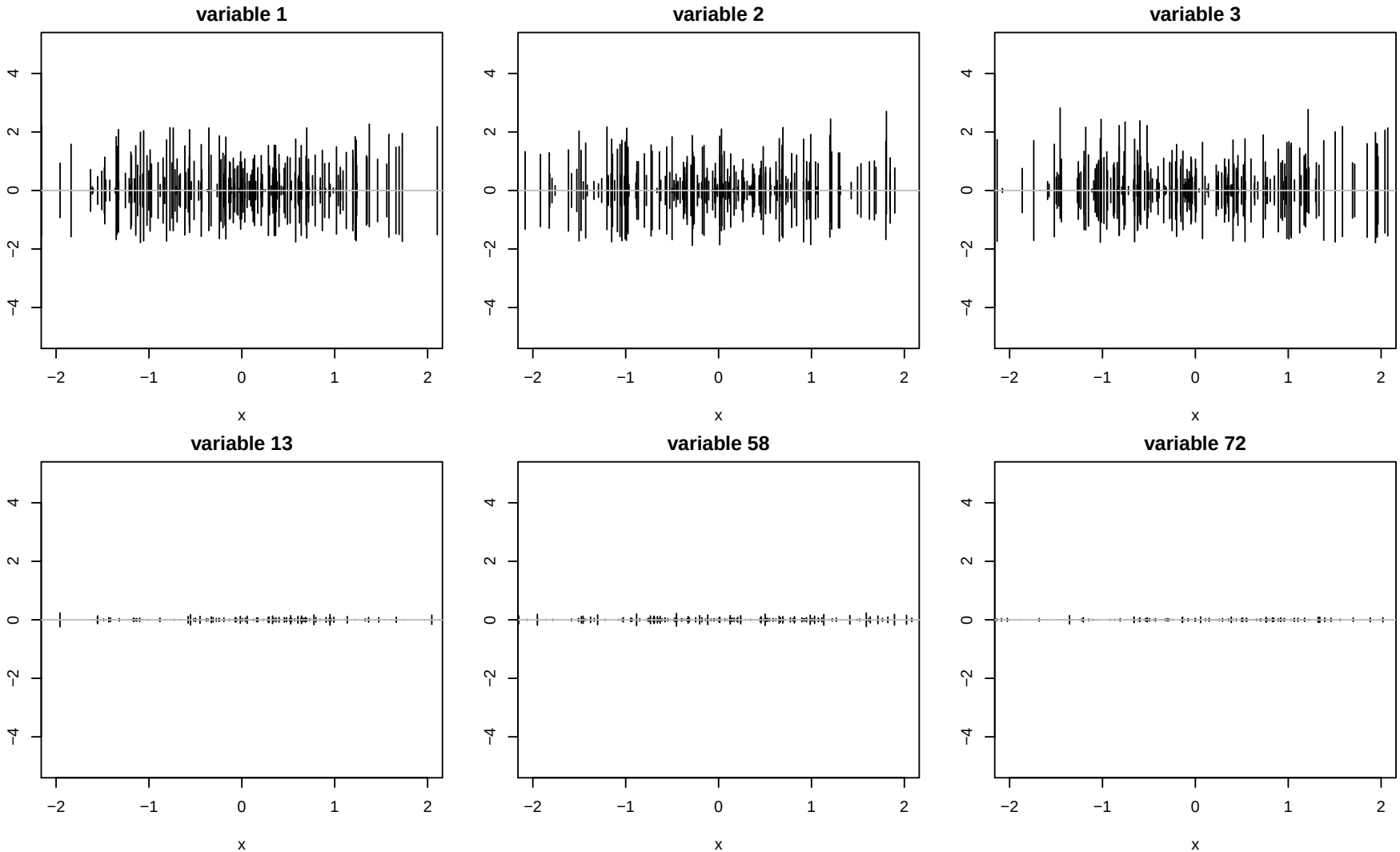
Uncorrelated case, Forward Stepwise



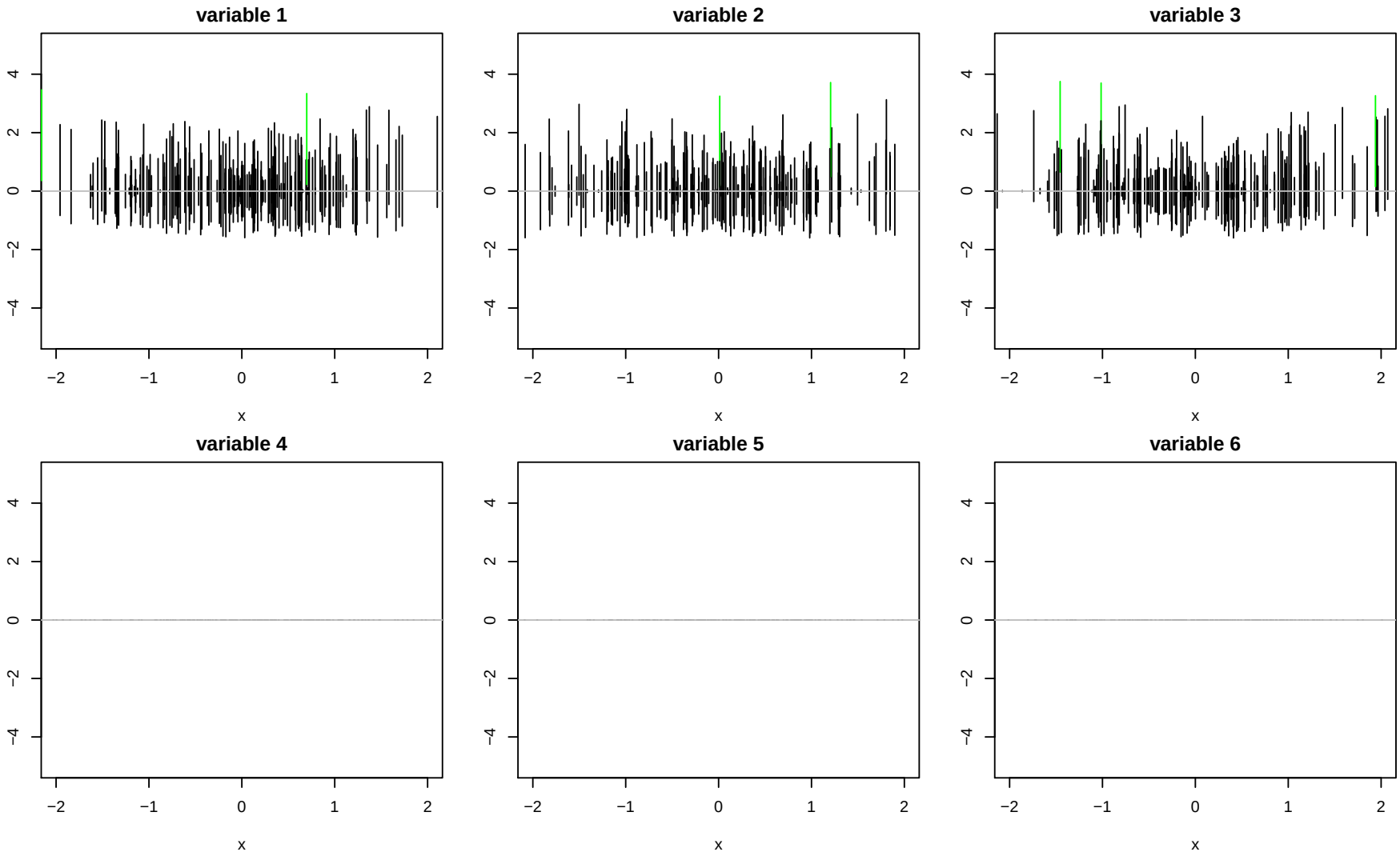
Correlated case, Lasso



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?
  - Valid in-sample prediction: can we provide $\hat{C}(X_i)$ for X_i's in the sample?
  - Higher order correction: can we produce prediction band with adaptive width?

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

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

Conclusion

- Conformalization: make fitting out-of-sample.
- More results, ongoing and future work
  1. Conformal prediction is also anti-conservative.
  2. Rigorous, more interpretable variable importance measure.
  3. Necessity and best practice of sample splitting.
  4. Other problems: clustering, classification, dimension reduction, time series, mixed effects model, etc.
  5. Combine with Bayesian methods: “best of both worlds”.

Thanks!

Questions?

Paper: "Distribution Free Predictive Inference for Regression"

arXiv:1604.04173

Slides:

http://www.stat.cmu.edu/~jinglei/conformal_pitt.pdf

Package: https://github.com/ryantibs/conformal