Stability Selection

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Meinshausen, Buhlmann (2010). Stability selection. JRSS-B, 72:417–473
- Shah, Samworth (2013). Variable selection with error control: another look at stability selection. JRSS-B, 75:55–80.

Stability path

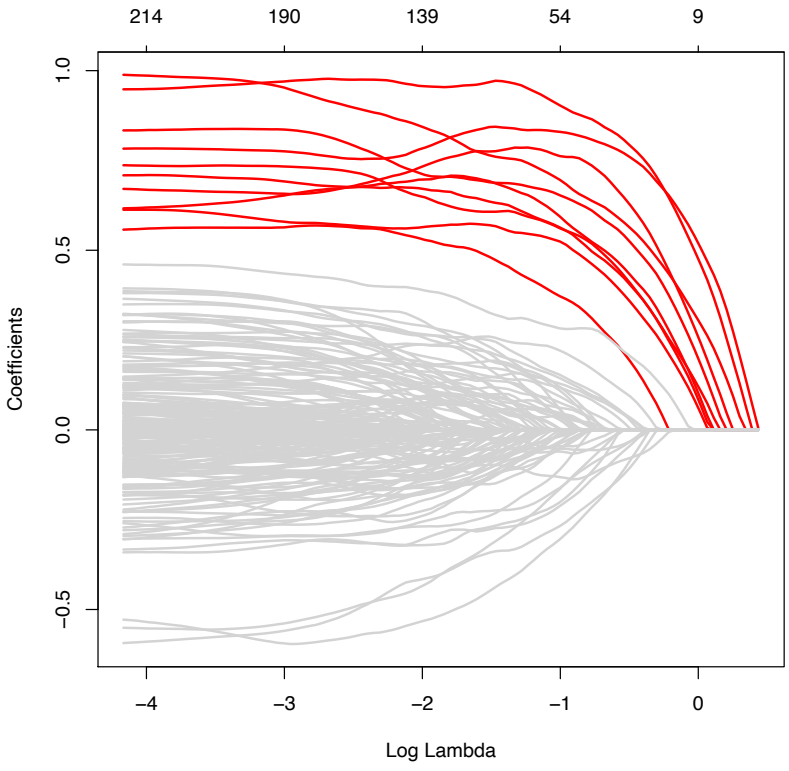
- The *regularisation path* of the lasso is

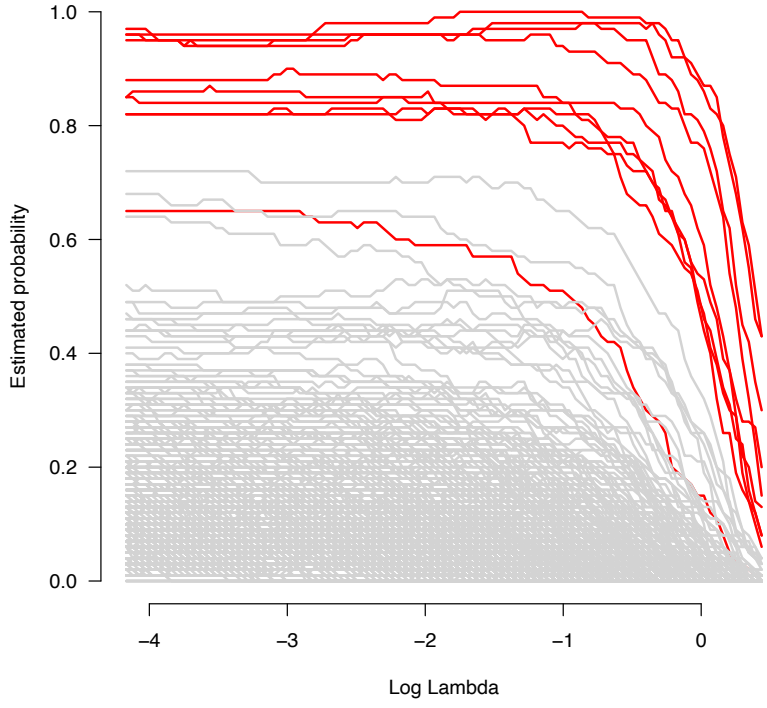
$$\{\hat{\beta}_j(\lambda), j = 1, \ldots, p, \lambda \in \Lambda\}$$

- The *stability path* is

$$\{\hat{\pi}_j(\lambda), j = 1, \ldots, p, \lambda \in \Lambda\}$$

  where $\hat{\pi}_j(\lambda)$ is the estimated probability for the jth predictor to be selected by the lasso(λ) when randomly resampling from the data


214
190
139
54
9
1.0
0.5
0.0
−0.5
Coefficients
−4
−3
−2
−1
0
Log Lambda


1.0
0.8
0.6
0.4
0.2
0.0
Estimated probability
−4
−3
−2
−1
0
Log Lambda

Algorithm 1 Stability Path Algorithm with the Lasso

Require: $B \in \mathbb{N}$, Λ grid, $\tau \in (0.5, 1)$

1: **for** $b = 1, \ldots, B$ **do**

2: Randomly select $n/2$ indices from $\{1, \ldots, n\}$;

3: Perform the lasso on the $n/2$ observations to obtain

$$\hat{S}_{n/2}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\} \qquad \forall\ \lambda \in \Lambda$$

4: **end for**

5: Compute the relative selection frequencies:

$$\hat{\pi}_j(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{j \in \hat{S}_{n/2}(\lambda)\} \qquad \forall\ \lambda \in \Lambda$$

6: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \max_{\lambda \in \Lambda} \hat{\pi}_j(\lambda) \geq \tau\}$$

Algorithm 2 (Complementary Pairs) Stability Selection

Require: A variable selection procedure $\hat{S}_n$, $B \in \mathbb{N}$, $\tau \in (0.5, 1)$

1: **for** $b = 1, \ldots, B$ **do**

2: Split $\{1, \ldots, n\}$ into (I^{2b-1}, I^{2b}) of size $n/2$, and for each get

$$\hat{S}^{2b-1}_{n/2} \subseteq \{1, \ldots, p\}, \qquad \hat{S}^{2b}_{n/2} \subseteq \{1, \ldots, p\}$$

3: **end for**

4: Compute the relative selection frequencies:

$$\hat{\pi}_j = \frac{1}{2B} \sum_{b=1}^{B} \left(\mathbb{1}\{j \in \hat{S}^{2b-1}_{n/2}\} + \mathbb{1}\{j \in \hat{S}^{2b}_{n/2}\} \right)$$

5: The set of *stable predictors* is given by

$$\hat{S}_{\text{stab}} = \{j : \hat{\pi}_j \geq \tau\}$$

- The relative selection frequency $\hat{\pi}_j$ is an unbiased estimator of

$$\pi_j^{n/2} = \mathrm{P}(j \in \hat{S}_{n/2})$$

  but, in general, a biased estimator of

$$\pi_j^n = \mathrm{P}(j \in \hat{S}_n) = \mathbb{E}(\mathbb{1}\{j \in \hat{S}_n\})$$

- The key idea of stability selection is to improve on the simple estimator $\mathbb{1}\{j \in \hat{S}_n\}$ of π_j^n through subsampling.
- By means of averaging involved in $\hat{S}_{\text{stab}}$, we hope that $\hat{\pi}_j$ will have reduced variance compared to $\mathbb{1}\{j \in \hat{S}_n\}$ and this increased stability will more than compensate for the bias incurred.

Theorem

Assume that

1. $\{\mathbb{1}\{j \in \hat{S}_{n/2}\}, j \in N\}$ *is exchangeable;*
2. *The variable selection procedure is not worse than random guessing, i.e.*

$$\frac{\mathbb{E}(|\hat{S}_{n/2} \cap S|)}{\mathbb{E}(|\hat{S}_{n/2} \cap N|)} \geq \frac{|S|}{|N|}.$$

Then, for $\tau \in (1/2, 1]$

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N|) \leq \frac{1}{(2\tau - 1)} \frac{q^2}{p}$$

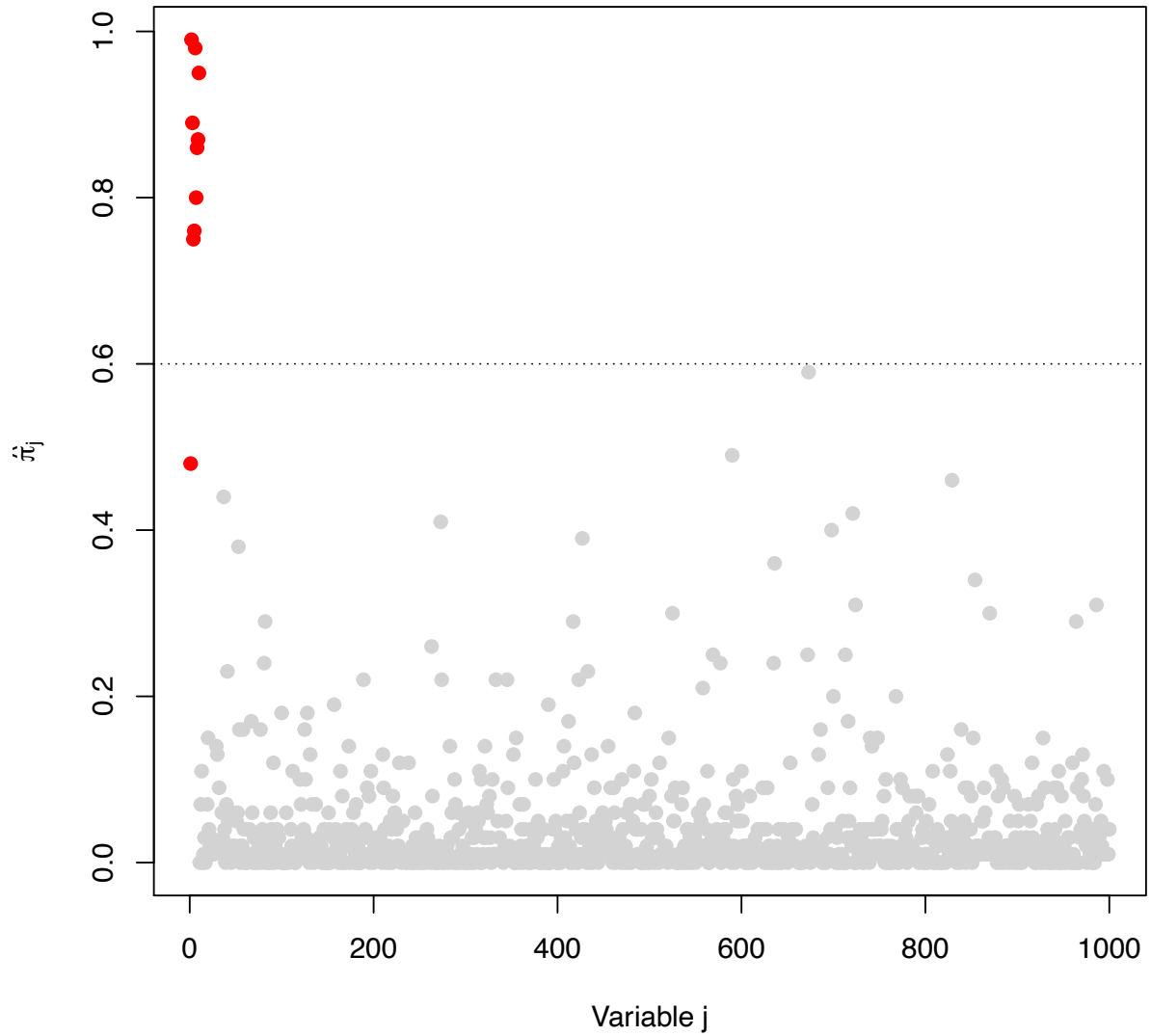
where $q = \mathbb{E}(|\hat{S}_{n/2}|)$

- The choice of the number of subsamples B is of minor importance
- It is possible to fix $q = \mathbb{E}(|\hat{S}_{n/2}|)$ and run the variable selection procedure until it selects q variables. However, if q is too small, one would select only a subset of the signal variables as

$$|\hat{S}_{\text{stab}}| \leq |\hat{S}_{n/2}| = q$$

- For example, with $p = 1000$, $q = 50$ and $\tau = 0.6$ then

$$\mathbb{E}(|\hat{S}_{\text{stab}} \cap N|) \leq 12.5$$


1.0
0.8
0.6
0.4
0.2
0.0
$\hat{\pi}_j$
0
200
400
600
800
1000
Variable j

The knockoff filter

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

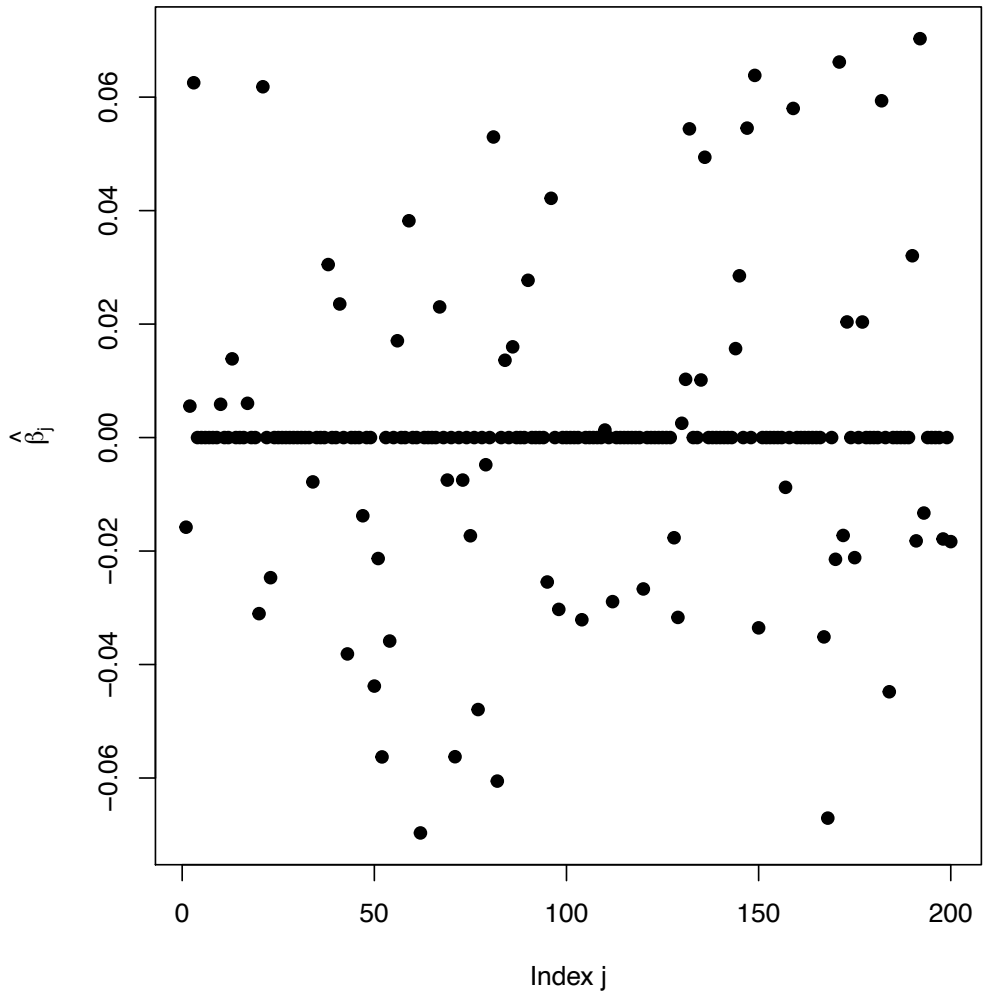
References

- Barber, Candès (2015) Controlling the False Discovery Rate via Knockoffs. Ann. Statist. 43:2504–2537
- Candès, Fan, Janson, Lv (2018). Panning for gold: model-X knockoffs for high dimensional controlled variable selection. JRSS-B 80:551–577.

There are two main approaches:

- *Fixed-X knockoffs*
  Requires that X is full rank with $n \geq 2p$
- *Model-X knockoffs*
  Requires assumptions on X but works with $p > n$

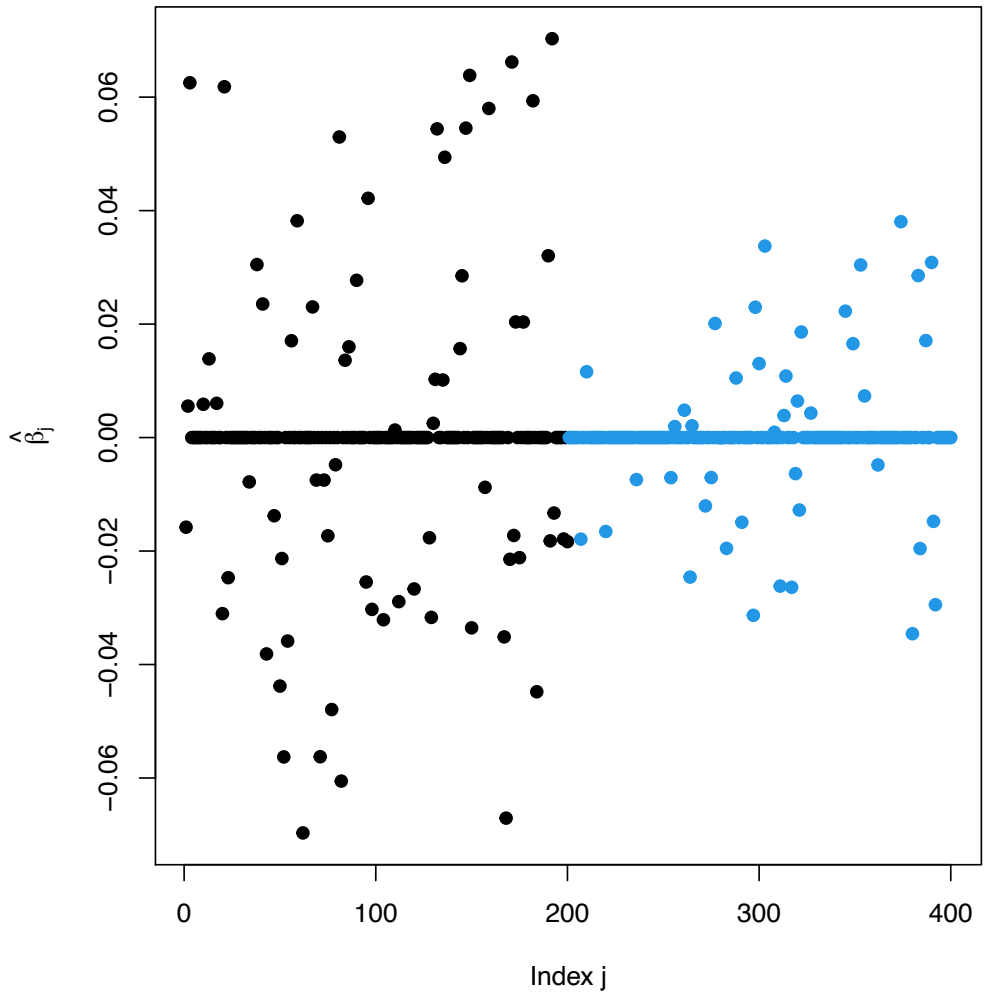
Fixed-X knockoffs



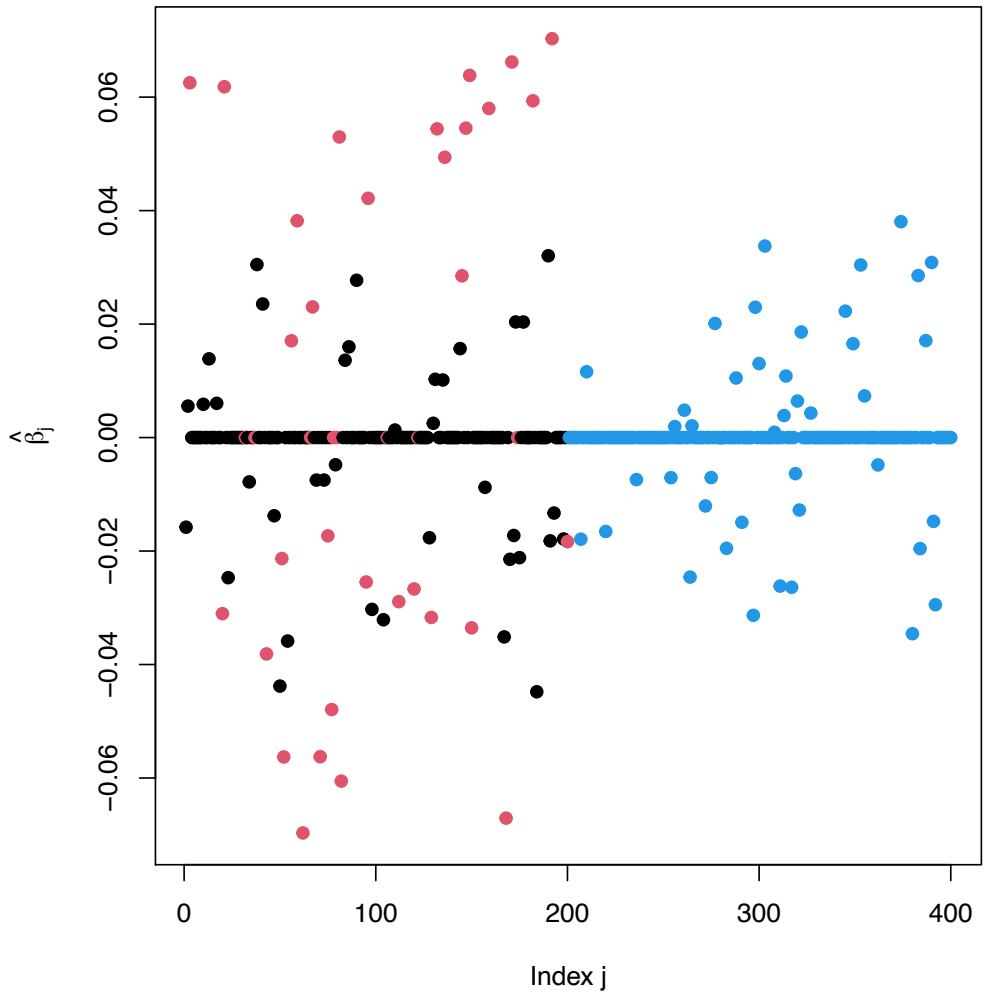
Lasso selects 67 features: $\mathrm{FDP}(\hat{S}) = ?/67$

Main idea

- For each feature X_j, construct a *knockoff* copy $\tilde{X}_j$
- Knockoffs $\tilde{X}_1, \ldots, \tilde{X}_p$ are independent of y and mimic the original variables $X_1, \ldots, X_p$ if they were null



Lasso selects 70 original and 43 knockoff: $\widehat{\mathrm{FDP}}(\hat{S}) = 43/70 \approx 61\%$



$$\text{True FDP}(\hat{S}) = 34/70 \approx 54\%$$

Knockoff construction

- Suppose without loss of generality that the features are centered and scaled such that $\|X_j\|_2^2 = 1$ for all j
- Let $\Sigma = X^t X$ be the correlation matrix of the features
- The method begins by augmenting the design matrix X with a second matrix $\tilde{X} \in \mathbb{R}^{n \times p}$ of knockoff variables, constructed to satisfy

$$
\begin{aligned}
G = [X\ \tilde{X}]^t [X\ \tilde{X}] &= \begin{bmatrix} X^t X & X^t \tilde{X} \\ \tilde{X}^t X & \tilde{X}^t \tilde{X} \end{bmatrix} \\
&= \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix}
\end{aligned}
$$

for some diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_p)$ such that G is positive definite

- The knockoffs have the same correlation structure as the original features

$$\tilde{X}^t \tilde{X} = X^t X = \Sigma$$

- The correlation between $\tilde{X}_k$ and X_j is

$$\tilde{X}_j^t X_k = X_j^t X_k \quad \forall\, k \neq j$$

- The correlation between $\tilde{X}_j$ and X_j is

$$\tilde{X}_j^t X_j = 1 - d_j$$

with d_j as close to 1 as possible

Equi-correlated knockoffs

Suppose we require $d_j = d$ for all j. Define

$$\tilde{X} = X(I_p - d\Sigma^{-1}) + UC$$

where

- $U \in \mathrm{R}^{n \times p}$ is an orthonormal matrix such that $U^t X = 0$
- $C \in \mathrm{R}^{p \times p}$ from the Cholesky decomposition of

$$C^t C = 4((d/2)I_p - (d/2)^2 \Sigma^{-1})$$

This approach corresponds to `method="equi"` in the `knockoff` package. A semidefinite programming approach is used to determine the values that minimize $\sum_{j=1}^{p}(1 - d_j)$ subject to the constraints (`method="sdp"`)

The knockoff statistics

- Fit the lasso to the augmented design matrix $[X\ \tilde{X}]$ for $\lambda \in \Lambda$
- Let $[\hat{\beta}(\lambda)\ \tilde{\beta}(\lambda)]$, $\lambda \in \Lambda$ denote the coefficient estimates
- Compute

$$Z_j = \sup\{\lambda \in \Lambda : \hat{\beta}_j(\lambda) \neq 0\} = \text{ first time } X_j \text{ enters the lasso path}$$
$$\tilde{Z}_j = \sup\{\lambda \in \Lambda : \tilde{\beta}_j(\lambda) \neq 0\} = \text{ first time } \tilde{X}_j \text{ enters the lasso path}$$

- Then define the statistics

$$W_j = \max(Z_j, \tilde{Z}_j) \cdot \text{sign}(Z_j - \tilde{Z}_j) = \begin{cases} Z_j & \text{if } X_j \text{ enters first } (Z_j > \tilde{Z}_j) \\ 0 & \text{if } Z_j = \tilde{Z}_j \\ -\tilde{Z}_j & \text{if } \tilde{X}_j \text{ enters first } (Z_j < \tilde{Z}_j) \end{cases}$$

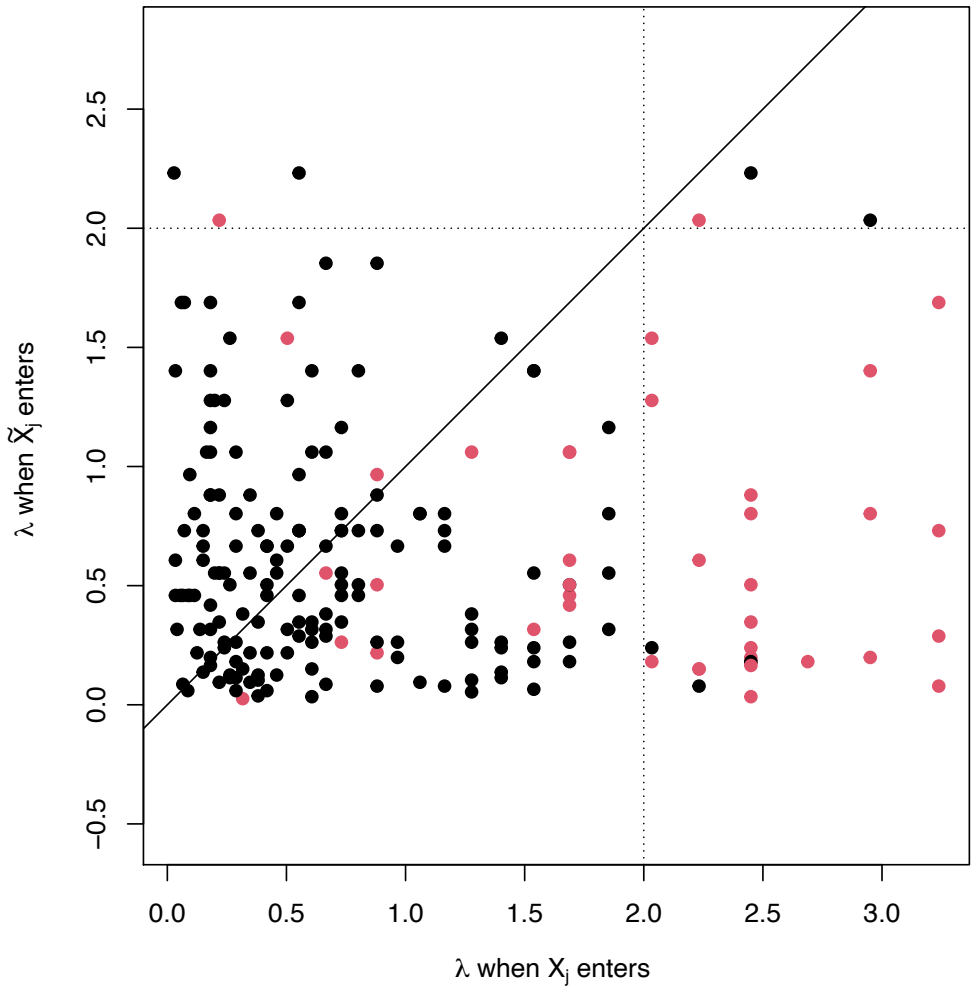
FDP estimate

- For some threshold $\tau \geq 0$, select

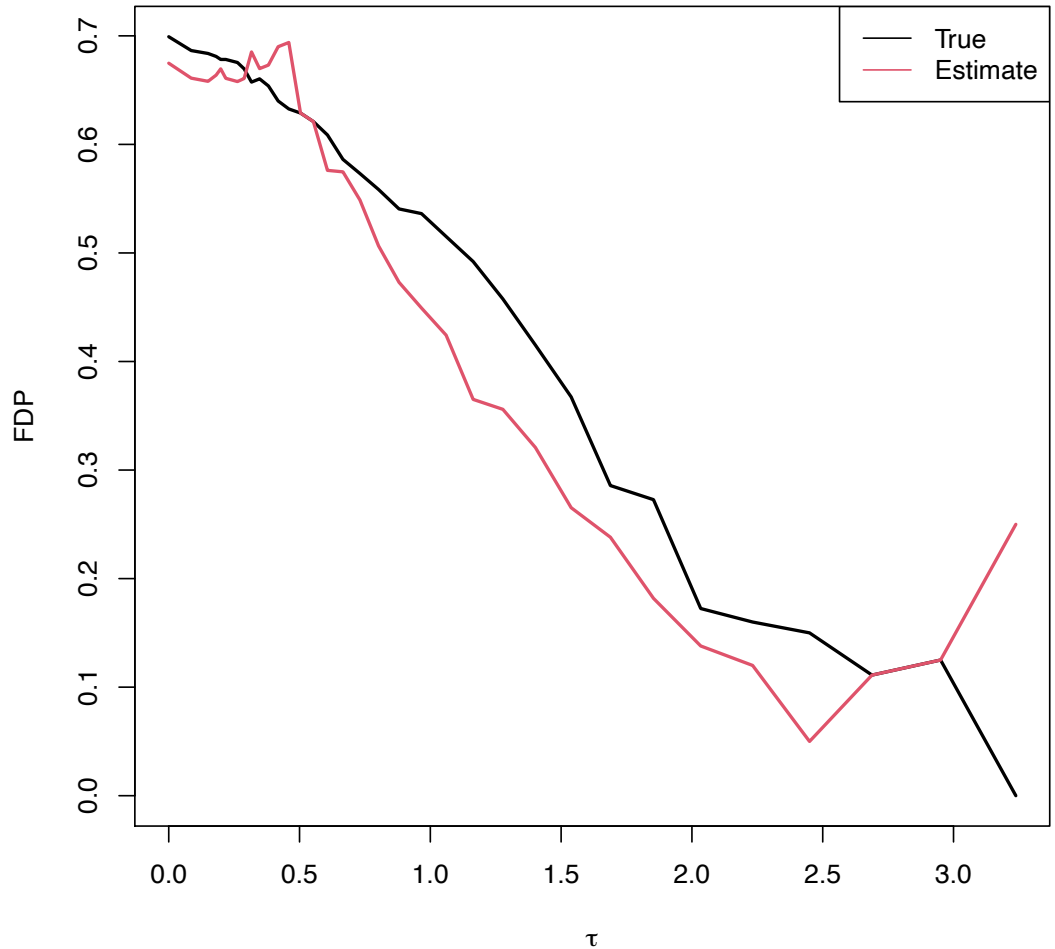
$$\hat{S}_\tau = \{j \in \{1, \ldots, p\} : W_j \geq \tau\}$$

- The knockoff estimate of the FDP is

$$\begin{aligned} \mathrm{FDP}(\hat{S}_\tau) &= \frac{\#\{j \in N : W_j \geq t\}}{\#\{j : W_j \geq t\}} \\ &\approx \frac{\#\{j \in N : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \\ &\leq \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} = \widehat{\mathrm{FDP}}(\hat{S}_\tau) \end{aligned}$$



For $\tau = 2$, $|\hat{S}_\tau| = 29$ with $\widehat{\mathrm{FDP}}(\hat{S}_\tau) = 4/29$ and $\mathrm{FDP}(\hat{S}_\tau) = 5/29$


True
Estimate
FDP
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.5
1.0
1.5
2.0
2.5
3.0
τ

The knockoff procedure chooses a data-dependent threshold

$$\hat{\tau} = \min\left\{\tau > 0 : \widehat{\mathrm{FDP}}(\hat{S}_\tau) \leq \alpha\right\}$$

with $\hat{\tau} = +\infty$ if no such τ exists.

Theorem

For any $\alpha \in (0, 1)$, the knockoff procedure selects

$$\hat{S}_{\hat{\tau}} = \{j \in \{1, \ldots, p\} : W_j \geq \hat{\tau}\}$$

with the guarantee that

$$\mathrm{FDR}(\hat{S}_{\hat{\tau}}) = \mathbb{E}\left(\frac{|N \cap \hat{S}_{\hat{\tau}}|}{|\hat{S}_{\hat{\tau}}|}\right) \leq \alpha$$

where the expectation is taken over ε in the Gaussian linear model $y = X\beta + \varepsilon$ while treating X and $\tilde{X}$ as fixed.

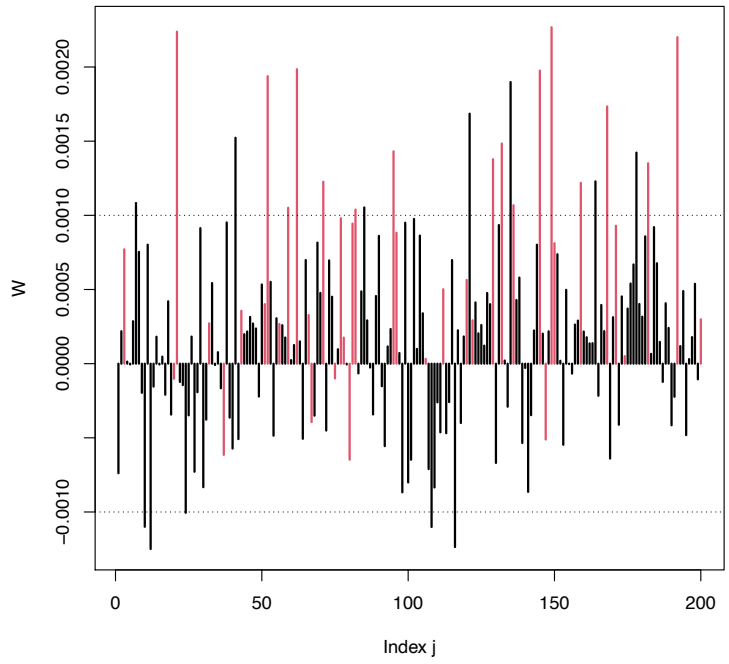
Variable importance statistics

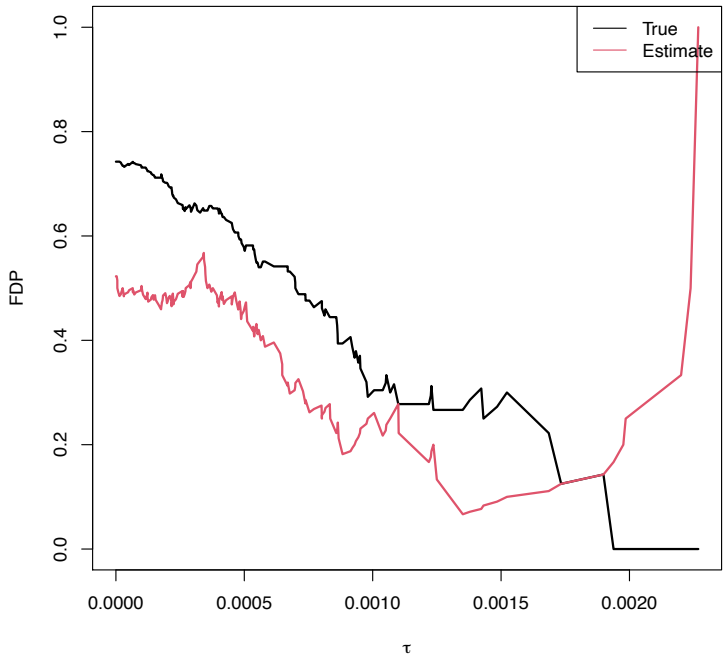
- Fit the Random Forest to the augmented design matrix $[X\ \tilde{X}]$
- Compute

$$Z_j = \text{VariableImportance}(X_j)$$
$$\tilde{Z}_j = \text{VariableImportance}(\tilde{X}_j)$$

  The importance of a variable is measured as the total decrease in node impurities from splitting on that variable, averaged over all trees
- Then define the statistics

$$W_j = \text{abs}(Z_j) - \text{abs}(\tilde{Z}_j)$$





For $\tau = 0.001$, $|\hat{S}_\tau| = 23$ with $\widehat{\mathrm{FDP}}(\hat{S}_\tau) = 4/23$ and $\mathrm{FDP}(\hat{S}_\tau) = 7/23$

Model-X knockoff

Modeling X

- X is treated as a random matrix with i.i.d. rows x_i
- (x_i, y_i), $i = 1, \ldots, n$ are i.i.d. from some unknown distribution
- Assume we know the *marginal distribution* of x_i, e.g.

$$x_i = (x_{i1}, \ldots, x_{ip}) \sim N_p(\mu, \Sigma)$$

- Null features given by *conditional independence*

$$N = \{j \in \{1, \ldots, p\} : y \perp\!\!\!\perp x_j | x_{-j}\}$$

where $x_{-j} = \{x_1, \ldots, x_p\} \setminus \{x_j\}$

Knockoffs in the Gaussian case

- The joint distribution of original features and knockoff copies satisfies

$$[x\ \tilde{x}] \sim N(M, V) \quad \text{with } M = \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \quad V = \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix}$$

  where $D = \text{diag}(d_1, \ldots, d_p)$ such that V is positive definite
- Draw a random $\tilde{x}_i$ from the conditional distribution $\tilde{x}_i | x_i$, which is normal with

$$\begin{aligned} \mathbb{E}(\tilde{x}_i | x_i) &= \mu + (\Sigma - D)\Sigma^{-1}(x_i - \mu) \\ \mathbb{V}\text{ar}(\tilde{x}_i | x_i) &= \Sigma - (\Sigma - D)\Sigma^{-1}(\Sigma - D) \end{aligned}$$

- If μ and Σ are unknown, replace by estimates $\hat{\mu}$ and $\hat{\Sigma}$

Conformal prediction

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari

References

- Lei, G'Sell, Rinaldo, Tibshirani, Wasserman (2018) Distribution-free predictive inference for regression. JASA,113:1094–1111

Suppose we have fitted a Gaussian linear model based on the training data $(\mathbf{y}, \mathbf{X})$, obtaining the estimates

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}, \quad \hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2/(n-p)$$

There are (at least) two levels at which we can make predictions

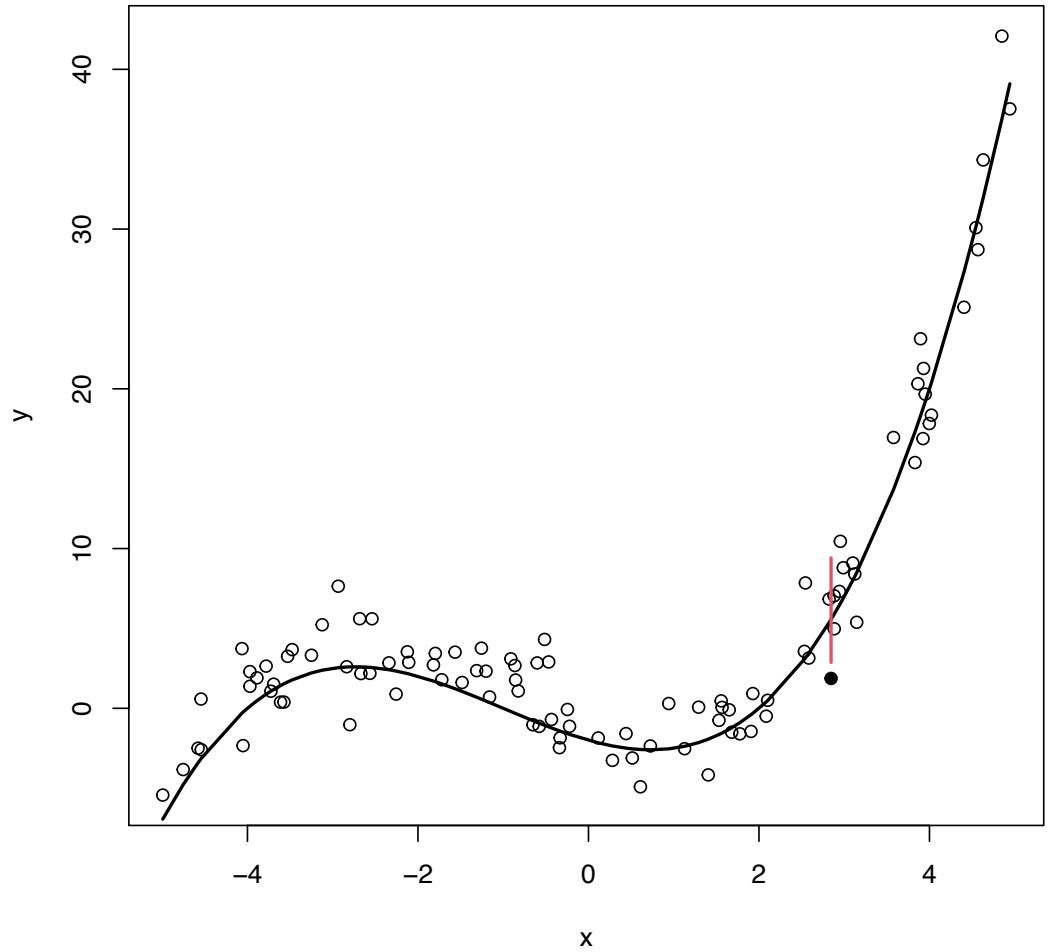
1. A *point prediction* is a single best guess about what a new Y will be when $X = x$
2. A *prediction interval*

   $$C_\alpha(x) = x^t\hat{\beta} \pm t_{n-p}^{1-\alpha/2}\hat{\sigma}\sqrt{x^t(\mathbf{X}^t\mathbf{X})^{-1}x + 1}$$

   for $Y|X = x$ with $(1-\alpha)$ *conditional coverage* guarantee, i.e.

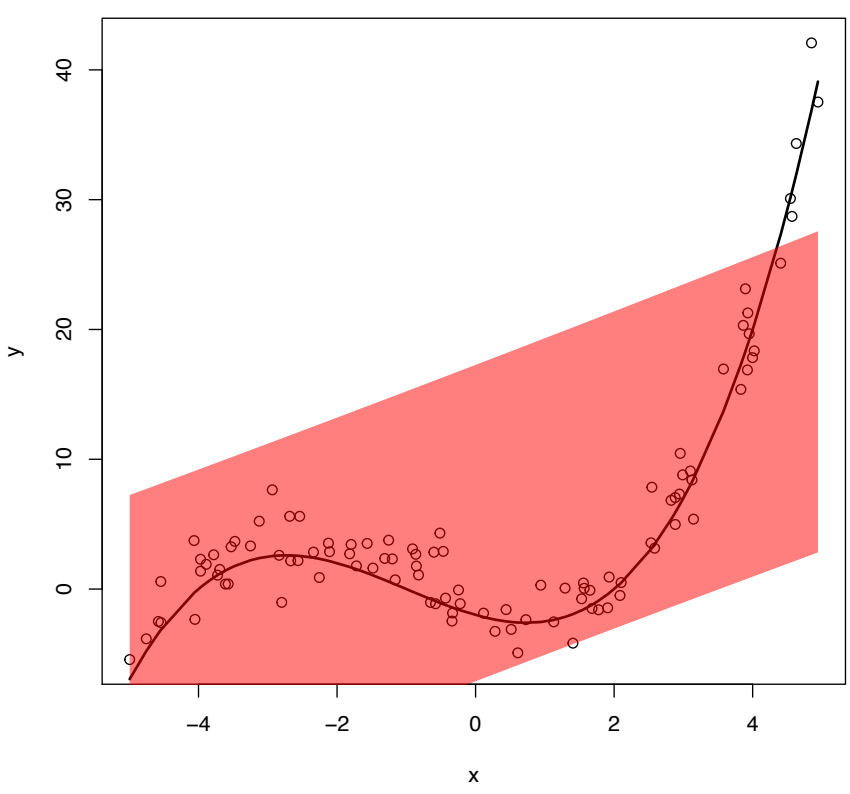
   $$\mathrm{P}(Y \in C_\alpha(x)|X = x) = 1 - \alpha$$

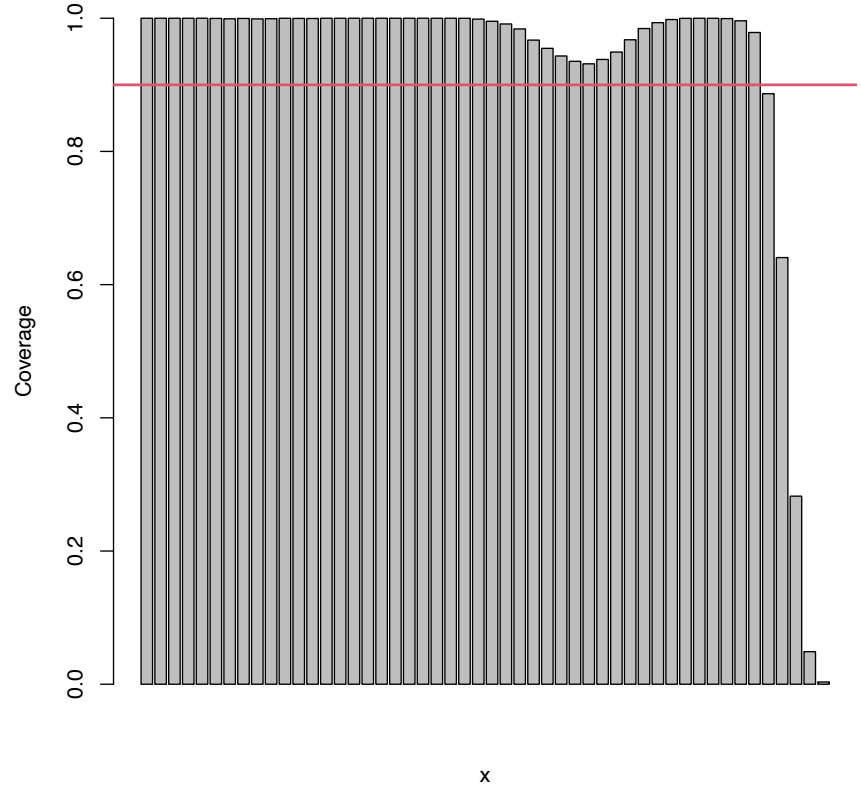
   where the probability is with respect to the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the new response Y at a fixed test point $X = x$



$$f(x) = \frac{1}{4}(x+4)(x+1)(x-2)$$

Model miss-specification





$1 - \alpha = 90\%$, marginal coverage $\approx 93\%$

Marginal and conditional coverage

- $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ follows some *unknown* joint distribution P_{XY}
- Training $(X_1, Y_1), \ldots, (X_n, Y_n)$ and test (X_{n+1}, Y_{n+1}) i.i.d. (X, Y)
- C_α satisfies *distribution-free marginal coverage* at level $1 - \alpha$ if

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha \qquad \forall\ P_{XY}$$

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X_{n+1}, Y_{n+1})
- C_α satisfies *distribution-free conditional coverage* at level $1 - \alpha$ if

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1}) | X_{n+1} = x) \geq 1 - \alpha \qquad \forall\ P_{XY},\ \ \forall\ x$$

  where the probability is w.r.t. $(X_1, Y_1), \ldots, (X_n, Y_n)$, and Y_{n+1} at a fixed test point $X_{n+1} = x$

Conformal prediction

Conformal prediction (Vovk, Gammerman, Saunders, Vapnik, 1996-1999) is a general framework for constructing prediction intervals by using *any* algorithm with finite sample and distribution-free *exact* marginal coverage, i.e.

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha \qquad \forall\ P_{XY}$$

Two main versions:

- *Full* conformal prediction
- *Split* conformal prediction

Algorithm 1 Full conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, test x_{n+1}, algorithm $\hat{\mu}$, level α, grid of values $\mathcal{Y} = \{y, y', y'', \ldots\}$

for $y \in \mathcal{Y}$ **do**
    Train $\hat{\mu}^y(x) = \hat{\mu}(x; (x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y))$
    Compute $R_i^y = |y_i - \hat{\mu}^y(x_i)|$ for $i = 1, \ldots, n$
    Sort $R_1^y, \ldots, R_n^y$ in increasing order: $R_{(1)}^y \leq \ldots \leq R_{(n)}^y$
    Compute $R_\alpha^y = R_{(k)}^y$ with $k = \lceil (1-\alpha)(n+1) \rceil$
    Compute $R^y = |y - \hat{\mu}^y(x_{n+1})|$
end for
$C_\alpha(x_{n+1}) = \{y \in \mathcal{Y} : R^y \leq R_\alpha^y\}$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$. This is the only assumption of the method
- The prediction interval

$$C_\alpha(x_{n+1}) = \{y \in \mathbb{R} : R^y \leq R^y_\alpha\},$$

  satisfies

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

  if and only if $\alpha \in \{1/(n+1), 2/(n+1), \ldots, n/(n+1)\}$
- Informally, the null hypothesis that the random variable Y_{n+1} will have the outcome y, i.e.

$$H_y : Y_{n+1} = y$$

  is rejected when $R^y > R^y_\alpha$

Algorithm 2 Split conformal prediction

Require: Training $(x_1, y_1), \ldots, (x_n, y_n)$, x_{n+1}, algorithm $\hat{\mu}$, validation sample size m, level α

1: Split $\{1, \ldots, n\}$ into L of size w and I of size $m = n - w$
2: Train $\hat{\mu}_L(x) = \hat{\mu}(x; (x_l, y_l), l \in L)$
3: Compute $R_i = |y_i - \hat{\mu}_L(x_i)|$ for $i \in I$
4: Sort $\{R_i, i \in I\}$ in increasing order: $R_{(1)} \leq \ldots \leq R_{(m)}$
5: Compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : |y - \hat{\mu}_L(x_{n+1})| \leq R_\alpha\} \\ &= [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha] \end{aligned}$$

- Assume that (X_i, Y_i), $i = 1, \ldots, n+1$ are i.i.d. from a probability distribution P_{XY} on the sample space $\mathbb{R}^p \times \mathbb{R}$
- The prediction interval

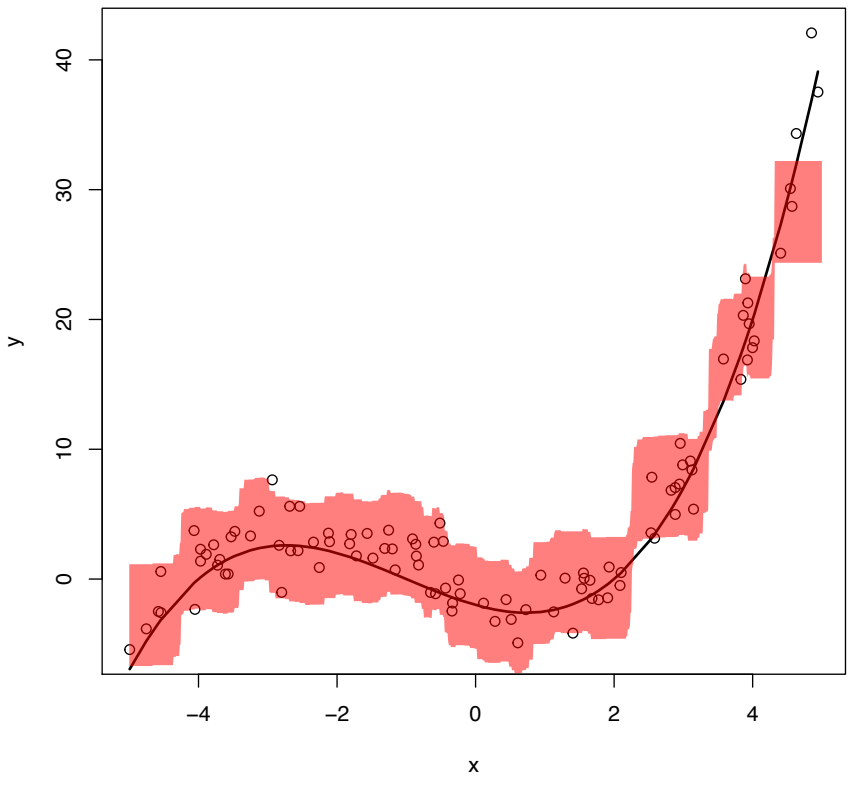
$$C_\alpha(x_{n+1}) = [\hat{\mu}_L(x_{n+1}) - R_\alpha, \hat{\mu}_L(x_{n+1}) + R_\alpha]$$

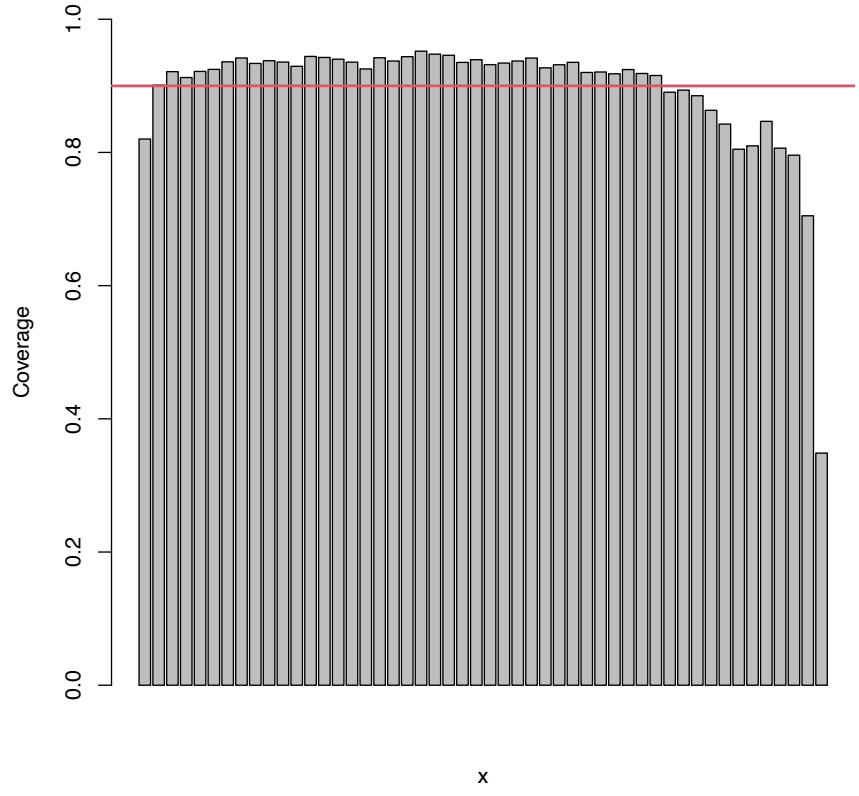
  satisfies

$$\mathrm{P}(Y_{n+1} \in C_\alpha(X_{n+1})) = 1 - \alpha$$

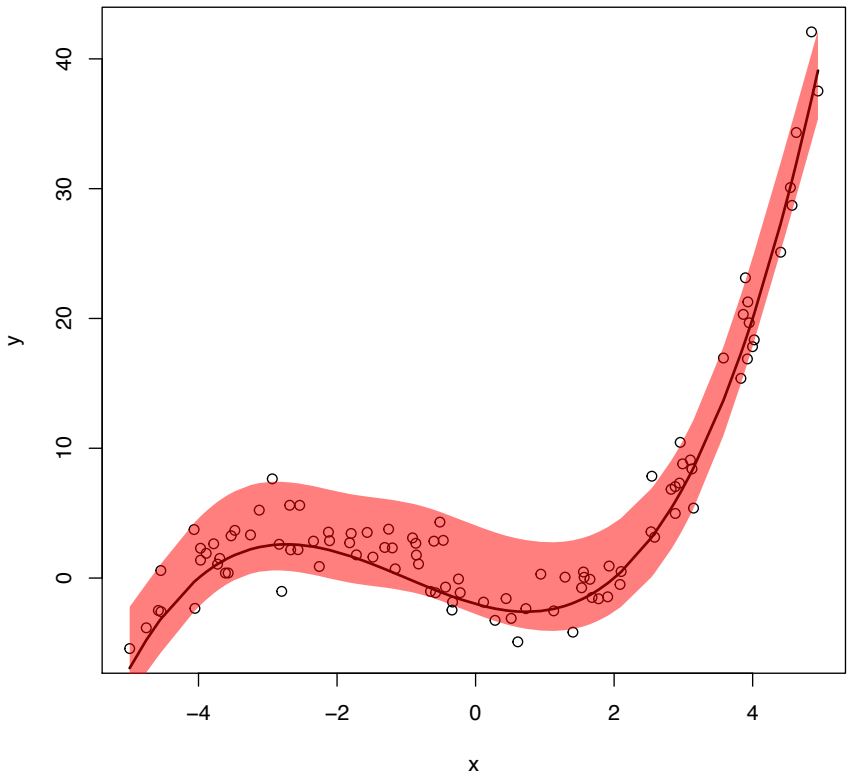
  if and only if $\alpha \in \{1/(m+1), 2/(m+1), \ldots, m/(m+1)\}$
- Note that in computing the critical value $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$, we need to have $k \leq m$, which happens if $\alpha \geq 1/(m+1)$ (otherwise if $k > m$ we need to set $R_\alpha = +\infty$)

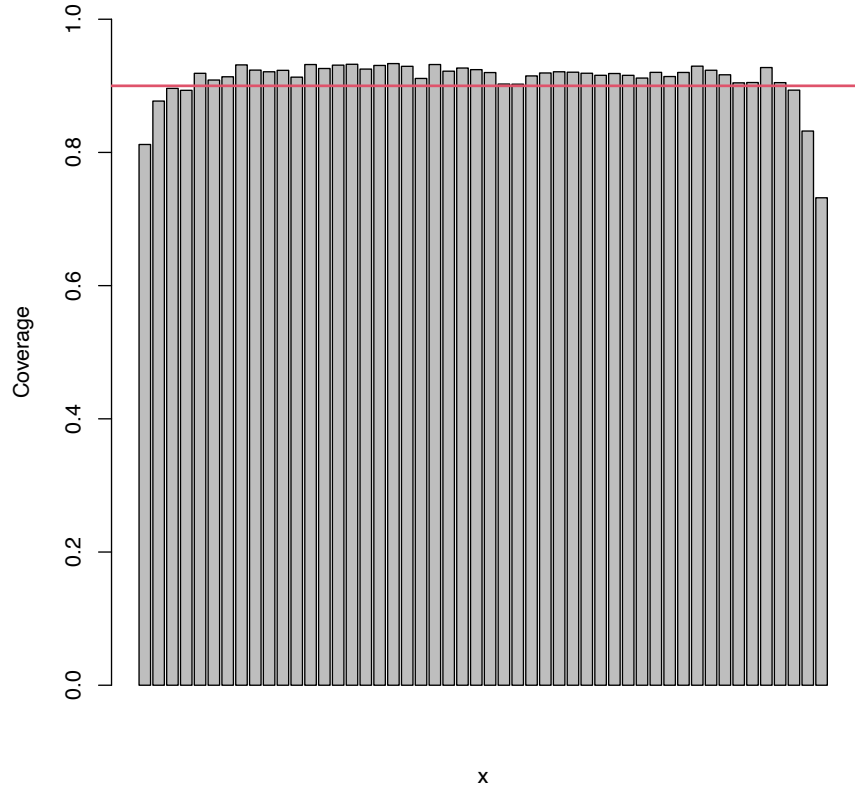
Random Forest





Smoothing splines





Conformity scores

- In the previous algorithm we used a statistic, called *conformity score*, which is the absolute value of the residual

$$R_i = |y_i - \hat{\mu}_L(x_i)|, \quad i \in I$$

  where $\hat{\mu}_L$ is an estimator of $\mathbb{E}(Y \mid X)$ based on $\{(X_i, Y_i), i \in L\}$
- The oracle knows the conditional distribution of $Y \mid X$. The oracle prediction interval

$$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$$

  where $q^\gamma(x)$ is the γ-quantile of $Y \mid X = x$, guarantees exact conditional coverage

$$\mathrm{P}(Y \in C^*_\alpha(X) | X = x) = 1 - \alpha \quad \forall\ x$$

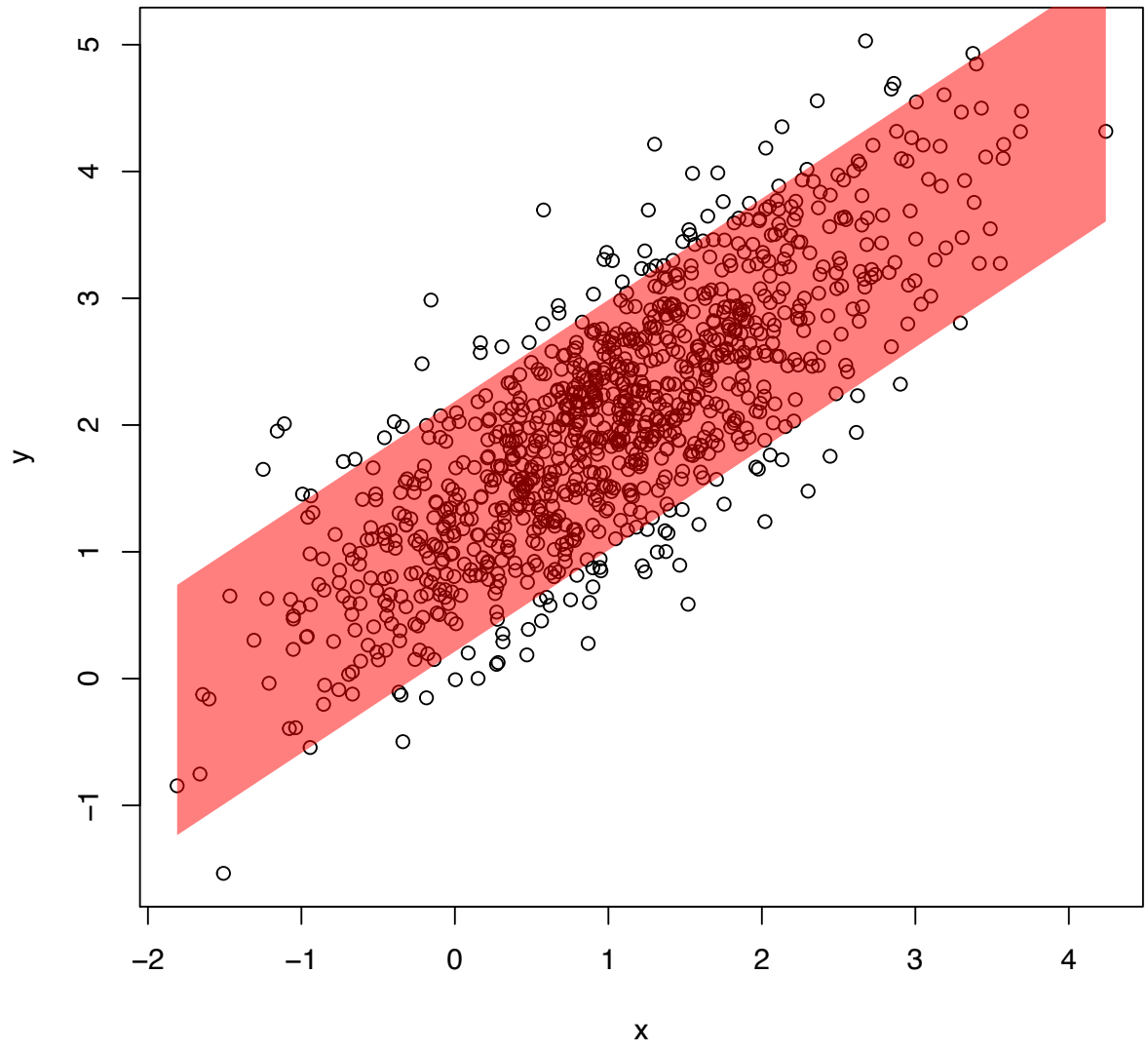
Suppose that

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right)$$

then the conditional distribution of $Y \mid X = x$ is

$$(Y|X = x) \sim N\Big(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1 - \rho^2)\Big)$$

from which we can compute the quantile $q^\gamma(x)$



$C^*_\alpha(x) = [q^{\alpha/2}(x), q^{1-\alpha/2}(x)]$ as a function of x

Conformal quantile regression

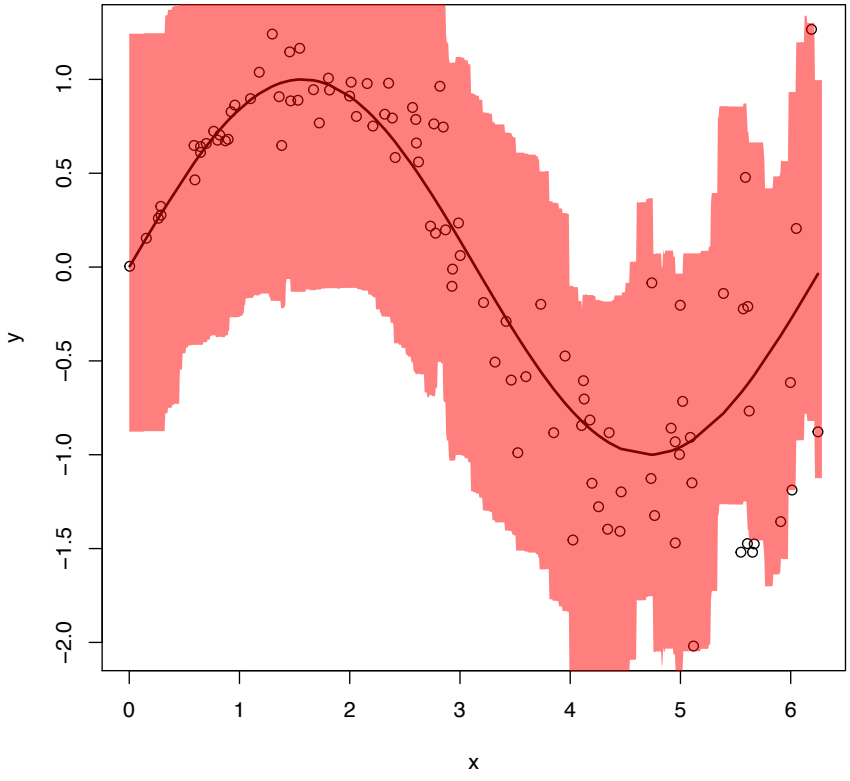
- Compute conformity scores

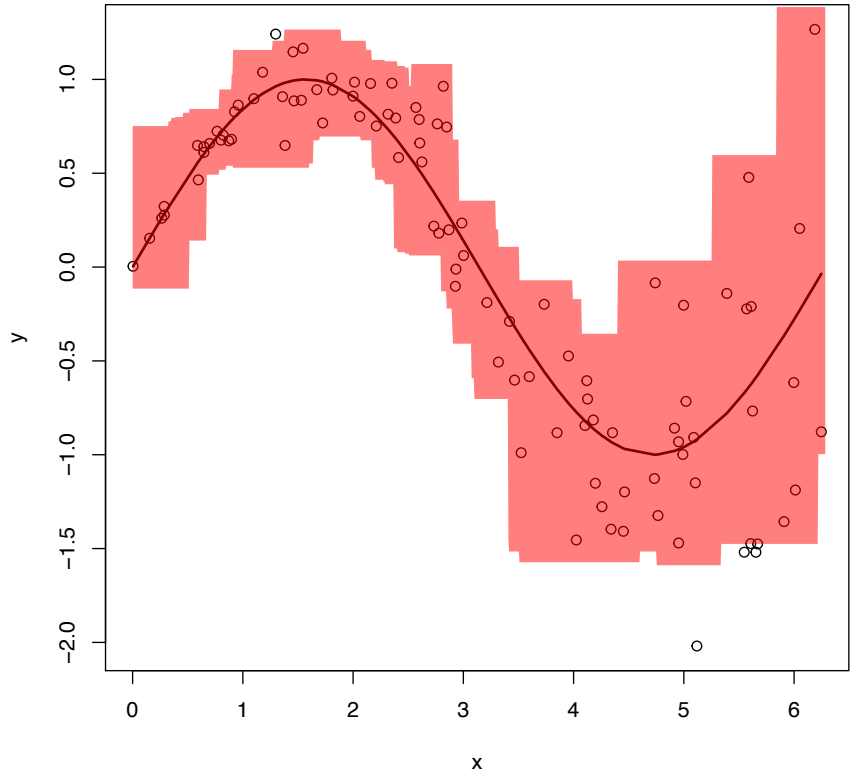
$$R_i = \max\left\{\hat{q}_L^{\gamma}(X_i) - Y_i, Y_i - \hat{q}_L^{1-\gamma}(X_i)\right\}, \quad i \in I$$

  where $\hat{q}_L^{\gamma}$ is an estimator of the γ-quantile of $Y \mid X$ based on $\{(X_i, Y_i), i \in L\}$
- Sort $\{R_i, i \in I\}$ in increasing order, obtaining $R_{(1)} \leq \ldots \leq R_{(m)}$, and compute $R_\alpha = R_{(k)}$ with $k = \lceil (1-\alpha)(m+1) \rceil$
- Compute the prediction interval

$$\begin{aligned} C_\alpha(x_{n+1}) &= \{y \in \mathbb{R} : \max\left\{\hat{q}_L^{\gamma}(x_{n+1}) - y, y - \hat{q}_L^{1-\gamma}(x_{n+i})\right\} \leq R_\alpha\} \\ &= [\hat{q}_L^{\gamma}(x_{n+1}) - R_\alpha, \hat{q}_L^{1-\gamma}(x_{n+1}) + R_\alpha] \end{aligned}$$

  or $C_\alpha(x_{n+1}) = \emptyset$ if $R_\alpha < (1/2)(\hat{q}_L^{\gamma}(x_{n+1}) - \hat{q}_L^{1-\gamma}(x_{n+1}))$





$$X_i \sim U(0, 2\pi), \epsilon_i \sim N(0, 1), Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i$$

Multi-split conformal prediction

Algorithm

1. Choose a number B of splits
2. Choose a threshold $\tau \in \{0, 1/B, 2/B, \ldots, (B-1)/B\}$
3. Compute B split conformal prediction intervals with coverage level $1 - \beta$

   $$C_\beta^{[1]}(x_{n+1}), \ldots, C_\beta^{[B]}(x_{n+1})$$

   where

   $$\beta = \alpha(1 - \tau)$$

4. Compute the aggregated prediction interval

   $$C_\alpha^\tau(x_{n+1}) = \{y \in \mathbb{R} : \Pi_\beta^y > \tau\}$$

   with

   $$\Pi_\beta^y = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{y \in C_\beta^{[b]}(x_{n+1})\}$$

- The multi-split prediction interval guarantees

$$\mathrm{P}(Y_{n+1} \in C_\alpha^\tau(X_{n+1})) \geq 1 - \alpha \qquad \forall\ P_{XY}$$

- The parameter τ can be regarded as a tuning parameter, and proper choice of τ is essential for good performance
- On the one hand, setting $\tau = 1 - 1/B$ gives the Bonferroni-intersection method with $C_\alpha^{(B-1)/B} = \bigcap_b C_{\alpha/B}^{[b]}$.
- On the other hand, setting $\tau = 0$ gives an unadjusted-union $C_\alpha^0 = \bigcup_b C_\alpha^{[b]}$.
- For B even, an intermediate choice $\tau = 1/2$ amounts to constructing B single split confidence intervals at level $\alpha/2$, that is $C_{\alpha/2}^{[b]}$, which is a small but not negligible price to pay for using multiple splits rather than just one split. In practice, however, $\tau = 1/2$ and $C_\alpha^{[b]}$ may give marginal coverage $\approx 1 - \alpha$