ESTYMACJA BŁĘDU PREDYKCJI I JEJ ZASTOSOWANIA

Jan Mielniczuk

Wisła, grudzień 2009

PLAN

- Błędy predykcji i ich podstawowe estymatory
- Estymacja błędu predykcji w modelu liniowym. Funkcje kryterialne
- Własności predykcyjne estymatorów postselekcyjnych
- Metoda minimalizacji ryzyka strukturalnego

BŁĘDY PREDYKCJI I ICH PODSTAWOWE ESTYMATORY

Rozpatrywane problemy

(i) Regresja wielokrotna z losowymi zmiennymi objaśniającymi

Wektor losowy $(\mathbf{X}, Y) \in R^p \times R$ o rozkładzie $P_{\mathbf{X},Y}$ takim, że

$$Y = f(\mathbf{X}) + \varepsilon, \qquad (*)$$

$f : R^p \to R$, ε zmienna niezależna od $\mathbf{X}$, $E(\varepsilon) = 0$.
Uwaga. Założenia implikują

$$f(\mathbf{X}) = E(Y|\mathbf{X}) \qquad i \qquad \varepsilon = Y - E(Y|\mathbf{X}).$$

Dla takich f i ε, $(*)$ spełniona oczywiście zawsze, tu zakładamy dodatkowo niezależność $\mathbf{X}$ i ε.

(ii) Klasyfikacja pod nadzorem

Tak jak w (i), zakładamy dodatkowo, że

$$Y \in \{1, 2, \ldots, K\}$$

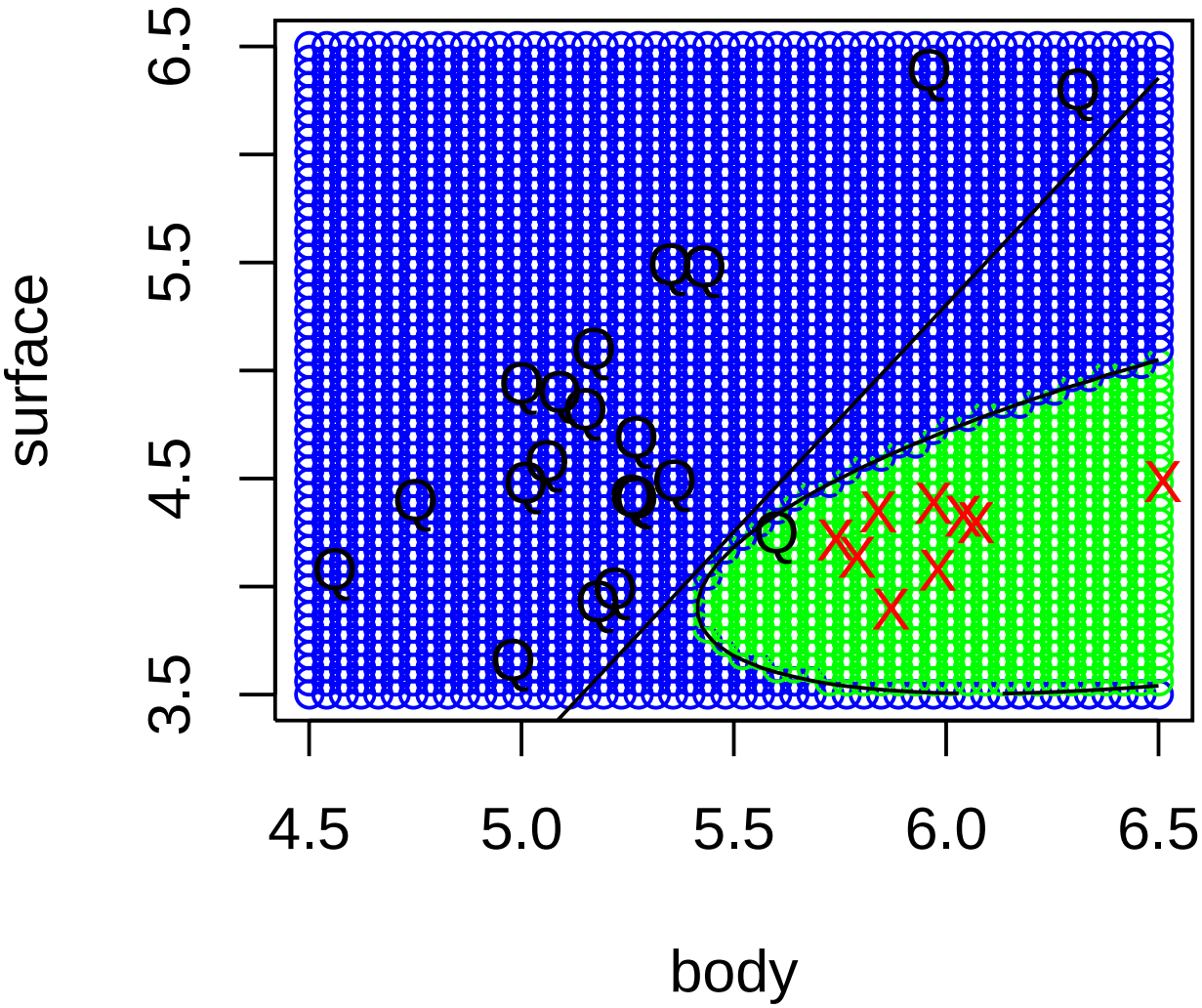
(Y - indeks przynależności do populacji).
Nie zakładamy niezależności $\mathbf{X}$ i ε.

Obserwowalna p.p.l. $\mathcal{U} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ z rozkładu $P_{\mathbf{X},Y}$.

Na jej podstawie estymator $\hat{f}(\mathbf{x})$ używany do prognozy zmiennej Y^0:

$$\hat{Y}^0 = \hat{f}(\mathbf{X}^0)$$

.


surface
6.5
5.5
4.5
3.5
4.5
5.0
5.5
6.0
6.5
body

(iii) Estymacja gęstości zmiennej losowej Y

$Y_1, \ldots, Y_n$ p.p.l. z rozkładu P_Y o gęstości f. Estymacja f w oparciu o model parametryczny $M = \{f_\theta\}_{\theta \in \Theta}$.

(iv) Estymacja gęstości warunkowej

$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ p.p.l. z rozkładu $P_{\mathbf{X},Y}$ na $R^p \times R$. Estymujemy gęstość warunkową Y pod warunkiem $\mathbf{X} = \mathbf{x}$ w oparciu o model parametryczny

$$M = \{f_{\theta(\mathbf{x})}\}_{\theta \in \Theta, \mathbf{x} \in R^p}$$

Problem prognozy

$(\mathbf{X}^0, Y^0) \sim P_{\mathbf{X},Y}$ niezależna od $\mathcal{U}$.

- Jak dalece $\hat{f}(\mathbf{X}^0)$ różni się od Y^0 ? (Problem (i) i (ii))

- Jak dalece gęstość $f_{\hat{\theta}}$ różni się od gęstości f? (problem (iii))

- Jak dalece gęstość $f_{\hat{\theta}(\mathbf{X}^0)}$ różni się od gęstości $f_{Y^0|\mathbf{X}^0}$? (problem (iv))

Błędy rozpatrywane w (iii) i (iv) będą związane z problemem prognozy.

Funkcje straty

Problemy (i) i (ii): $L(y, \hat{f}(\mathbf{x})) \geqslant 0$:

$(y - \hat{f}(\mathbf{x}))^2$, $|y - \hat{f}(\mathbf{x})|$, $I\{y \neq \hat{f}(\mathbf{x})\}$ (funkcja straty 0-1 dla problemu (ii)).

Problemy (iii) i (iv): $L(y, \hat{\theta}(\mathbf{x}))$:

$$L(y, \hat{\theta}) = -2 \log f_{\hat{\theta}}(y)$$

$$L(y, \hat{\theta}(\mathbf{x})) = -2 \log f_{\hat{\theta}(\mathbf{x})}(y)$$

Błąd predykcji (uogólnienia)

$(\mathbf{X}^0, Y^0) \sim P_{\mathbf{X},Y}$ niezależna od $\mathcal{U}$.

$$Err_{\mathcal{U}} = E(L(Y^0, \hat{f}(\mathbf{X}^0))|\mathcal{U}) \qquad (i), (ii)$$

$$Err_{\mathcal{U}} = E(L(Y^0, \hat{\theta}(\mathbf{X}^0))|\mathcal{U}) \qquad (iii), (iv)$$

Warunkowy błąd predykcji (w (ii): prawdopodobieństwo błędnej klasyfikacji), $Err_{\mathcal{U}} = G(\mathcal{U})$ jest funkcją próby $\mathcal{U}$.

Dla ustalonej próby uczącej $\mathcal{U}$ konstruujemy predyktor $\hat{f}$ i oceniamy jego działanie dla nowej obserwacji $(\mathbf{X}^0, Y^0)$.

Miara jakości odpowiadająca faktycznej funkcji predyktora.

Bezwarunkowy błąd predykcji

$$Err = E(Err_{\mathcal{U}})$$

'Uśredniamy losowość próby uczącej, na podstawie której skonstruowano predyktor' (Hastie et al., ESL).

Wewnątrzpróbkowy błąd predykcji (in-sample error)

$Y_i^0 \sim P_{Y|\mathbf{X}=\mathbf{X}_i}$ niezależne, $i = 1, 2, \ldots, n$.

$\mathbf{Y}^0 = (Y_1^0, \ldots, Y_n^0)'$

$$Err_{in} = \frac{1}{n} \sum_{i=1}^{n} E_{\mathbf{Y}^0}(L(Y_i^0, \hat{f}(\mathbf{X}_i))|\mathcal{U})$$

Losujemy indeks $I \in \{1, \ldots, n\}$, nowa obserwacja Y_i^0 w punkcie $\mathbf{X}_i$ dla $i = I$, dla niej oczekiwany błąd predykcji.

Estymatory błędu predykcji

- estymator metodą próby testowej
- estymator metodą powtórnego podstawienia $\bar{err}$
- estymator kroswalidacyjny $\hat{Err}_{CV}$
- estymator metodą bootstrap $\hat{Err}_{boot}$

Estymator metodą próby testowej

Oprócz próby uczącej $\mathcal{U}$ dysponujemy również w pełni obserwowalną próbą testową $\mathcal{T}$ $(\mathcal{D} = \mathcal{U} \cup \mathcal{T}, \mathcal{U} \cap \mathcal{T} = 0)$, $|\mathcal{T}| = m$

$$\hat{Err}_{\mathcal{U}} = \frac{1}{m} \sum_{i=1}^{m} I\{(\mathbf{X}_i, Y_i) \in \mathcal{T} : \hat{f}(\mathbf{X}_i) \neq Y_i\}.$$

Mamy

$$E_{\mathcal{T}}(\hat{Err}_{\mathcal{U}} | \mathcal{U}) = Err_{\mathcal{U}}$$

Nieobciążone oszacowanie prawdopodobieństwa błędnej klasyfikacji.
Ale często w powyższym schemacie będziemy używać estymatora $\hat{f}_{\mathcal{D}}$ a nie $\hat{f}_{\mathcal{U}}$ - wtedy zbyt konserwatywnie z reguły oceniamy prawdopodobieństwo błędnej klasyfikacji tego klasyfikatora.

Trudny problem: $Err_{\mathcal{U}} = 0.05$,

$$2\sqrt{\frac{Err_{\mathcal{U}}(1 - Err_{\mathcal{U}})}{m}} = 2\sqrt{\frac{0.05 \times 0.95}{m}} \leqslant 0.01$$

$$m \geqslant 1900 \text{ !!}$$

Dwie możliwości zastąpienia próby testowej:

- metoda analityczna - uwzględnienie obciążenia estymatora metodą powtórnego podstawienia
- metoda repróbkowania - efektywne ponowne użycie próby uczącej

Estymator metodą powtórnego podstawienia

$$\bar{err} = \frac{1}{n}\sum_{i=1}^{n} L(\hat{f}(\mathbf{X}_i), Y_i).$$

$\bar{err}$ jest naturalnym estymatorem $Err_{\mathcal{U}}$. **Ale**: elementy próby $\mathcal{U}$ pełnią funkcję niezależnych od $\mathcal{U}$ obserwacji (X, Y). Jest estymatorem 'optymistycznym' w tym sensie, że z reguły możemy sie spodziewać, że

$$\bar{err} < Err_{\mathcal{U}}.$$

Możemy porównać

$$E_{\mathbf{Y}}(\bar{err}|\mathbf{X}) \quad \text{z} \quad E_{\mathbf{Y}}(Err_{in}|\mathbf{X})$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)', \ \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)'.$

Twierdzenie 1. Dla kwadratowej funkcji straty w (i) i 0-1 w (ii)

$$E_{\mathbf{Y}}(\bar{err}|\mathbf{X}) = E_{\mathbf{Y}}(Err_{in}|\mathbf{X}) - \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(\hat{Y}_i, Y_i|\mathbf{X}).$$

Im bardziej elastyczna metoda predykcji (silnej skorelowane $\hat{Y}_i$ i Y_i) tym większy optymizm $\bar{err}$. Człon

$$\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(\hat{Y}_i, Y_i|\mathbf{X})$$

kara kowariancyjna.

Dowód Porównujemy przy ustalonym $\mathbf{X}$ kwadraty wyrażeń

$$(Y_i - \hat{f}(\mathbf{X}_i)) \quad (1) \qquad \text{z} \qquad (Y_i^0 - \hat{f}(\mathbf{X}_i)) \quad (2)$$

$$(1) = (Y_i - f(\mathbf{X}_i)) - (\hat{f}(\mathbf{X}_i) - E(\hat{f}(\mathbf{X}_i)|\mathbf{X}) + (f(\mathbf{X}_i) - E(\hat{f}(\mathbf{X}_i)|\mathbf{X}) \\ =: A + B + C$$

$$(2) = A' + B + C \qquad A' := Y_i^0 - f(\mathbf{X}_i)$$

Liczymy $E((1)^2|\mathbf{X})$ i $E((2)^2|\mathbf{X})$

$$2E(A \times B|\mathbf{X}) \neq 2E(A' \times B|\mathbf{X})$$

$$E(A'B|\mathbf{X}) = E_{\mathbf{Y}}(E(A'B|\mathbf{X}, \mathbf{Y})) = E_{\mathbf{Y}}(BE(A'|\mathbf{X}, \mathbf{Y})) = 0$$

$$E(AB|\mathbf{X}) = \mathrm{Cov}(Y_i, \hat{Y}_i|\mathbf{X})$$

Model liniowy

Przypadek szczególny (i):

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i$$

$\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)'$ macierz eksperymentu $n \times p$ z wierszami $= \mathbf{X}_1', \mathbf{X}_2', \ldots \mathbf{X}_n'$.

Z definicji $Err_{in} = n^{-1} \sum_{i=1}^{n} E_{\mathbf{Y}^0}(Y_i^0 - \hat{f}(\mathbf{X}_i))^2 | \mathcal{U})$

$$E_{\mathbf{Y}}(Err_{in}|\mathbf{X}) = \sigma^2 + \frac{1}{n} E(||\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}||^2 | \mathbf{X})$$

Fakt Dla modelu liniowego

$$\sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i|\mathbf{X}) = p\sigma^2,$$

jeśli rząd $\mathbf{X} = p$ i $\sigma^2 = \mathrm{Var}(\varepsilon)$.

Dowód $\hat{\mathbf{Y}} = \mathbf{HX}$, gdzie $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$$\sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i|\mathbf{X}) = E((\mathbf{HY} - \mathbf{X}\beta)'\varepsilon)|\mathbf{X}) =$$

$$= E(\varepsilon'\mathbf{H}\varepsilon) = p\sigma^2.$$

Tylko w przybliżeniu prawdziwe dla innych modeli i metod !

Dla modelu liniowego i kwadratowej funkcji straty estymator $E_{\mathbf{Y}}(Err_{in}|\mathbf{X})$ ma postać

$$\bar{err} + 2\frac{p}{n}\hat{\sigma}^2 = \frac{1}{n}RSS + 2\frac{p}{n}\hat{\sigma}^2 \quad (*)$$

Jeśli $E\hat{\sigma}^2 = \sigma^2$ to $(*)$ nieobciążony estymator $E(Err_{in}|\mathbf{X})$.

W takiej sytuacji nieobciążony estymator $\frac{1}{n}E(||\mathbf{X}\hat{\beta} - \mathbf{X}\beta||^2|\mathbf{X})$ ma postać

$$C_p = \frac{1}{n}RSS + 2\frac{p}{n}\hat{\sigma}^2 - \hat{\sigma}^2.$$

Funkcja kryterialna Mallowsa.

Estymator kroswalidacyjny (rotacyjny) oczekiwanego błędu prognozy Err

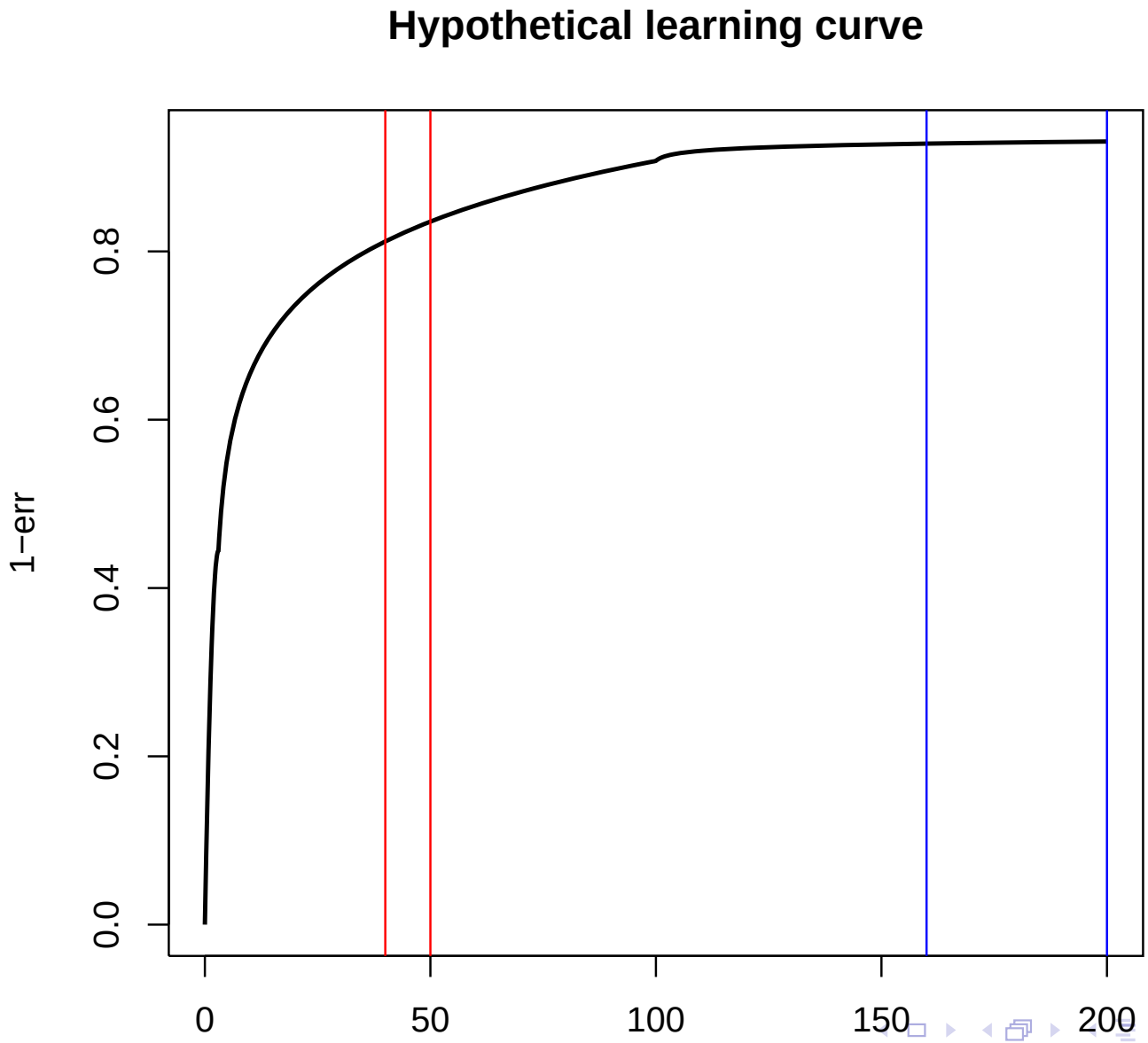
Modele (i) i (ii). Próbę $\mathcal{U}$ dzielimy na $K > 1$ w przybliżeniu równych części. $\kappa : \{1, \ldots, n\} \to \{1, \ldots, K\}$ (funkcja przynależności do elementu podziału). $\hat{Y}_j^{-i} = \hat{f}^{-i}(\mathbf{X}_j)$ prognoza Y_j na podstawie próby z usuniętą i-tą częścią.

$$CV = \frac{1}{n} \sum_{j=1}^{n} L(Y_j, \hat{Y}_j^{-\kappa(j)}).$$

Y_j prognozowana na podstawie próby niezależnej od Y_j zgodnie z definicją Err.
Co estymujemy? $\mathcal{U}^{-i}$: próba z usuniętą i-tą częścią.
Oszacowanie błędu

$$CV = \frac{1}{K} \sum_{i=1}^{K} \hat{E}(L(\mathbf{X}, Y)|\mathcal{U}^{-i}) = \hat{E}(Err_{\mathcal{U}^{-1}})$$

Nie estymujemy **warunkowego** błędu klasyfikatora $\hat{d}_{\mathcal{U}}$ dla danej próby $\mathcal{U}$ tylko **oczekiwany** błąd klasyfikatora $\hat{d}_{\mathcal{U}^{-1}} = Err_{(n(K-1)/K)}$

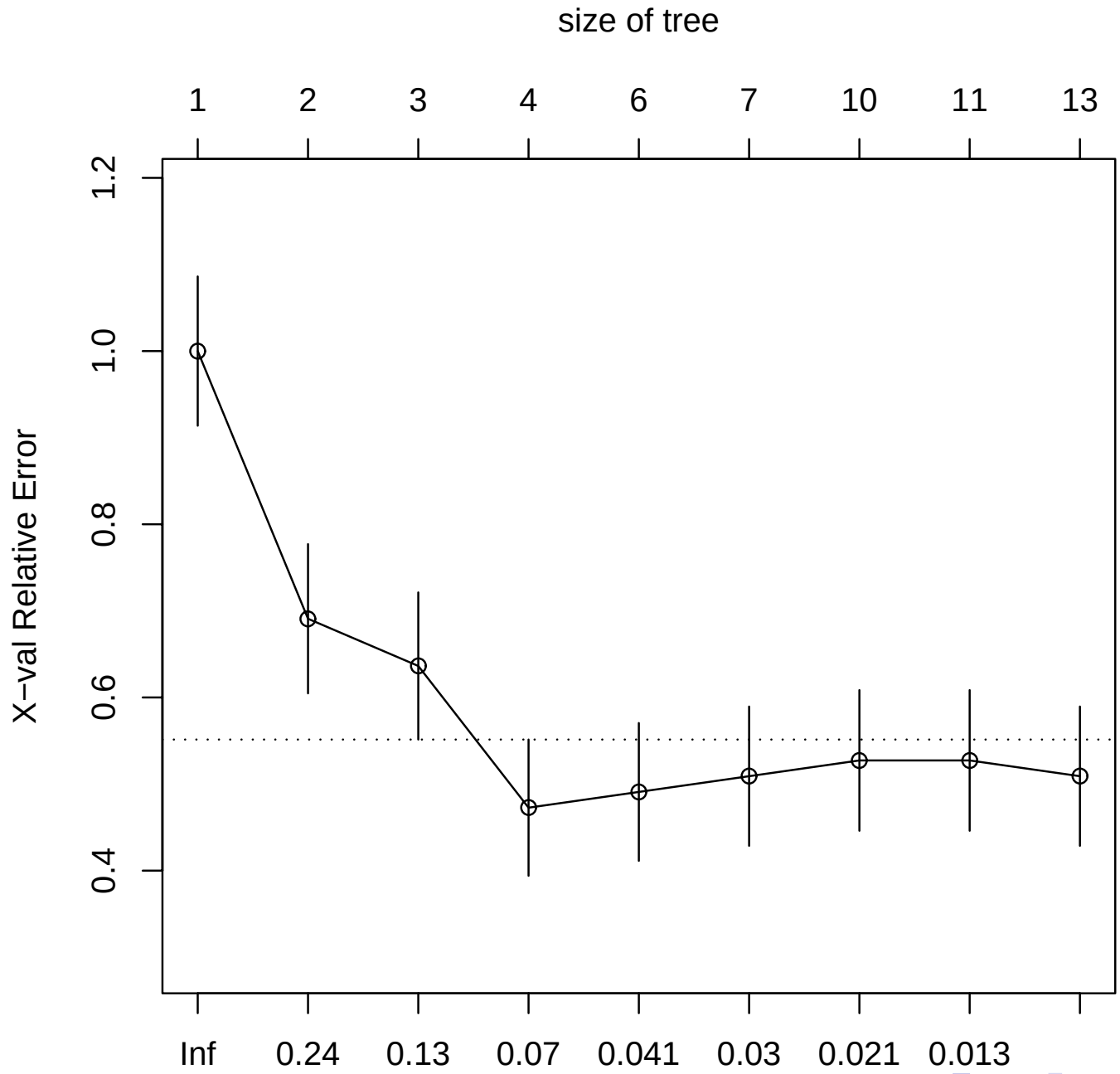


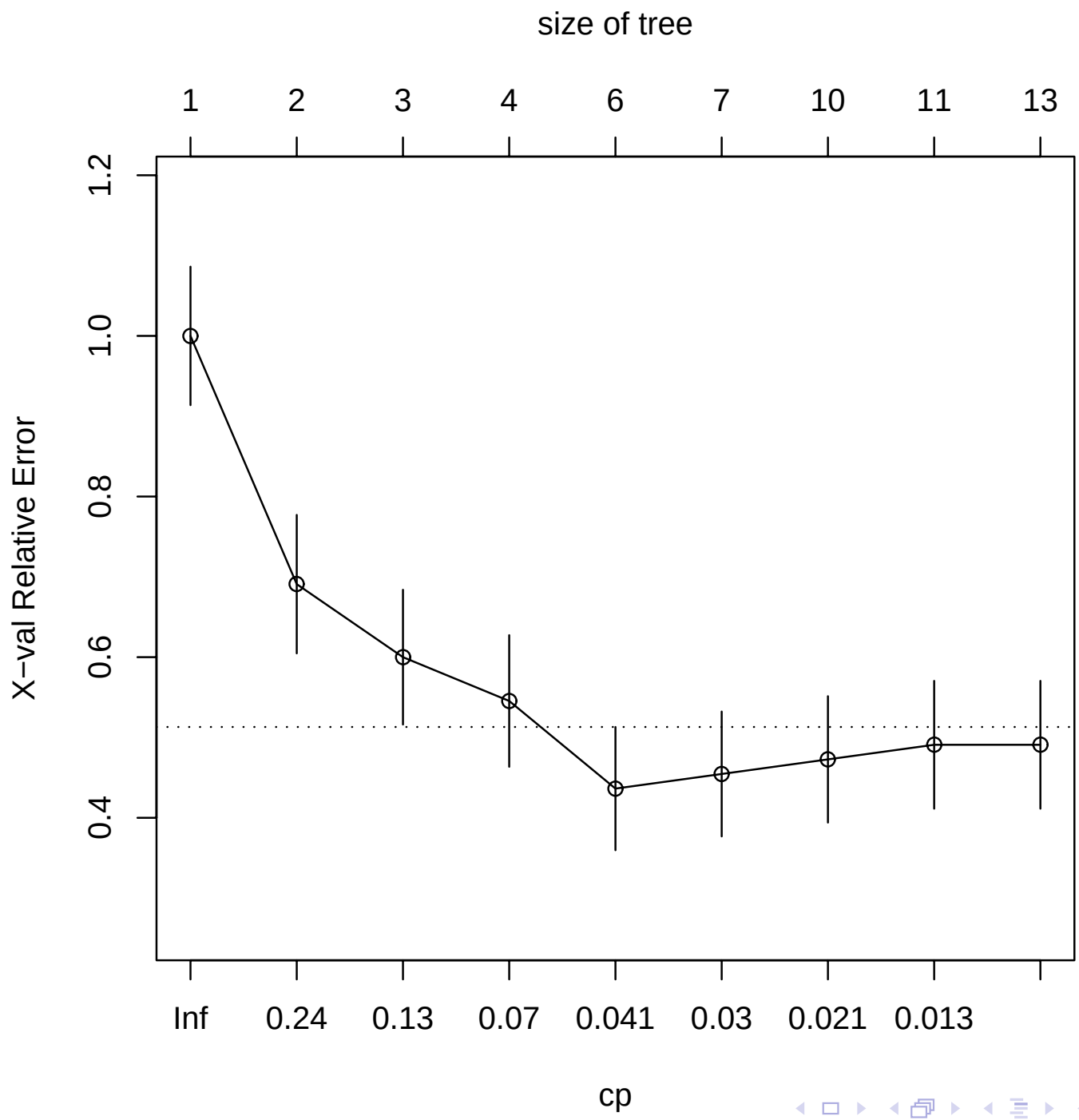
Uwagi o estymatorze kroswalidacyjnym

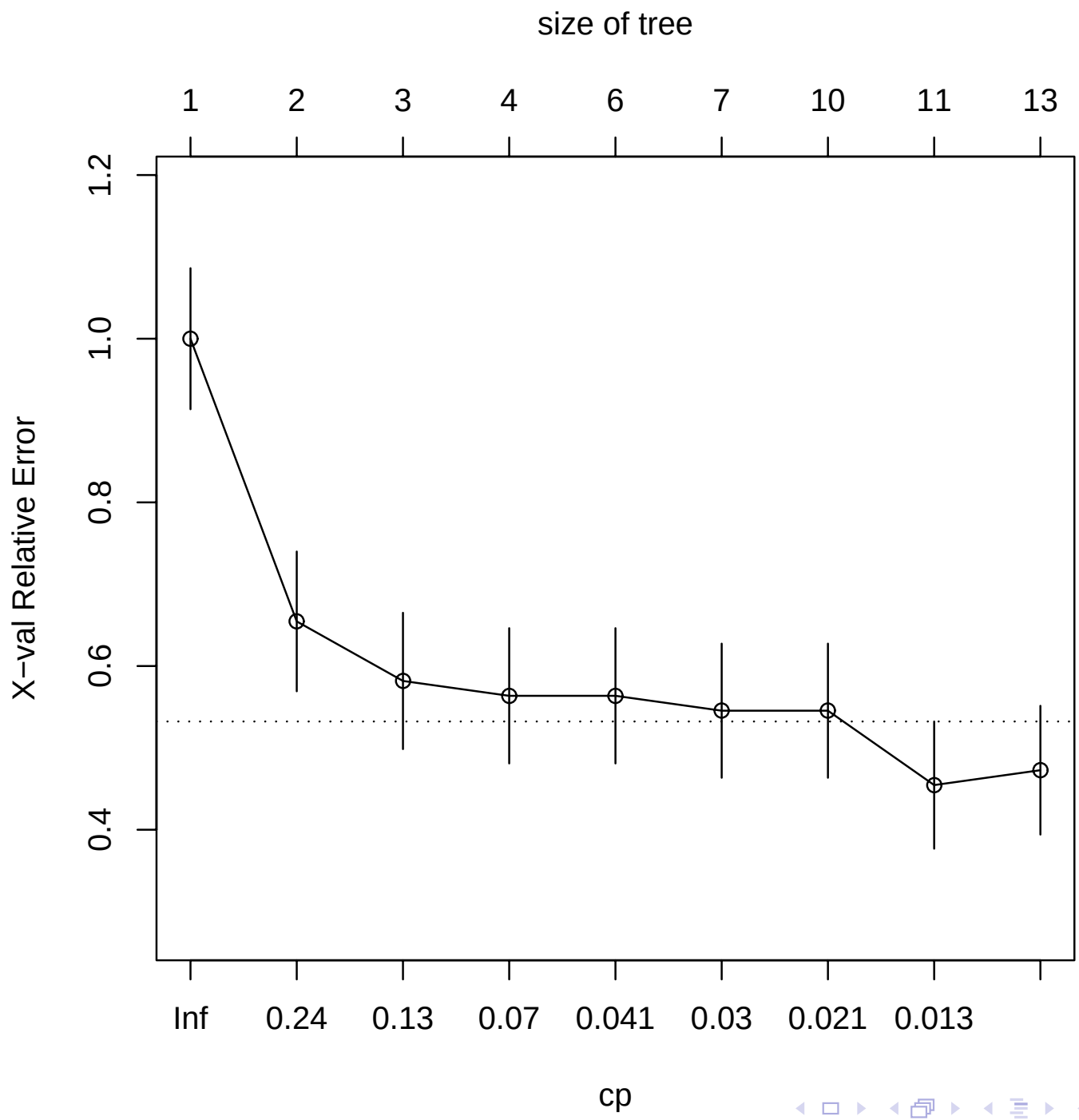
1. Liczba podziałów $K = 5, 10$ i $K = n$. (leave-one out). W ostatnim przypadku najmniejsze obciążenie estymatora (jako oszacownania Err_n dla liczności próby n), największa wariancja (skorelowanie czynników w definicji CV.

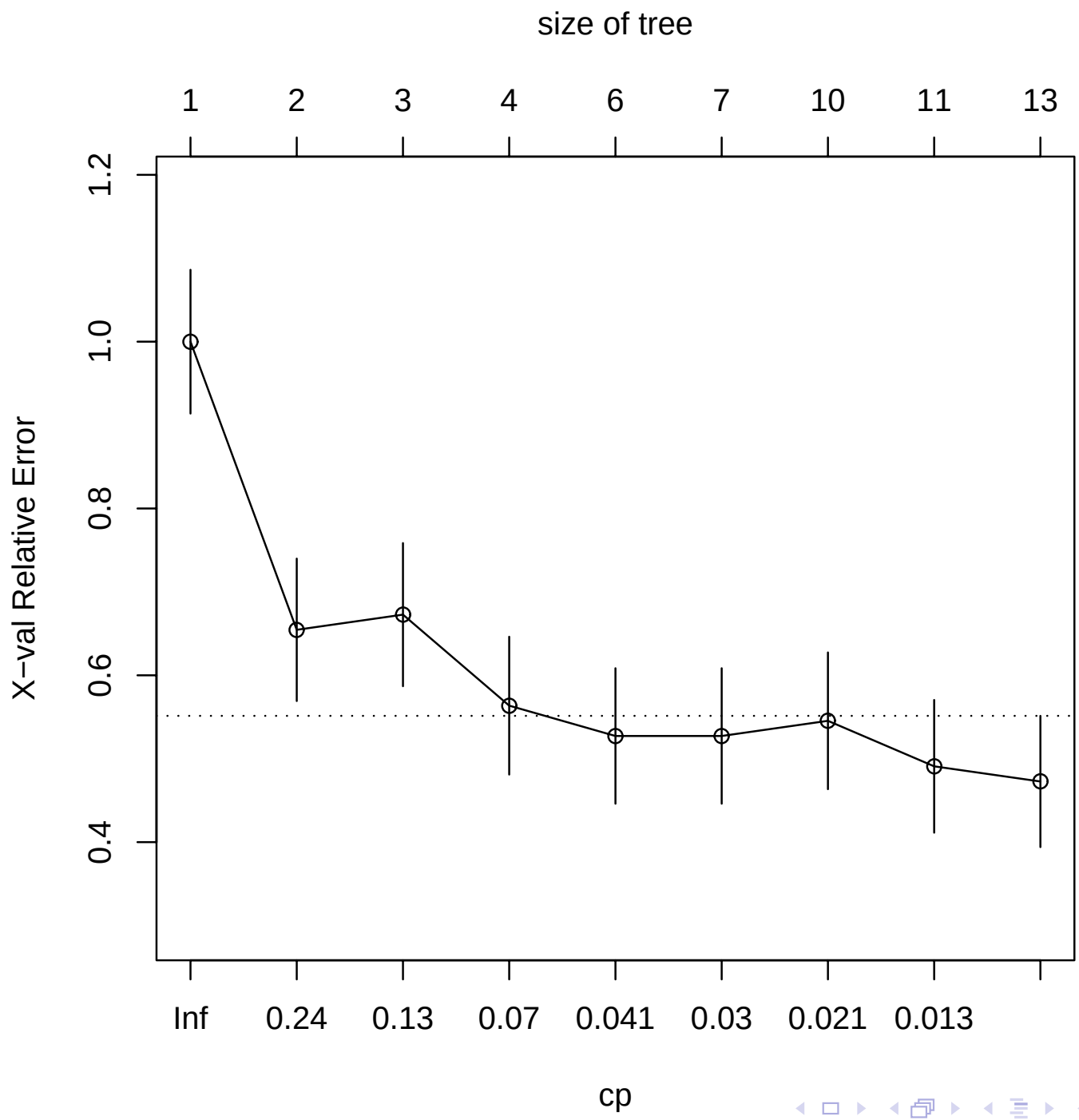
2. Obciażenie estymatora kroswalidacyjnego jako estymatora Err_n dla $K = 5, 10$ może być znaczne.

3. Niestabilność CV związana z losowością podziału na K części.




size of tree
1
2
3
4
6
7
10
11
13
X−val Relative Error
0.4
0.6
0.8
1.0
1.2
Inf
0.24
0.13
0.07
0.041
0.03
0.021
0.013
cp


size of tree
1
2
3
4
6
7
10
11
13
X−val Relative Error
0.4
0.6
0.8
1.0
1.2
Inf
0.24
0.13
0.07
0.041
0.03
0.021
0.013
cp


size of tree
1
2
3
4
6
7
10
11
13
X−val Relative Error
0.4
0.6
0.8
1.0
1.2
Inf
0.24
0.13
0.07
0.041
0.03
0.021
0.013
cp

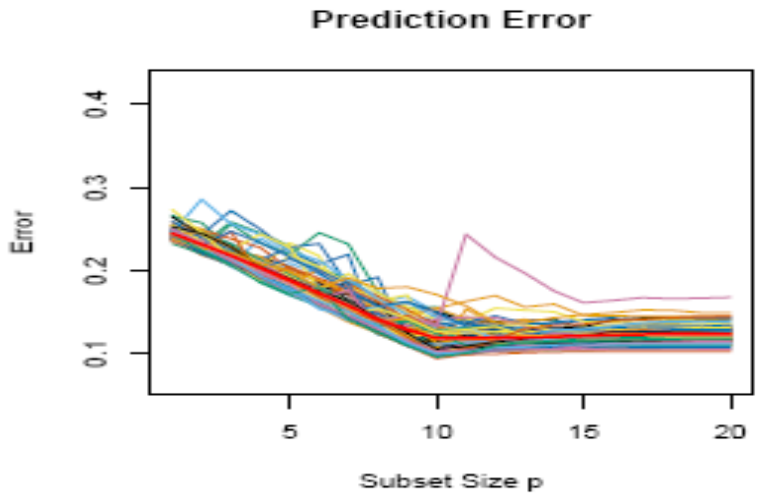
CV jako metoda estymacji $Err_{\mathcal{U}}$ i Err

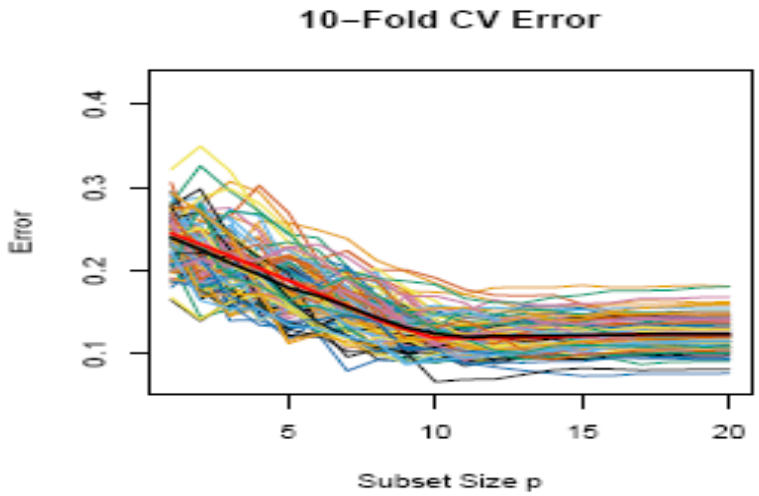
Przykład (ESL, rozdział VII)

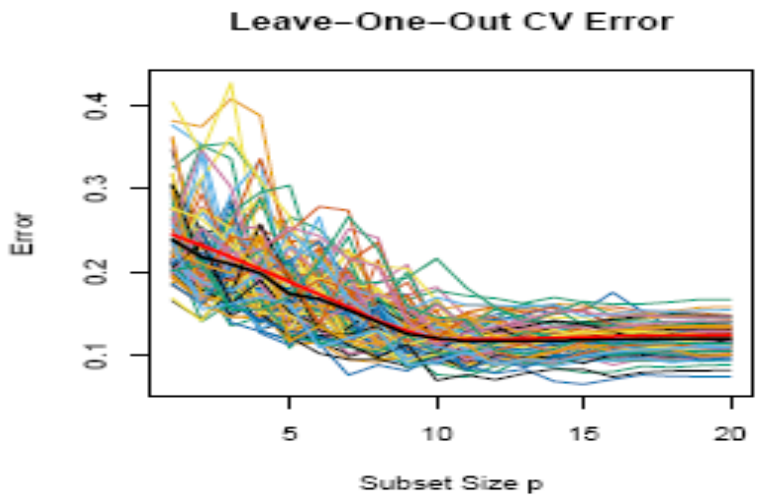
$(Y, \mathbf{X}) \in \{0,1\} \times R^{20}$: $Y = 1$ jeśli $\sum_{i=1}^{10} X_i > 5$, 0 w przeciwnym przypadku.
CV-10 i CV leave-one-out jak estymatory $Err_{\mathcal{U}}$ i Err liczone dla 100 prób uczących ($n = 80$).
Metoda estymacji: regresji liniowa dla najlepszego podzbioru predyktorów danej liczności ($\operatorname{argmin}_{M:|M|=p} RSS_M$)

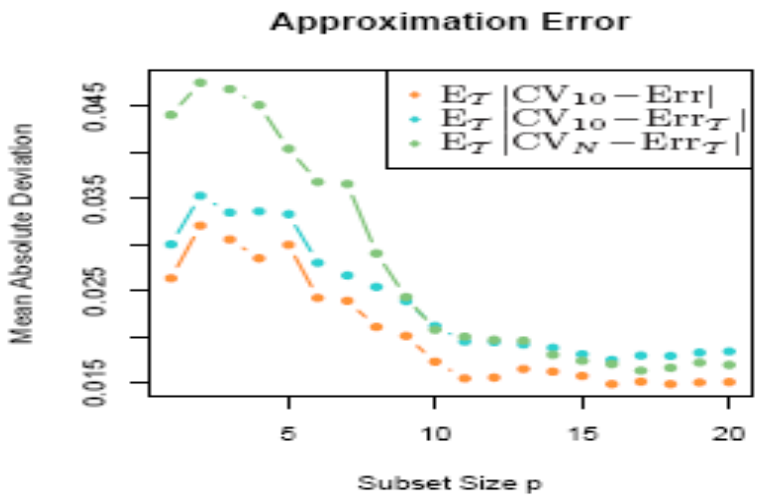
Czerwona linia - Err

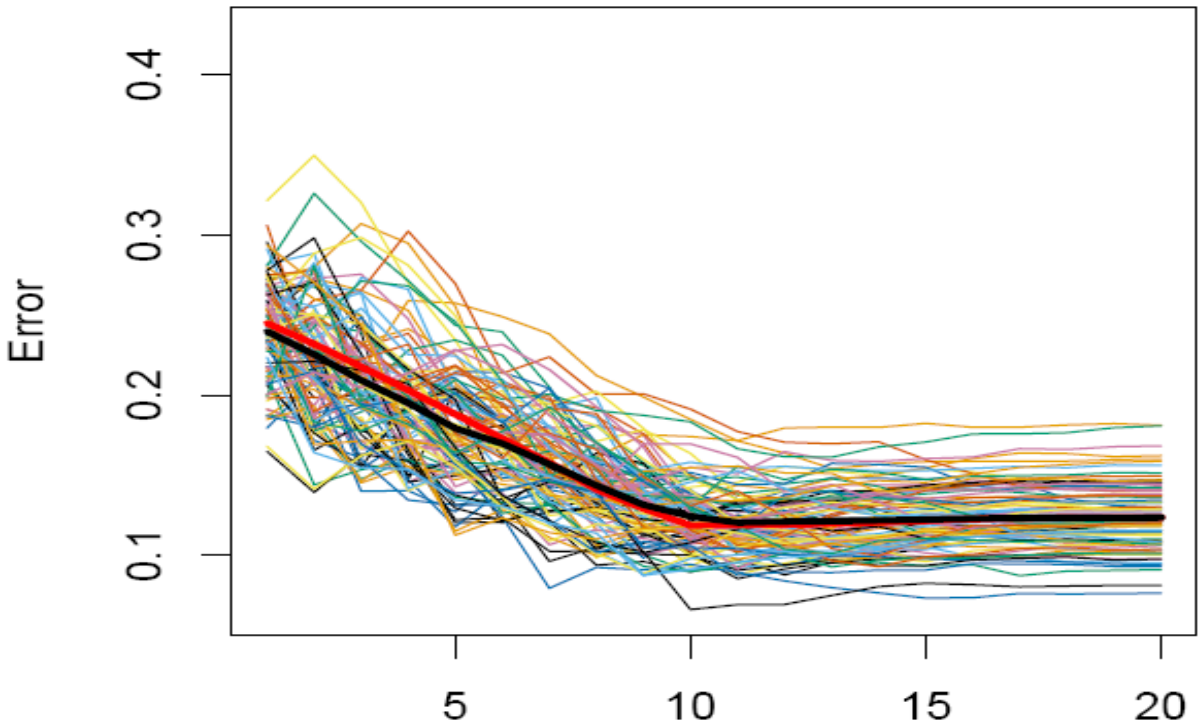
Czarna linia - ECV_{10} i ECV_n

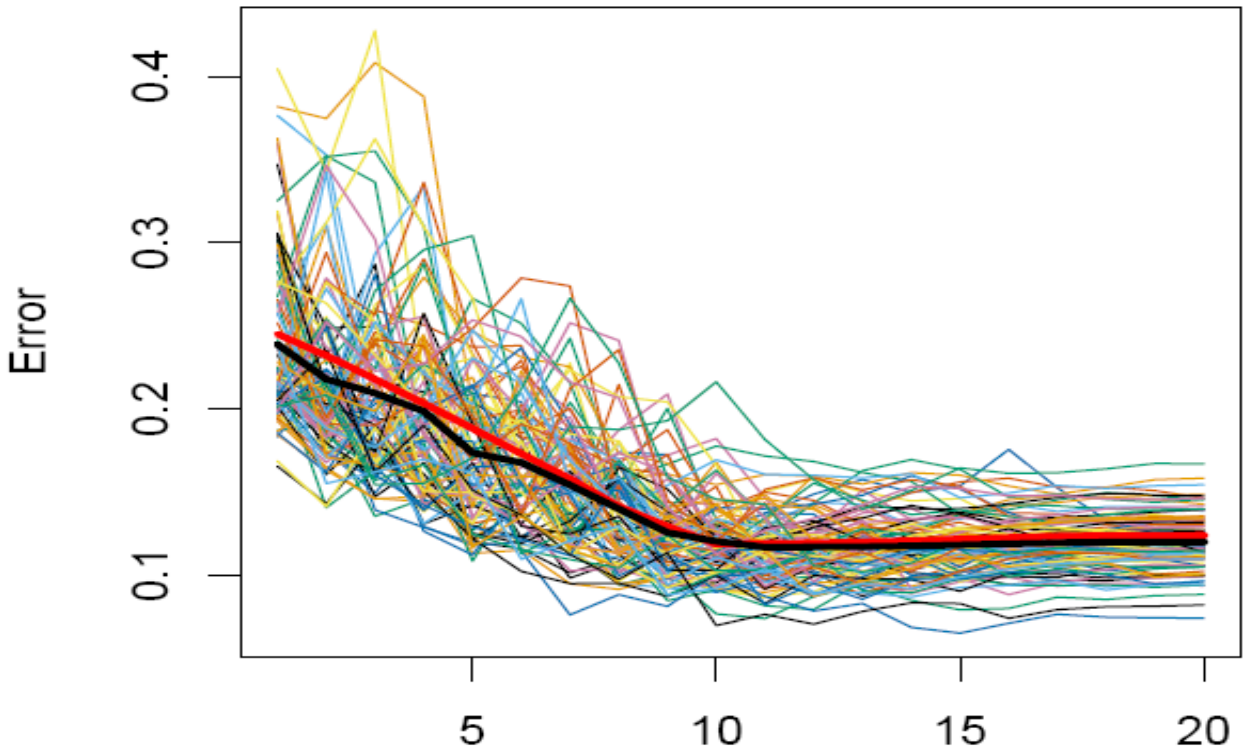

Prediction Error
Error
0.1
0.2
0.3
0.4
5
10
15
20
Subset Size p

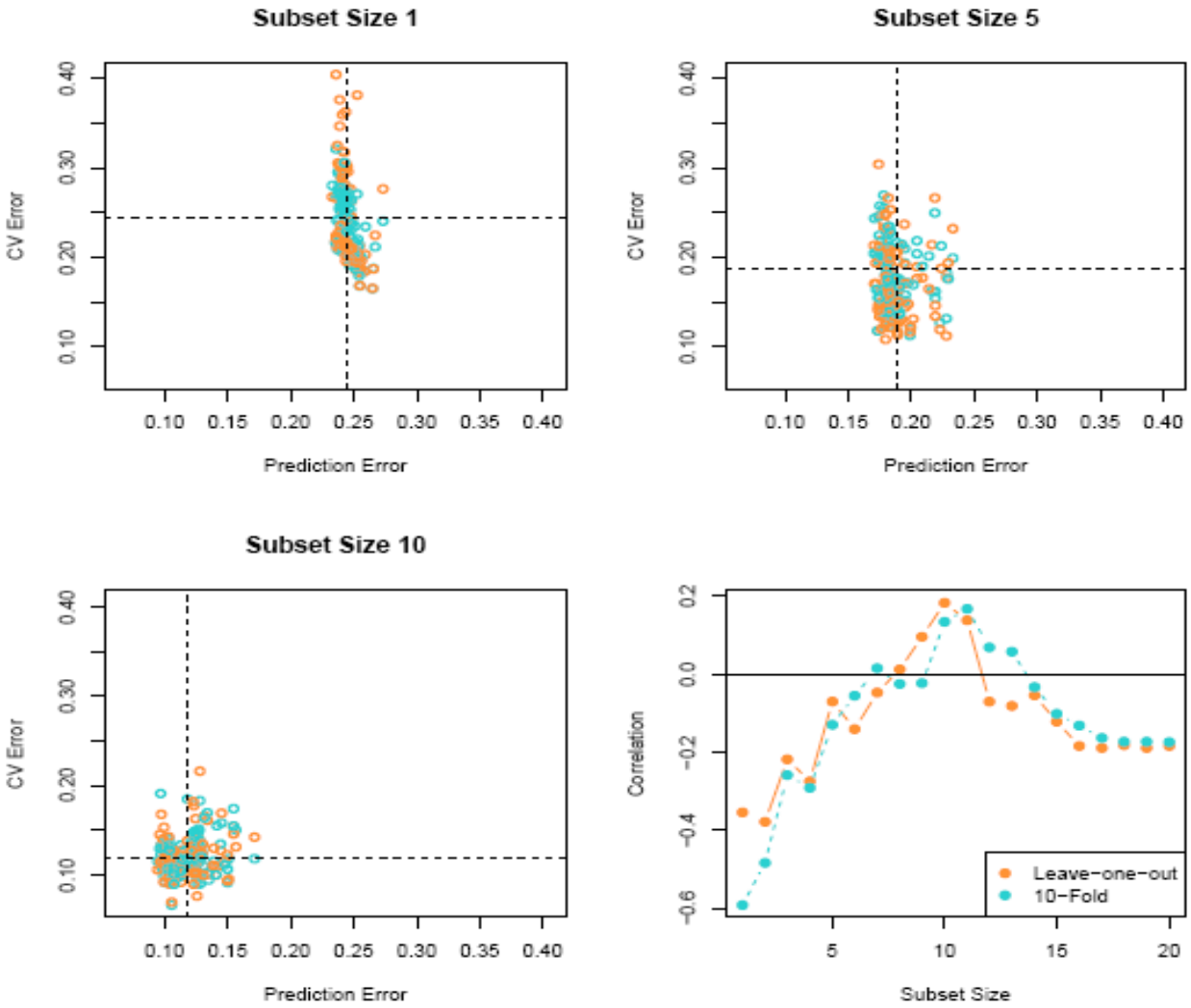

10−Fold CV Error
Error
0.1
0.2
0.3
0.4
5
10
15
20
Subset Size p

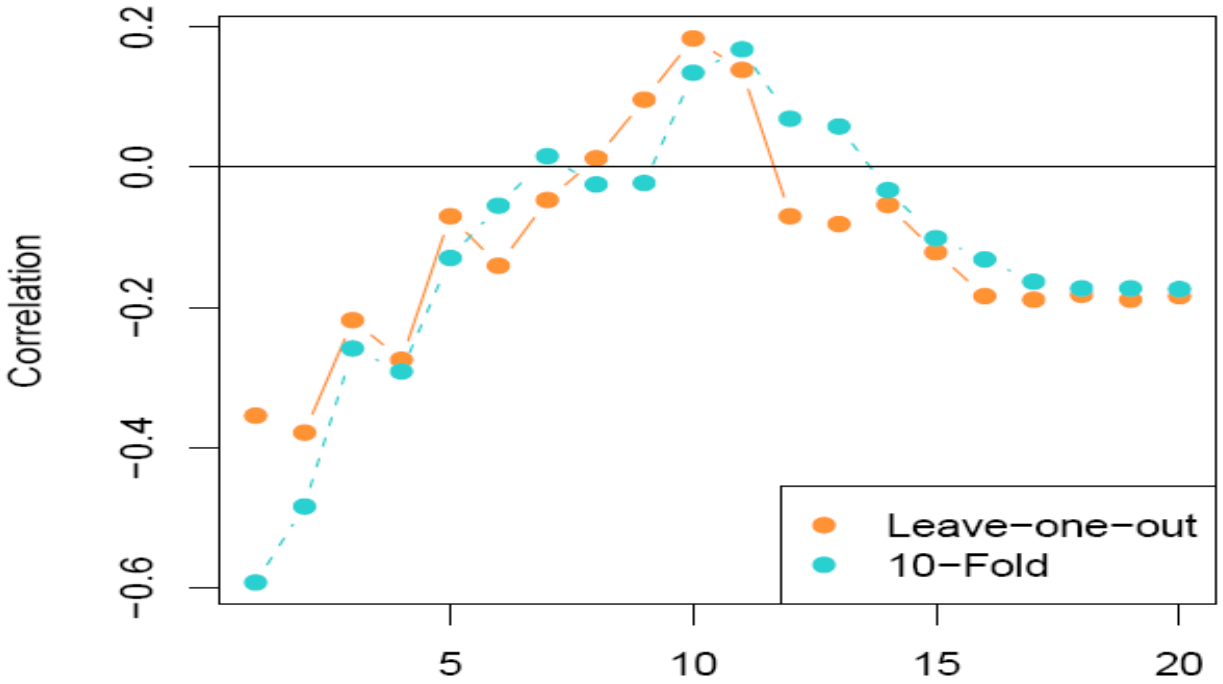

Leave−One−Out CV Error
Error
0.1
0.2
0.3
0.4
5
10
15
20
Subset Size p


Approximation Error
Mean Absolute Deviation
0.015
0.025
0.035
0.045
$E_{\mathcal{T}}|CV_{10} - Err|$
$E_{\mathcal{T}}|CV_{10} - Err_{\mathcal{T}}|$
$E_{\mathcal{T}}|CV_{N} - Err_{\mathcal{T}}|$
5
10
15
20
Subset Size p


Error
0.1
0.2
0.3
0.4
5
10
15
20


Error
0.1
0.2
0.3
0.4
5
10
15
20


Subset Size 1
CV Error
Prediction Error
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Subset Size 5
CV Error
Prediction Error
Subset Size 10
CV Error
Prediction Error
Correlation
Subset Size
5
10
15
20
−0.6
−0.4
−0.2
0.0
0.2
Leave−one−out
10−Fold


Correlation
0.2
0.0
−0.2
−0.4
−0.6
5
10
15
20
Leave−one−out
10−Fold