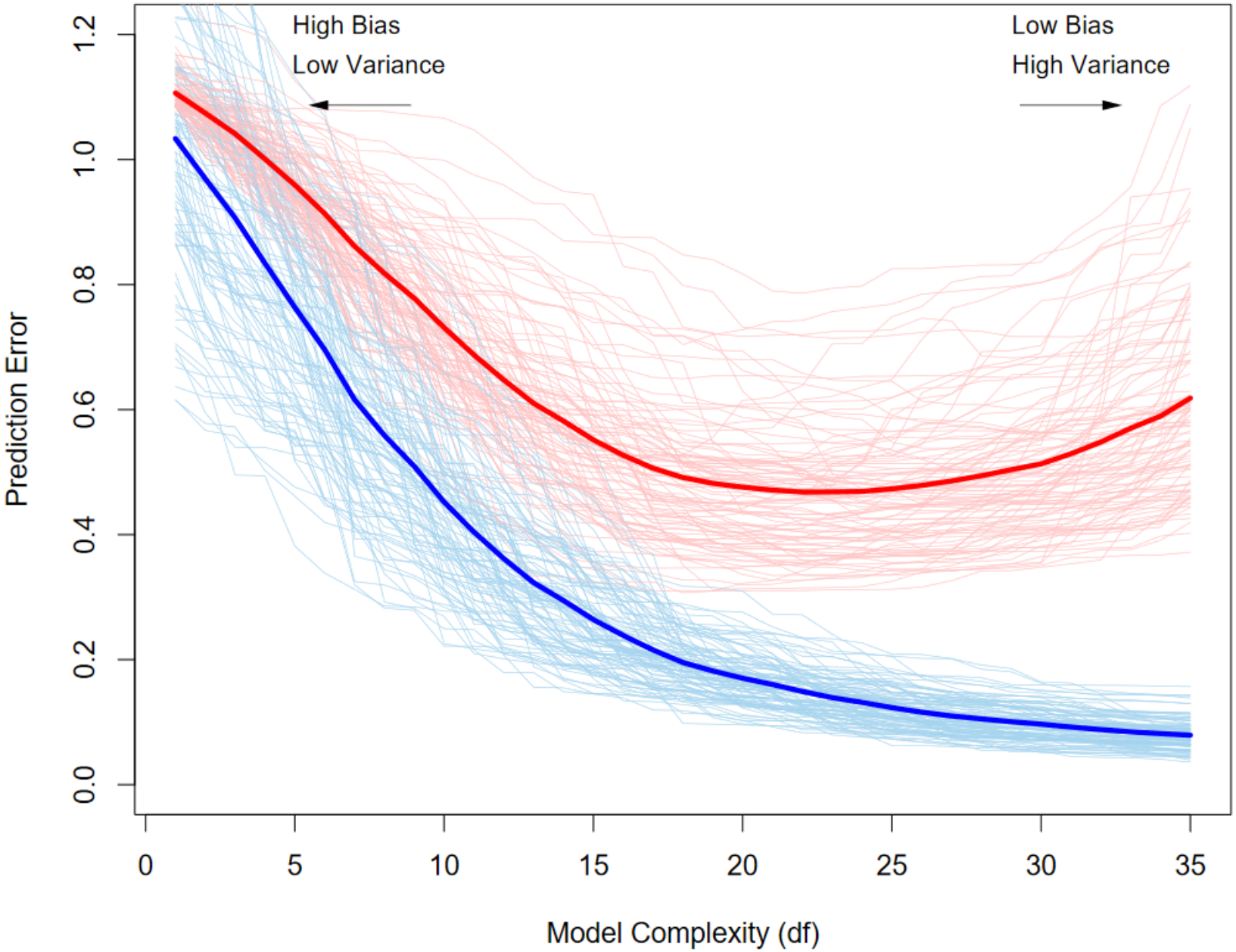
Spis treści

Obciążenie, wariancja i złożoność modelu

Mamy zmienną objaśnianą Y, macierz zmiennych objaśniających X, model predykcyjny $\hat{f}(X)$, zbiór treningowy $\mathcal{T}_1$ oraz zbiór testowy $\mathcal{T}_2$.
Najpopularniejszymi funkcjami straty ($L(Y, \hat{f}(X))$) są:

1. $(Y - \hat{f}(X))^2$ - błąd kwadratowy
2. $|Y - \hat{f}(X)|$ - błąd bezwzględny

Zdefiniujmy błąd treningowy: $\hat{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i))$, gdzie x_i pochodzą ze zbioru treningowego. Błąd testowy: $Err_{\mathcal{T}_2} = E[L(Y, f(\hat{X})|X \subset \mathcal{T}_2)]$. Oraz oczekiwany błąd predykcji $Err = E[Err_{\mathcal{T}_2}]$


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
35
Model Complexity (df)

Ta sama idea w modelach klasyfikacyjnych

Mamy kategoryczną zmienną objaśnianą G. Tutaj nasz model zwraca prawdopodobieństwa, że $p_k(X) = P(G = k|X)$ dla $k = 1, \ldots, K$. Wynikiem modelu jest k o największej wartości z tych prawdopodobieństw: $argmax_k \hat{p_k}(X)$. Najpopularniejszymi funkcjami straty ($L(G, \hat{G}(X))$) są:

1. $I(G, \hat{G}(X))$ - strata 0-1
2. $-2\sum_{k=1}^{K} I(G = k) log \hat{p_k}(X) = -2log \hat{p_G}(X)$ - -2 x logwiarogodność

Błędy treningowe i testowe tak samo, tylko na innych funkcjach straty.

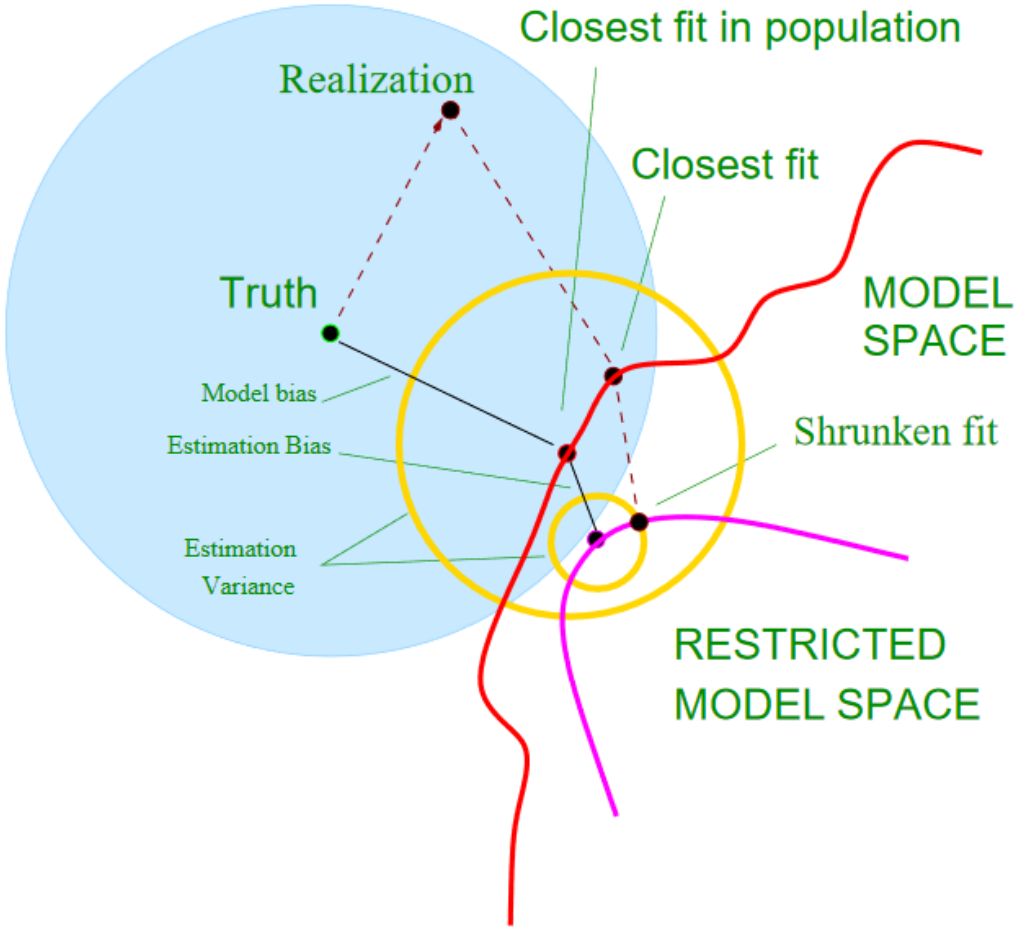
Podział zbioru zmiennych objaśniających

Sytuacja idealna (duży zbiór danych):

Bias-variance tradeoff

Zakładamy, że $Y = f(X) + \epsilon$, gdzie $E[\epsilon] = 0$ i $Var(\epsilon) = \sigma_\epsilon^2$. Możemy teraz rozpisać oczekiwany błąd predykcji (na kwadratowej funkcji straty) jako:
$Err(Y, \hat{f}(X)) = E[(Y - \hat{f}(X))^2] =$
$\sigma_\epsilon^2 + (E\hat{f}(X) - f(X))^2 + E[(\hat{f}(X) - E\hat{f}(X))^2] = \sigma_\epsilon^2 + Bias^2 + Var$
W szczególności dla modelu liniowego mamy:
$Bias^2 = (f(X) - X^T(X^TX)^{-1}X^TY$ i $Var = ||X(X^TX)^{-1}X||^2\sigma_\epsilon^2$

Bias-variance tradeoff na wykresie



Optymizm błędu treningowego

Niech $ERR_{\mathcal{T}} = E_{X^0,Y^0}[L(Y^0, \hat{(f)}(X^0)|\mathcal{T}]$, punkt (X^0, Y^0) pochodzi z rozkładu (X, Y). Niech $ERR = E_{\mathcal{T}}[ERR_{\mathcal{T}}]$. Zazwyczaj empiryczny błąd treningowy to $\hat{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) < ERR_{\mathcal{T}}$.

Zdefiniujmy błąd in-sample $Err_{in} = \frac{1}{n}\sum_{i=1}^{n} E_{Y^0}[L(Y_i^0, \hat{f}(x_i))|\mathcal{T}]$

Definiujemy optymizm jako $opt = Err_{in} - \hat{err}$

Oraz średni optymizm: $\omega = E_Y[opt]$.

W ogólności dla większości funkcji straty (między innymi dla straty kwadratowej, 0-1 itd.) $\omega = \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$.

I na koniec zyskujemy ważną równość:

$$E_Y[Err_{in}] = E_Y[\hat{err}] + \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$$

Co w przypadku, gdy mamy zależność liniową pomiędzy objaśnianą i objaśniającymi upraszcza się do:
$E_Y[Err_{in}] = E_Y[\hat{err}] + \frac{2d}{n}\sigma_\epsilon^2$ (gdzie d to liczba zmiennych użytych w modelu)

Metody estymacji błędu in-sample

Uogólniona formułą na estymator błędu in-sample: $\hat{Err}_{in} = \hat{opt} + \hat{err}$

Statystyka C_p Mallow'a: $C_p = \hat{err} + \frac{2d}{n} \cdot \hat{\sigma}_\epsilon^2$, np. $\hat{\sigma}_\epsilon^2 = \frac{||Y - X\hat{Beta}^{LS}||^2}{n-d}$

AIC: $-2E[logP_{\hat{\theta}}(Y)] \approx \frac{-2}{n} E[loglik] + \frac{2d}{n}$, gdzie $loglik = \sum_{i=1}^{n} P_{\hat{\theta}}(y_i)$

AIC: $\frac{-2}{n} loglik + \frac{2d}{n}$

Kryterium BIC

BIC: $BIC = -2loglik + log(n) \cdot d$ i
AIC: $AIC = -2loglik + 2d$
Dla $n > e^2 \approx 7.4$ BIC częściej odrzuca modele.

MDL (minimum description length)

Message	z_1	z_2	z_3	z_4
Code	0	10	110	111

Twierdzenie Shannona: $I_i = -\log_2 P(z_i)$

Wtedy zachodziłaby następująca zależność $-\sum_{i=1}^{n} P(z_i) log_2 P(z_i) \leqslant E[I]$

Dla $P(z_1) = 1/2, P(z_2) = 1/4, P(z_3) = P(z_4) = 1/8$ optymalnym kodowaniem byłoby kodowanie z poprzedniego slajdu.

Dla ciągłych: $-\int P(z) log_2 P(z) dz$

Od teraz zmieniamy notację zamiast $log_2 P(Z)$ będzie po prostu $logP(Z)$

MDL przy wyborze modelu

Mamy model M z parametrami θ. Mamy $P(Y|M,\theta,X)$. Wtedy dlugością kodowania jest $I = -logP(Y|M,\theta,X) - logP(\theta|M)$
Przykład: mamy Y z $N(\theta,\sigma^2)$, θ z $N(0,1)$ i nie mamy zmiennych wejściowych. Wtedy powyższy zapis uprości się do
$I = c + log\sigma + \frac{(Y-\theta)^2}{\sigma^2} + \frac{\theta^2}{2}$

Kroswalidacja

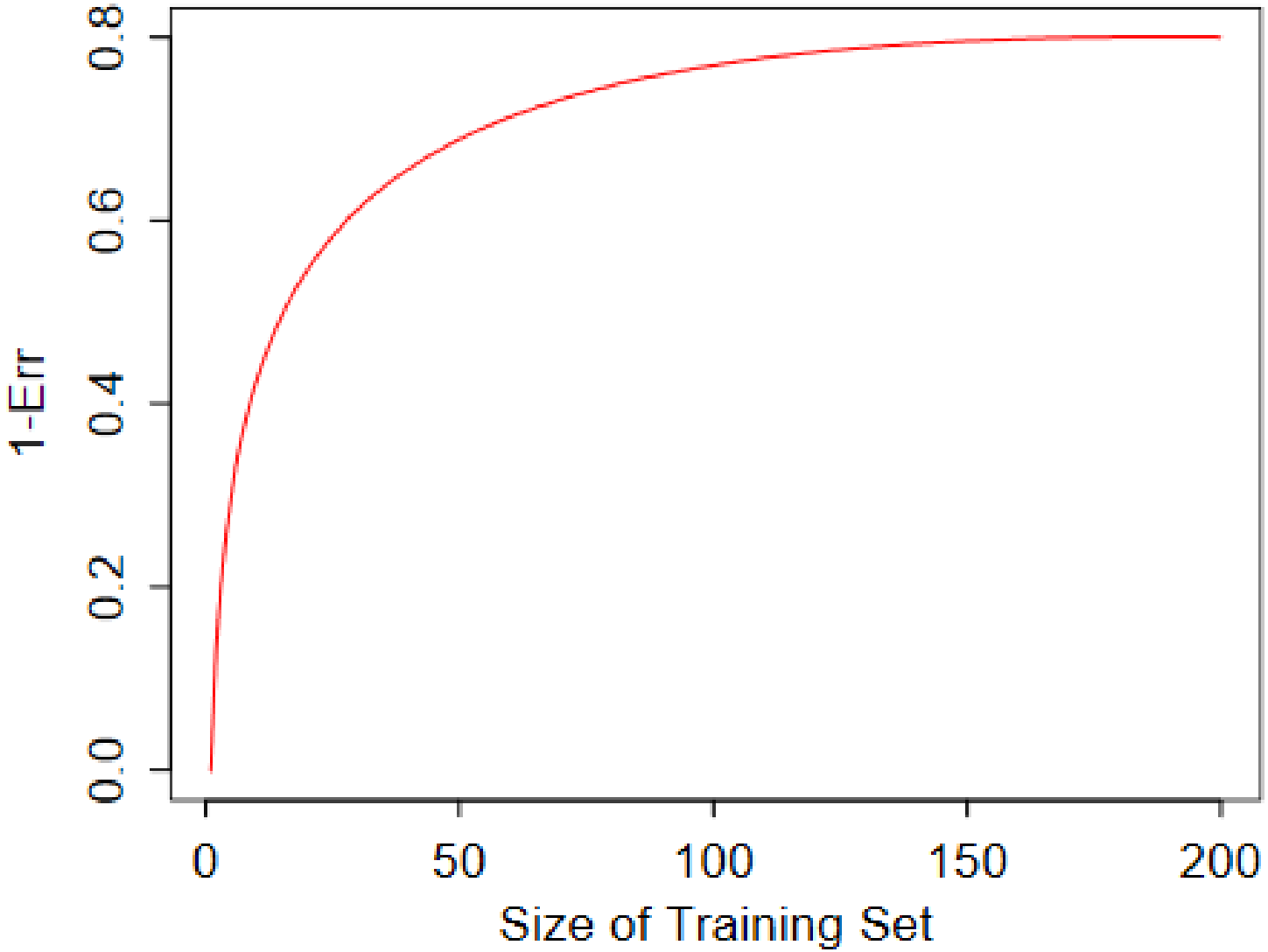
Niech $\mathcal{K} : \{1, \ldots, n\} \to \{1, \ldots, k\}$. Mamy $\hat{f}^i(x)$

$$CV(\hat{f}) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, \hat{f}^{\mathcal{K}(j)}(x_j))$$

Leave-one-out cv: $\mathcal{K}(i) = i$. Można ten zapis rozwinąć, ze względu na wybór modelu z wektorem parametrów α, wtedy:

$$CV(\hat{f,\alpha}) = \frac{1}{n} \sum_{j=1}^{n} L(y_j, \hat{f}^{\mathcal{K}(j)}(x_j, \alpha))$$

Jakość estymatora błędu predykcji

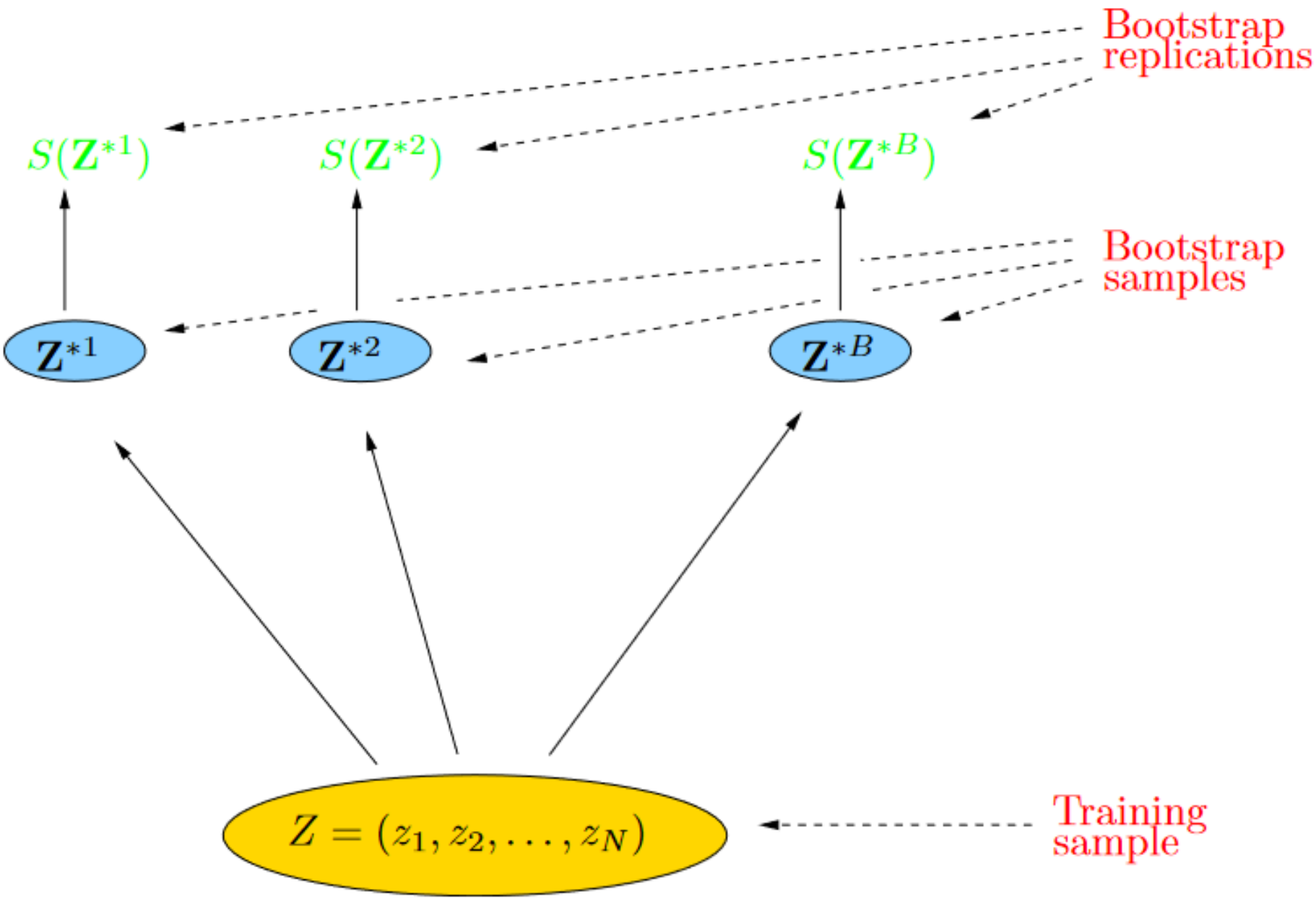


Uogólniona kroswalidacja

Mamy jakieś liniowe dopasowanie: $\hat{y} = Sy$.
Wtedy

$$GCW(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{f}(x_i)}{1 - s_{ii}})^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{f}(x_i)}{1 - trace(S)/n})^2$$

Bootstrap


Bootstrap replications
$S(\mathbf{Z}^{*1})$
$S(\mathbf{Z}^{*2})$
$S(\mathbf{Z}^{*B})$
Bootstrap samples
$\mathbf{Z}^{*1}$
$\mathbf{Z}^{*2}$
$\mathbf{Z}^{*B}$
$Z = (z_1, z_2, \ldots, z_N)$
Training sample

$$\hat{var}(\hat{S}(Z)) = \frac{1}{B-1}\sum_{b=1}^{B}(S(Z^{*b}) - \overline{S^*})^2$$

Estymator błędu predykcji (gorszy): $\hat{Err}_{boot} = \frac{1}{n}\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, f^{*b}(\hat{x_i}))$.

Przykład: prawdziwy błąd predykcji 0.5. Za pomocą bootsrapu: $P(\text{obserwacja } i \in b) = 1 - (1 - \frac{1}{n})^n \approx 1 - e^{-1} = 0.632$, więc błąd predykcji z bootstrapu to $1/2 * 0.368$ czyli zaniżony...

Estymator błędu predykcji (lepszy): $\hat{Err}_{boot} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\#C^{-i}}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$