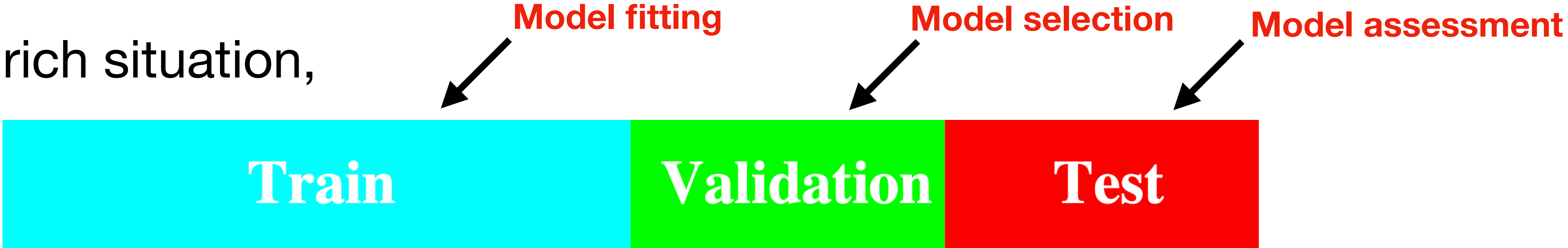


7. Model Assessment and Selection

Index

- **Introduction**
- **Bias, Variance and Model Complexity**
- **Estimates of In-Sample Error (C_p , AIC, BIC)**
- **Directly estimates Extra-Sample Error (Cross validation, Bootstrap method)**
- **Conditional or Expected Test Error?**

Introduction

- **Model selection:** Estimating the performance of different models in order to choose the best one.
- **Model assessment:** having chosen a final model, estimation its prediction error on new data
- If we are in a data-rich situation,

The diagram illustrates the data splitting process for a data-rich situation. It features a horizontal bar divided into three colored segments: cyan for 'Train', green for 'Validation', and red for 'Test'. Above the 'Train' segment is the label 'Model fitting' with a black arrow pointing to the segment. Above the 'Validation' segment is the label 'Model selection' with a black arrow pointing to the segment. Above the 'Test' segment is the label 'Model assessment' with a black arrow pointing to the segment.
- The methods in this chapter are designed for situations where there is **insufficient** data to split it into three parts.

Introduction

- Consider first the case of a quantitative response.
- Y : target variable, X : input vector, \hat{f} : prediction model estimated from training set $\mathcal{T} = \{(x_i, y_i) : i = 1, \dots, N\}$ and typical loss functions are squared error, and absolute error.
- **(Test error or generalization error)** $Err_{\mathcal{T}} = \mathbb{E} \left[L(Y, \hat{f}(X)) \mid \mathcal{T} \right]$
- **(Expected prediction error)** $Err = \mathbb{E} \left[Err_{\mathcal{T}} \right]$ is more amenable to statistical analysis.
- **(Training error)** $e\bar{r}r = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$

Introduction

- For qualitative or categorical response $G \in \mathcal{G} = \{1, \dots, K\}$
- Model $p_k(X) = P(G = k | X)$ and $\hat{G}(X) = \operatorname{argmax}_k \hat{p}_k(X)$. Typical loss functions are 0-1 loss, and $-2 \times \log$ -likelihood (deviance).
 - Ex. $P_{\theta(X)}(Y)$ is the density of Y , then $L(Y, \theta(X)) = -2 \log P_{\theta(X)}(Y)$
 - The “-2” makes the log-likelihood loss for the Gaussian distribution match squared-error loss.
- **(Misclassification error)** $Err_{\mathcal{T}} = \mathbb{E} \left[L(G, \hat{G}(X) | \mathcal{T}) \right]$
- **(Expected misclassification error)** $Err = \mathbb{E} \left[Err_{\mathcal{T}} \right]$

Bias, Variance and Model Complexity

- Assume that $Y = f(X) + \epsilon$ where $\mathbb{E}(\epsilon) = 0$, $Var(\epsilon) = \sigma_\epsilon^2$, using squared error loss:

- **(Expected prediction error)**

$$Err(x_0) = \mathbb{E} \left[\left(Y - \hat{f}(x_0) \right)^2 \mid X = x_0 \right]$$

$$= \mathbb{E} \left[\left(Y - f(x_0) + f(x_0) - \hat{f}(x_0) \right)^2 \mid X = x_0 \right]$$

$$= \mathbb{E} \left[\left(Y - f(x_0) \right)^2 + \left(f(x_0) - \mathbb{E}\hat{f}(x_0) + \mathbb{E}\hat{f}(x_0) - \hat{f}(x_0) \right)^2 \mid X = x_0 \right]$$

$$= \sigma_\epsilon^2 + \left\{ f(x_0) - \mathbb{E}\hat{f}(x_0) \right\}^2 + Var(\hat{f}(x_0))$$

Irreducible
Error at x_0

(Bias)²

Variance

Bias, Variance and Model Complexity

- Consider a linear model fit $\hat{f}_p(x) = x^T \hat{\beta}$, where β is fit by least squares.

- (Bias term):**
$$\mathbb{E} \left[f(x_0) - \mathbb{E} \hat{f}_p(x_0) \right]^2 = \mathbb{E} \left[f(x_0) - x_0^T \beta \right]^2 + \mathbb{E} \left[x_0^T \beta - \mathbb{E} x_0^T \hat{\beta} \right]^2$$
$$= \text{Ave} [\text{Model Bias}]^2 + \text{Ave} [\text{Estimation Bias}]^2$$

The error between the true function
and the best-fitting linear approximation

The error between the average estimate
and the best-fitting linear approximation

→ 0 in ordinary least squares

- (Variance term):**
$$\text{Var} \left[x_0^T \hat{\beta} \right] = x_0^T \text{Var} \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \right] x_0$$
$$= \sigma_\epsilon^2 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 = \sigma_\epsilon^2 \sum_{i=1}^N \frac{1}{d_i^2} z_i^2$$

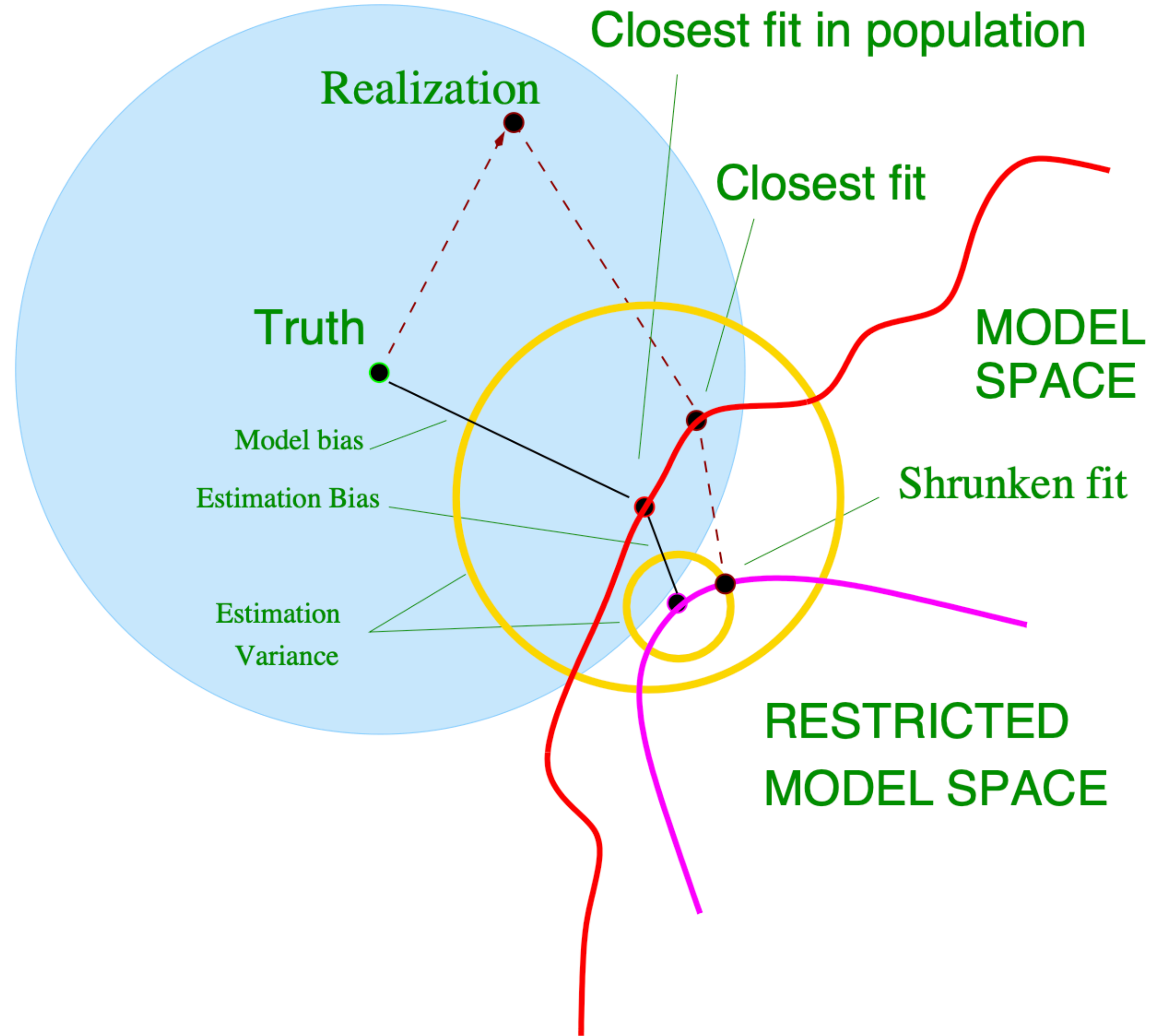
where $\mathbf{z} = (z_1, \dots, z_N) = V^T x_0$, $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T$, $\mathbf{D} = \text{diag}(d_1, \dots, d_N)$

Bias, Variance and Model Complexity

- For ridge regression fit \hat{f}_α , where $\hat{\beta}_\alpha = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$, $\alpha > 0$
- Since $\mathbb{E}\hat{f}_\alpha(x_0) \neq x_0^T\beta$, the average squared estimation bias term is non-zero.
$$\mathbb{E} \left[f(x_0) - \mathbb{E}\hat{f}_\alpha \right] \leq \mathbb{E} \left[f(x_0) - \mathbb{E}\hat{f}_p \right]$$
- $$\begin{aligned} Var \left[x_0^T \hat{\beta}_\alpha \right] &= x_0^T Var \left[(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \right] x_0 \\ &= \sigma_\epsilon^2 x_0^T V D^{*-1} D^2 D^{*-1} V^T x_0 \quad \text{where } D^* = \text{diag}(d_1^2 + \alpha, \dots, d_N^2 + \alpha) \\ &= \sigma_\epsilon^2 \sum_{i=1}^N \frac{d_i^2}{(d_i^2 + \alpha)^2} z_i^2 \leq \sigma_\epsilon^2 \sum_{i=1}^N \frac{1}{d_i^2} z_i^2 = Var \left[x_0^T \hat{\beta} \right] \end{aligned}$$

We trade it off with the benefits of a reduced variance

Bias, Variance and Model Complexity



Bias, Variance and Model Complexity

Dietterich, Thomas G., and Eun Bae Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University, 1995.

- Suppose we repeatedly draw training sets $\mathcal{T}_1, \dots, \mathcal{T}_l$, each of size N .

- Then we can estimate $\mathbb{E}\hat{f}(x_0)$ by $\bar{\hat{f}}(x_0) = \lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l \hat{f}_{\mathcal{T}_i}(x_0)$

- **For classification with 0-1 loss**, estimate misclassification error

$$Err(x_0) = \mathbb{E} \left[I \left(G \neq \hat{G}(x_0) \right) \mid X = x_0 \right] \text{ by } \bar{p} = \lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l I \left[G \neq \hat{G}_{\mathcal{T}_i}(x_0) \right], \text{ and}$$

estimate $\mathbb{E}\hat{G}(x_0)$ by $\bar{\hat{G}}(x_0) = \text{Mode} \left(\hat{G}_{\mathcal{T}_1}(x_0), \dots, \hat{G}_{\mathcal{T}_l}(x_0) \right)$.

Bias, Variance and Model Complexity

- Now consider for $\bar{\hat{p}} > 0.5$, which means we expect on the average at x_0 will be misclassified. Thus, we define the **bias** by $I(\bar{\hat{p}} > 0.5) = I(G \neq \mathbb{E}\hat{G}(x_0))$.

- And define the variance at x_0 to be the difference between the *Err* and the bias, because of **bias-variance decomposition** (excluding noise term):

$$(\text{Variance}) = \begin{cases} \bar{\hat{p}} & \text{if } \bar{\hat{p}} \leq 0.5 \\ 1 - \bar{\hat{p}} & \text{if } \bar{\hat{p}} > 0.5 \end{cases} \quad \text{or by analogy to the sum of squares,}$$

$$(\text{Variance}) = P\left(\hat{G}(x_0) \neq \mathbb{E}\hat{G}(x_0)\right) \text{ and estimate by } \lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l I\left(\hat{G}_{\mathcal{T}_i} \neq \bar{\hat{G}}(x_0)\right)$$

Bias, Variance and Model Complexity

- $Err(x_0)$ is also can be written as $P \left(G \neq \hat{G}(x_0) \right)$
- If bias is 0, then by bias-variance decomposition (excluding noise term),
 $Err(x_0) = P \left(G \neq \hat{G}(x_0) \right) = 0 + (\text{Variance}) = P \left(\hat{G}(x_0) \neq \mathbb{E}\hat{G}(x_0) \right)$
- Else if bias is 1, $Err(x_0) = P \left(G \neq \hat{G}(x_0) \right) = 1 - P \left(Y = \hat{G}(x_0) \right)$
 - Since bias is 1, $G \neq \mathbb{E}\hat{G}(x_0)$. Thus $\hat{G}(x_0) \neq \mathbb{E}\hat{G}(x_0)$ if $G = \hat{G}(x_0)$.
- $Err(x_0) = 1 - P \left(\hat{G}(x_0) \neq \mathbb{E}\hat{G}(x_0) \right) = (\text{Bias}) - (\text{Variance})$

Bias, Variance and Model Complexity

- For binary case, let $Err_B(x_0) = P(G \neq G(x_0))$ which is irreducible bayes error at x_0 , $f(x_0) = P(G = 1 | X = x_0)$. Then,
$$Err(x_0) = Err_Bx_0 + |2f(x_0) - 1| P(G(x_0) \neq \hat{G}(x_0) | X = x_0)$$
- If $G(x_0) = 1$, then
$$\begin{aligned} Err(x_0) &= (1 - f(x_0)) \left(1 - P(G \neq \hat{G}(x_0) | X = x_0) \right) + f(x_0) P(G \neq \hat{G}(x_0) | X = x_0) \\ &= 1 - f(x_0) + (2f(x_0) - 1) P(G \neq \hat{G}(x_0) | X = x_0) \\ &= P(G \neq G(x_0)) + (2f(x_0) - 1) P(G \neq \hat{G}(x_0) | X = x_0) \end{aligned}$$

Bias, Variance and Model Complexity

- Now assume that $\hat{f}(x_0) \sim \mathcal{N} \left(\mathbb{E}\hat{f}(x_0), \text{Var}(\hat{f}(x_0)) \right)$.

- Then,

$$\text{Err}(x_0) = P \left(G \neq \hat{G}(x_0) \mid X = x_0 \right) \approx \Phi \left(\frac{\text{sign} (0.5 - f(x_0)) \left(\mathbb{E}\hat{f}(x_0) - 0.5 \right)}{\sqrt{\text{Var}(\hat{f}(x_0))}} \right)$$

- Bias and variance combine in a **multiplicative** rather than additive fashion.

Bias, Variance and Model Complexity

\mathbf{X} is uniformly distributed in the hyper cube $[0,1]^{20}$

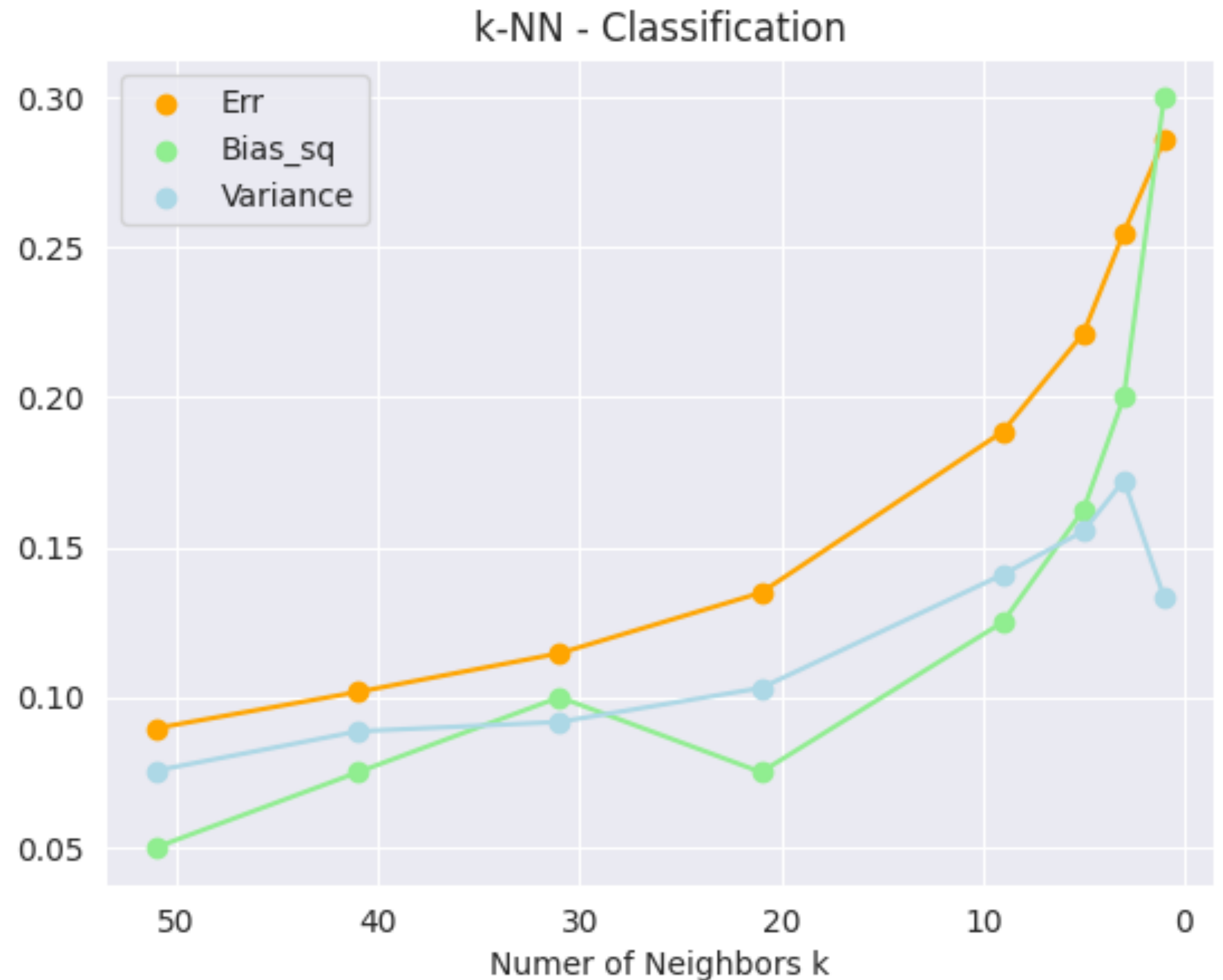
Y is 0 if $X_1 \leq 0.5$ and 1 if $X_1 > 0.5$ with 0-1 loss for a 200-simulated example.

$$\bar{\hat{G}}(x_0) = \text{Mode} \left(\hat{G}_{\mathcal{T}_1}(x_0), \dots, \hat{G}_{\mathcal{T}_l}(x_0) \right)$$

$$\mathbb{E} [\text{Bias}] = \text{Avg} \left[I \left(G \neq \bar{\hat{G}}(x_0) \right) \right]$$

$$\mathbb{E} [\text{Varaince}] = \text{Avg} \left[I \left(\hat{G}_{\mathcal{T}_i} \neq \bar{\hat{G}}(x_0) \right) \right]$$

$$Err = \text{Avg} \left[I \left(G \neq \hat{G}_{\mathcal{T}_i}(x_0) \right) \right]$$



Bias, Variance and Model Complexity

- (In-sample error)** $Err_{in} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{Y^0} \left[L(Y_i^0, \hat{f}(x_i) \mid \mathcal{T}) \right]$ where Y^0 indicates that N -new response at each of the training points x_i .

Note. The test input vectors don't need to coincide with the training input vectors. Thus, $Err_{\mathcal{T}}$ can be thought as an extra-sample error.
- (Optimism)** $op = Err_{in} - \bar{err}$, $w = \mathbb{E}_{\mathbf{y}}(op)$ ($\because \mathcal{T}$ is fixed.)
- For squared-error loss, $w = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{y}} \left[\mathbb{E}_{Y^0} \left(Y_i^0 - f(x_i) + f(x_i) - \mathbb{E}_{\mathbf{y}} \hat{Y}_i + \mathbb{E}_{\mathbf{y}} \hat{Y}_i - \hat{Y}_i \right)^2 - \left(Y_i - f(x_i) + f(x_i) - \mathbb{E}_{\mathbf{y}} \hat{Y}_i + \mathbb{E}_{\mathbf{y}} \hat{Y}_i - \hat{Y}_i \right)^2 \right]$

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{y}} \left[\mathbb{E}_{Y^0} \left(Y_i^0 - f(x_i) \right)^2 + 2 \mathbb{E}_{Y^0} \left(Y_i^0 - f(x_i) \right) \left(\hat{Y}_i - \mathbb{E}_{\mathbf{y}} \hat{Y}_i \right) - \left(Y_i - f(x_i) \right)^2 \right]$$

$$= \frac{2}{N} \sum_{i=1}^N \mathbb{E}_{\mathbf{y}} \left[\left(Y_i^0 - \mathbb{E}_{Y^0} Y_i^0 \right) \left(\hat{Y}_i - \mathbb{E}_{\mathbf{y}} \hat{Y}_i \right) \right] = \frac{2}{N} \sum_{i=1}^N Cov(\hat{Y}_i, Y_i)$$

Bias, Variance and Model Complexity

- If \hat{Y}_i is obtained by linear fit with d inputs, i.e. $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ where $rank(\mathbf{H}) = \underline{tr(\mathbf{H})} = d$
Effective number of parameters
- Then, $w = \frac{2}{N}tr[Cov(\mathbf{H}\mathbf{y}, \mathbf{y})] = \frac{2d}{N}\sigma_\epsilon^2$ for additive error model $Y = f(X) + \epsilon$
- Therefore, $\mathbb{E}_{\mathbf{y}}Err_{in} = \mathbb{E}_{\mathbf{y}}err + \frac{2d}{N}\sigma_\epsilon^2$
 - This equation holds approximately for other models, such as binary data & entropy loss

Bias, Variance and Model Complexity

- Our goal is to estimate **the prediction error** $Err_{\mathcal{J}}$ or Err ($= \mathbb{E}Err_{\mathcal{J}}$)
 1. Estimate in-sample error by **estimate the optimism** and add it to $e\bar{r}r$, such as C_p , AIC, BIC.
 2. **Directly estimates the extra-sample error** Err , such as Cross validation, Bootstrap

Q. Why we estimates in-sample error instead the prediction error? -> In-sample error is convenient and **the relative (rather than absolute) size of the error** is what matters in model comparison.

Estimates of In-Sample Prediction Error

- The general form of the in-sample estimates is $E\hat{r}_{in} = e\bar{r}r + \hat{w}$
- $C_p = e\bar{r}r + \frac{2d}{N}\hat{\sigma}_\epsilon^2$, $AIC = -\frac{2}{N}\text{loglik} + 2\frac{d}{N}$, $BIC = -2\text{loglik} + d \log N$ where loglik is the maximized log-likelihood
- Under Gaussian model, assuming the variance σ_ϵ^2 is known, $-2 \cdot \text{loglik}$ equals up to a constant $N\frac{e\bar{r}r}{\sigma_\epsilon^2}$ for squared error loss.
- Hence, $AIC = \frac{1}{\sigma_\epsilon^2} \left[e\bar{r}r + \frac{2d}{N}\sigma_\epsilon^2 \right]$, $BIC = \frac{N}{\sigma_\epsilon^2} \left[e\bar{r}r + (\log N) \frac{d}{N}\sigma_\epsilon^2 \right]$

Estimates of In-Sample Prediction Error

- BIC will select correct model as $N \rightarrow \infty$. This is not the case for AIC, which tends to choose models which are too complex as $N \rightarrow \infty$.
- For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity.
- Then which method is better?

Estimates of In-Sample Prediction Error

Foster, D. & George, E. "The Risk Inflation Criterion for Multiple Regression." Ann. Statist. 22 (4) 1947 - 1975, December, 1994

- Assume that Gaussian model with σ_ϵ^2 is known, $\mathbf{y} = \mathbf{X}\beta + \epsilon$ where $\mathbf{X} \in \mathbb{R}^{N \times p}$ is fixed
- Define $\gamma = (\gamma_1, \dots, \gamma_p)$ where $\gamma_1 = 1, \gamma_i \in \{0, 1\}$ for $i = 2, \dots, p$

$$\hat{\beta}_\gamma = ((\mathbf{X}D_\gamma)^T \mathbf{X}D_\gamma)^{-1} (\mathbf{X}D_\gamma)^T \mathbf{y} \text{ where } D_\gamma = \text{diag}[\gamma]$$

- **(Oracle estimator)** $\eta(\beta) = (1, \eta_2, \dots, \eta_p)$ where $\eta_i = I[\beta_i \neq 0]$

$$\hat{\beta}_\eta = ((\mathbf{X}D_\eta)^T \mathbf{X}D_\eta)^{-1} (\mathbf{X}D_\eta)^T \mathbf{y}$$

- **(Risk inflation)**

$$RI(\gamma) = \sup_{\beta} \left\{ \frac{R(\beta, \hat{\beta}_\gamma)}{R(\beta, \hat{\beta}_\eta)} \right\} \text{ where } R(\beta, \hat{\beta}) = \mathbb{E}_{\beta} \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2 \text{ is called the risk of } \hat{\beta}$$

Estimates of In-Sample Prediction Error

- Since $R(\beta, \hat{\beta}_\eta) = \mathbb{E}_\beta \|\mathbf{X}\hat{\beta}_\eta - \mathbf{X}\beta\|^2 = \text{tr} \left[\text{Cov}(\mathbf{X}\hat{\beta}_\eta) \right] = |\eta| \sigma_\epsilon^2$,

$$RI(\gamma) = \sup_{\beta} \frac{R(\beta, \hat{\beta}_\gamma)}{|\eta| \sigma_\epsilon^2}$$

- **The general form of above methods:** $\gamma_\Pi = \arg \min_{\gamma} \left[e\bar{r}r(\gamma) + |\gamma| \sigma_\epsilon^2 \Pi \right]$ where $\Pi \geq 0$

is a pre-specified constant, $|\gamma|$ is the number of nonzero components of γ

- If $\Pi = 2$, then C_p, AIC . And if $\Pi = \log n$, then BIC .

Estimates of In-Sample Prediction Error

- Now assume that all predictors are independent, i.e. $\mathbf{X}^T \mathbf{X}$ is a diagonal matrix.
- $$\begin{aligned}\gamma_{\Pi} &= \arg \min_{\gamma} \|\mathbf{y} - \mathbf{X}\hat{\beta}_{\gamma}\|^2 + |\gamma| \sigma^2 \Pi \\ &= \arg \min_{\gamma} \|\mathbf{X}\hat{\beta}_{\gamma_{LS}} - \mathbf{X}\hat{\beta}_{\gamma}\|^2 + |\gamma| \sigma^2 \Pi \\ &= \arg \min_{\gamma} \sum_{i=1}^p \left((\hat{\beta}_{\gamma_{LS}})_i - (\hat{\beta}_{\gamma})_i \right)^2 X_i^T X_i + |\gamma| \sigma^2 \Pi \\ &= \arg \min_{\gamma} \sum_{i=1}^p \left[I(\gamma_i = 0) \frac{(X_i^T \mathbf{y})^2}{X_i^T X_i} + I(\gamma_i = 1) \sigma^2 \Pi \right] = \{1, \gamma_2^* \dots \gamma_p^*\} \text{ where } \gamma_i^* = I \left[\frac{(X_i^T \mathbf{y})^2}{X_i^T X_i} > \sigma^2 \Pi \right] \text{ for } i = 2, \dots, p\end{aligned}$$
- Note that $X_i^T \epsilon = X_i^T (\mathbf{y} - \sum_{j=1}^p X_j \beta_j) = X_i^T \mathbf{y} - X_i^T X_i \beta_i$ ($\because X_i^T X_j = 0$ for $i \neq j$)

Estimates of In-Sample Prediction Error

- Thus, $\gamma_i^* = I \left[\frac{(X_i^T \mathbf{y})^2}{X_i^T X_i} > \sigma^2 \Pi \right] = I \left[\frac{(X_i^T \epsilon + X_i^T X_i \beta_i)^2}{X_i^T X_i} > \sigma^2 \Pi \right] = I \left[\left(\frac{(X_i^T \epsilon)}{|X_i|} + |X_i| \beta_i \right)^2 > \sigma^2 \Pi \right]$
- And the risk is

$$R(\beta, \hat{\beta}_\gamma) = \mathbb{E}_\beta \|\mathbf{X} \hat{\beta}_\gamma - \mathbf{X} \beta\|^2$$

$$= \mathbb{E}_\beta \sum_{i=1}^p \left[I(\gamma_i = 0) \beta_i^2 X_i^T X_i + I(\gamma_i = 1) \left(\frac{X_i^T \mathbf{y}}{X_i^T X_i} - \beta_i \right)^2 X_i^T X_i \right]$$

$$= \mathbb{E}_\beta \sum_{\gamma_i=0} (|X_i| \beta_i)^2 + \mathbb{E}_\beta \sum_{\gamma_i=1} \left(\frac{X_i^T (\mathbf{y} - X_i \beta_i)}{X_i^T X_i} \right)^2 X_i^T X_i$$

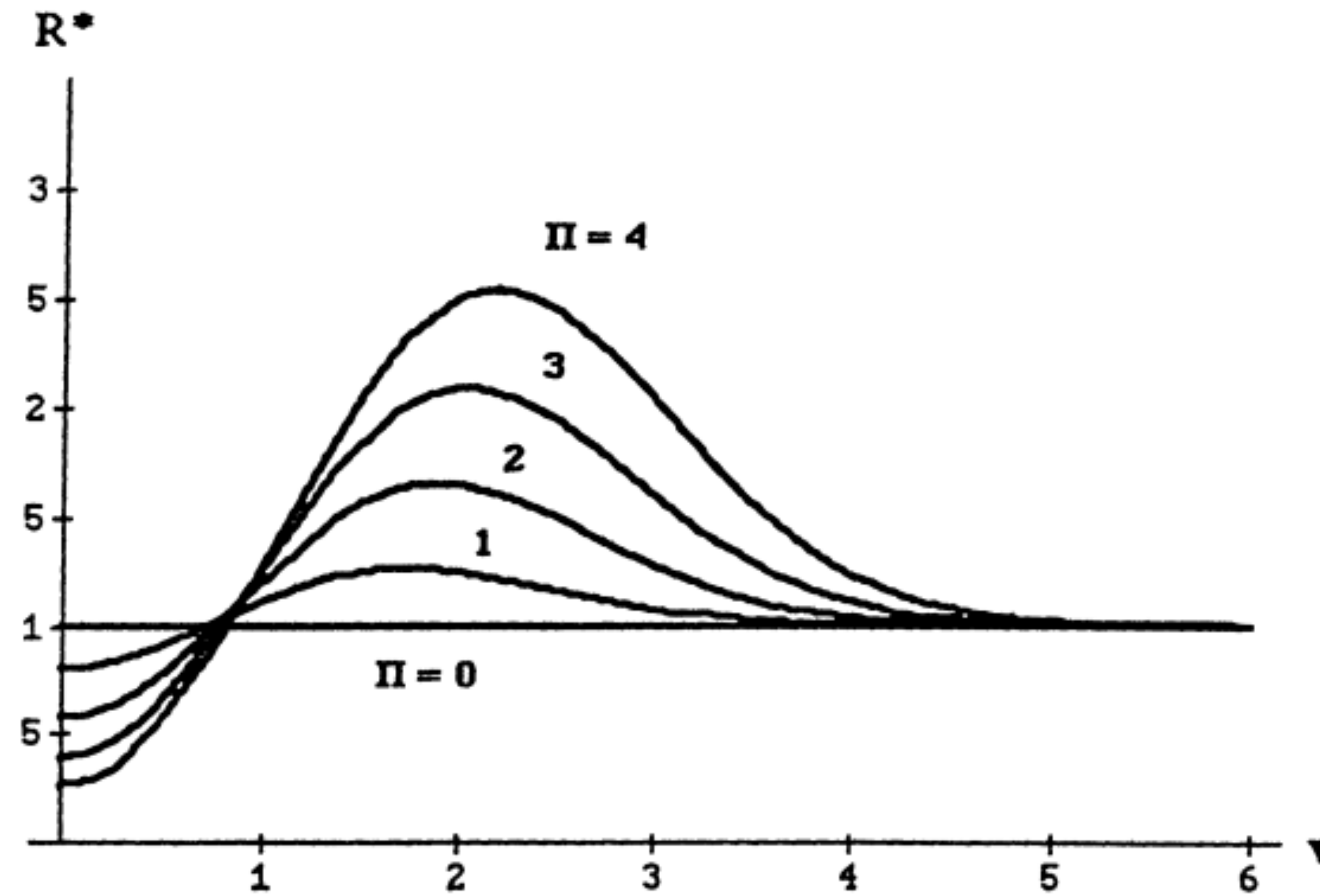
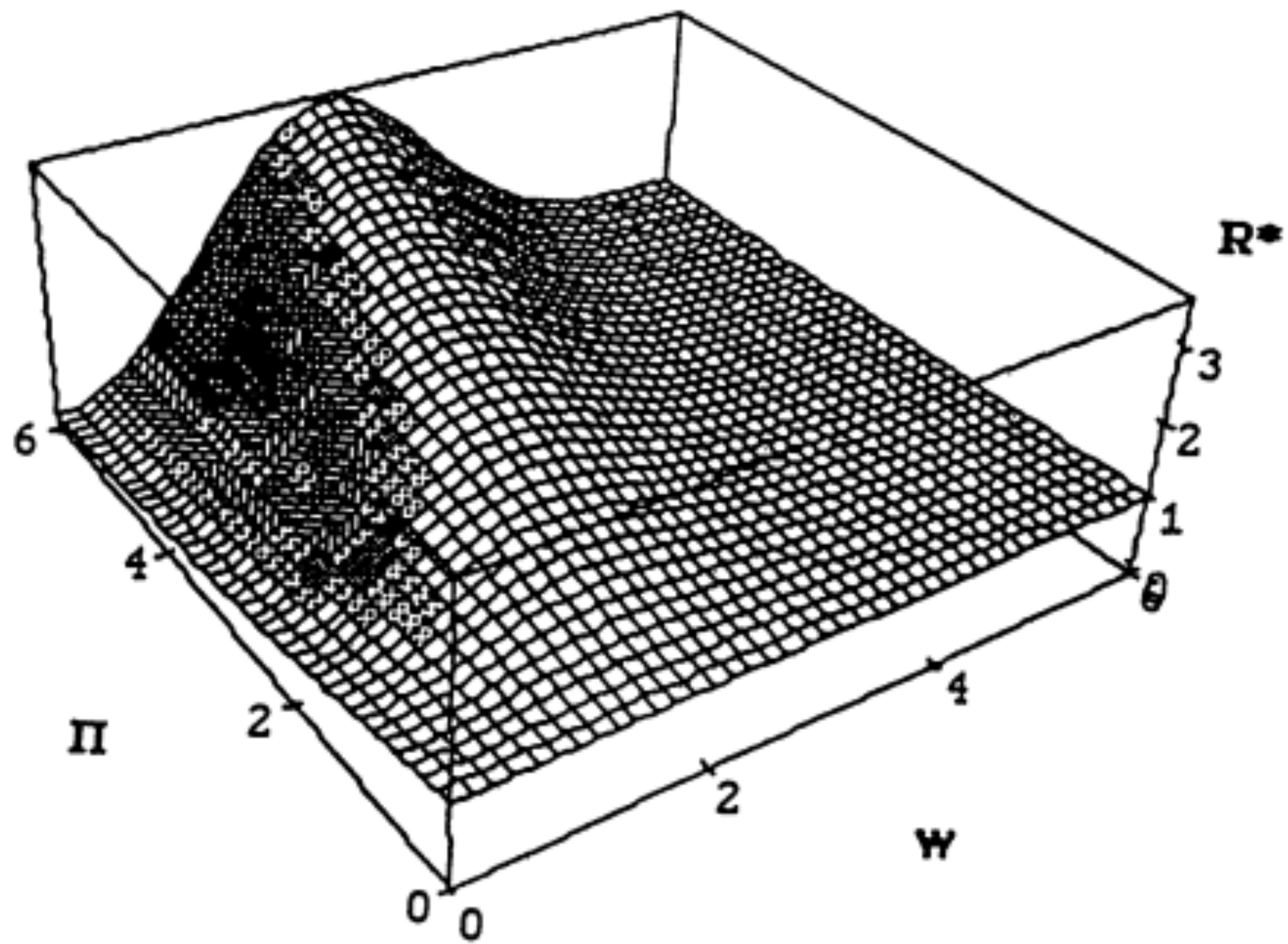
$$= \mathbb{E}_\beta \sum_{\gamma_i=0} (|X_i| \beta_i)^2 + \mathbb{E}_\beta \sum_{\gamma_i=1} \frac{(X_i^T \epsilon)^2}{|X_i|^2}$$

$$= \left[\sum_{i=2}^p (|X_i| \beta_i)^2 P[\gamma_i = 0] \right] + \left[\sigma^2 + \sum_{i=2} \mathbb{E}_\beta \frac{(X_i^T \epsilon)^2}{|X_i|^2} I(\gamma_i = 1) \right]$$

Estimates of In-Sample Prediction Error

- By gaussian assumption, $(X_i^T \epsilon) | X_i| = \sigma_\epsilon Z$ where $\mathbf{Z} \sim \mathcal{N}(0,1)$
- Then for $\hat{\beta}_{\gamma_\Pi}$,
$$(|X_i| \beta_i)^2 P[\gamma_i^* = 0] = (|X_i| \beta_i)^2 P \left[(\sigma^2 Z + |X_i| \beta_i)^2 \leq \sigma \Pi \right]$$
$$\mathbb{E}_\beta \frac{(X_i^T \epsilon)^2}{|X_i|^2} I(\gamma_i^* = 1) = \sigma^2 \mathbb{E} \left[Z^2 I \left[(\sigma^2 Z + |X_i| \beta_i)^2 > \sigma^2 \Pi \right] \right]$$
- Therefore, $R(\beta, \hat{\beta}_{\gamma_\Pi}) = \sigma^2 + \sigma^2 \sum_{i=2}^p R^*\left(\frac{|X_i| \beta_i}{\sigma}, \Pi\right)$ where
$$R^*(w, \Pi) = w^2 P \left[(w + Z)^2 \leq \Pi \right] + \mathbb{E} \left[Z^2 I \left[I(w + Z)^2 > \Pi \right] \right]$$

Estimates of In-Sample Prediction Error



Estimates of In-Sample Prediction Error

- **(Partial risk inflation)** The maximization risk over the set of β 's with exactly j non-zero components.

$$RI(j, \gamma) = \sup_{\beta \in B_j} \frac{R(\beta, \hat{\beta}_\gamma)}{j\sigma^2} \text{ where } B_j = \{\beta : |\eta| = j\}$$

- $$RI(j, \gamma_\Pi) = \sup_{\beta \in B_j} \frac{1}{j\sigma^2} \left[\sigma^2 + \sigma^2 \sum_{i=2}^p R^*(w_i, \Pi) \right]$$
$$= \sup_{\beta \in B_j} \frac{1}{j} \left[1 + \sum_{i=2}^p \{ I(\beta_i = 0) R^*(w_i = 0, \Pi) + I(\beta_i \neq 0) R^*(w_i, \Pi) \} \right]$$
$$= \frac{1}{j} \left[1 + (p - j) R^*(0, \Pi) + (j - 1) \sup_w R^*(w, \Pi) \right]$$

Estimates of In-Sample Prediction Error

- **The risk inflation of γ_Π :**

$$RI(\gamma_\Pi) = \max_j RI(j, \gamma_\Pi) = \max [RI(1, \gamma_\Pi), RI(p, \gamma_\Pi)]$$

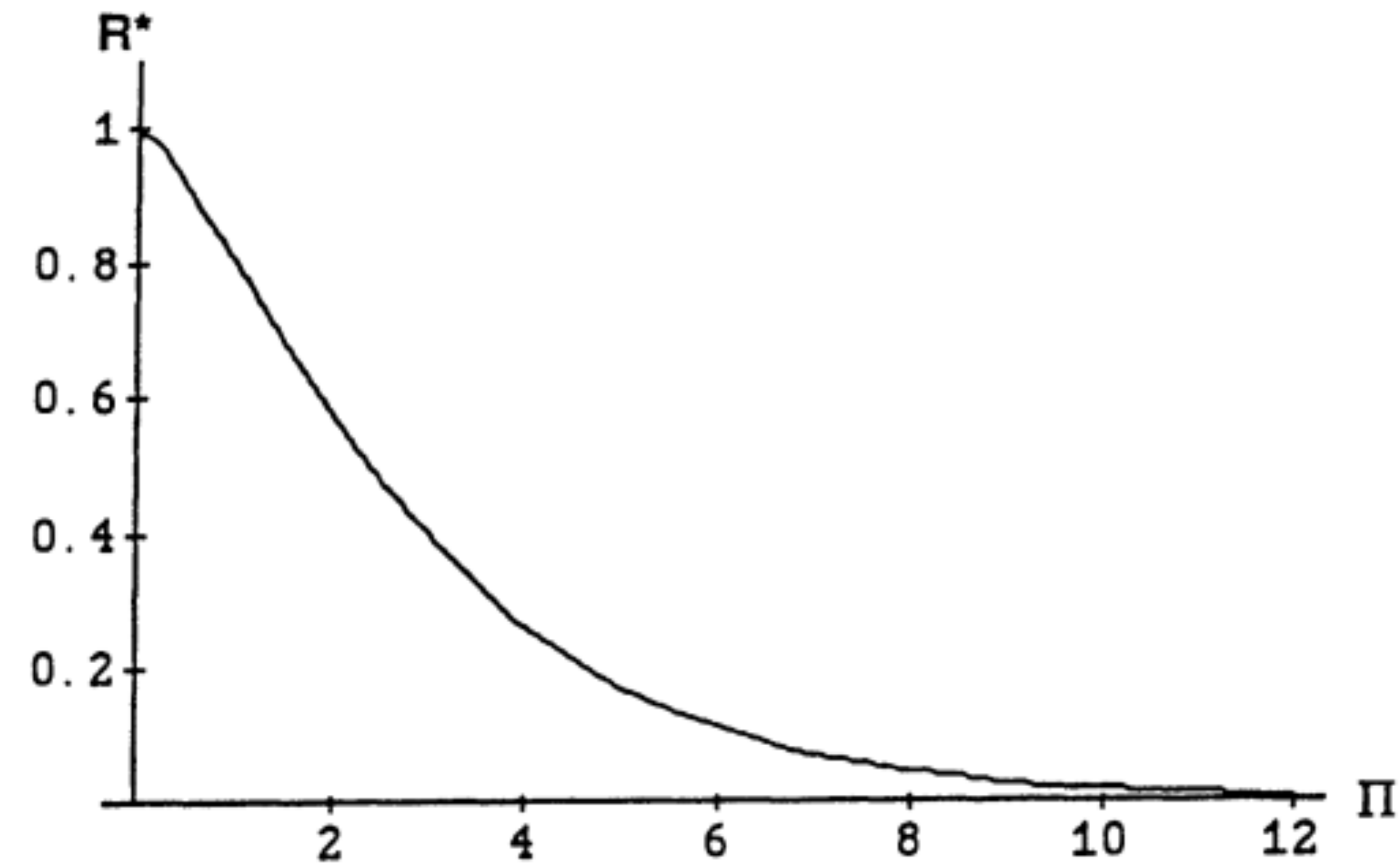
$$= \max \left[1 + (p - 1)R^*(0, \Pi), 1/p + (1 - 1/p) \sup_w R^*(w, \Pi) \right]$$

$$\approx \max \left[pR^*(0, \Pi), \sup_w R^*(w, \Pi) \right] \text{ for large } p$$

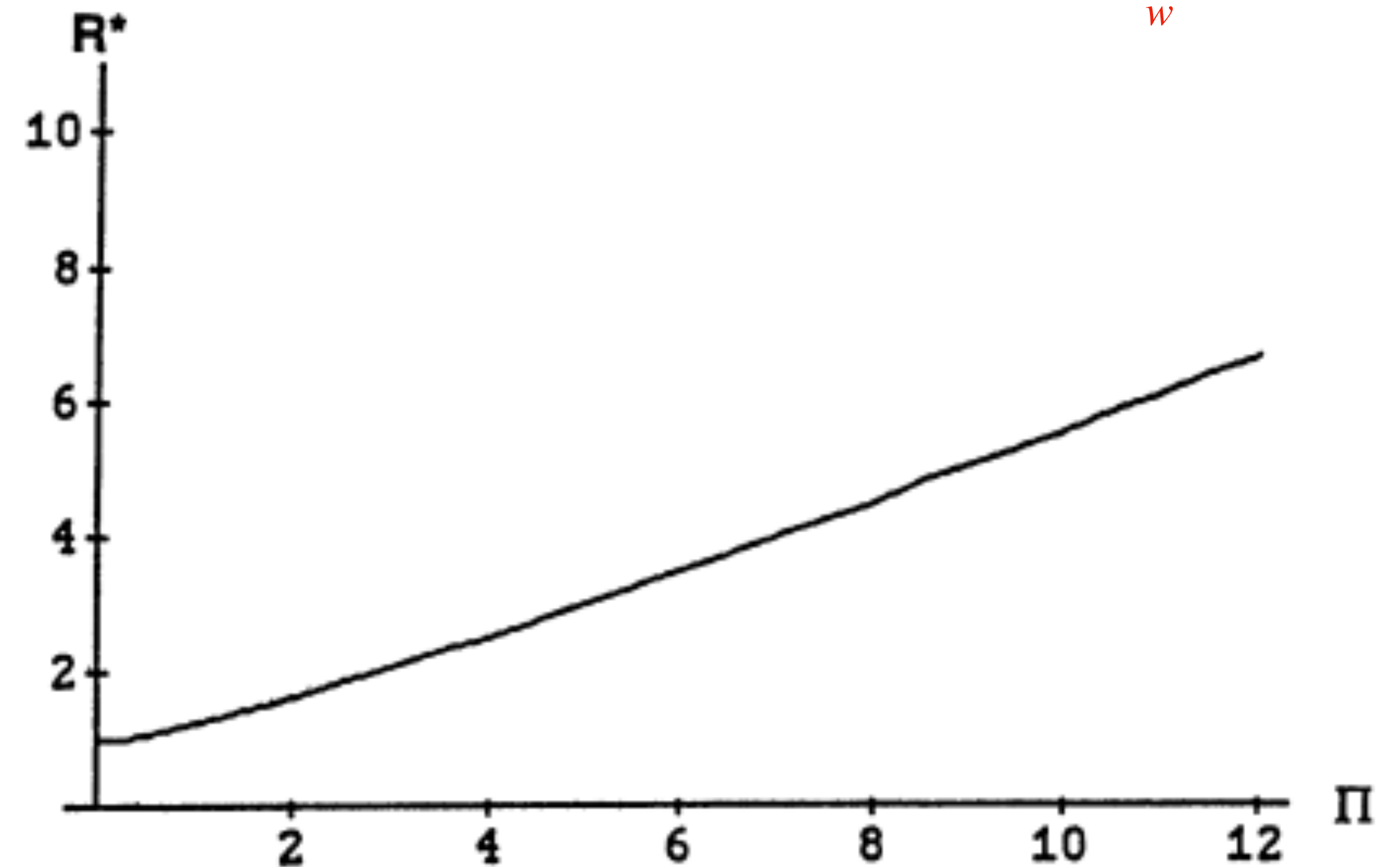
- $R^*(0, \Pi) = \mathbb{E} [Z^2 I(Z^2 > \Pi)] = 2 \left[\sqrt{\Pi} \phi(\sqrt{\Pi}) + \Phi(-\sqrt{\Pi}) \right] \approx 2\sqrt{\Pi} \phi(\sqrt{\Pi})$ for large Π
- $\Pi - o(\Pi) < \sup_w R^*(w, \Pi) < \Pi + 1$
- Hence, $RI(\gamma_\Pi) \approx \max \left[p2\sqrt{\Pi} \phi(\sqrt{\Pi}), \Pi \right]$

Estimates of In-Sample Prediction Error

The component risk at $\beta_j = 0$: $R^*(0, \Pi)$



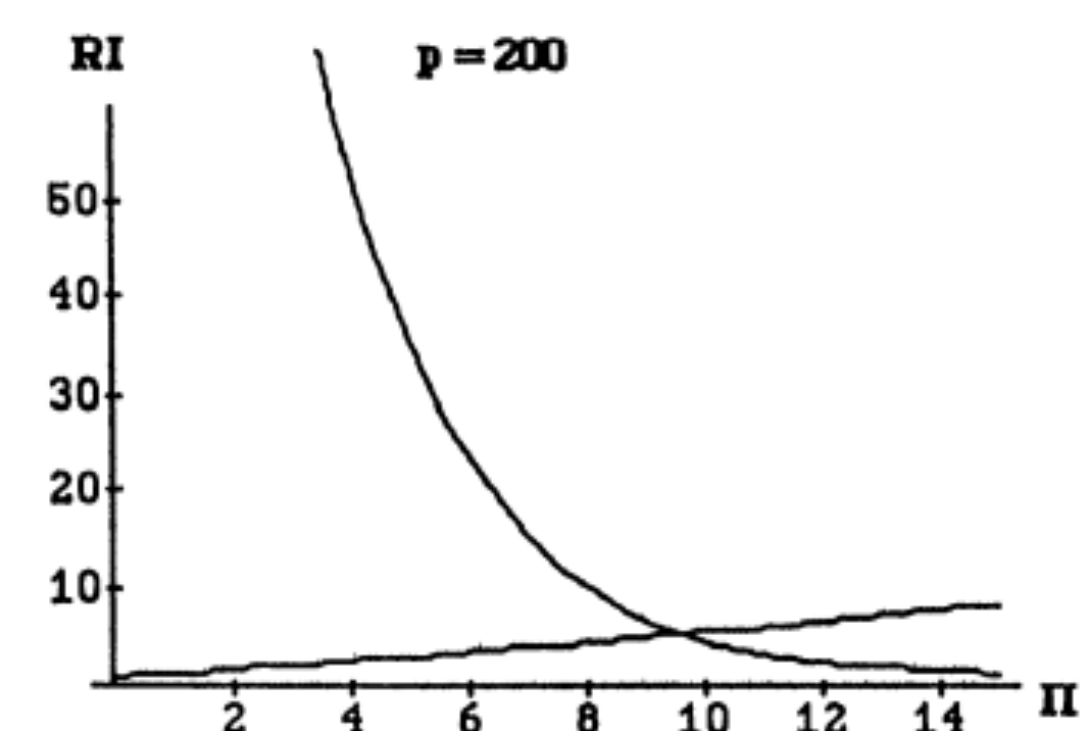
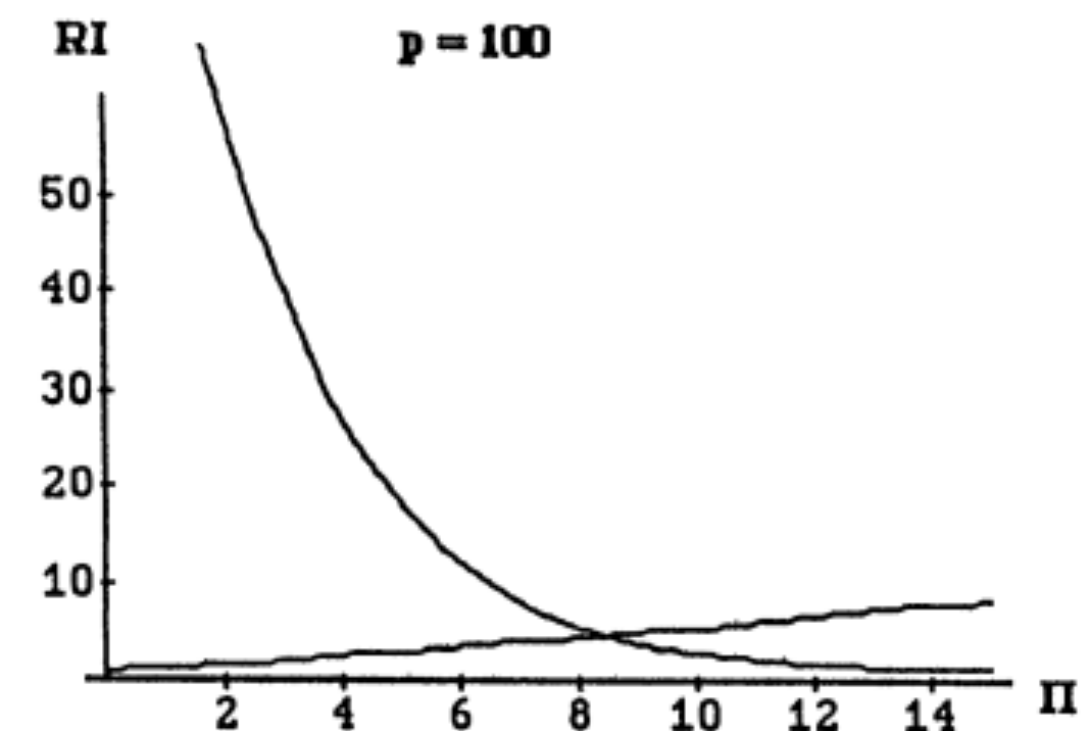
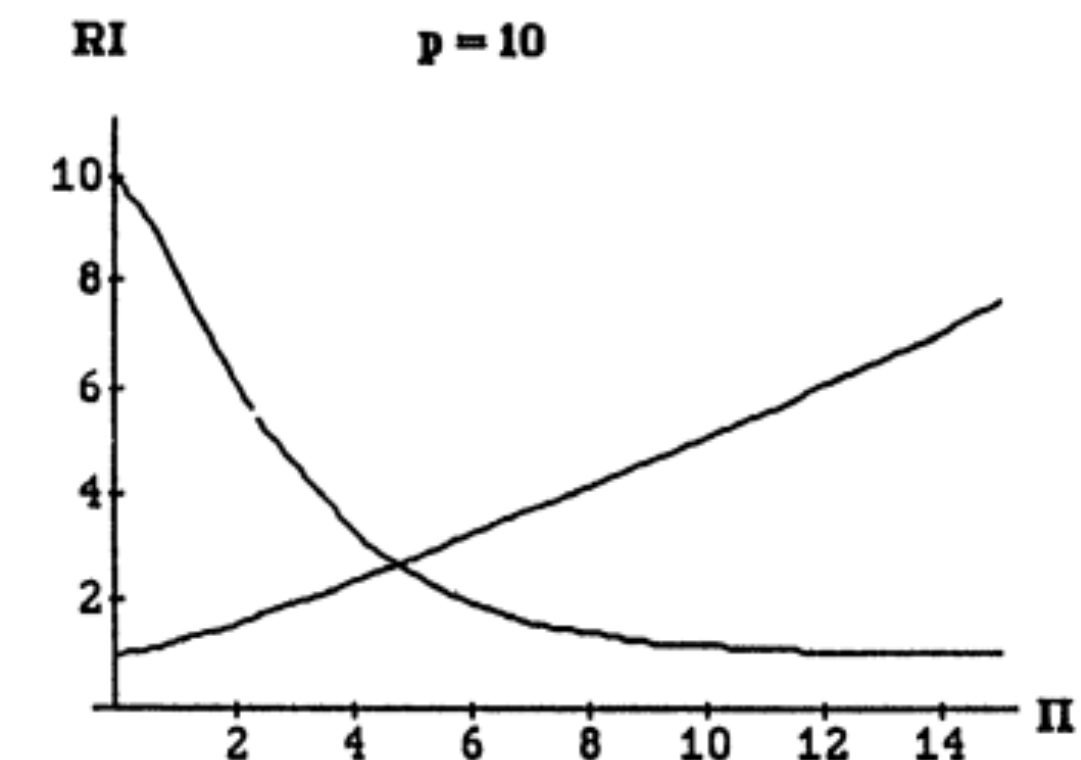
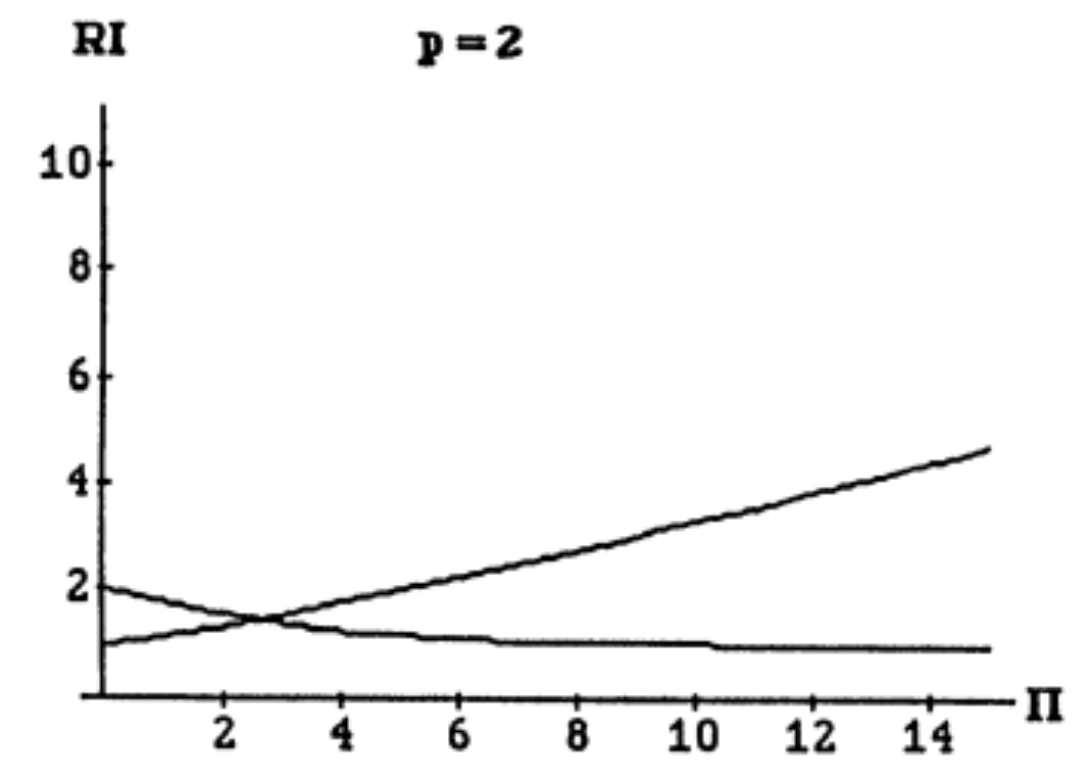
The worst possible component risk: $\sup_w R^*(w, \Pi)$



Estimates of In-Sample Prediction Error

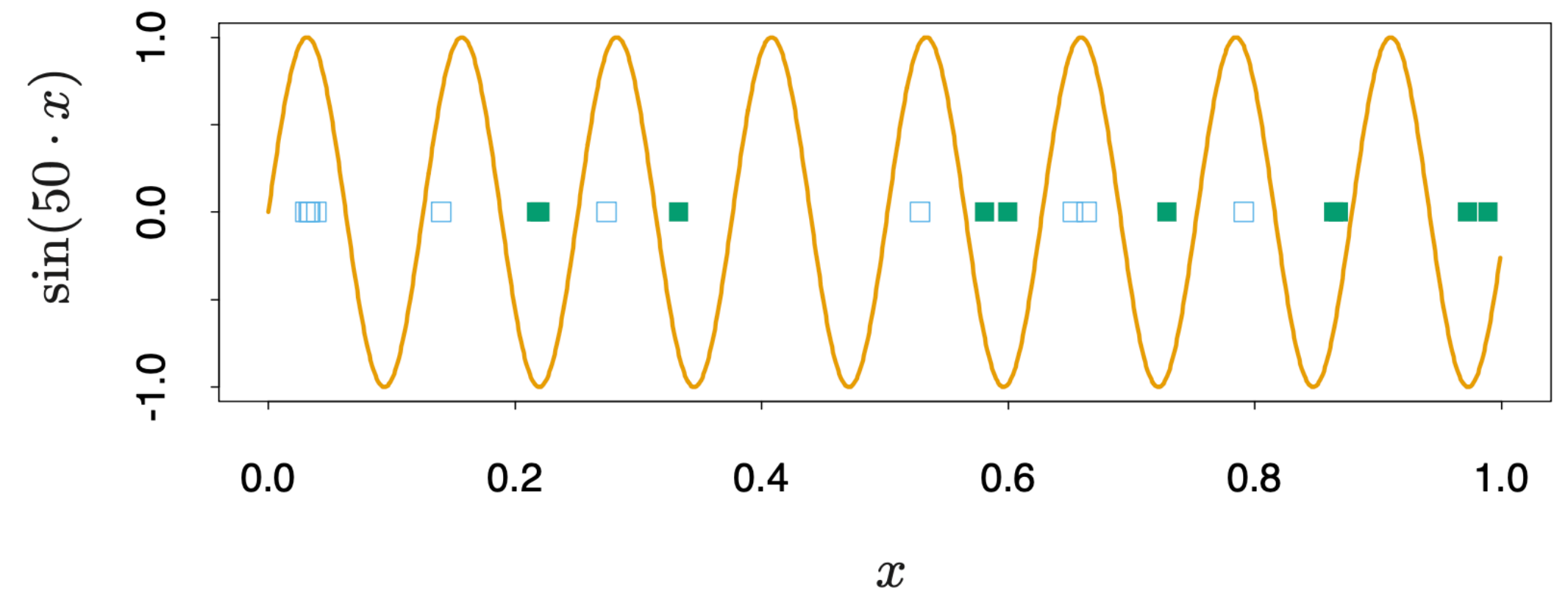
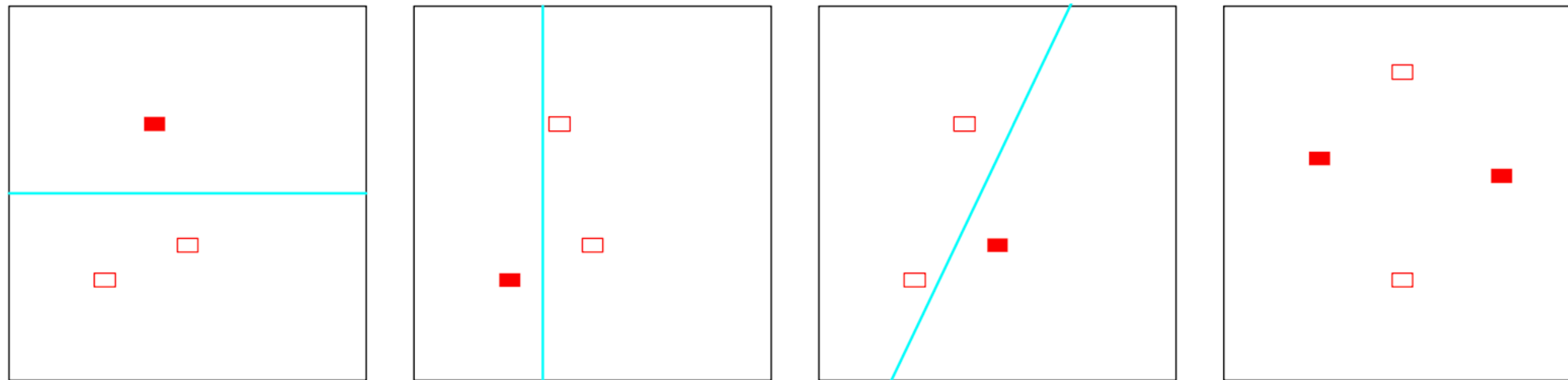
The risk inflation of various procedures (The $X'X$ diagonal case)

Method	Π	$R^*(0, \Pi)$	$\sup_w R^*(w, \Pi)$	Risk inflation
LS	0	1	1	p
max adj R^2	1	.801	1.26	$\approx p(.801)$
AIC/ C_p	2	.573	1.65	$\approx p(.573)$
BIC	$\log n$	$\approx \sqrt{(2 \log n)/(\pi n)}$	$\approx \log n$	$\approx \log n$ if $p \ll \sqrt{n}$ $\approx \sqrt{(2 \log n)/(\pi n)}$ if $p \gg \sqrt{n}$
General γ_Π	Π	$\approx 2\sqrt{\Pi}\phi(\Pi)$	$\approx \Pi$	$\approx 2\sqrt{\Pi}\phi(\Pi) \vee \Pi$
$\gamma_{2 \log p}$	$2 \log p$	$\approx \sqrt{(4 \log p)/(\pi p)^2}$	$\approx 2 \log p$	$\approx 2 \log p$
General γ				$\geq 2 \log p - o(\log p)$



Estimates of In-Sample Prediction Error

- Suppose we have a class of functions $\{f(x, \alpha)\}$ index by a parameter α , with $x \in \mathbb{R}^p$
- Consider for $f(x, \alpha) = I(\alpha_0 + \alpha_1^T x)$ and $f(x, \alpha) = I(\sin \alpha x)$
- How to specify the number of parameters (or the complexity) in general?



Estimates of In-Sample Prediction Error

- **(The Vapnik-Chervonenkis dimension)** The VC dimension of the class $\{f(x, \alpha)\}$ is defined to be the largest number of points that can be shattered by members of $\{f(x, \alpha)\}$
- Examples)
 - $\{f(x, \alpha)\}$ is a constant classifier, its VC dimension is 0.
 - $\{f(x, \alpha)\}$ is a single-parametric threshold classifier on \mathbb{R} , its VC dimension is 1.
 - $\{f(x, \alpha)\}$ is a linear indicator function on \mathbb{R}^p , its VC dimension is $p + 1$.
 - $\{I(\sin \alpha x) > 0\}$ has infinite VC dimension.

Estimates of In-Sample Prediction Error

- (Estimates of **extra-sample** prediction error using VC dimension) Suppose that we fit N -training points using a class of functions $\{f(x, \alpha)\}$ having VC dimension h .

$$P \left[Err_{\mathcal{T}} \leq e\bar{r}r + \frac{\epsilon}{2} \left(1 + \sqrt{1 + \frac{4e\bar{r}r}{\epsilon}} \right) \right] \geq 1 - \eta \text{ (binary classification)}$$

$$P \left[Err_{\mathcal{T}} \leq \frac{e\bar{r}r}{\left(1 - c\sqrt{\epsilon}\right)_+} \right] \geq 1 - \eta \text{ (regression) where } \epsilon = a_1 \frac{h \left[\log(a_2 N/h) + 1 \right] - \log(\eta/4)}{N}$$

$$\text{and } 0 < a_1 \leq 4, 0 < a_2 \leq 2$$

Estimates of In-Sample Prediction Error

- We note that above upper bounds are often very loose, but the **relative** (not absolute) size of the test error is important.
- Vapnik's structural risk minimization (SRM) : For nested sequence of models of **increasing VC dimensions** $h_1 < h_2 < \dots$, and then chooses the model with the smallest value of the upper bound.
 - Example) Polynomials of increasing degrees

Cross Validation

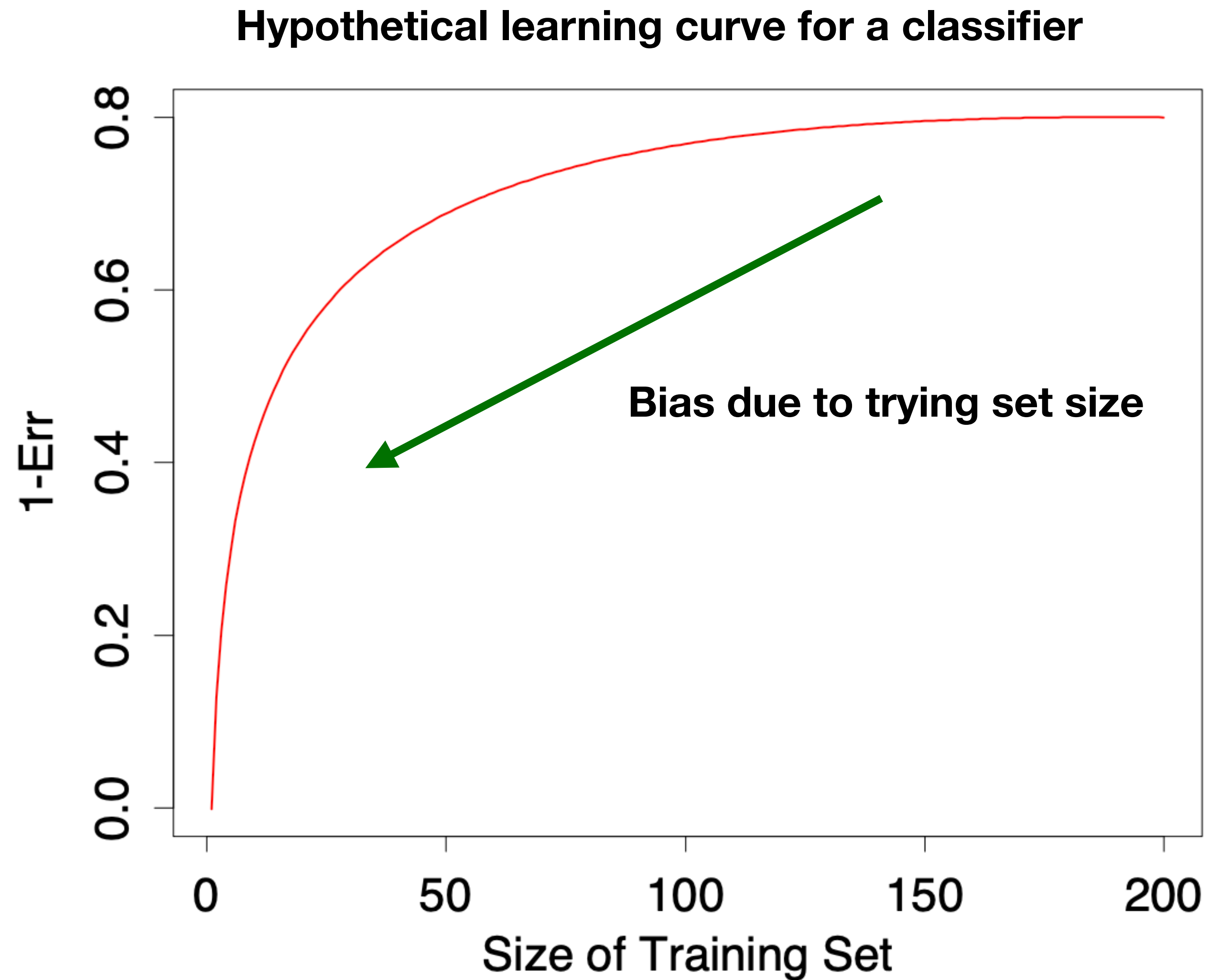
- This method directly estimates the expected extra-sample error Err
- **(K-Fold Cross-Validation)** Given a set of models $\{f(x, \alpha)\}$. Let $I : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by the **randomization**. Denote by $\hat{f}^{(-k)}(x, \alpha)$ the set of models indexed by parameter α , computed with k -th part of the data removed.

(Estimate of the test error curve)

$$CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^N L \left(y_i, \hat{f}^{-I(i)}(x_i) \right)$$

- Find the tuning parameter $\hat{\alpha}$ that minimizes $CV(\hat{f}, \alpha)$

Cross Validation



Cross Validation

- **(Generalized cross-validation)** For linear fitting methods $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$,

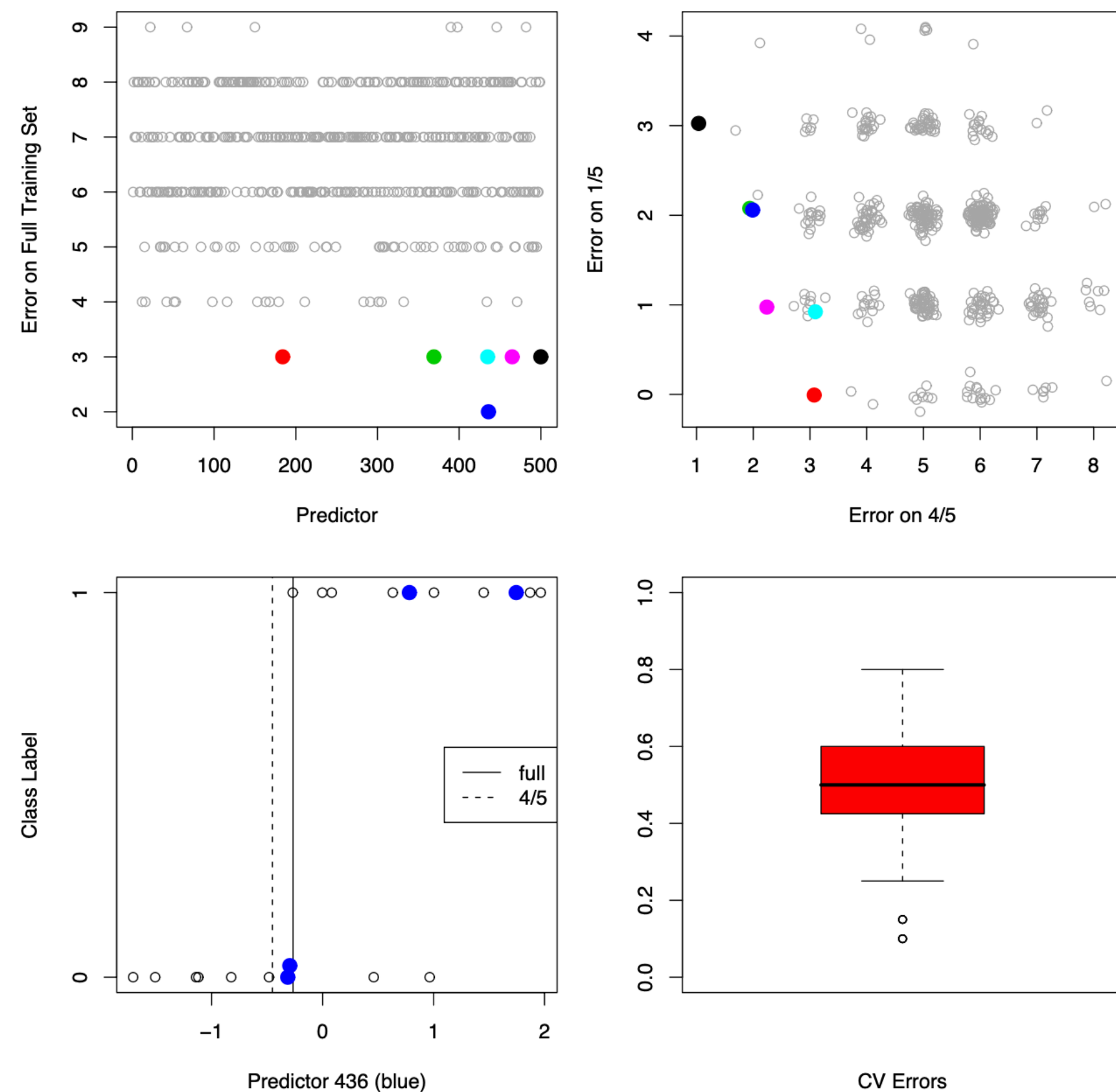
$$\text{(LOOCV)} \quad \frac{1}{N} \sum_{i=1}^N \left[y_i - \hat{f}^{-i}(x_i) \right]^2 = \frac{1}{N} \sum_{i=1}^N \left[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}} \right]^2$$

$$\text{The GCV approximation is } \frac{1}{N} \sum_{i=1}^N \left[\frac{y_i - \hat{f}(x_i)}{1 - \text{tr}(\mathbf{S})/N} \right]^2$$

$$\begin{aligned} \text{Use the approximation } 1/(1-x)^2 \approx 1 + 2x, \quad & \frac{1}{N} \sum_{i=1}^N \left[\frac{y_i - \hat{f}(x_i)}{1 - \text{tr}(\mathbf{S})/N} \right]^2 \\ & \approx \frac{1}{N} \sum_{i=1}^N \left(y_i - \hat{f}(x_i) \right)^2 \left(1 + 2\text{tr}(\mathbf{S})/N \right) = e\bar{r}r + \frac{2\text{tr}(\mathbf{S})}{N} \hat{\sigma}_\epsilon^2 \end{aligned}$$

Cross Validation

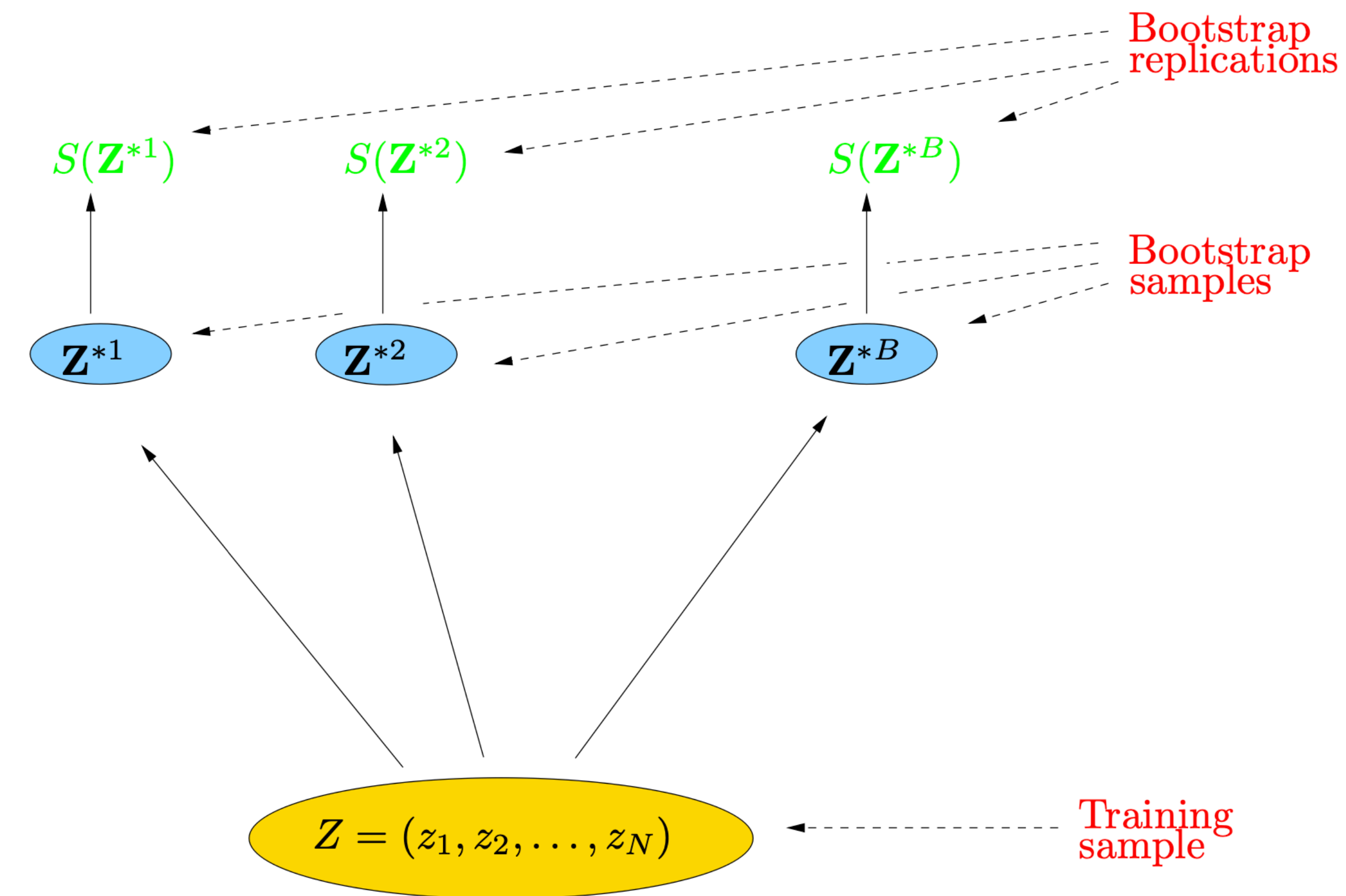
- Consider a scenario with $N = 20$ samples in two equal-sized classes, and $p = 500$ quantitative predictors having a standard Gaussian distribution that are **independent of the class labels**. Then the true error rate of any classifier is 50 %. And consider a single split model that single split minimizes the misclassification error



5-Fold CV must base its split on the 4/5ths data, and this incurs two errors out of four samples.

As we would hope, the average cross-validation error is around 50%, which is the true expected prediction error for this classifier.

Bootstrap methods



$$\hat{Err}_{boot} = \frac{1}{BN} \sum_{b=1}^M \sum_{i=1}^N L \left(y_i, \hat{f}^{*(b)}(x_i) \right)$$

Consider for example a 1-nearest neighbor classifier applied to a **two-class classification** problem with the same number of observation in each class, in which the **predictors and class labels are independent**. Then true error rate is 0.5.

$$P(z_i \in Z^*) = 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - e^{-1} = 0.632$$

$$\hat{Err}_{boot} = P(z_i \in Z^*) \cdot 0 + P(z_i \notin Z^*) \cdot 0.5 = 0.184$$

which under-estimate the test error due to overlap

Bootstrap methods

- **(LOOCV Bootstrap Estimate)**

$$\hat{Err}^{(1)} = \frac{1}{N} \sum_{i=1}^N \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

where C^{-i} is the the set of indices of the bootstrap samples b that do not contain

$|C^{-i}|$ must **non-zero**

Option1. Choose B large enough to ensure that $|C^{-i}|$ is non-zero for all i

Option2. Leave out terms s.t. $|C^{-i}| = 0$

- LOOCV Bootstrap contains 63 % of training samples in each boot strap samples which occurs bias due to training set size. Thus, it dose not work well at light-fitting situations.

- **(.632 Estimate)** $\hat{Err}^{(.632)} = .368e\bar{rr} + .632\hat{Err}^{(1)}$ Note. $e\bar{rr} < \hat{Err}^{(1)}$ in general.

For above example, $e\bar{rr} = 0$, $\hat{Err}^{(1)} = 0.5$. Thus, $\hat{Err}^{(.632)} = 0.632 * 0.5 = 0.316$ which under-estimate the true error rate

Bootstrap methods

- **(No-information error rate)** Error rate of our prediction rule if all predictors are independent to class labels. Denote by γ .
- $\hat{\gamma} = \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N L(y_i, \hat{f}(x_{i'}))$ which evaluates the prediction rule on all possible combinations.
- **(Relative over-fitting rate)** $\hat{R} = \frac{\hat{Err}^{(1)} - e\bar{r}r}{\hat{r} - e\bar{r}r} \in (0,1)$

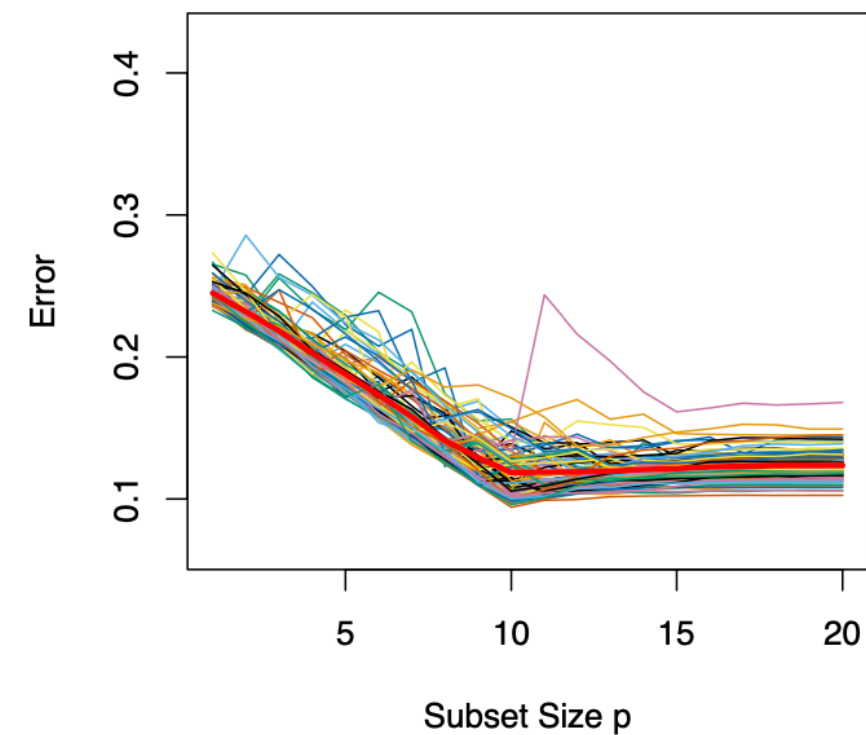
If there is no over-fitting, then $e\bar{r}r = \hat{Err}^{(1)}$, $\hat{R} = 0$
If overfitting is the overfitting equals the no-information value, $\hat{R} = 1$
- **(.632+ Estimate)** $\hat{Err}^{(.632+)} = (1 - \hat{w})e\bar{r}r + \hat{w}\hat{Err}^{(1)}$ where $\hat{w} = \frac{.632}{1 - .368\hat{R}} \in (.632,1)$

Conditional or Expected Test Error?

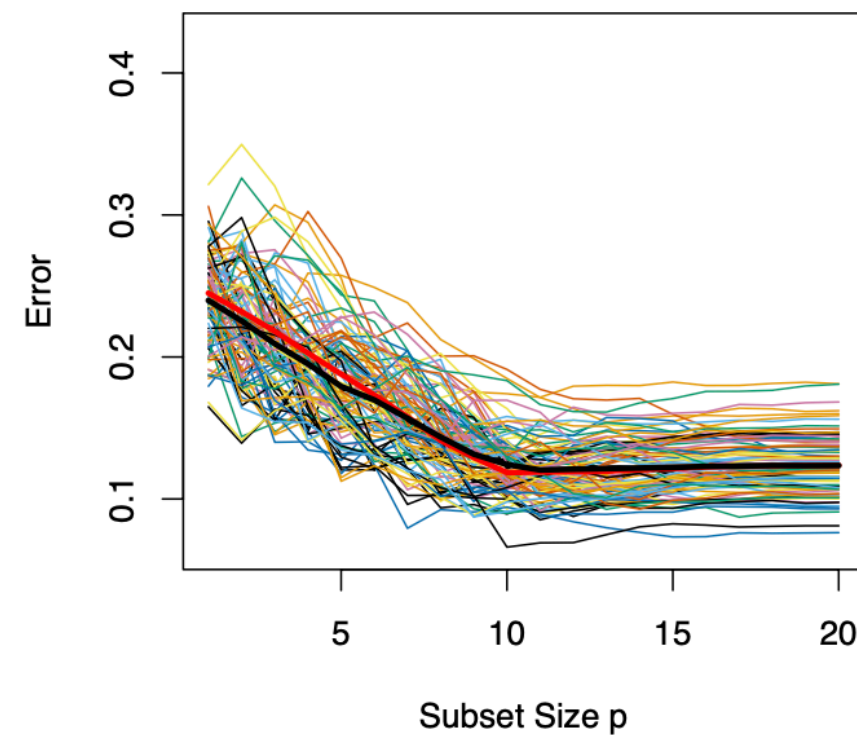
\mathbf{X} is uniformly distributed in the hyper cube $[0,1]^{20}$

Y is 1 if $\sum_{j=1}^{10} X_j > 5$ and 0 otherwise with squared-error loss for a 200-simulated example using linear regression model

Prediction Error



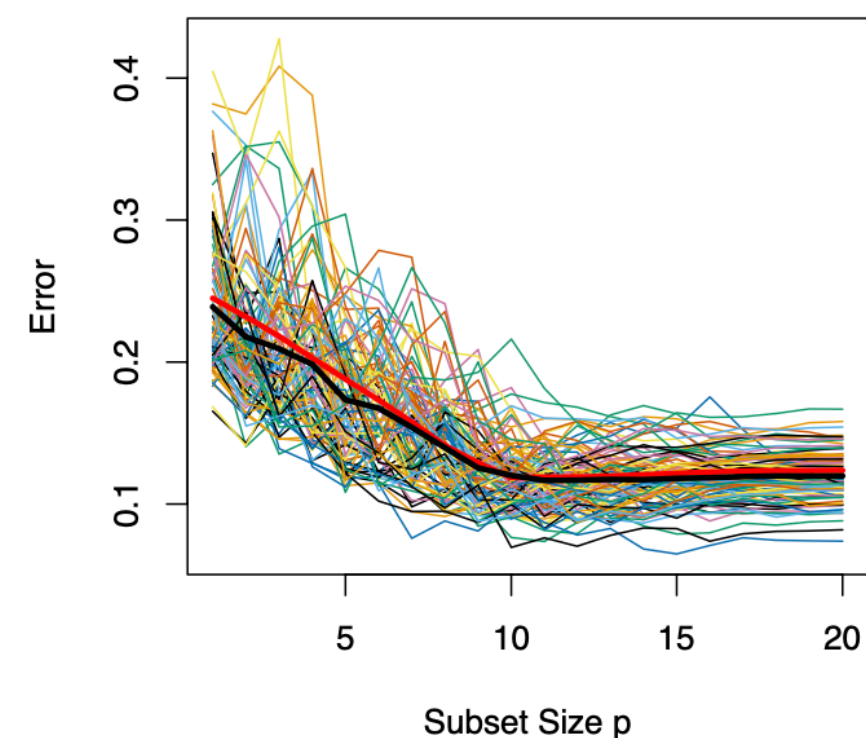
10-Fold CV Error



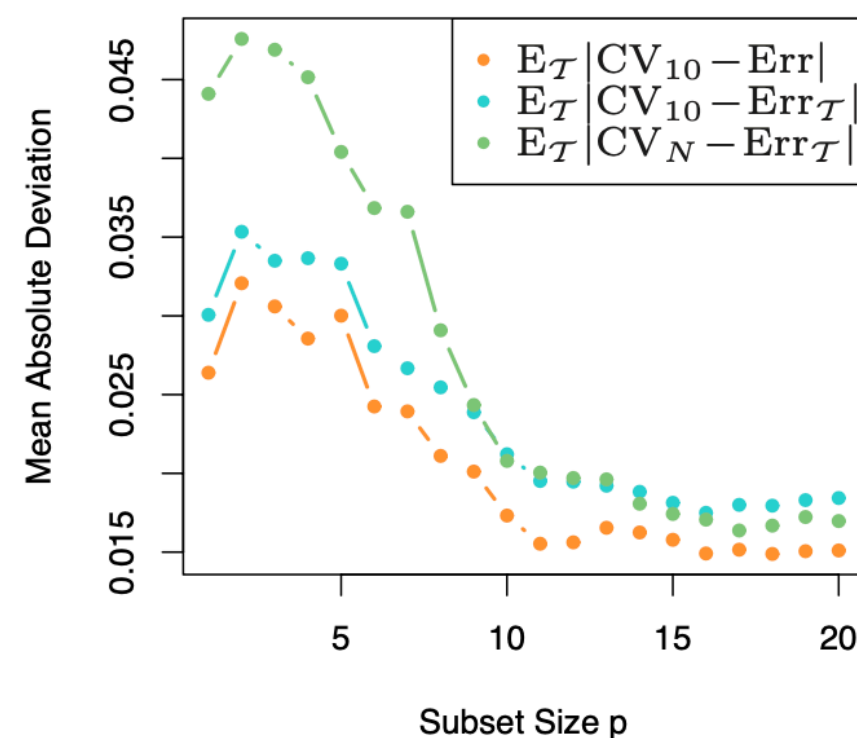
The thick red curve in each plot is Err , while the thick black curves are the expected cross-validation curves.

The similarity of the two black curves with the red curve suggests both CV curves are approximately unbiased for Err , with 10-fold having less variance

Leave-One-Out CV Error

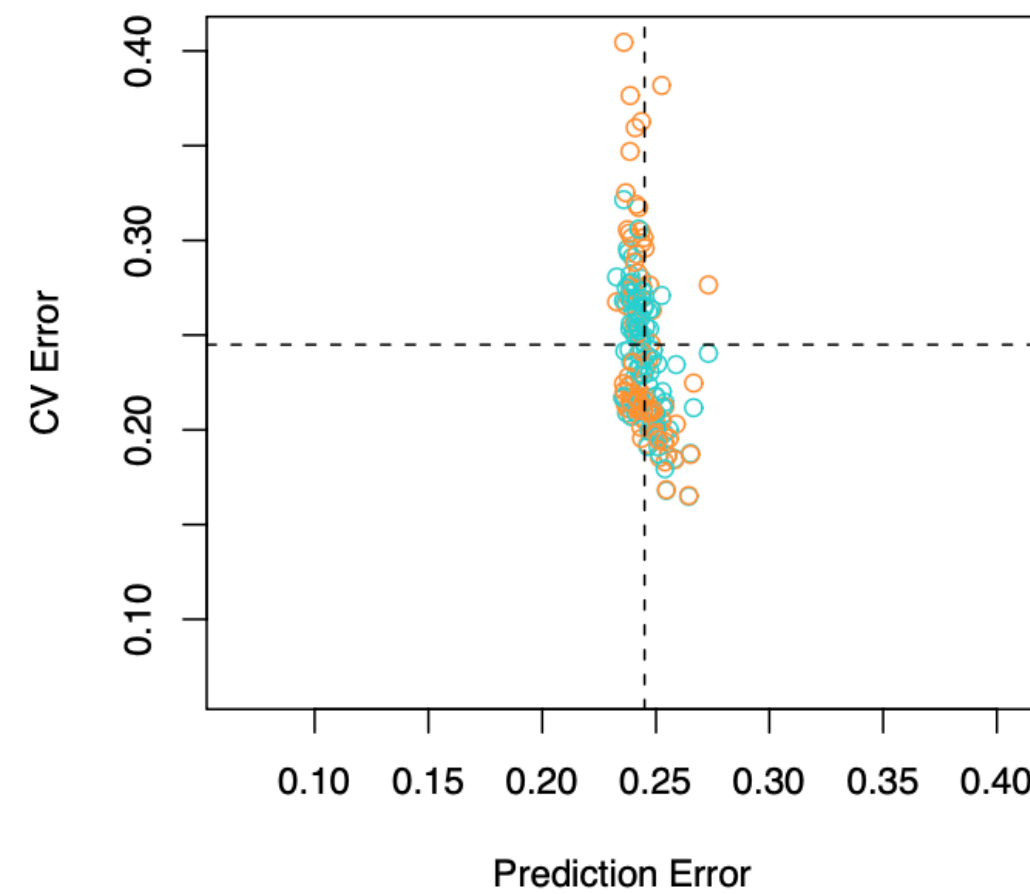


Approximation Error

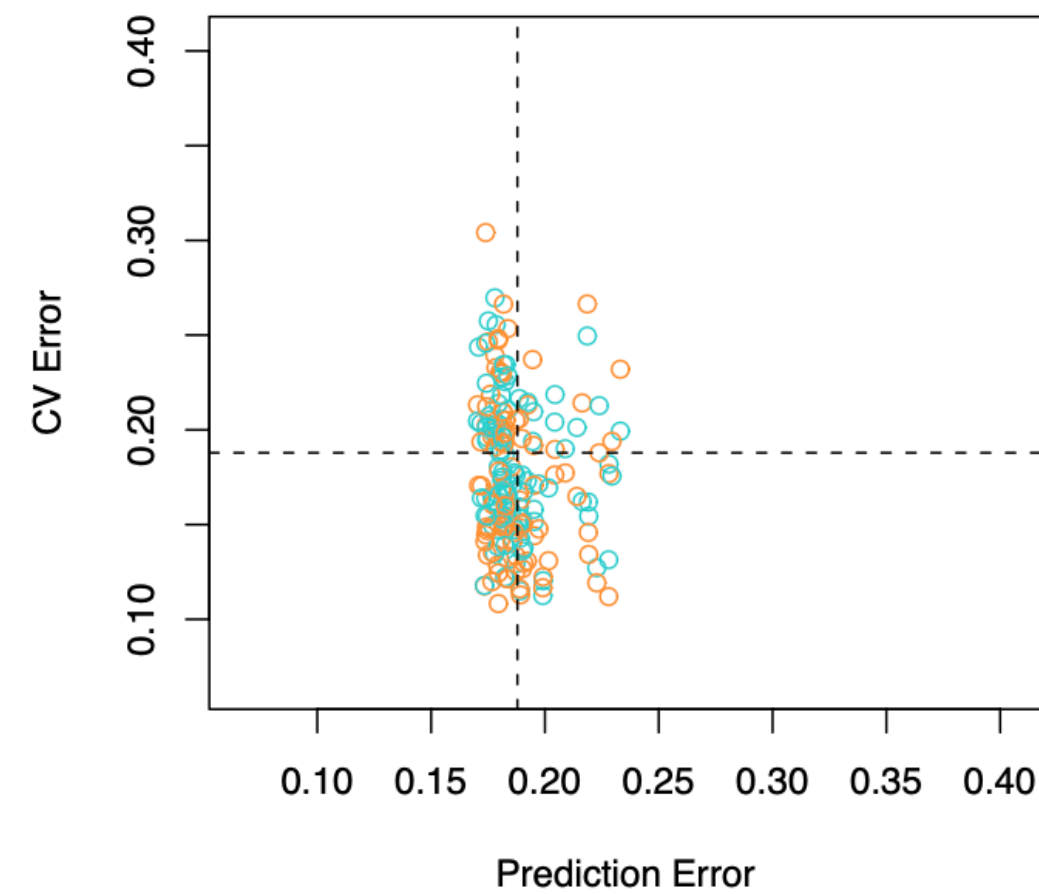


Conditional or Expected Test Error?

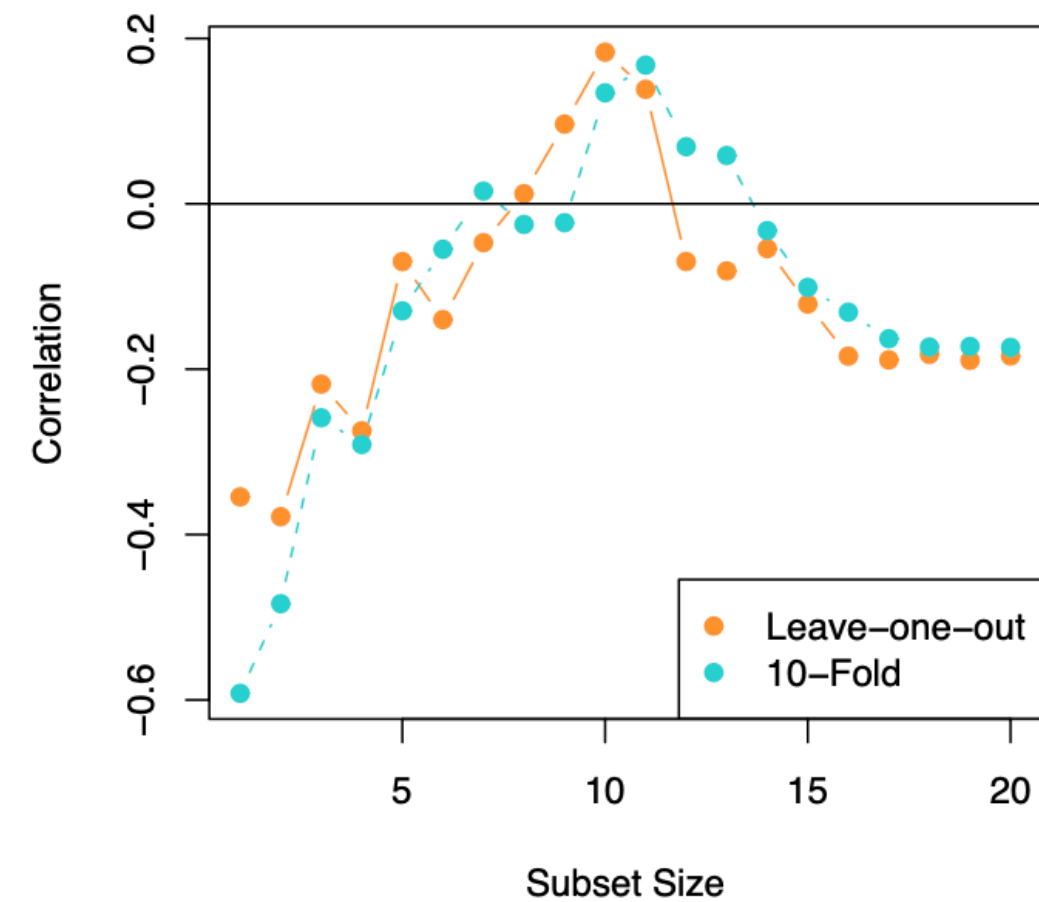
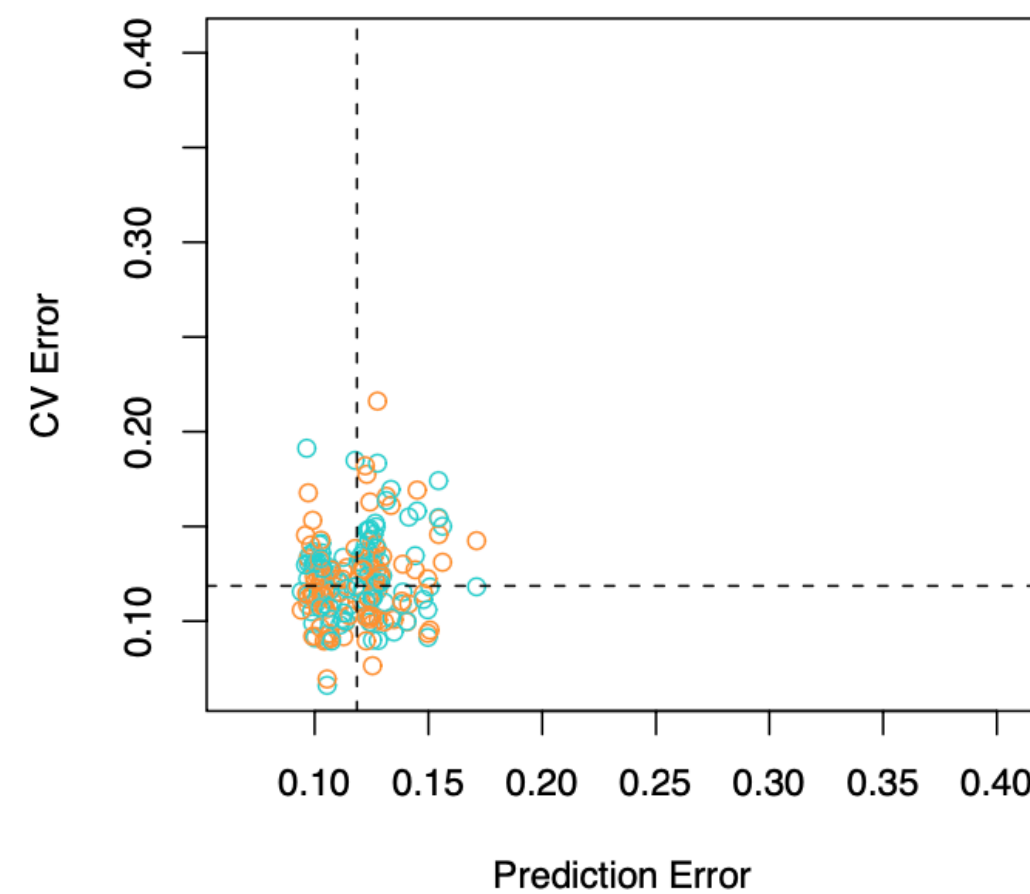
Subset Size 1



Subset Size 5



Subset Size 10



Correlation between CV estimates and $Err_{\mathcal{T}}$

This negative correlation explains why neither form of CV estimates $Err_{\mathcal{T}}$ well.

The broken lines in each plot are drawn at $Err(p)$ for each **best subset**. We see again that both forms of CV are approximately unbiased for Err , but the variation in test error for different training sets is quite substantial.

This phenomenon also occurs for bootstrap estimates of error, and we would guess, for any other estimate of conditional prediction error.

Conditional or Expected Test Error?

- Estimation of test error for a **particular training set** is not easy in general, **given just the data from that same training set**.
- Instead, cross-validation and related methods may provide reasonable estimates of **the expected error *Err***.