# MSEP for model selection

matthieu.lesnoff@cirad.fr
Montpellier, 21 May 2019

# Preliminary

**Difficulties in the literature …**

- Non-trivial concepts behind the “push-button” rules
  - various theoretical frames (quadratic/non quadratic risks, discrepancy, Bayes, information theory, etc.)
- Non consistent vocabulary
- Fixed *vs.* Random parts (several random sources)
  - not often precisely detailed
  - “implicit” notations (ambiguity for non specialists)

## Examples of introductory references

- **Hastie, T., Tibshirani, R.J., 1990**. *Generalized Additive Models*, Monographs on statistics and applied probablity. Chapman and Hall/CRC, New York, USA.

- **Eubank, R.L., 1999**. *Nonparametric Regression and Spline Smoothing*, 2nd ed, Statistics: Textbooks and Monographs. Marcel Dekker, Inc., New York, USA.

- **Hastie, T., Tibshirani, R., Friedman, J., 2009**. *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer, New York, USA.

- … (huge)

**Two separate goals for MSEP statistics**

- **Model selection:** estimating the performance of different models in order to choose the best one

- **Model assessment:** having chosen a final model, estimating its prediction (generalization) error on new data

(Hastie et al. 2009)

This presentation focuses on model selection

# Statistical model

**Joint distribution of the data** $(x, y) \sim F_{x,y}$

$x$ = set of covariables
$y$ = output (scalar, category)

$y|x = E_{x,y}(y|x) + \varepsilon$ Average relation between $y$ and $x$
Relation not perfect

$= g(x, \gamma) + \varepsilon$ $g(x, \gamma)$: "True" deterministic model
(**unknown form**; $\gamma$ may be very large; complex reality)

- $\varepsilon$ independent of $x$
- $\mathrm{E}(\varepsilon) = 0$
- $Var(\varepsilon) = Var_{\varepsilon}(y|x) = \sigma^2$
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$ $i, j$: 2 realizations of $F_{x,y}$

Hypothesis: No error measure on $\boldsymbol{x}$

**Training set** $F_{x,y} \rightarrow \tau = (\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), ..., (x_n, y_n)\}$

$x_i$ = set of covariables for observation $i$
$y_i$ = output for observation $i$

**New observation** $F_{x,y} \rightarrow (x^*, y^*)$

$y^*|x^* = g(x^*, \gamma) + \varepsilon^*$

Same as for $\varepsilon$

- $E(\varepsilon^*) = 0$
- $Var(\varepsilon^*) = Var_{\varepsilon*}(y^*|x^*) = \sigma^2$
- $Cov(\varepsilon_i^*, \varepsilon_j^*) = 0 \quad i \neq j$

**Test set** $F_{x,y} \rightarrow \tau^* = (\boldsymbol{x}^*, \boldsymbol{y}^*) = \{(x_1^*, y_1^*), ..., (x_m^*, y_m^*)\}$

**Let's $\mathcal{M}$ be a given hypothetical model** $\mathcal{M} : f(x, \theta)$

Training set $\tau = (\boldsymbol{x}, \boldsymbol{y}) \rightarrow$ Estimate $\widehat{\mathcal{M}} : f(x, \hat{\theta})$

Predictions $\hat{y}|x_i = f(x_i, \hat{\theta})$

*Residual* $e\,|x_i = y|x_i - f(x_i, \hat{\theta})$ *calibration error*

New observation $y^*|x^*$ non observable

Prediction $\hat{y}^*|x^* = f(x^*, \hat{\theta}) = \hat{y}|x^*$

*Prediction error* $e^*|\,x^* = y^*|x^* - f(x^*, \hat{\theta})$ non-observable

– **Residual** $e|x_i = y|x_i - f(x_i, \hat{\theta})$

$= (g(x_i, \gamma) + \varepsilon_i) - f(x_i, \hat{\theta})$ 1 variation source ($\boldsymbol{\varepsilon}$)

- $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\} \rightarrow \hat{\theta}$
  infinity of training sets $\tau$ of size $n$, with $\boldsymbol{x}$ fixed

– **Prediction error** $e^*|x^* = y^*|x^* - f(x^*, \hat{\theta})$

$= (g(x^*, \gamma) + \varepsilon^*) - f(x^*, \hat{\theta})$ 2 variation sources ($\boldsymbol{\varepsilon}$, $\varepsilon^*$)

- $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_n\} \rightarrow \hat{\theta}$
  infinity of training sets $\tau$ of size $n$, with $\boldsymbol{x}$ fixed

- $\varepsilon^*$

**Expected values and variances of the residual** $e|x_{\mathrm{i}}$

(1 variation source = $\varepsilon$)

- $E_{\varepsilon}(e|x_i)$ $= E_{\varepsilon}(y|x_i - f(x_i, \hat{\theta}))$

  $= E_{\varepsilon i}(y|x_i) - E_{\varepsilon}(f(x_i, \hat{\theta})$

  $= g(x_i, \gamma) - \mathrm{E}_{\varepsilon}(f(x_i, \hat{\theta}))$ Bias term

- $Var_{\varepsilon}(e|x_i)$ $= E_{\varepsilon}(y|x_i - f(x_i, \hat{\theta}))$

  $= Var_{\varepsilon i}(y|x_i) + Var_{\varepsilon}(f(x_i, \hat{\theta})) - 2\times Cov_{\varepsilon}(y|\mathrm{x}_i, f(x_i, \hat{\theta}))$

  $= \sigma^2 + Var_{\varepsilon}(f(x_i, \hat{\theta})) - 2\times Cov_{\varepsilon}(y|\mathrm{x}_i, f(x_i, \hat{\theta}))$

**Expected values and variances of the prediction error $e^*|x_0$ on a given point $x_0$**
(2 variation sources = $\boldsymbol{\varepsilon}$, $\varepsilon^*$)

- Conditional to the training set $\tau = (\boldsymbol{x}, \boldsymbol{y})$ ($\tau$, and then $\hat{\theta}$, fixed)

  - $E_{\varepsilon^*}((e^*|x_0)|\tau) = g(x_0, \gamma) - f(x_0, \hat{\theta}) = \nu(x_0)$ Bias term

  - $Var_{\varepsilon^*}((e^*|x_0)|\tau) = \sigma^2$

- Marginalized over an infinity of training sets $\tau$ of size $n$, with $\boldsymbol{x}$ fixed

  - $E_{\boldsymbol{\varepsilon}, \varepsilon^*}(e^*|x_0) = E_{\boldsymbol{\varepsilon}} E_{\varepsilon^*}((e^*|x_0)|\tau)$

    $= g(x_0, \gamma) - E_{\boldsymbol{\varepsilon}}(f(x_0, \hat{\theta})) = \alpha(x_0)$ Bias term

  - $Var_{\boldsymbol{\varepsilon}, \varepsilon^*}(e^*|x_0) = E_{\boldsymbol{\varepsilon}} Var_{\varepsilon^*}((e^*|x_0)|\tau) + Var_{\boldsymbol{\varepsilon}} E_{\varepsilon^*}((e^*|x_0)|\tau)$

    $= \sigma^2 + Var_{\boldsymbol{\varepsilon}}(f(x_0, \hat{\theta})) = \sigma_*^2(x_0)$

$$Var_{\varepsilon,\varepsilon^*}(e^*|x_0) = \sigma_*^2(x_0) \qquad = \qquad \sigma^2 \quad + \quad Var_\varepsilon(f(x_0, \hat{\theta}))$$

*irreducible error*

**If $x_0$ is set to $x_i$**

$$E_{\varepsilon}(e\,|x_i) = g(x_i, \gamma) - \mathrm{E}_{\varepsilon}(f(x_i, \hat{\theta}))$$

$$E_{\varepsilon,\,\varepsilon^*}(e^*|x_i) = g(x_i, \gamma) - \mathrm{E}_{\varepsilon}(f(x_i, \hat{\theta}))$$

$$Var_{\varepsilon}(e|x_i) = \sigma^2 + Var_{\varepsilon}(f(x_i, \hat{\theta})) - 2\times Cov_{\varepsilon}(y|\mathrm{x}_i, f(x_i, \hat{\theta}))$$

$$Var_{\varepsilon,\,\varepsilon^*}(e^*|x_i) = \sigma_*{}^2(x_i) = \sigma^2 + Var_{\varepsilon}(f(x_i, \hat{\theta}))$$

**Expected square prediction error** **= expected value of** $(e^*|x_0)^2$

Conditional

- $E_{\varepsilon*}((e^*|x_0)^2|\,\tau) = \sigma^2 + (g(x_0, \gamma) - f(x_0, \hat{\theta}))^2$

$= \sigma^2 + \nu(x_0)^2$

Marginal

- $E_{\varepsilon}E_{\varepsilon*}((e^*|\, x_0)^2|\,\tau) = MSEP(x_0) = PR(x_0) = EPE\,(x_0) = PSE(x_0) = \ldots$

= *Mean square error of prediction*, *Predictive risk*,
*Expected square error*, *Predictive square error*, …

$= \sigma^2 + E_{\varepsilon}((g(x_0, \gamma) - f(x_0, \hat{\theta}))^2)$

$= \sigma^2 + MSE(x_0) \quad = \sigma^2 + Risk(x_0)$

$= \sigma^2 + Var_{\varepsilon}(f(x_0, \hat{\theta})) + (g(x_0, \gamma) - E_{\varepsilon}(f(x_0, \hat{\theta}))^2$

$= \sigma_*^2(x_0) + \alpha(x_0)^2$

- $MSEP(x_0) = \sigma_*^2(x_0) + \alpha(x_0)^2$

The bias term $\alpha(x_0)^2$ can be split into two terms representing a

*model bias* ($f(x_0, \theta)$ *vs.* $g(x_0, \gamma)$) and a *statistical bias* ($f(x_0, \hat{\theta})$ *vs.* $f(x_0, \theta)$)

# Model performances

**Loss function**

In this presentation: theoretical framework based on a *loss function* $L$

- $L(y|x\ , f(x, \hat{\theta}))$ : Quantity of loss when prediction $\hat{y}|x = f(x, \hat{\theta})$ is used instead of the realization $y|x$ from $F_{x,y}$

Example: quadratic loss function

- $L(y|x\ , f(x, \hat{\theta})) = (y|x - f(x, \hat{\theta}))^2$

**Two main types of performance measures**

Definitions and notations of Hastie *et al.* 2009 p.220

- *Conditional test (or generalization) error* $Err_{\tau}$
- *Expected prediction (or test) error* $Err$

$Err$ also used in Efron 1983

- *Efron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316–331*

Most statistical methods estimate $Err$

Estimating $Err$ **from the training set** $\rightarrow$ model selection: model(s) with the lowest $Err$ estimate(s)

Conditional test error

- $Err_\tau \quad = E_{x^*, y^*}\,(L(y^*|x^*, f(x, \hat{\theta}))\ |\ \tau)$

$F_{x,y}$

New observations

Fixed training set

Here the training set $\tau$ is fixed (therefore $\hat{\theta}$ also)
Test error refers to the error for this specific training set
(Hastie *et al.* 2019 p. 220)

Expected prediction error

- $Err$ $= E_\tau(Err_\tau)$ Marginal expectation over an infinity of $\tau$ (size $n$)

$$= E_\tau E_{x^*, y^*}\left(L(y^*|x^*, f(x, \hat{\theta})) \mid \tau\right)$$

$F_{x,y}$ ~ new training set of size $n$

$F_{x,y}$ ~ new observations

$$= E_{\mathbf{x},\mathbf{y}} E_{x^*, y^*}\left(L(y^*|x^*, f(x^*, \hat{\theta})) \mid \tau\right)$$

$$= E_{\mathbf{x},\boldsymbol{\varepsilon}} E_{x^*, y^*}\left(L(y^*|x^*, f(x^*, \hat{\theta})) \mid \tau\right)$$

**Average training error $\overline{err}$ as an estimate of $Err$?**

- $\overline{err} = \frac{1}{n} \sum_{i=1}^{n} L(y|x_i, f(x_i, \hat{\theta}))$

Example for a quadratic loss function

- $\overline{err} = \frac{1}{n} \sum_{i=1}^{n} (y|x_i - f(x_i, \hat{\theta}))^2 \qquad = ASR = RSS/n$

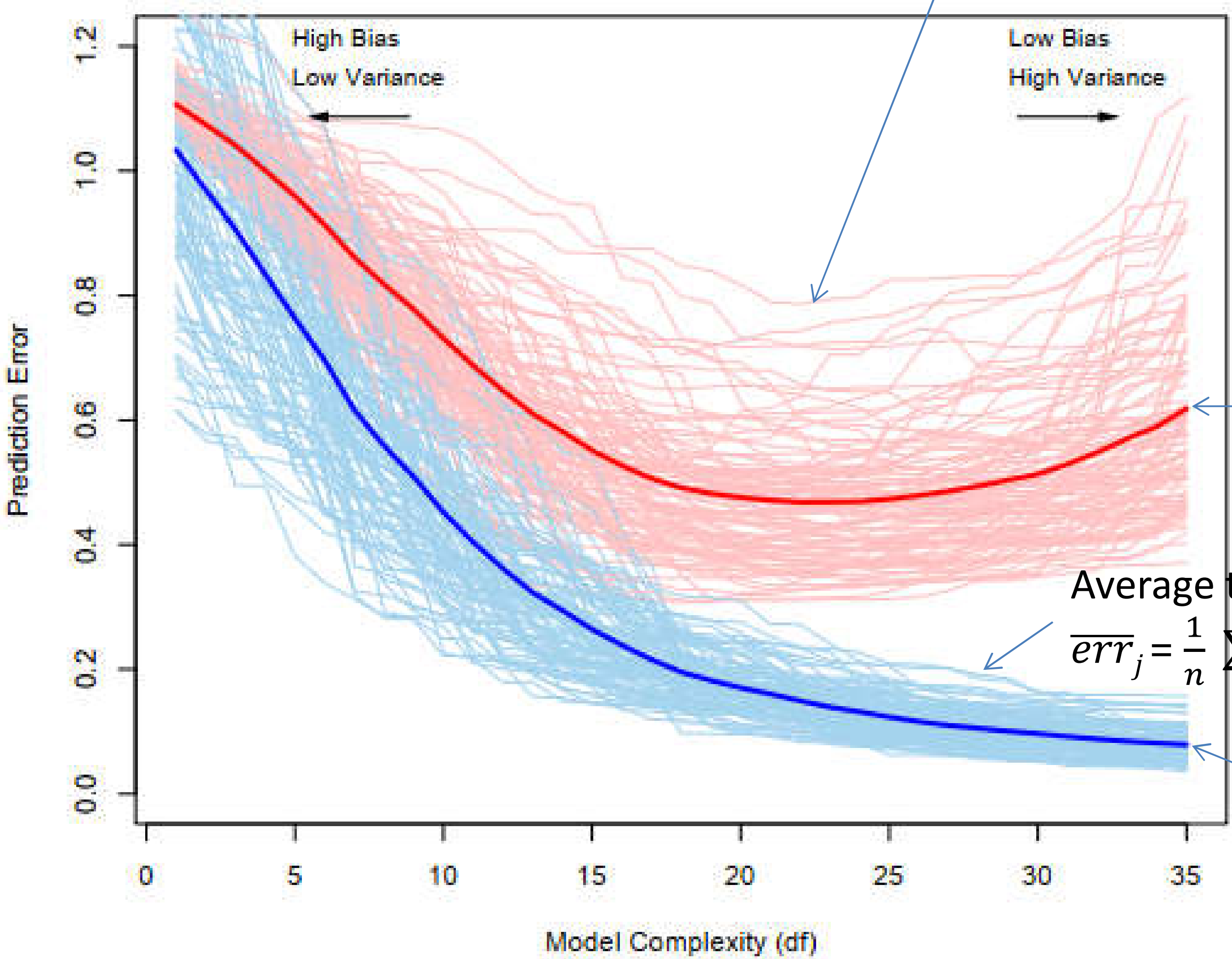
Unfortunately, $\overline{err}$ is generally biased downward as an estimate of $Err$

**Hastie *et al* 2009 Fig 7.1 p. 220**

Simulation of 100 training set $\tau_j$ $j = 1, \ldots, 100$

Known true model $g(x, \gamma)$

$Err_{\tau j} = E_{x^*, y^*}(L(y^*|x^*, f(x, \hat{\theta})) \mid \tau_j)$




$Err = E_\tau(Err_{\tau j})$

Average training error

$\overline{err}_j = \frac{1}{n} \sum_{i=1}^{n} [L(y_{(j)}|x_i, f(x_{i(j)}, \hat{\theta}_j))]$

$E_\tau(\overline{err})$

**Estimation of $Err$ from the training set $\tau$**

- Formal: *Parsimony criterions* (Mallows's $Cp$, Akaike $AIC$, etc.)

  $\rightarrow$ Estimates of a particular case of $Err$ ($E_{\tau}(Err_{\text{in}})$ see later)

- Simulations

  $\rightarrow$ Direct estimates of $Err$

  - Bootstrap (e.g. ".632 estimate")
  - Cross-validation
    - K-Fold
    - LOO
      For linear smoothers: (1) LOO can be calculated without simulation, (2) Alternative: Generalized cross-validation (approximation of LOO)

# Examples of formal estimation of $Err$

**Background**

- Conditional test error $Err_\tau$ $= E_{x^*, y^*}\,(L(y^*|x^*, f(x, \hat{\theta})) \mid \tau)$

  $= E_{x^*, \varepsilon^*}\,(L(y^*|x^*, f(x, \hat{\theta})) \mid \tau)$

Here the test input vector $\boldsymbol{x}^*$ does not need to coincide with the training input vector $\boldsymbol{x}$

$\Rightarrow$ Conditional *extra-sample* test error
(Hastie *et al.* 2009 p. 228)

- $Err_\tau$ $= E_{x^*,\, \varepsilon^*}\left(L(y^*|x^*, f(x, \hat{\theta})) \mid \tau\right)$

  = Conditional *extra-sample* test error

For model selection, we need to estimate $Err = E_\tau(Err_\tau)$ from the training set $\tau$ = $(\boldsymbol{x}, \boldsymbol{y})$ $\rightarrow$ In $Err_\tau$, we force the test input vector $\boldsymbol{x}$ to coincide with $\boldsymbol{x}$

- $Err_{\text{in}}$ $= \frac{1}{n} \sum_{i=1}^{n} E_{\varepsilon^*}\left(L(y^*|x_i, f(x_i, \hat{\theta})) \mid \tau\right)$

  = Conditional *in-sample* test error
  (Hastie et al. 2009 p. 228)

  = Plug-in of $Err_\tau$ on $\boldsymbol{x}$ : $F_x \sim x^*$ is replaced by the empirical distribution $\boldsymbol{x}$

$Err_{\text{in}}$ is a special case of $Err_\tau$

We are going to estimate $E_\tau(Err_{\text{in}})$ in place of $E_\tau(Err_\tau) = Err$

- $Err_{\text{in}} \quad = \frac{1}{n}\sum_{i=1}^{n} E_{\varepsilon *}(L(y^*|x_i, f(x_i, \hat{\theta})) \mid \tau)$

- $E_\tau(Err_{\text{in}}) = \frac{1}{n}\sum_{i=1}^{n} E_\tau E_{\varepsilon *}(L(y^*|x_i, f(x_i, \hat{\theta})) \mid \tau)$

The estimation of $E_\tau(Err_{\text{in}})$ is simplified under another hypothesis:
For the variations of $\tau$, the training input $\boldsymbol{x}$ is set fixed (only $\boldsymbol{\varepsilon}$ varies)

$$= \frac{1}{n}\sum_{i=1}^{n} E_\varepsilon E_{\varepsilon *}(L(y^*|x_i, f(x_i, \hat{\theta})) \mid \tau)$$

= the $Err$ criterion to be estimated

**Case of a quadratic loss function**

$$E_{\varepsilon}(Err_{\mathrm{in}}) = \frac{1}{n}\sum_{i=1}^{n} E_{\varepsilon} E_{\varepsilon*}\left(\left(y^*|x_i - f(x_i, \hat{\theta})\right)^2 \mid \tau\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E_{\varepsilon} E_{\varepsilon*}\left((e^*|x_i)^2 \mid \tau\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} MSEP(x_i)$$

$= MSEP(\boldsymbol{x})$ To be estimated

- $MSEP(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} MSEP(x_i)$

$= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} MSE(x_i) = \sigma^2 + MSE(\boldsymbol{x})$

$= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} Var_\varepsilon(f(x_i, \hat{\theta})) + \frac{1}{n}\sum_{i=1}^{n} \alpha(x_i)^2$ with $\alpha(x_i) = g(x_i, \gamma) - E_\varepsilon(f(x_i, \hat{\theta}))$

$= \overline{\sigma}_*^2(\boldsymbol{x}) + \frac{1}{n}\alpha(\boldsymbol{x})'\alpha(\boldsymbol{x})$ with $\overline{\sigma}_*^2(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \sigma_*^2(x_i)$

Minimizing $MSEP(\boldsymbol{x})$ is the same as minimizing $MSE(\boldsymbol{x})$

Minimization of a “variance-bias” compromise

- When the dimension of $\hat{\theta}$ (model) increases, the bias term decreases but the mean variance of the prediction errors $\overline{\sigma}_*^2(\boldsymbol{x})$ increases

- When $n$ increases, the training set $\tau$ allows higher $\hat{\theta}$ dimensions

**An example of $MSEP(\boldsymbol{x})$ estimation: The Mallows's $Cp$ approach**

- $MSEP(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} E_{\varepsilon} E_{\varepsilon*}((y|x_i - f(x_i, \hat{\theta}))^2 |\ \tau)$

- $\overline{err} = \frac{1}{n} \sum_{i=1}^{n} (y|x_i - f(x_i, \hat{\theta}))^2 = RSS/n$

As before, we could consider $\overline{err} = RSS/n$ as an estimate of $MSEP(\boldsymbol{x})$

But in general $RSS/n$ is biased for $MSEP(\boldsymbol{x})$ : $E_{\varepsilon}(RSS/n) < MSEP(\boldsymbol{x})$

Important hypothesis for the $Cp$ approach

For all the next calculations, we consider that the models $f(x, \hat{\theta})$ are linear in their parameters

$$\hat{\boldsymbol{y}} \mid \boldsymbol{x} = f(\boldsymbol{x}, \hat{\theta}) = \boldsymbol{H}\,\boldsymbol{y}$$

where $\boldsymbol{H}$ does not depend on $\boldsymbol{y}$

$\in$ *Linear smoothers* (Hastie & Tibshirani 1990)
(linear models, ridge regression, PCR, cubic splines, …)

Ex: Usual OLS $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

$$MSEP(\boldsymbol{x}) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} Var_{\varepsilon}(f(x_i, \hat{\theta})) + \frac{1}{n}\sum_{i=1}^{n} \alpha(x_i)^2$$

$$= \sigma^2 + \frac{1}{n}Tr(\boldsymbol{H}\boldsymbol{H}')\sigma^2 + \frac{1}{n}\alpha(\boldsymbol{x})'\alpha(\boldsymbol{x})$$

$$E_{\varepsilon}(RSS/n) = \sigma^2 + \frac{1}{n}Tr(\boldsymbol{H}\boldsymbol{H}')\sigma^2 + \frac{1}{n}\alpha(\boldsymbol{x})'\alpha(\boldsymbol{x}) - \frac{2}{n}Tr(\boldsymbol{H})\sigma^2$$

$$= MSEP(\boldsymbol{x}) - \underbrace{\frac{2}{n}Tr(\boldsymbol{H})\sigma^2}$$

Bias of $RSS/n$ for $MSEP(\boldsymbol{x})$
Increases with the model dimension

$\Rightarrow$ One approach for estimating $MSEP$ is correcting $RSS/n$ by its bias

$$MSEP(\boldsymbol{x}) \qquad = E_{\varepsilon}(RSS/n) + \frac{2}{n}Tr(\boldsymbol{H})\sigma^2$$

$$M\hat{S}EP(\boldsymbol{x}) \qquad = \frac{1}{n}\ RSS + \frac{2}{n}Tr(\boldsymbol{H})\sigma^2 \qquad = \text{Mallows's } Cp \text{ approach}$$

This approach estimates $MSEP(\boldsymbol{x})$ without estimating the bias $\boldsymbol{\alpha}$
(which is very useful since $g(x, \gamma)$ is unknown)

… but an estimate of $\sigma^2$ is still needed

Usual recommendation:

Estimating $\sigma^2$ from an over-parameterized model ("little smoothing")
$\rightarrow$ having low bias

Let $\mathcal{M}_0$ be a model with low bias (in practice, often the maximal model)

$$E_{\varepsilon}(RSS_0/n) = \sigma^2 - \frac{1}{n}Tr(2\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H}_0{}')\sigma^2 + \frac{1}{n}\boldsymbol{\alpha}_0{}'\boldsymbol{\alpha}_0$$

$$\Rightarrow E_{\varepsilon}(RSS_0) = n\ \sigma^2 - Tr(2\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H}_0{}')\sigma^2 + \boldsymbol{\alpha}_0{}'\boldsymbol{\alpha}_0$$

Low bias $\Rightarrow \boldsymbol{\alpha}_0{}'\boldsymbol{\alpha}_0 \approx 0$

$$\Rightarrow E_{\varepsilon}(RSS_0) \approx n\ \sigma^2 - Tr(2\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H}_0{}')\sigma^2$$

$$\Rightarrow \widehat{\sigma}_0{}^2 = RSS_0 \ / \ (n - Tr(2\boldsymbol{H}_0 - \boldsymbol{H}_0\boldsymbol{H}_0{}'))$$

$\rightarrow$ Final estimate

$$M\hat{S}EP(\boldsymbol{x}) = \underbrace{\frac{1}{n}\ RSS}_{\overline{err}} + \underbrace{\frac{2}{n} Tr(\boldsymbol{H})\ \widehat{\sigma}_0{}^2}$$

Expected *model optimism*
($\omega$ ; Efron 1983, and Hastie *et al.* 2009 p. 229)

Penalty increasing with the model dimension

$Tr(\boldsymbol{H})$ = *model df*
*Effective number of parameters* (Hastie *et al.* 2009 p. 231)
Indication on the quantity of smoothing generated by $\boldsymbol{H}$

$$M\hat{S}EP(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{2}{n}Tr(\boldsymbol{H})\ \widehat{\sigma}_0{}^2$$

If $\boldsymbol{H}$ is idempotent ($\boldsymbol{HH} = \boldsymbol{H}$ projector)

$$M\hat{S}EP(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{2}{n}r(\boldsymbol{H})\ \widehat{\sigma}_0{}^2$$

For OLS models $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ ,
$\boldsymbol{H}$ symmetric (orthogonal projector) $\Rightarrow$ $Tr(\boldsymbol{H}) = r(\boldsymbol{H}) = r(\boldsymbol{X}) = p$

$$M\hat{S}EP(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{2}{n}p\ \widehat{\sigma}_0{}^2$$

Original expression of the Mallows's $Cp$

- *Mallows, C.L., 1973. Some Comments on Cp. Technometrics 15, 661–675*

$MSEP(\boldsymbol{x}) = \sigma^2 + MSE(\boldsymbol{x}) \quad \Rightarrow \quad MSE(\boldsymbol{x}) = MSEP(\boldsymbol{x}) - \sigma^2$

$\Rightarrow \quad \frac{n}{\sigma^2} MSE(\boldsymbol{x}) = \frac{n}{\sigma^2} MSEP(\boldsymbol{x}) - n$ *"Scaled risk"* Mallows 1973

$Cp = \frac{n}{\widehat{\sigma}_0{}^2} M\hat{S}E(\boldsymbol{x})$ $Cp$ is an estimate of the scaled risk

$= \frac{n}{\widehat{\sigma}_0{}^2} M\hat{S}EP(\boldsymbol{x}) - n$

$= \frac{1}{\widehat{\sigma}_0{}^2} RSS + 2Tr(\boldsymbol{H}) - n$

$= \frac{1}{\widehat{\sigma}_0{}^2} RSS + 2p - n$ (OLS models) Eq.3 in Mallows 1973

## Examples of other parsimony criterions than $Cp$

$AIC = n \, log(RSS/n) + 2\,p$ *Akaike criterion*

Maximum likelihood estimation

Information theory

Akaike 1974, Burnham & Anderson 1998

$FPE = \frac{n+p}{n-p} RSS$ *Final prediction error*

Akaike 1970, Shibata 1984

$Cp$ is very similar (and asymptotically equivalent) to $AIC$ and $FPE$


- Akaike, H., 1970. Statistical predictor identification. Ann Inst Stat Math 22, 203–217
- Akaike, H., 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19, 716–722
- Burnham, K.P., Anderson, D.R., 1998. Model selection and inference. A practical information-theoretic approach. Springer, New York.
- Shibata, R., 1984. Approximate efficiency of a selection procedure for the number of regression variables. Biometrika 71, 43–49.

Ex:

For a linear model with $p$ independent parameters

$$Cp(p) = \frac{1}{\widehat{\sigma}_0{}^2} RSS(p) + 2k - n$$

$$\frac{FPE(p)}{\widehat{\sigma}_p{}^2} = \frac{1}{\widehat{\sigma}_p{}^2} RSS(p) + 2p$$

$Cp$ and $FPE$ simply use two different estimates of $\sigma^2$

- $Cp$ uses $\widehat{\sigma}_0{}^2$ : estimate from the maximal (low biased) model
- $FPE$ uses $\widehat{\sigma}_p{}^2$ : estimate from the model under evaluation

$Cp$ tends to overfit (non null asymptotic probability of overfitting)

$$M\hat{S}EP(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{2}{n} p\ \widehat{\sigma}_0{}^2$$

- *Zhang, P., 1992. On the Distributional Properties of Model Selection Criteria. Journal of the American Statistical Association 87, 732–737*

An approach is to increase the penalty $\rightarrow$ generalized indicators

- *Shibata, R., 1984. Approximate efficiency of a selection procedure for the number of regression variables. Biometrika 71, 43–49.*

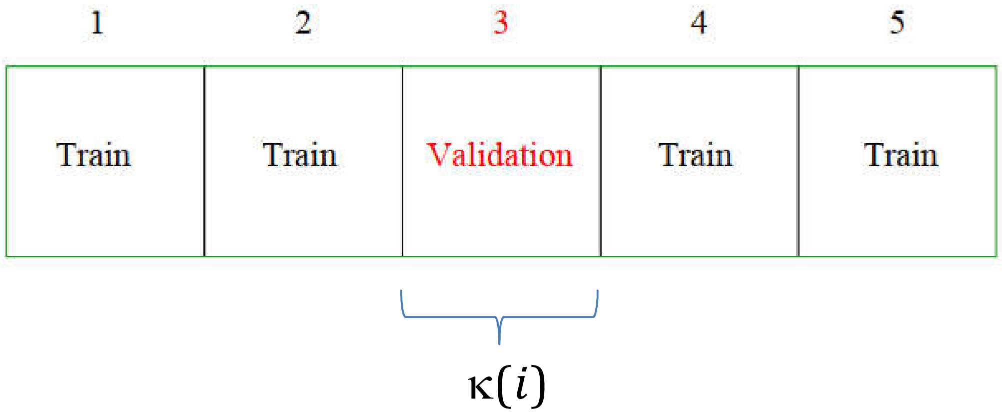
Ex:

$$M\hat{S}EP_g(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{a}{n} p\ \widehat{\sigma}_0{}^2 \quad \text{with } a > 2$$

$$M\hat{S}EP_g(\boldsymbol{x}) = \frac{1}{n}\ RSS + \frac{\log(n)}{n} p\ \widehat{\sigma}_0{}^2 \quad BIC \text{ approach}$$

# Estimating $Err$ by cross validation

## K-Fold CV and LOO CV




Ex: $K = 5$
From Hastie *et al.* 2009

$$\hat{C}V_{\text{K-Fold}} = \frac{1}{n} \sum_{i=1}^{n} L(y|x_i, f(x_i, \hat{\theta}^{-\kappa(i)}))$$

$$K = n \quad \Rightarrow \quad \hat{C}V_{\text{LOO}} = \frac{1}{n} \sum_{i=1}^{n} L(y|x_i, f(x_i, \hat{\theta}^{-i}))$$

Quadratic loss

$$\hat{C}V_{\text{K-Fold}} = \frac{1}{n} \sum_{i=1}^{n} \left(y|x_i - f(x_i, \hat{\theta}^{-\kappa(i)}\right)^2$$

$$= M\hat{S}EP_{\text{CV}}$$

Both $\hat{C}V_{\text{LOO}}$ and $\hat{C}V_{\text{K-Fold}}$ estimate the expected prediction error $Err$ (directly, since here the training set is varied, artificially)

$$Err \;= E_{\tau}(Err_{\tau}) \qquad \text{Marginal expectation over an infinity of } \tau \text{ (size } n\text{)}$$

$$= E_{\tau} E_{x^*, y^*} \left(L(y^* | x^*, f(x, \hat{\theta})) \mid \tau\right)$$

but CV loses the training size constraint $n \rightarrow$ potential bias

- $\hat{C}V_{\text{LOO}} = \hat{E}rr_{\text{LOO}}$ Approximately unbiased but high variance (almost uses the full training sample to fit a new test point)

- $\hat{C}V_{\text{K-Fold}} = \hat{E}rr_{\text{K-Fold}}$ Lower variance but potentially biased Over-estimation of $Err$ if the CV training is set too small

Hastie *et al.* 2009 p. 243

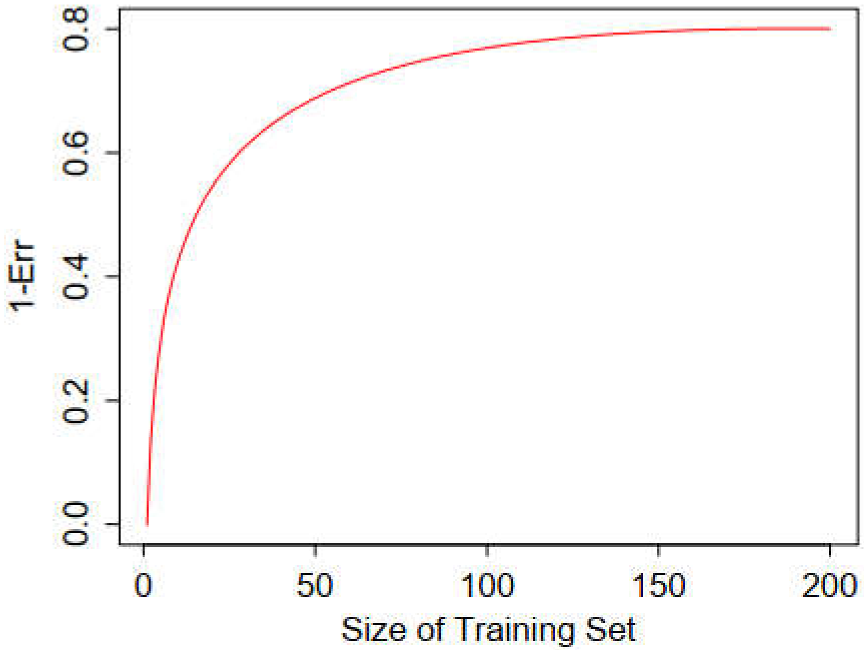



**FIGURE 7.8.** *Hypothetical learning curve for a classifier on a given task: a plot of* $1 - \mathrm{Err}$ *versus the size of the training set* $N$*. With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset of 50 observations fivefold cross-validation would use training sets of size 40, and this would result in a considerable overestimate of prediction error.*

Usual recommendations: $K$ = 5-10 (Hastie *et al.* 2009)

$K \geq 20$ (Kohavi 1995)

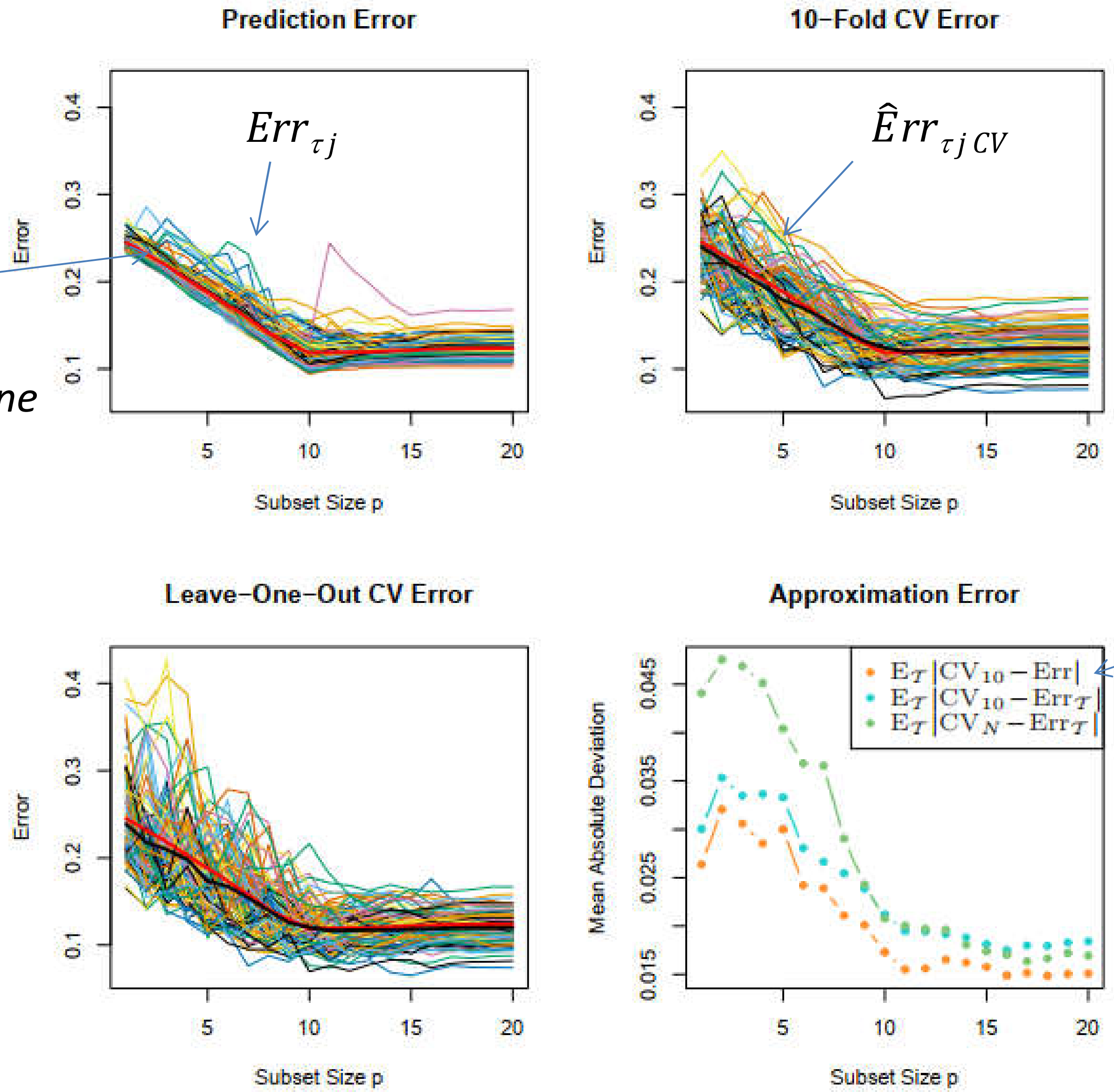
… (lot of references with simulation studies)

No definitive rules


- *Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143*

**Hastie *et al* 2009   Fig 7.14 p. 220**

Simulation of 100 training set $\tau_j$   $j = 1, \ldots, 100$   with known true model $g(x, \gamma)$




$Err$

$= E_\tau(Err_\tau)$

*Thick red line*

$Err$

*Thick red line*

$\text{Mean}(\hat{E}rr_{\tau j\, CV})$

*Thick black line*

Mean absolute deviations:

$\text{Mean}|\hat{E}rr_{\tau j\, CV} - Err|$

$\text{Mean}|\hat{E}rr_{\tau j\, CV} - Err_\tau|$

**On the example of Fig. 7.14**

- $\mathrm{Mean}(\hat{E}rr_{\tau j\,CV})$ very different from $Err_{\tau j}$ $\quad$ ($\rightarrow$ bias)

  $\Rightarrow$ CV does not estimate $Err_{\tau}$

  Surprisingly, even worst for $\hat{C}V_{\mathrm{LOO}}$ than for $\hat{C}V_{\text{K-Fold}}$ (lower right panel).

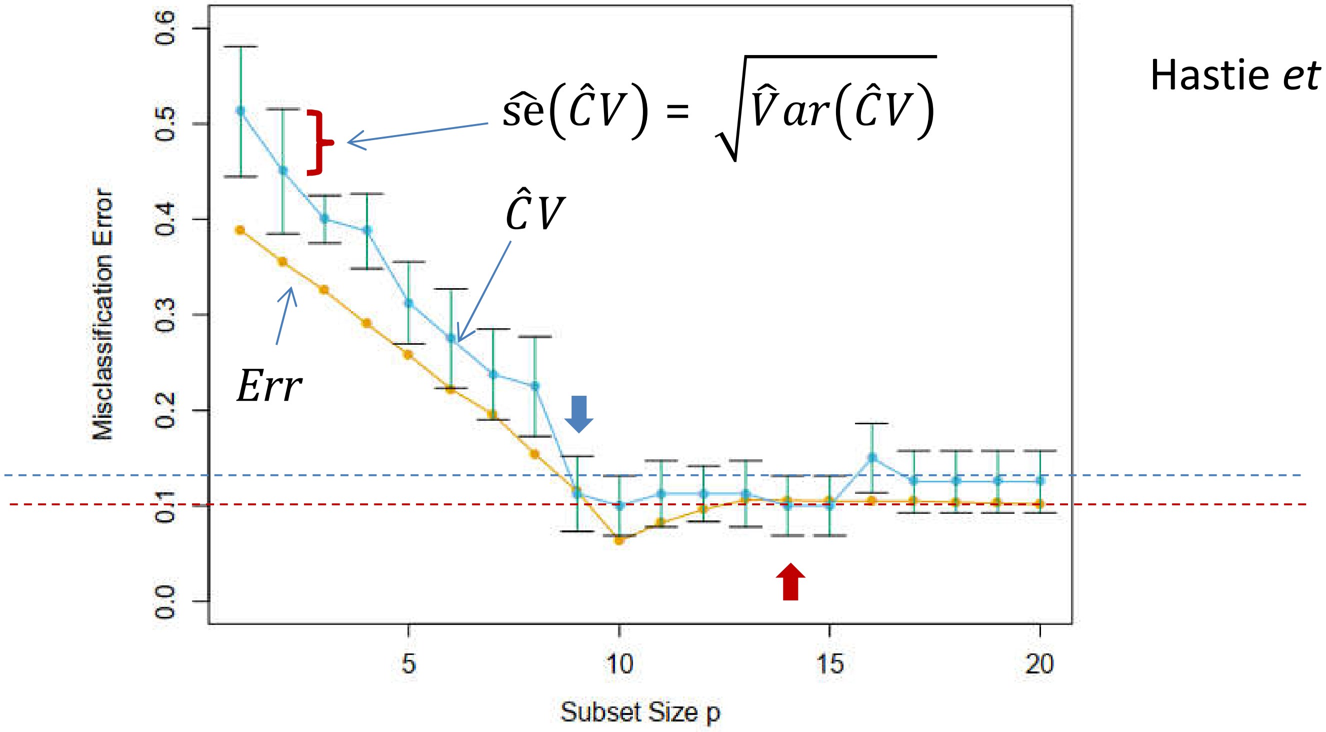
- $\mathrm{Mean}(\hat{E}rr_{\tau j\,CV}) \approx Err$ $\quad$ (see the red and black thick curves)

  $\Rightarrow$ $\hat{C}V_{\mathrm{LOO}}$ and $\hat{C}V_{\text{K-Fold}}$ are approximately unbiased estimates of $Err$

- The variance of $\hat{E}rr_{\tau j\,CV}$ is globally higher for $\hat{C}V_{\mathrm{LOO}}$ than for $\hat{C}V_{\text{K-Fold}}$

## The "one standard-error" rule

Final selection of the most parsimonious model whose error is no more than one standard error above the error of the best model.




Hastie *et al* 2009 p. 244

**FIGURE 7.9.** *Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3.*

Tibshirani *et al.* 2019 does not details the calculation of $\widehat{\mathrm{se}}(CV)$

- One approach is proposed in lecture notes of Tibshirani Jr 2013:

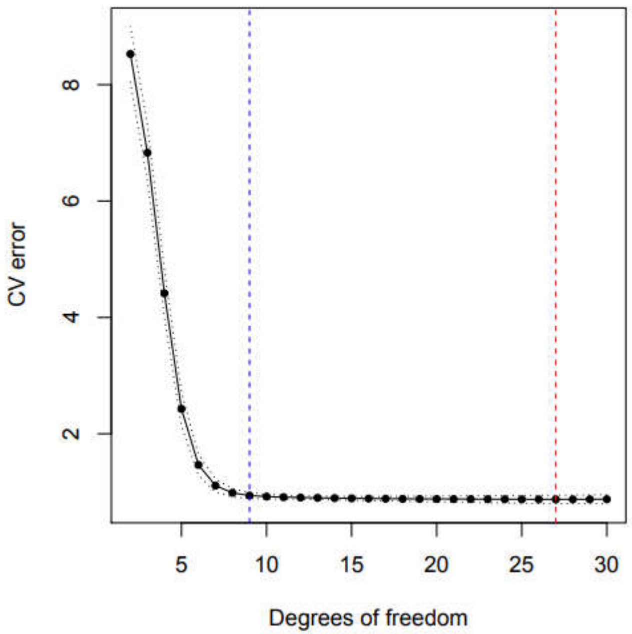
  *http://www.stat.cmu.edu/~ryantibs/datamining/lectures/18-val1.pdf*
  *http://www.stat.cmu.edu/~ryantibs/datamining/lectures/19-val2.pdf*

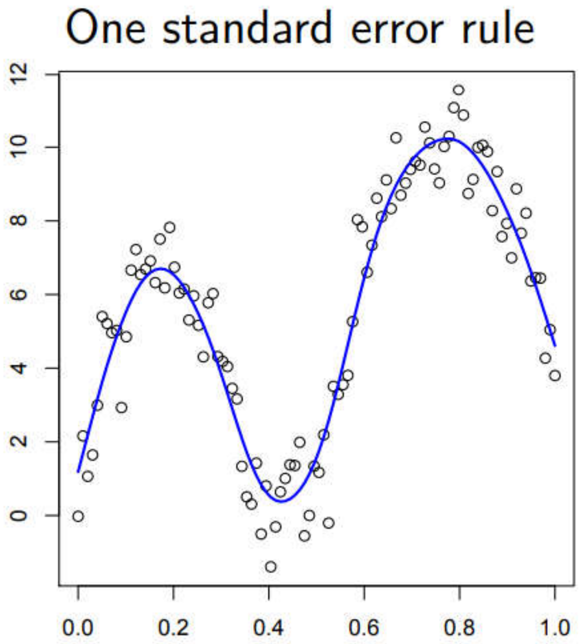
  $$\widehat{\mathrm{se}}\big(\hat{C}V\big) = \sqrt{\hat{V}ar\big(\{\hat{C}V(1), \ldots, \hat{C}V(K)\}\big)/K}$$

  See also (p. 162): *Filzmoser, P., Liebmann, B., Varmuza, K., 2009. Repeated double cross validation. Journal of Chemometrics 23, 160–171.*

- Alternative idea: For a quadratic loss function, using the *Chi2* approximation (same principle as for the test set)

From Tibshirani Jr 2013        Smoothing spline




The one standard error rule selects a model with 9 degrees of freedom

**The LOO "short-cut"**

LOO-CV is very time consuming (or even impracticable) for large training set

But for some models, LOO-CV does not require simulation

- In particular, for models linear in their parameters

  $\hat{\boldsymbol{y}} \mid \boldsymbol{x} = f(\boldsymbol{x}, \hat{\theta}) = \boldsymbol{H}\,\boldsymbol{y}$ where $\boldsymbol{H}$ does not depend on $\boldsymbol{y}$

  and *constant preserving*

  $\boldsymbol{H}\,\boldsymbol{1} = \boldsymbol{1}$ $\quad \sum_{j=1}^{n} h_{ij} = 1$ $\quad h_{ij}$ weight $j$ for observation (row) $i$

Then

$$\hat{C}V_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y|x_i - f(x_i, \hat{\theta})}{1-h_{ii}}\right)^2$$ **= *LOO short-cut***

For quadratic loss

$$E_{\varepsilon}(\hat{C}V_{\mathrm{LOO}}) \approx \underbrace{MSEP(\boldsymbol{x})}_{\sim Err} + \underbrace{\frac{2}{n}\sum_{i=1}^{n} h_{ii}\,\alpha_i^2}_{\text{Bias term } (> 0)}$$

We see again that $\hat{C}V_{\mathrm{LOO}}$ is a (low biased) estimate of $Err$

**Generalized cross validation**

Models linear in their parameters and constant preserving
Quadratic loss

GCV can be considered as a simplification of CV-LOO

$$\hat{C}V_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y|x_i - f(x_i, \hat{\theta})}{1-h_{ii}}\right)^2$$

$$G\hat{C}V = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y|x_i - f(x_i, \hat{\theta})}{1-Tr(\boldsymbol{H})/n}\right)^2 \qquad Tr(\boldsymbol{H})/n = \sum_{i=1}^{n} h_{ii}/n$$

$h_{ii}$ is replaced by the average of the values $h_{ii}$ $\quad i = 1, \ldots, n$

But $G\hat{C}V$ can also be used outside of the LOO short-cut

$$G\hat{C}V = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y|x_i - f(x_i, \hat{\theta}^{-i})\right)^2}_{\hat{C}V_{\mathrm{LOO}}} \underbrace{\left(\frac{1-hii}{1-Tr(\boldsymbol{H})/n}\right)^2}_{\text{weight}}$$

Eubank 1999

$G\hat{C}V$ can be considered as a weighted version of $\hat{C}V_{\mathrm{LOO}}$

$G\hat{C}V$ often gives close results to $\hat{C}V_{\mathrm{LOO}}$

Relation between $G\hat{C}V$ and $Cp$

$$\mathrm{M}\hat{S}EP(\boldsymbol{x}) \quad = \frac{1}{n}\ RSS + \frac{2}{n} Tr(\boldsymbol{H})\ \widehat{\sigma}_0{}^2 \qquad\qquad Cp \text{ approach}$$

$$G\hat{C}V \qquad = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y|x_i - f(x_i, \widehat{\theta})}{1-Tr(\boldsymbol{H})/n}\right)^2$$

$$\approx \frac{1}{n}\ RSS + \frac{2}{n} Tr(\boldsymbol{H})\ RSS/n \qquad\qquad \text{Hastie \& Tibshirani 1990}$$

$Cp$ uses a low-biased estimate of $\sigma^2$, while $GCV$ uses $RSS/n$

$Cp$ and $GCV$ gives close results (Hastie & Tibshirani 1990)

# Next

- Model selection bias (optimistic $Var(\hat{\theta})$) , model selection uncertainty, model averaging

  - *Burnham, K.P., Anderson, D.R., 1998. Model selection and inference. A practical information-theoretic approach. Springer, New York*
  - *Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical Inference. Journal of the Royal Statistical Society: Series A (Statistics in Society) 158, 419–444*
  - *Zucchini, W., 2000. An Introduction to Model Selection. Journal of Mathematical Psychology 44, 41–61*
  - *Zhang, P., 1992. Inference after variable selection in linear regression models. Biometrika 79, 741–746*

- Repeated double CV: model selection + uncertainty

  - *Filzmoser, P., Liebmann, B., Varmuza, K., 2009. Repeated double cross validation. Journal of Chemometrics 23, 160–171*
  - *Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics 6, 10*

- Degrees of freedom for PLSR $\hat{\boldsymbol{y}} \mid \boldsymbol{x} = \boldsymbol{H}_{\boldsymbol{y}} \boldsymbol{y}$
  - *Denham, M.C., 2000. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. Journal of Chemometrics 14, 351–361*
  - *Efron, B., 2004. The Estimation of Prediction Error. Journal of the American Statistical Association 99, 619–632*
  - *Krämer, N., Sugiyama, M., 2011. The Degrees of Freedom of Partial Least Squares Regression. Journal of the American Statistical Association 106, 697–705*

- KNN-LWPLSR
  - Automatization of model selection ($k$, $ncomp$, $h$) for each observation to predict