Model selection and cross validation

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

January 24, 2023

Outline

Contents

- model selection aspects
- bias and variance
- model selection
  - model selection scores (AIC, AICc, BIC)
  - cross-validation (as a resampling method)

Model selection

Factors in model selection

objective: obtain a good model at a low cost

1. **quality of the model:** defined by a measure of the goodness, e.g., the mean-squared error, log-likelihood
   - MSE consists of a *bias* and a *variance* contribution
   - a complex model has small bias but higher variance (than a simple model)
   - the best model structure is therefore a trade-off between *flexibility* and *parsimony*
2. **price of the model:** an estimation method (which typically results in an optimization problem) highly depends on the model structures, which influences:
   - algorithm complexity
   - properties of the loss function
3. **intended use of the model:** prediction, controller design, inference

Bias-variance decomposition

assume that the observation Y obeys

$$Y = f(X) + \nu, \quad \mathbf{E}\nu = 0, \quad \mathbf{cov}(\nu) = \sigma^2$$

the mean-squared error of a regression fit $\hat{f}(X)$ at $X = x_0$ is

$$\begin{aligned}
\mathsf{MSE} &= \mathbf{E}[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma^2 + [\mathbf{E}\hat{f}(x_0) - f(x_0)]^2 + \mathbf{E}[\hat{f}(x_0) - \mathbf{E}\hat{f}(x_0)]^2 \\
&= \sigma^2 + \mathsf{Bias}^2(\hat{f}(x_0)) + \mathsf{Var}(\hat{f}(x_0))
\end{aligned}$$

- this relation is known as **bias-variance decomposition**
- no matter how well we estimate $f(x_0)$, σ^2 represents *irreducible error*
- typically, the more complex we make model $\hat{f}$, the lower the bias, but the higher the variance

Proof of bias-variance decomposition

note that

- the true f is deterministic
- $\mathbf{var}(Y|X=x) = \sigma^2$ and $\mathbf{E}[Y|X=x] = f(x)$
- $\hat{f}(x)$ is random

we will omit the notation of conditioning on $X = x$

$$
\begin{aligned}
\mathbf{E}[(Y - \hat{f}(X))^2] &= \mathbf{E}[Y^2] + \mathbf{E}[\hat{f}(x)^2] - \mathbf{E}[2Y\hat{f}(x)] \\
&= \mathbf{var}(Y) + \mathbf{E}[Y]^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \mathbf{var}(Y) + f(x)^2 + \mathbf{var}\,\hat{f}(x) + \mathbf{E}[\hat{f}(x)]^2 - 2f(x)\mathbf{E}[\hat{f}(x)] \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (f(x) - \mathbf{E}[\hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + (\mathbf{E}[f(x) - \hat{f}(x)])^2 \\
&= \sigma^2 + \mathbf{var}\,\hat{f}(x) + [\mathrm{Bias}(\hat{f}(x))]^2
\end{aligned}
$$

Bias and variance in linear models

two nested linear regression models: predictor X in $\mathcal{M}_1$ is also contained in $\mathcal{M}_2$

$$\mathcal{M}_1 : y = X\beta \quad \text{VS} \quad \mathcal{M}_2 : y = \begin{bmatrix} X & \tilde{x} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} \triangleq Z\gamma$$

setting: two models are estimated by LS method, denoted by $\hat{\beta}$ and $\hat{\gamma}$

1. $\mathcal{M}_2$ has lower MSE in predicting y than the MSE of $\mathcal{M}_1$
2. $\mathbf{cov}(\hat{\beta})$ of $\mathcal{M}_2$ is larger than $\mathbf{cov}(\hat{\beta})$ of $\mathcal{M}_1$
3. variance of $\hat{y}$ from $\mathcal{M}_2$ is higher than that of $\mathcal{M}_1$

$\mathcal{M}_2$ (complex model) has less bias but more variance both in estimator and prediction

our proof will use subscript 1 for $\mathcal{M}_1$ and and 2 for $\mathcal{M}_2$

Inverse of block matrices

the inverse of a block matrix

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \succ 0$$

can be obtained in block using Schur complement: $S = (D - CA^{-1}B)^{-1} \succ 0$

$$X^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \quad (1)$$

we often encounter the difference of two quadratic forms

$$\begin{bmatrix} u \\ v \end{bmatrix}^T \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^{-1} \begin{bmatrix} u \\ v \end{bmatrix} - u^T A^{-1} u = (v - B^T A^{-1} u)^T S^{-1} (v - B^T A^{-1} u) \geq 0 \quad (2)$$

which is always non-negative

proof of $\text{MSE}_2 \leq \text{MSE}_1$

- let P_1 and P_2 be orthogonal projection of y onto $\mathcal{R}(X)$ and $\mathcal{R}(Z)$, resp
- it can be shown that $\text{MSE}_1 = \|y\|_2^2 - y^T P_1 y$ and $\text{MSE}_2 = \|y\|_2^2 - y^T P_2 y$
- it is left to show that $y^T P_2 y \geq y^T P_1 y$

$$P_2 = Z(Z^T Z)^{-1} Z^T = \begin{bmatrix} X & \tilde{x} \end{bmatrix} \begin{bmatrix} X^T X & X^T \tilde{x} \\ \tilde{x}^T X & \tilde{x}^T \tilde{x} \end{bmatrix}^{-1} \begin{bmatrix} X^T \\ \tilde{x}^T \end{bmatrix}, \quad P_1 = X(X^T X)^{-1} X^T$$

- apply the inverse of block matrix

$$P_2 - P_1 = (\tilde{x} - X(X^T X)^{-1} X^T \tilde{x}) S^{-1} (\tilde{x} - X(X^T X)^{-1} X^T \tilde{x})^T \succeq 0$$

where $S = \tilde{x}^T \tilde{x} - \tilde{x}^T X (X^T X)^{-1} X^T \tilde{x}$

proof of $\mathbf{cov}(\hat{\beta}_2) \succeq \mathbf{cov}(\hat{\beta}_1)$

- $\mathbf{cov}(\hat{\beta}_2)$ is the leading (1,1) block of $\mathbf{cov}(\hat{\gamma})$, while $\mathbf{cov}(\hat{\beta}_1) = (X^T X)^{-1}$
- use $\mathbf{cov}(\hat{\gamma}) = (Z^T Z)^{-1}$ and the inverse of block matrix

$$(Z^T Z)^{-1} = \begin{bmatrix} X^T X & X^T \tilde{x} \\ \tilde{x}^T X & \tilde{x}^T \tilde{x} \end{bmatrix}^{-1} \triangleq \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1} B S^{-1} B^T A^{-1} & \times \\ \times & \times \end{bmatrix}$$

  where $S = D - B^T A^{-1} B \succeq 0$
- $\mathbf{cov}(\hat{\beta}_2)$ is bigger than $\mathbf{cov}(\hat{\beta}_1)$ because

$$\mathbf{cov}(\hat{\beta}_2) - \mathbf{cov}(\hat{\beta}_1) = A^{-1} + A^{-1} B S^{-1} B^T A^{-1} - A^{-1} = A^{-1} B S^{-1} B^T A^{-1} \succeq 0$$

proof of $\mathbf{var}(\hat{y}_2) \geq \mathbf{var}(\hat{y}_1)$

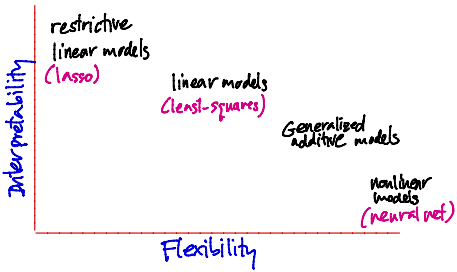
- suppose $\hat{y}_1 = u^T\hat{\beta}$ and $\hat{y}_2 = w^T\hat{\gamma}$ where $w = (u, v)$
- we test prediction of y from new regressors u and (u, v)
- since the model is simply linear, the variance can be obtained by

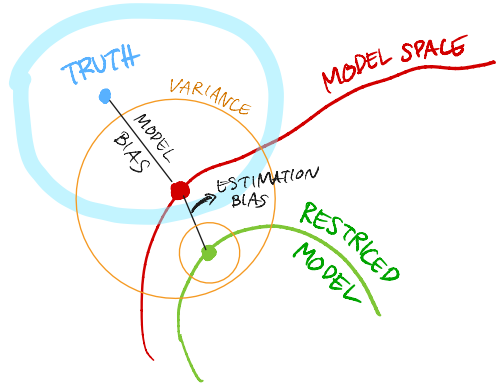
$$\begin{aligned}\mathbf{var}(\hat{y}_2) - \mathbf{var}(\hat{y}_1) &= w^T\,\mathbf{cov}(\gamma)w - u^T\,\mathbf{cov}(\beta)u \\ &= \begin{bmatrix}u\\v\end{bmatrix}^T \begin{bmatrix}X^TX & X^T\tilde{x}\\ \tilde{x}^TX & \tilde{x}^T\tilde{x}\end{bmatrix}^{-1} \begin{bmatrix}u\\v\end{bmatrix} - u^T(X^TX)^{-1}u\end{aligned}$$

- the difference is non-negative (using result on page 9)

Model properties

consider bias and variance of model with different structures



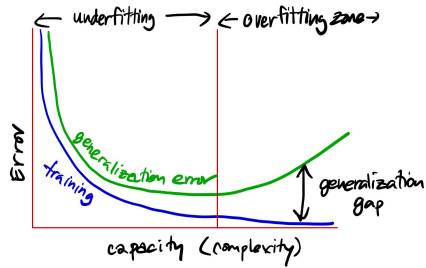


(T. Hastie *et.al. The Elements of Statistical Learning*, Springer, 2010 page 225)

a simple model has less flexibility (more bias) but easy to interpret and has less variance

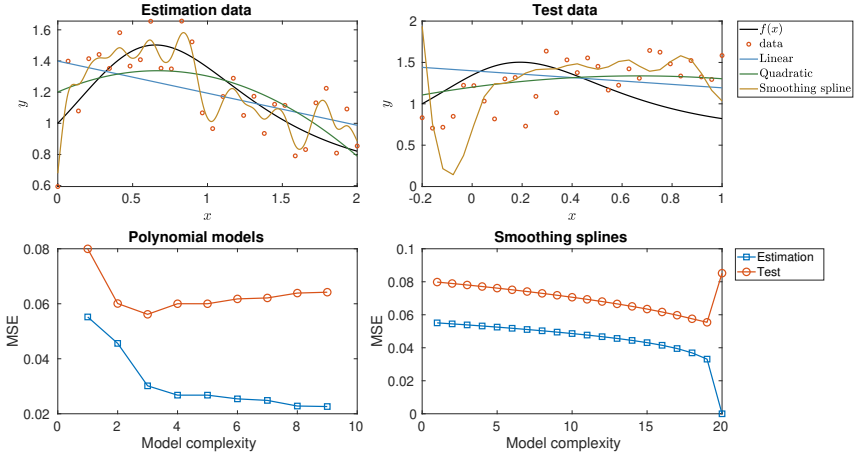
U-shape of generalization error

models are estimated on training data set and evaluated on test set (unseen data)



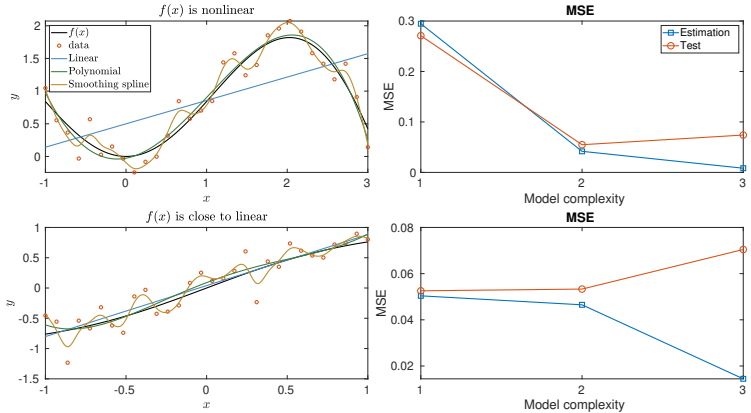
- training errors always decrease as model complexity increase
- generalization error initially decreases as model picks up relevant features of data
- however, if the model complexity exceeds a certain degree, the generalization error can rise up again – this is when we observe overfitting

Observe overfitting on test error



- too complex models cannot generalize well on test (unseen) data
- overfitting occurs when MSE on test set decreases but starts to rise again

Does overfitting always occur?



- when the true description is highly nonlinear, test MSE does not significantly increase
- overfitting is apparent when the estimated model is more complex (than it should be) in order to explain a simpler ground-truth model

Model selection criterion

parsimony principle: among competing models which all explain the data well, the model with the smallest number of parameters should be chosen

a model selction criterion consists of two parts:

$$\text{loss function} + \text{model complexity}$$

- the first term is to assess the quality of the model, e.g., likelihood function, RSS, MSE, Fit Percent $(1 - \frac{\|y-\hat{y}\|}{\|y-\bar{y}\|}) \times 100\%$
- the second term is to penalize the model order and grows as the number of parameters increases
- we choose the best model as the one with the lowest model selection score

Model selection scores

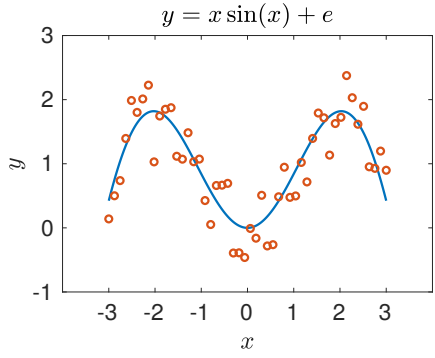
model quality: $\mathcal{L}$: log-likelihood, V: loss function

model complexity: d: effective number of parameters

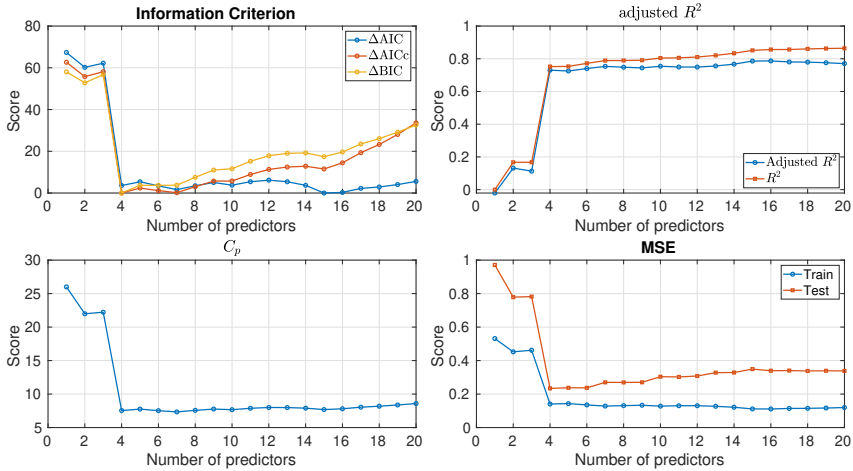
- Akaike information criterion (AIC): $\text{AIC}(\alpha) = -2\mathcal{L}(\alpha) + 2d$
- corrected Akaike information (AICc): $\text{AICc}(\alpha) = -2\mathcal{L}(\alpha) + 2d + \frac{2d(d+1)}{N-d-1}$
- Bayesian information criterion (BIC): $\text{BIC}(\alpha) = -2\mathcal{L}(\alpha) + d\log N$
- Akaike's final prediction-error criterion (FPE): $\text{FPE}(\alpha) = V(\hat{\theta})\left(\frac{1+d/N}{1-d/N}\right)$
- Mallow's C_p: $C_p(\alpha) = \frac{1}{N}\left[\text{RSS}(\alpha) + 2d\hat{\sigma}^2\right]$
- adjusted R^2: $1 - \frac{\text{RSS}(\alpha)/(N-d-1)}{\text{TSS}/(N-1)}$

Variable selection in linear regression

model: $\hat{y} = \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx)$ for $n = 1, 2, \ldots, 20$ and $N = 50$



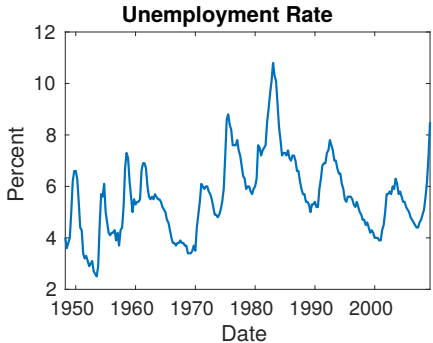
- aim to choose the number of basis function (n)
- set the effective number of parameters $d = 2n$ (the number of $\sin(kx), \cos(kx)$)
- compute ΔAIC, ΔAICc, ΔBIC (subtracted by its minimum), C_p, adjusted R^2



- AIC and adjusted R^2 chose a complex model, while AICc and BIC picked 4 basis functions (simpler), and C_p chose 7 basis functions
- train MSE always decreases, as well as, R^2 always increases but the curves have a knee around $n = 4$

Choosing AR lag order

fitting AR model of order $p = 1, 2, \ldots, 20$ to unemployment rate time series



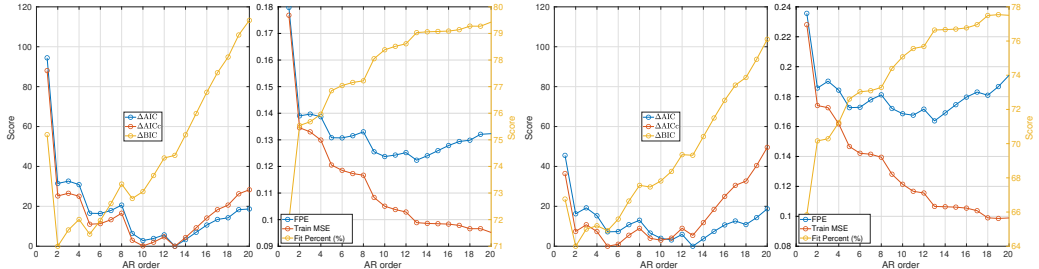
p-order autoregressive (AR) model

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_p y(t-p) + e(t)$$

parameter: $\beta = (a_1, a_2, \ldots, a_p)$
fitting: least-squares

- the effective number of parameters is chosen as $d = p$
- compute ΔAIC, ΔAICc, ΔBIC, FPE, train MSE, and Fit Percent
- data samples: $N = 245$, examine two cases: (i) use all data (ii) use only half

left: use all data right: use half of data



- left: AIC, AICc and FPE tend to choose a higher order model $(p = 13)$ but BIC prefers a simpler model $(p = 2)$
- right: AICc chose a lower order model when N is halved (sample size was corrected)
- both *train* MSE and Fit Percent are not good indicators for model selection

Log-likelihood based scores (AIC, AICc)

AIC, AICc, BIC use negative log-likelihood to indicate model quality

$$\begin{aligned}
\text{AIC}(\alpha) &= -2\mathcal{L}(\alpha) + 2d \\
\text{AICc}(\alpha) &= -2\mathcal{L}(\alpha) + 2d + \frac{2d(d+1)}{N-d-1} \\
\text{BIC}(\alpha) &= -2\mathcal{L}(\alpha) + d\log N
\end{aligned}$$

- AIC is an approximation of Kullback-Leibler (KL) divergence between the true density ($f(x)$ and the model ($g(x|\hat{\theta})$)

$$\begin{aligned}
I(f,g) &= \int f(x)\log(f(x)/g(x|\theta))dx \\
-\mathcal{L}(\hat{\theta}) + d &\approx \mathbf{E}_{\hat{\theta}}[I(f(x), g(x|\hat{\theta}))] + \text{constant}
\end{aligned}$$

- AICc penalizes more on complexity for small N (as quadratic term in d); it approaches AIC for large samples (large N)

Log-likelihood based score (BIC)

- BIC penalizes more on complexity than AIC (as indicated by $\log N > 2$)
- when model candidates contain a true model, BIC is consistent (probability of choosing the correct model $\to 1$ as $N \to \infty$)
- model with minimum BIC $\Leftrightarrow$ model with *highest* posterior density

$$\text{posterior odds} = \frac{P(\mathcal{M}_m|\text{data})}{P(\mathcal{M}_l|\text{data})} = \underbrace{\frac{P(\mathcal{M}_m)}{P(\mathcal{M}_l)}}_{\text{prior}} \cdot \underbrace{\frac{P(\text{data}|\mathcal{M}_m)}{P(\text{data}|\mathcal{M}_l)}}_{\text{Bayes factor}}$$

  model prior tells which model is more likely to be preferred (by users)
- when prior is not available (all models have equal probabilities), Bayes factor directly affects the posterior odds
- BIC (with -2 factor) is an approximate of Bayes factor (see Hastie et al. book)

- for nested models $\mathcal{M}_1$ (complex), $\mathcal{M}_2$ (simple) with $d(\mathcal{M}_1) = d(\mathcal{M}_2) + m$
  - AIC picks complex model if $\mathcal{L}(\mathcal{M}_1) - \mathcal{L}(\mathcal{M}_2) > 2m$ (it's worth to use complex model since model quality improved much more)
  - BIC picks complex model if $\mathcal{L}(\mathcal{M}_1) - \mathcal{L}(\mathcal{M}_2) > m \log N$
- improved gap of log-likelihood required by AIC is less than that of BIC; hence, AIC is prone to choosing a complex model more easily than BIC
- for LR (log-likelihood ratio) test, with test statistic

$$2(\mathcal{L}(\mathcal{M}_1) - \mathcal{L}(\mathcal{M}_2)) \sim \mathcal{X}^2(m)$$

  - LR test picks $\mathcal{M}_1$ (complex) if $2\mathcal{L}(\mathcal{M}_1) > 2\mathcal{L}(\mathcal{M}_2)$ by $\mathcal{X}^2_{0.05}(m)$
  - for $m < 7$, we have $2m < \mathcal{X}^2_{0.05}(m)$; hence, AIC tends to pick a complex model more easily than LR test in this case

Akaike's final prediction (FPE)

denote $V(\hat{\theta})$ a loss function used in prediction error method (*e.g.*, det or trace of error covariance)

$$\text{FPE}(\alpha) = V(\hat{\theta}) \left(\frac{1 + d/N}{1 - d/N} \right)$$

- model complexity is cooperated in *multiplicative form* (as compared to additive form in AIC, BIC)
- when model output is scalar, $V(\hat{\theta})$ is simply MSE and FPE reduces to

$$\text{FPE} = \frac{1}{N} \sum_{t=1} \varepsilon^2(t, \hat{\theta}) \cdot \frac{1 + d/N}{1 - d/N}$$

- it was shown in Ljung book that FPE is a way to approximate of $\lim_{N \to \infty} \mathbf{E}[V(\theta)]$ (population), which can be estimated using $V(\hat{\theta})$ evaluated on *estimation data*

Mallow's C_p

C_p is mostly used in linear regression with d predictors and homoskedastic noise

$$C_p(\alpha) = \frac{1}{N}\left[\mathrm{RSS}(\alpha) + 2d\hat{\sigma}^2\right]$$

- C_p uses *quadratic loss* to measure model quality
- $\hat{\sigma}^2$ is an estimate of noise variance using **full** model
- RSS/N always decreases when d increases; penalty on complexity is put on $2d\hat{\sigma}^2$
- in Hastie et al. book, it showed that C_p is an estimate of test MSE
- other form of C_p exists: $C_p = \mathrm{RSS}/\hat{\sigma}^2 + 2d - N$ but result in choosing the same d

Adjusted R^2

R^2 (coefficient of determination) is based on the decomposition:

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\text{zero if model has a constant}}$$

R^2 is the proportion of the total variation in Y that can be linearly predicted by X

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \quad \text{adjusted } R^2 = 1 - \frac{\text{RSS}(\alpha)/(N-d-1)}{\text{TSS}/(N-1)}$$

- for linear model, $0 \leq R^2 \leq 1$ and always increases for larger models
- the presence of d penalizes the criterion for the number of predictor variables
- adjusted R^2 increases if the added predictor variables decrease RSS enough to compensate for the increase in d

References

1. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition, Springer, 2009
2. G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013
3. T. Söderström and P. Stoica, *Chapter 11: System Identification*, Prentice Hall, 1989
4. L. Ljung, *Chapter 16: System Identification: Theory for the User*, 2nd edition, Prentice Hall, 1999
5. K.P. Burnham and D.R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, Springer, 2002

Resampling method: Cross validation

Resampling methods

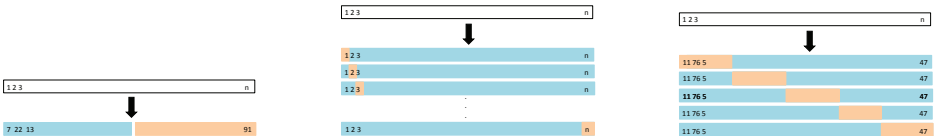
- a process of *repeatedly* drawing samples from a training set and refitting a model on each sample
- we seek for information that would not be obtained from fitting the model only *once* using the original training sample
- resampling approaches can be computationally expensive but with nowaday technology, it becomes less prohibitive
  - cross-valiation: used in estimation of test error or model flexibility
  - bootstrap: a measure of accuracy of a parameter estimate

Cross validation

- training error rate: the average error that results from using a trained model (or method) back on the training data set
- test error rate: the average error that results from using a statistical learning method to predict the response on a **new observation**
- training error can be quite different from the test error rate
- **cross validation** can be used to estimate *test error rate* using available data: split into training and validation sets
  - validation set approach
  - leave-one-out cross validation
  - k-fold cross validation

Splitting data

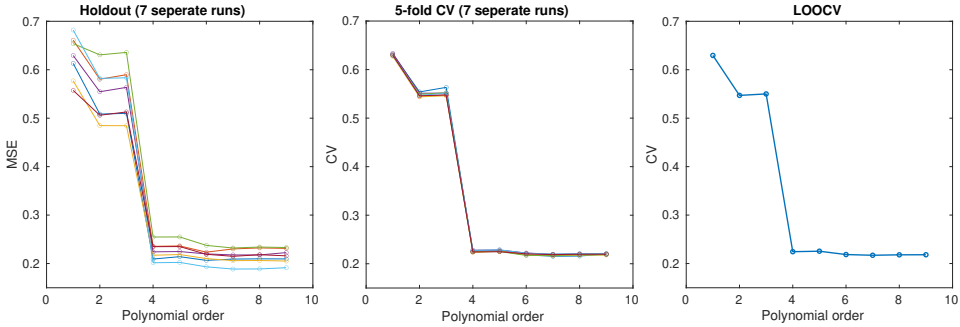
- training set: used for fitting a model
- validation set: used for predicting the response from the fitted model



- validation set approach or hold out (left): randomly split data
- leave-one-out or LOOCV (middle): leave 1 sample for validation set
- k-fold (right): randomly split data into k folds; leave 1 fold for validation
  - repeat k times where each time a different fold is regarded as validation set and compute MSE_1, $\text{MSE}_2, \ldots, \text{MSE}_k$
  - the test error rate is estimated by **averaging** the k MSE's

Cross validation on polynomial order

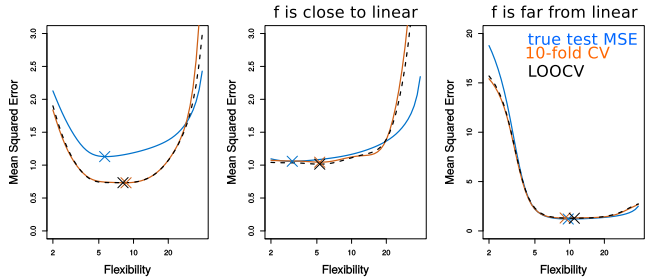
$N = 500$, show 7 runs of holdout, and 5-fold



- result of holdout has high variation since it depends on random splitting
- 5-fold results has less variation because MSE is averaged over k folds
- LOOCV requires N loops (high computation cost); MSE_i's are highly correlated

Estimate a true test MSE by CV

accuracy of test error rate (on simulation data set): using model of smoothing splines



compute the *true test MSE* (assume to know true f) as a function of complexity

- (left): cv estimates have the correct general U shape but underestimate test MSE
- (center): cv gives overestimate of test MSE at high flexibility
- (right): the true test MSE and the cv estimates are almost identical

Usage of cross-validation

most of the times we may perform cv on

- a number of statistical methods: and to see which method has the lowest test MSE
- a single statistical method but different flexibilities: and to see which model complexity yield the lowest test MSE

though sometimes cv method underestimate the true test MSE, they can select the correct level of flexibility

Trade-off for k-fold

examine the unbiasedness and variance of test MSE

method	validation set	loocv	k-fold
computation	less	high	feasible
training samples	ratio e.g. 70:30	$n-1$	$(k-1)n/k$
unbiasedness	low	approximately unbiased	intermediate
variance		high	less

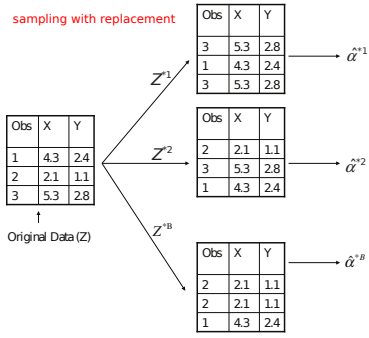
- test MSE is calculated by taking the **average** of many MSE's:
- most of MSE's from *loocv* are highly correlated while MSE's of k-fold are less correlated (since loocv uses more overlapped data in training – hence, fitted models are almost identical)
- fact: the sample mean of highly correlated entries has **more variance** than the sample mean of less correlated entries

conclusion: trade-off between bias and variance when choosing k in k-fold

Resampling method: Bootstrap

Bootstrap

a scheme of obtaining distinct data sets by **repeatedly** sampling with **replacement** from the original data set



use each of new sampled data set to compute a new estimate of α (a quantity)

Illustrated example of the Bootstrap

suppose $\alpha, 1-\alpha$ are fractions of investment we put in yield returns of X and Y

- we want to minimize $\mathbf{var}(\alpha X + (1-\alpha)Y)$
- one can show that the solution α that minimizes the variance is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

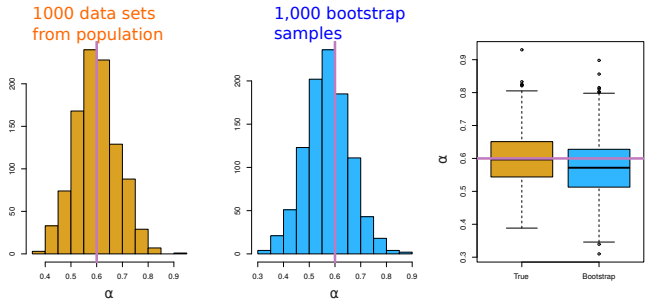
- we estimate the value of α by using $\hat{\sigma}_Y^2, \hat{\sigma}_X^2, \hat{\sigma}_{XY}$
- we generate 100 paired observations of X and Y and repeat 1000 times to get

$$\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \ldots, \hat{\alpha}^{(1000)}$$

(so we have $1,000$ data sets from population)

Example

1,000 data sets from population VS 1,000 bootstrap samples



- histograms of $\hat{\alpha}$ from two approaches are similar and the sample means are close
- standard deviations of $\hat{\alpha}$ are 0.083 (1,000 data sets) and 0.087 (bootstrap)
- the box plots of $\hat{\alpha}$ are also quite similar (true α is 0.6)
- we can use bootstrap when we cannot generate new samples from population

MATLAB example

boostrap for estimating the histogram and SE of correlation

- we have only 15 samples of GPA and LSAT scores of law-school students
- we want to compute the correlation between GPA and LSAT

```
load lawdata
rng default  % For reproducibility
[bootstat,bootsam] = bootstrp(1000,@corr,lsat,gpa);
figure
histogram(bootstat)
se = std(bootstat)

    0.1285

% 1000 is the number of bootstrap samples -- specified by
%
```

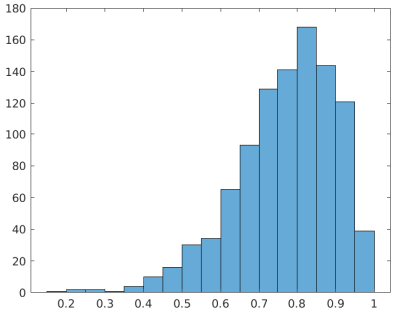


figure shows the histogram of correlation coefficient between LSAT and GPA

References

some figures and examples are taken from

1. Chapter 5 in G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer
2. Chapter 7 in T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, Springer, 2009