5. Resampling methods

- cross validation
- bootstrap
  - basic: estimate variability of estimator
  - moving blocks bootstrap
  - jackknife

Resampling methods

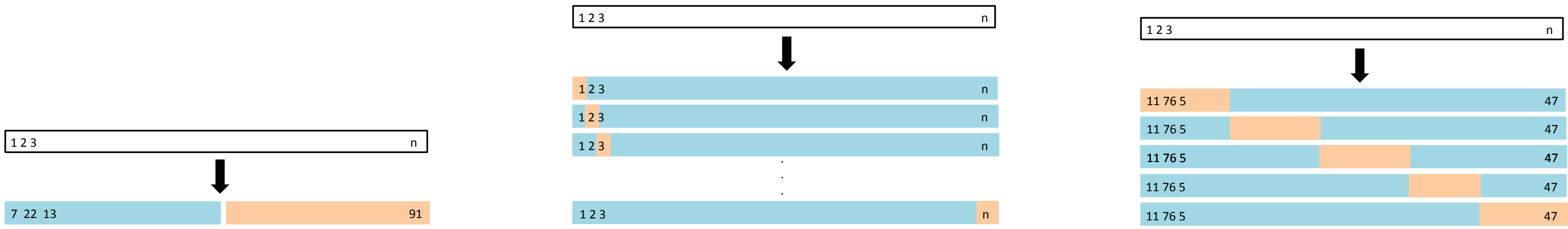
- a process of *repeatedly* drawing samples from a training set and refitting a model on each sample
- we seek for information that would not be obtained from fitting the model only *once* using the original training sample
- resampling approaches can be computationally expensive but with nowaday technology, 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

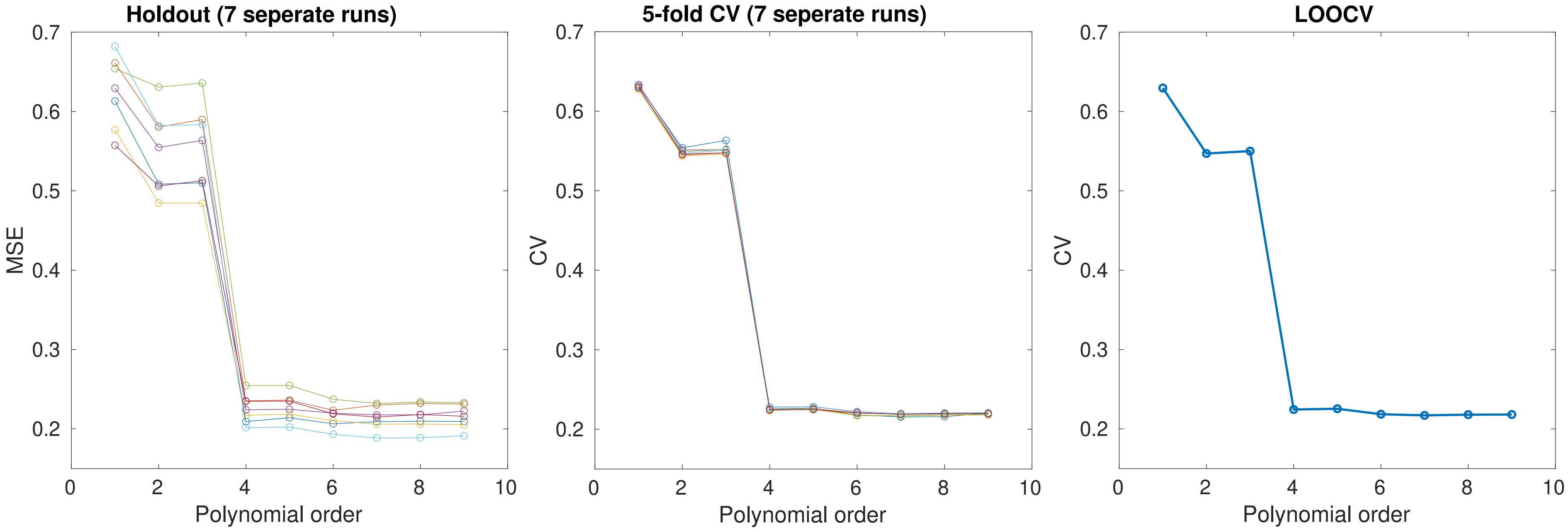
- training set: used for fitting a model
- validation set: used for predicting the response from the fitted model



- 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 $\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_k$
  - the test error rate is estimated by **averaging** the k MSE's

Cross validation on polynomial order

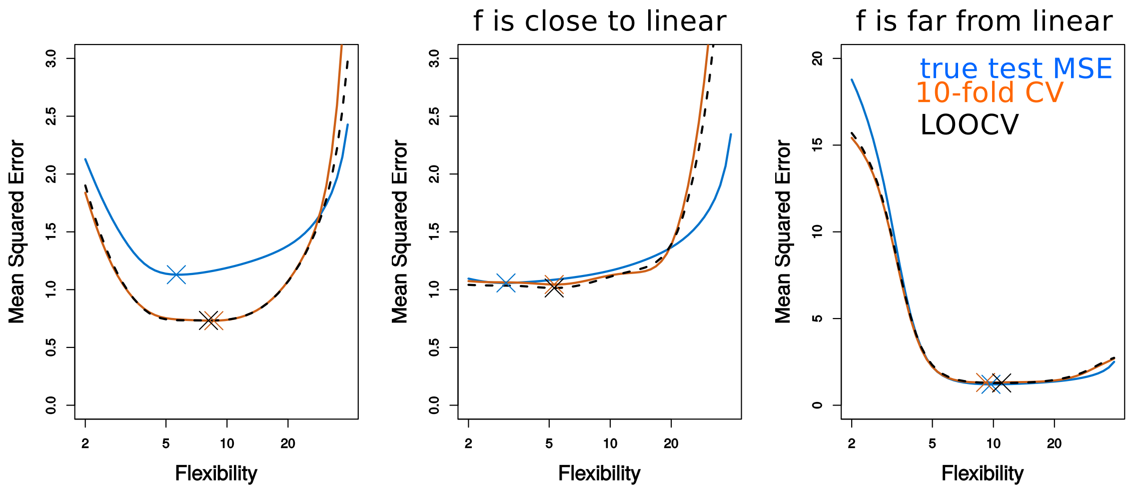
$N = 500$, show 7 runs of holdout, and 5-fold



- 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

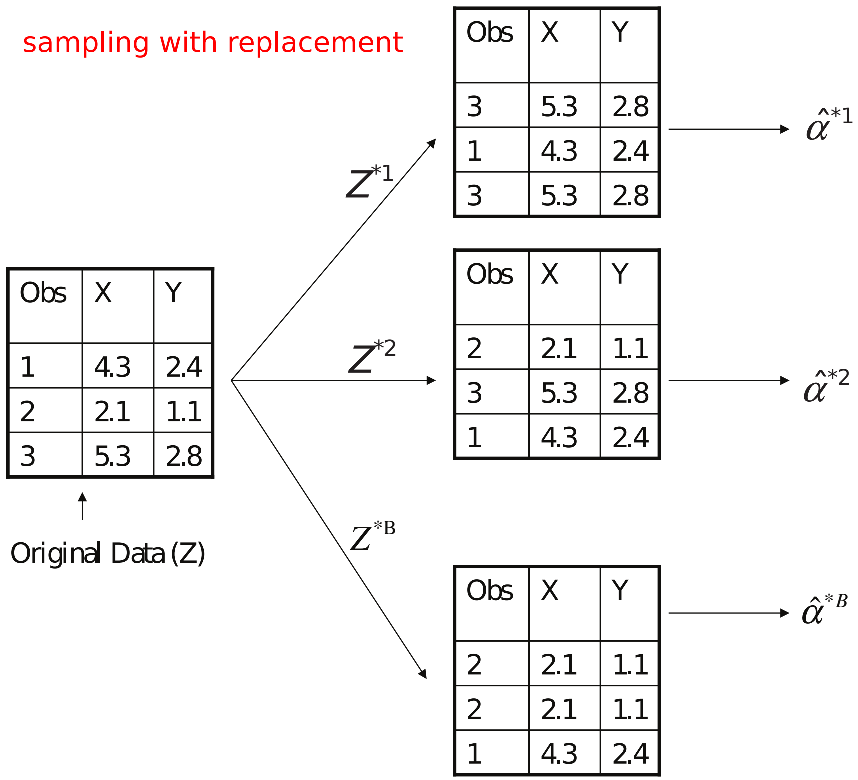
conclusion: trade-off between bias and variance when choosing k in k-fold

Resampling methods

- cross validation
- **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

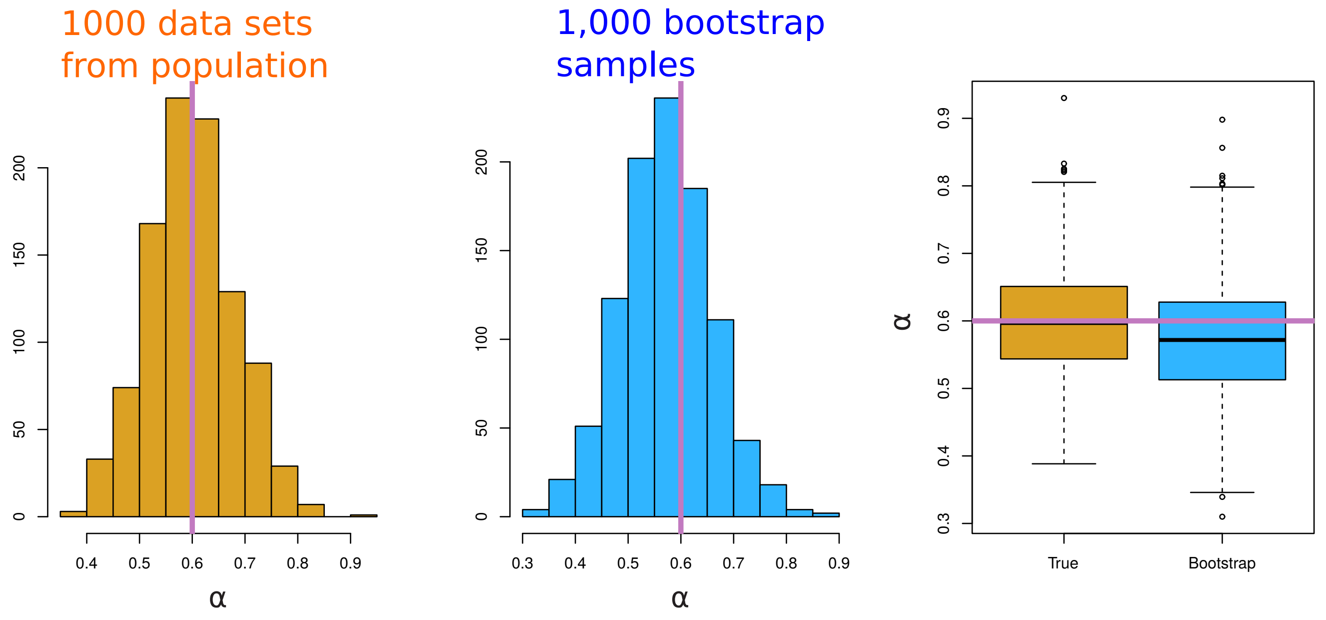
$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

- 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)

$1,000$ 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

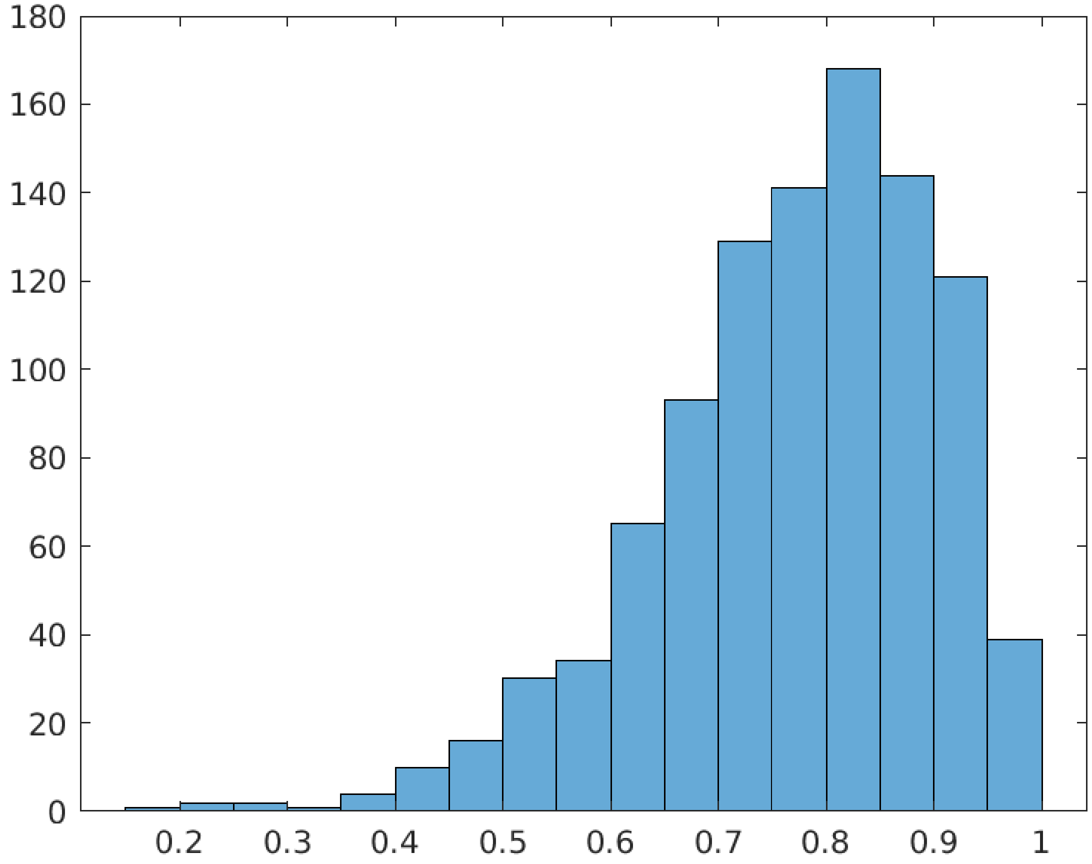
- we have only 15 samples of GPA and LSAT scores of law-school students
- we want to compute the correlation between GPA and LSAT

```
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 user)

histogram of correlation coefficient between LSAT and GPA



References

Some figures and examples are taken from Chapter 5 in

G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2015

Chapter 7 in

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, Springer, 2009