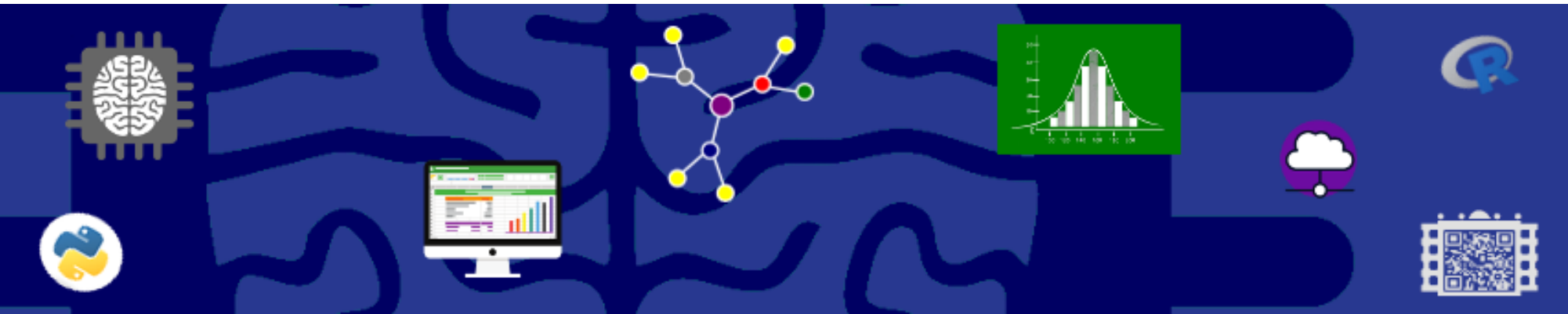
Introduction to Data Science

Winter Semester 2023/24

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

Contents II

Contents III

Contents

Resampling Methods

- **Resampling methods** refers to a set of statistical tools which involve refitting a model on different subsets of a given data set in order to assess the variability of the resulting models.
- These methods are computationally more demanding, but now feasible due to increased computing resources.
- Resampling is useful for **model assessment**, i.e., the process of evaluating a model's performance, as well as **model selection**, i.e., the process of selecting the proper level of model flexibility.
- In this chapter we introduce the resampling methods **cross validation** and the **bootstrap**.

Contents

Resampling Methods

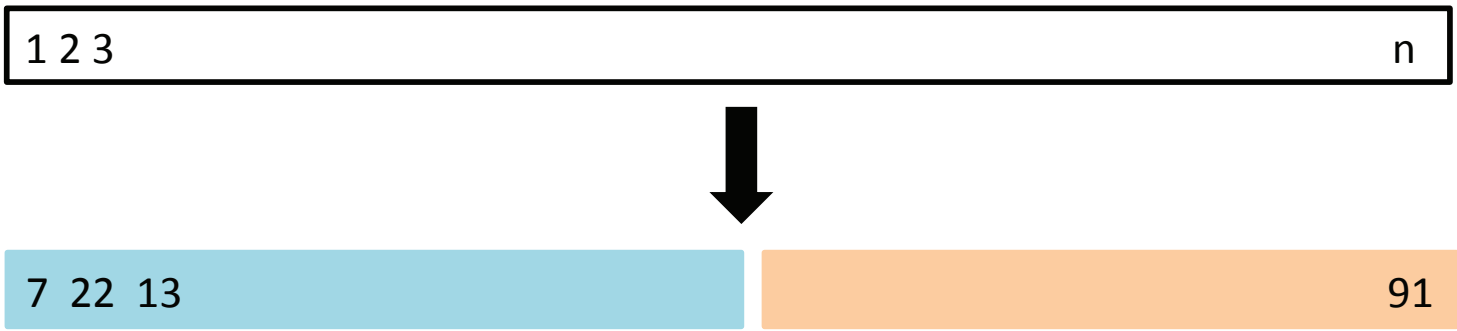
Validation set approach

- Chapter 2: training set error vs. test set error.
- Training set error easily calculated, but usually overoptimistically low.
- Predictive value of model rests on low test set error.

Resampling Methods

Validation set approach

- Chapter 2: training set error vs. test set error.
- Training set error easily calculated, but usually overoptimistically low.
- Predictive value of model rests on low test set error.
- **Validation set approach**: divide available observations into *training set* and **validation set** or *hold-out set* and use latter as test set data.



Validation set approach schematic: n observations randomly split into training set (beige) and validation set (blue).

Resampling Methods

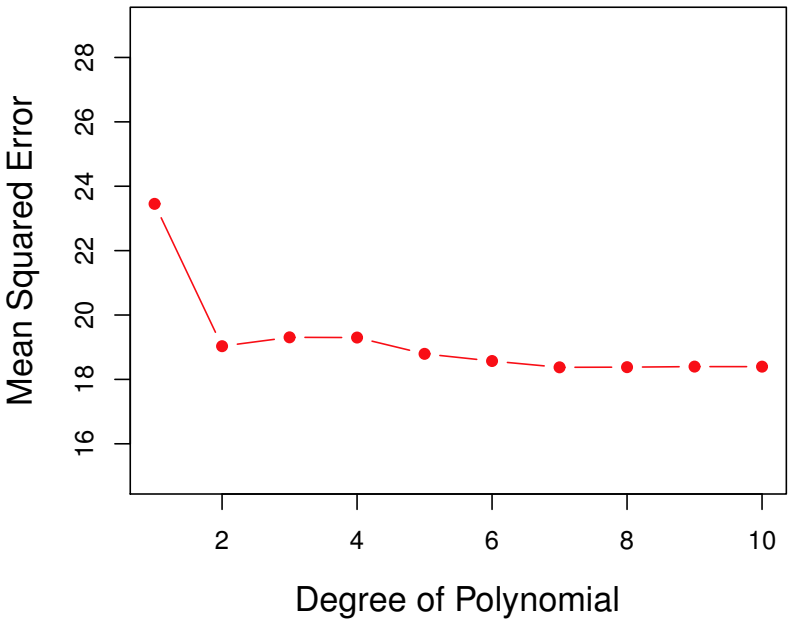
Validation set approach

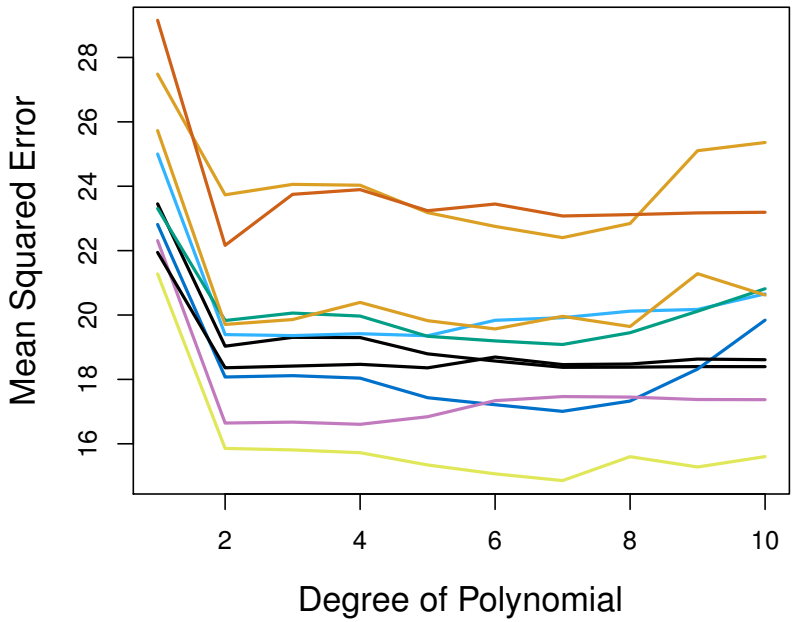
- Recall `Auto` data set (Chapter 3): model predicting `mpg` using `horsepower` and $\texttt{horsepower}^2$ better than linear model.
- Q: would model using higher order polynomial terms yield better prediction results?

Resampling Methods

Validation set approach

- Recall `Auto` data set (Chapter 3): model predicting `mpg` using `horsepower` and $\texttt{horsepower}^2$ better than linear model.
- Q: would model using higher order polynomial terms yield better prediction results?
- Validation set approach: partition the 392 observations into two sets of 196 each, use as training and validation sets, compute test MSE for various polynomial regression models. Compare different random partitions.





- All 10 partitions agree: adding quadratic term leads to lower validation set MSE, no benefit for higher degree terms.
- Different validation set MSE sequence for each partition.

- All 10 partitions agree: adding quadratic term leads to lower validation set MSE, no benefit for higher degree terms.
- Different validation set MSE sequence for each partition.

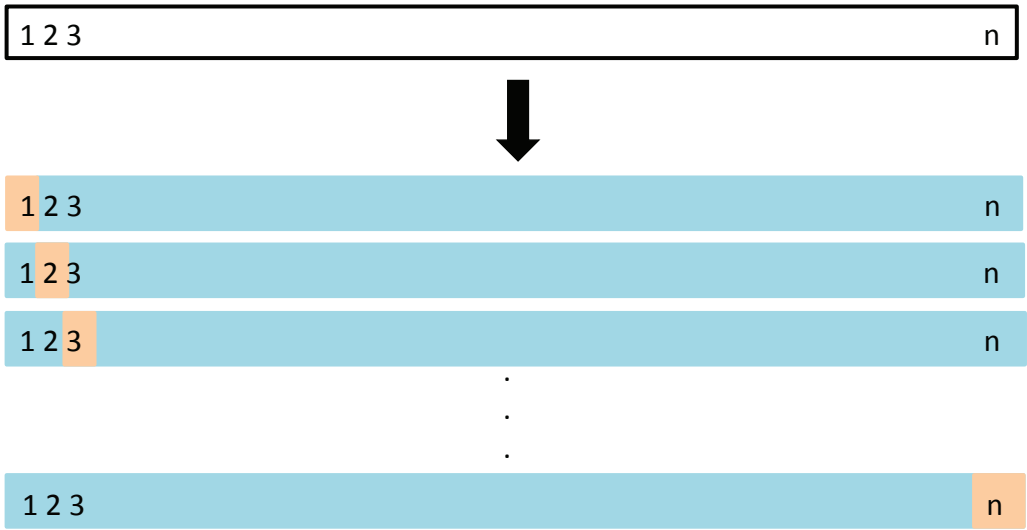
Two principal shortcomings of validation set approach:

1. High variability of validation set MSE with changing partitions.
2. Valuable data not used to fit model, we expect this results in *overestimating* the test error rate (when all the data is used for fitting).

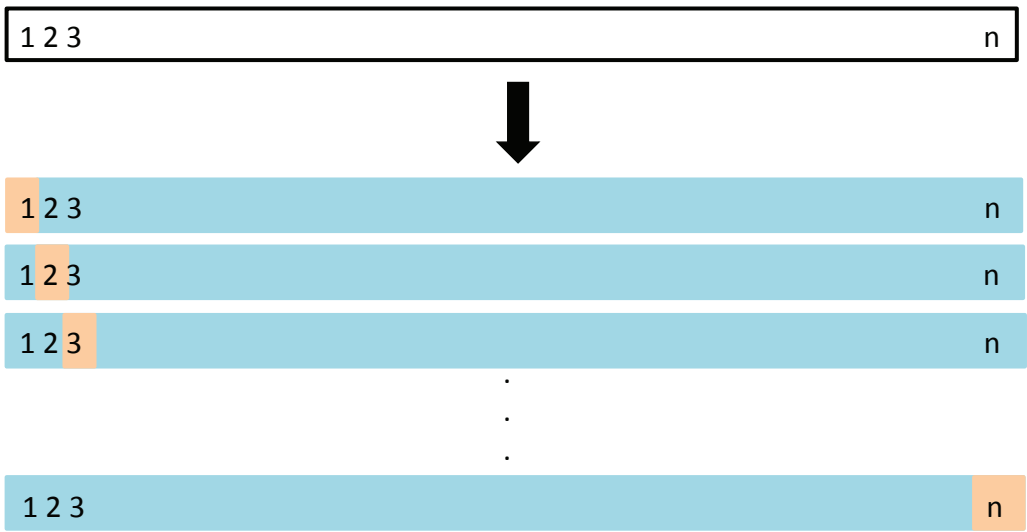
Resampling Methods

Leave-one-out cross-validation (LOOCV)

- **Leave-one-out cross-validation** (LOOCV): for n observations, use n one-element validation sets, fit model using $(n-1)$-element training sets.



- **Leave-one-out cross-validation** (LOOCV): for n observations, use n one-element validation sets, fit model using $(n-1)$-element training sets.



- MSE_i, $i = 1, \ldots, n$: test MSE when validation set consists of i-th observation.
- LOOCV estimate:

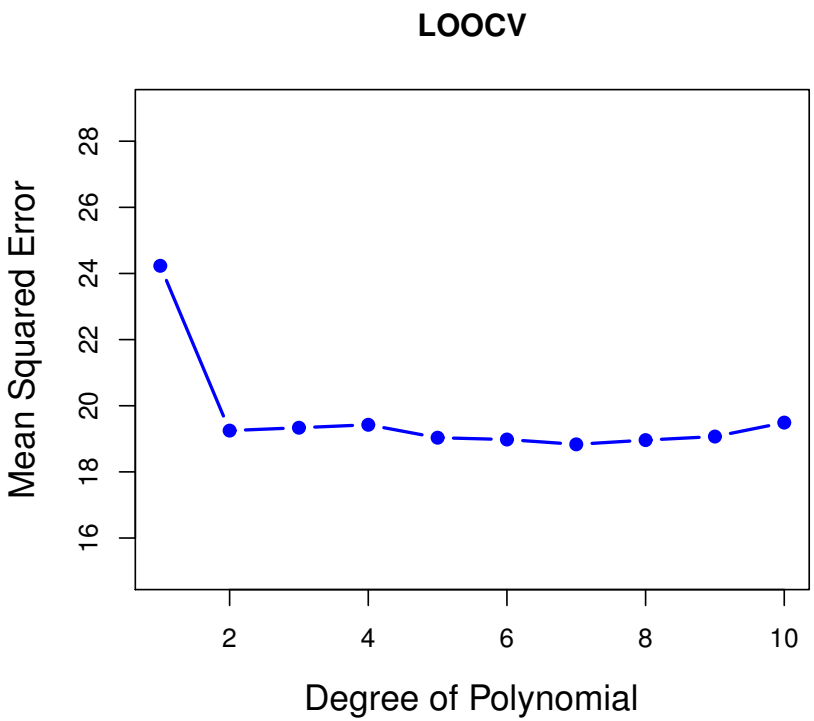
$$\mathrm{CV}_{(n)} := \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i \,.$$

Advantages of LOOCV:

1. Less bias, since each fit uses nearly all observations, less overestimation of test error rate.
2. Well-defined approach, no arbitrariness in partitioning the data as in validation set approach.

Advantages of LOOCV:

1. Less bias, since each fit uses nearly all observations, less overestimation of test error rate.
2. Well-defined approach, no arbitrariness in partitioning the data as in validation set approach.



LOOCV error curve for `Auto` data set: predicting `mpg` as a polynomial function of `horsepower` for varying polynomial degrees.

Resampling Methods

Leave-one-out cross-validation (LOOCV)

- LOOCV requires n fits of $n-1$ observations rather than one for for n observations. Potentially expensive for large n.

- LOOCV requires n fits of $n-1$ observations rather than one for for n observations. Potentially expensive for large n.
- **Magic formula**:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2, \qquad h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}. \tag{5.1}$$

$h_i \in (1/n, 1)$ is the leverage statistic of observation i as defined in (3.31).

- LOOCV requires n fits of $n-1$ observations rather than one for for n observations. Potentially expensive for large n.
- **Magic formula**:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2, \qquad h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}. \tag{5.1}$$

  $h_i \in (1/n, 1)$ is the leverage statistic of observation i as defined in (3.31).
- CV estimate is weighted MSE.
- Makes LOOCV cost same as single fit!

- LOOCV requires n fits of $n-1$ observations rather than one for for n observations. Potentially expensive for large n.
- **Magic formula**:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2, \qquad h_i = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}. \tag{5.1}$$

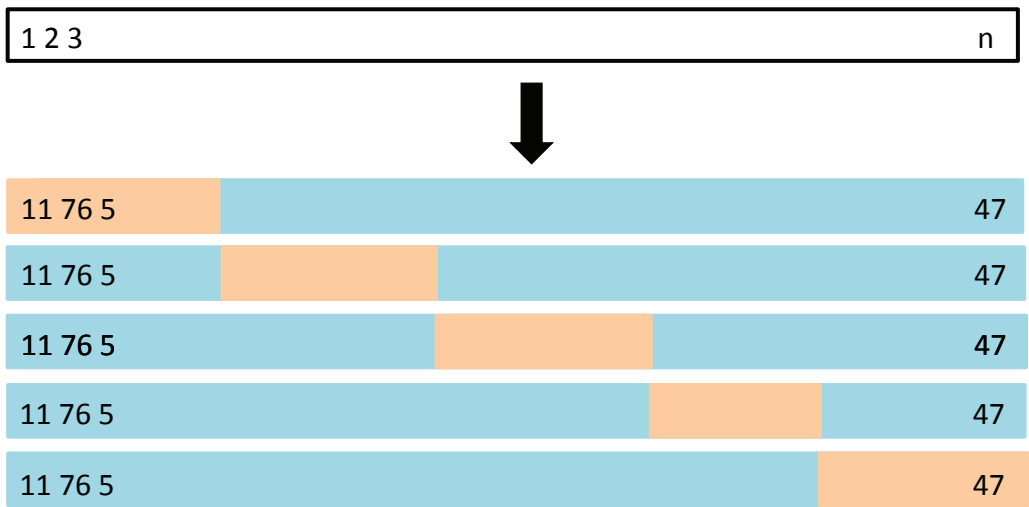
  $h_i \in (1/n, 1)$ is the leverage statistic of observation i as defined in (3.31).
- CV estimate is weighted MSE.
- Makes LOOCV cost same as single fit!
- LOOCV widely applicable (logistic regression, LDA, ...), but (5.1) does not hold in general.
- For linear fitting methods such as linear regression, where $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $h_i = [\boldsymbol{S}]_{i,i}$ the popular **generalized cross validation** method (GCV) replaces h_i by $\mathrm{tr}(\boldsymbol{S})/n$. Note that $\mathrm{tr}(\boldsymbol{S})$ is the *effective number of parameters/DOF*.

Resampling Methods

k-fold cross validation

- Alternative to LOOCV: **k-fold CV**.
- Randomly partition observations into k groups or **folds**, $\approx$ equal in size.
- Use first fold as validation set and fit using remaining observations.
- Mean-squared error MSE_1 computed using first fold.
- Repeat $k-1$ more times with remaining folds, to obtain $\mathsf{MSE}_2, \ldots, \mathsf{MSE}_k$, and set

$$\mathsf{CV}_{(k)} := \frac{1}{k} \sum_{i=1}^{k} \mathsf{MSE}_i \,. \tag{5.2}$$



Resampling Methods

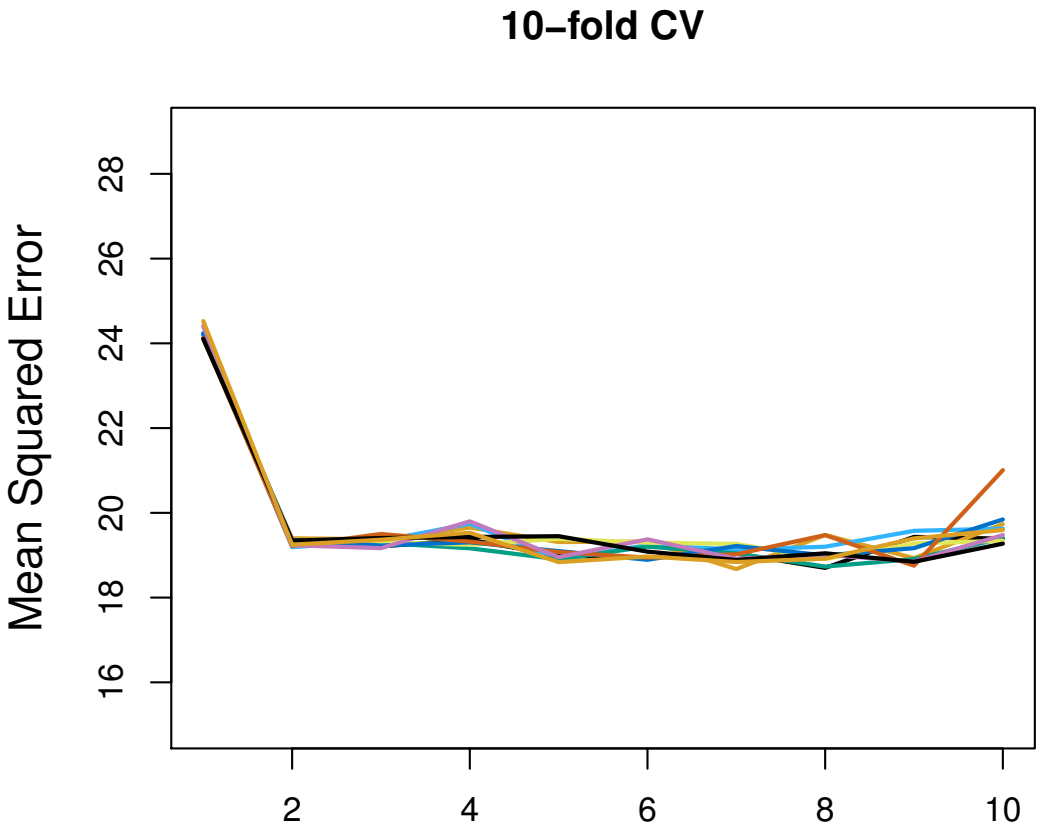
k-fold cross validation

- LOOCV special case of k-fold CV with $k = n$.
- $k = 5$ or $k = 10$ commonly used.
- Appeal: computationally cheaper when magic formula cannot be used.

Resampling Methods

k-fold cross validation

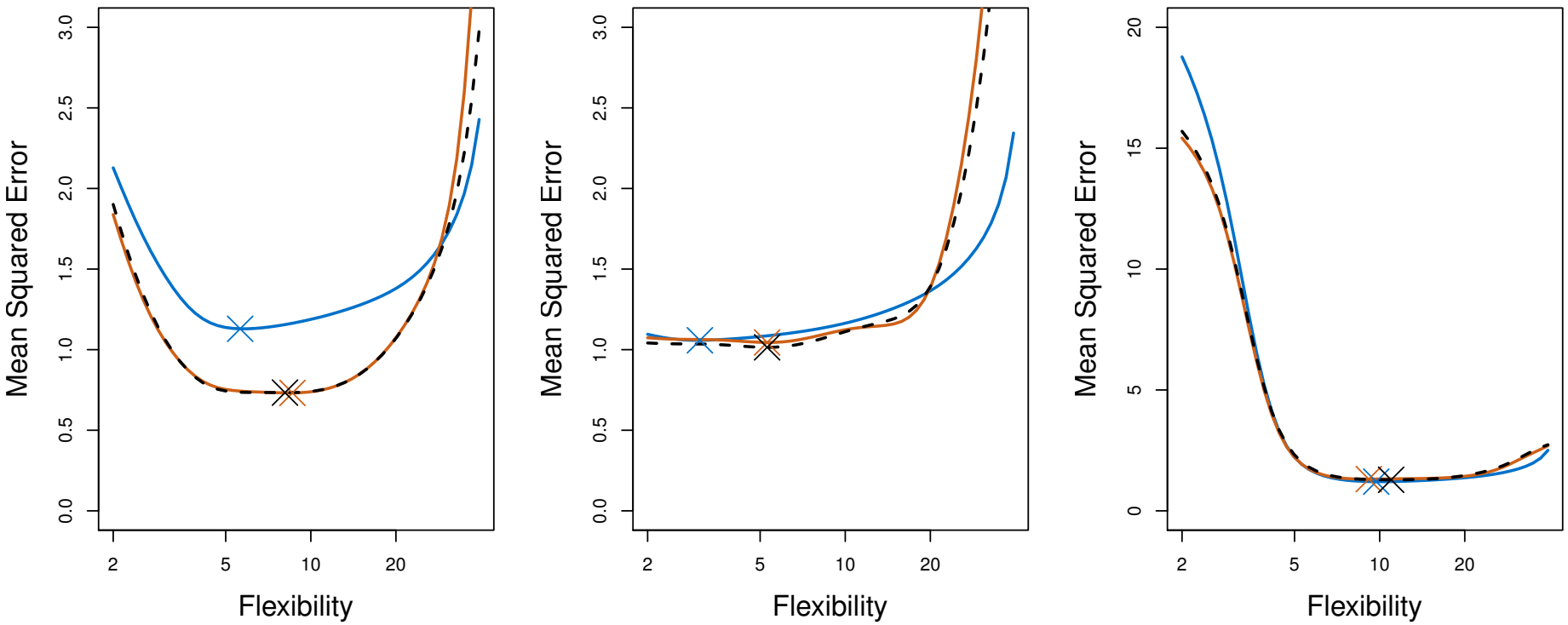
- LOOCV special case of k-fold CV with $k = n$.
- $k = 5$ or $k = 10$ commonly used.
- Appeal: computationally cheaper when magic formula cannot be used.



Nine 10-fold CV estimates for `Auto` data set, each resulting from a different random partition into 10 folds.
Some variability visible, much less than for validation set approach.

Resampling Methods

CV applied to example from Chapter 2



CV estimates for smoothing splines applied to simulated data sets from Chapter 2: LOOCV (black dashed), 10-fold CV (orange solid) beside true test MSE (blue). Crosses denote minimum of each curve.

- Besides its computational advantage, k-fold CV often gives more accurate test MSE estimates than LOOCV.
- **Bias :** LOOCV gives approximately unbiased estimates, since it uses $n-1$ observations to fit. Validation set approach: most bias, since fewest observations used. k-fold CV: intermediate, as $(k-1)n/k$ observations in each training set.
- **Variance:** LOOCV has higher variance than k-fold CV with $k < n$. Reason: LOOCV gives average of n fitted models, each trained on nearly identical set of models, hence outputs highly correlated.
- For k-fold CV with $k < n$, average outputs of k fitted models whose outputs are less correlated (since overlap between training sets smaller).
- Mean of many highly correlated quantities has higher variance than mean of many quantities which are not as highly correlated, test error estimate resulting from LOOCV tends to have higher variance than test error estimate resulting from k-fold CV.

- For classification (Y qualitative) replace MSE by *number of misclassification* and set

$$\mathrm{CV}_{(n)} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i \qquad \mathrm{Err}_i := \mathbb{1}_{\{y_i \neq \hat{y}_i\}}. \tag{5.3}$$

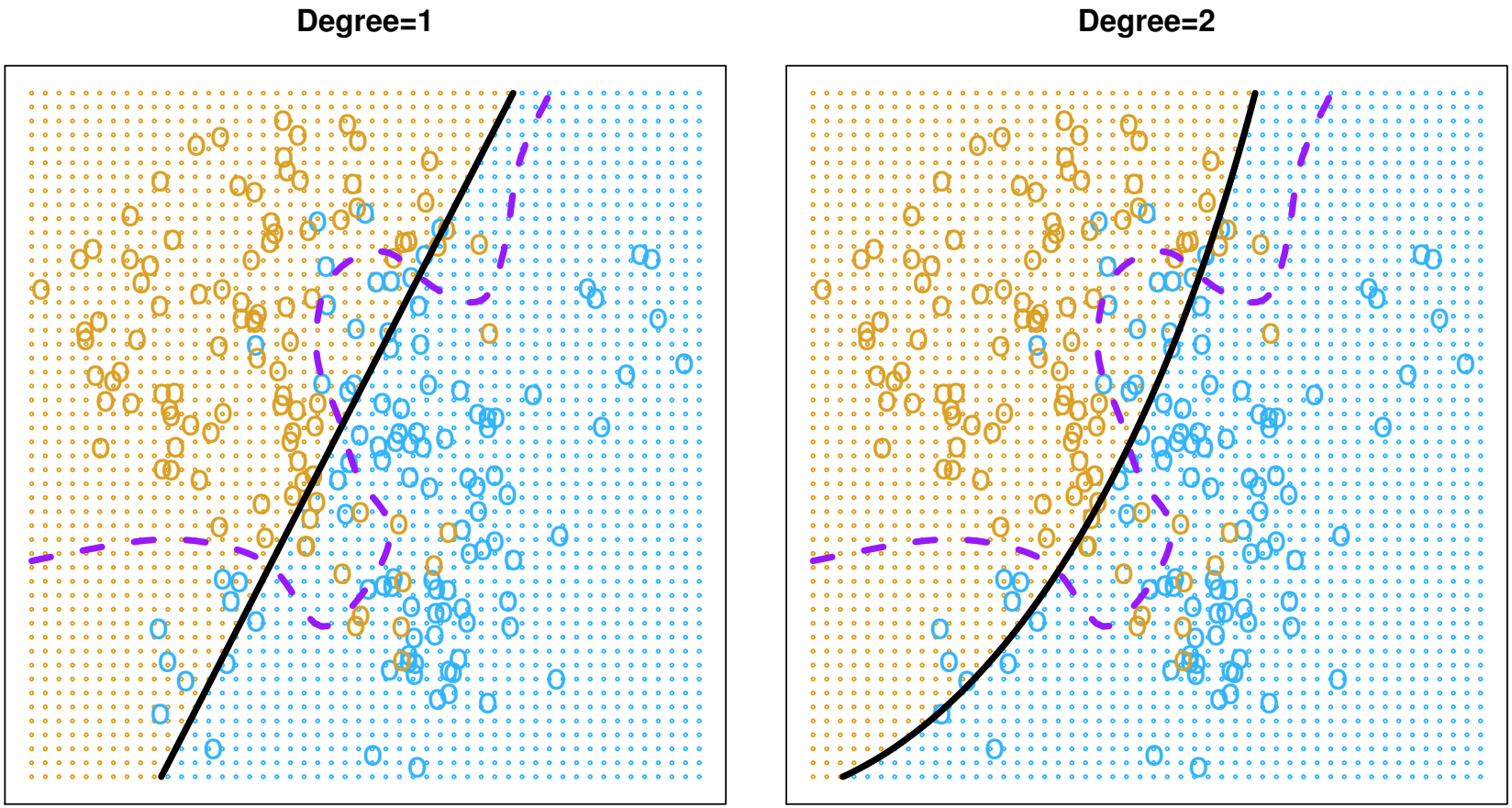
  k-fold CV and validation set error rates defined analogously.
- Can use CV e.g. to perform logistic regression.
- As in linear regression setting, can use polynomial functions in predictor variables:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2. \tag{5.4}$$

- Consider classification problem from Chapter 2 (Slide 66)

Resampling Methods

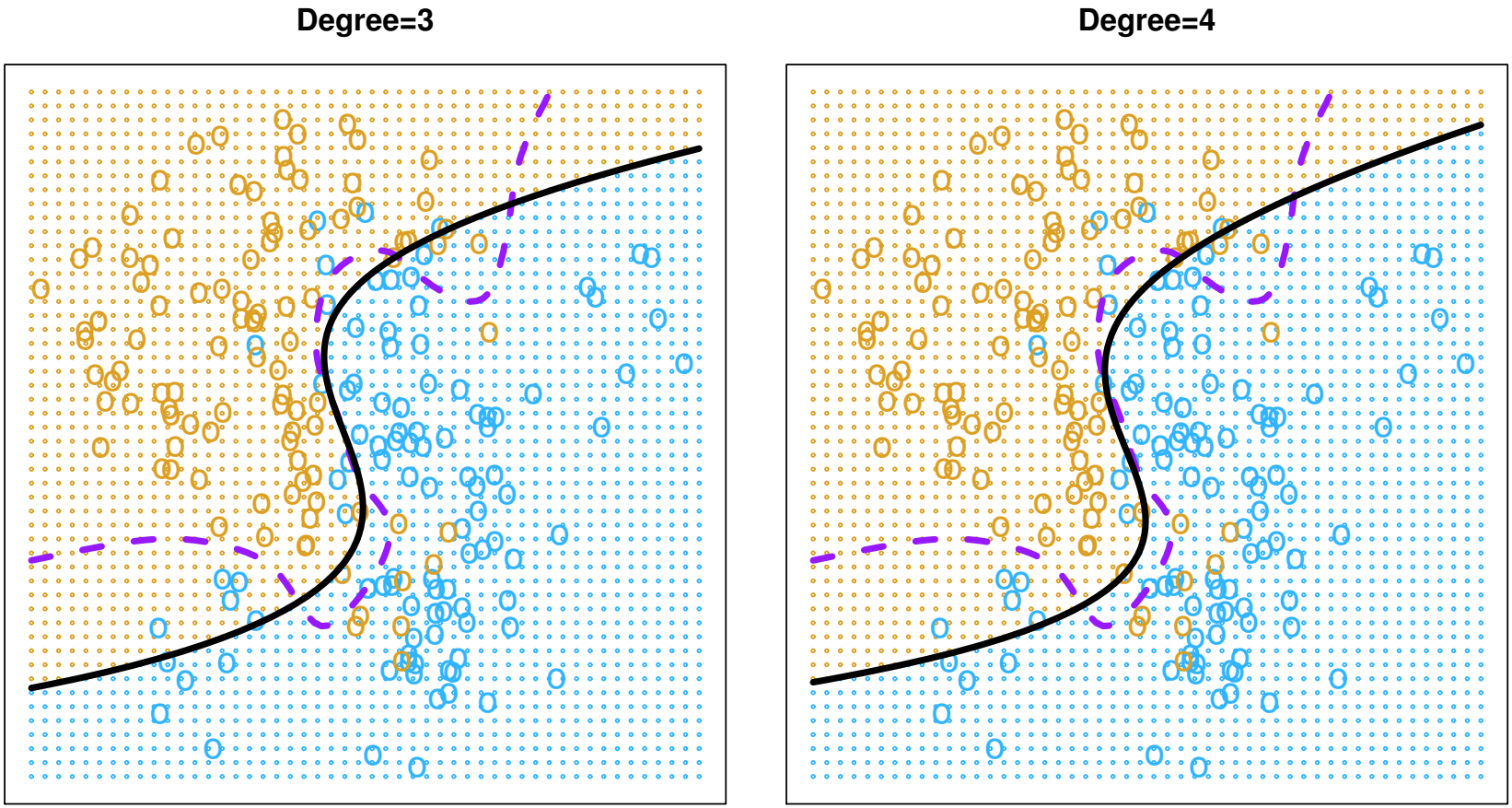
CV in classification setting



Logistic regression fit of 2D classification problem from Slide 66: Bayes decision boundary (purple dashed) and estimated decision boundary (solid black). Left: linear fit. right: quadratic fit. Bayes error rate: 0.133. (True) test error rates: 0.201 and 0.197.

Resampling Methods

CV in classification setting

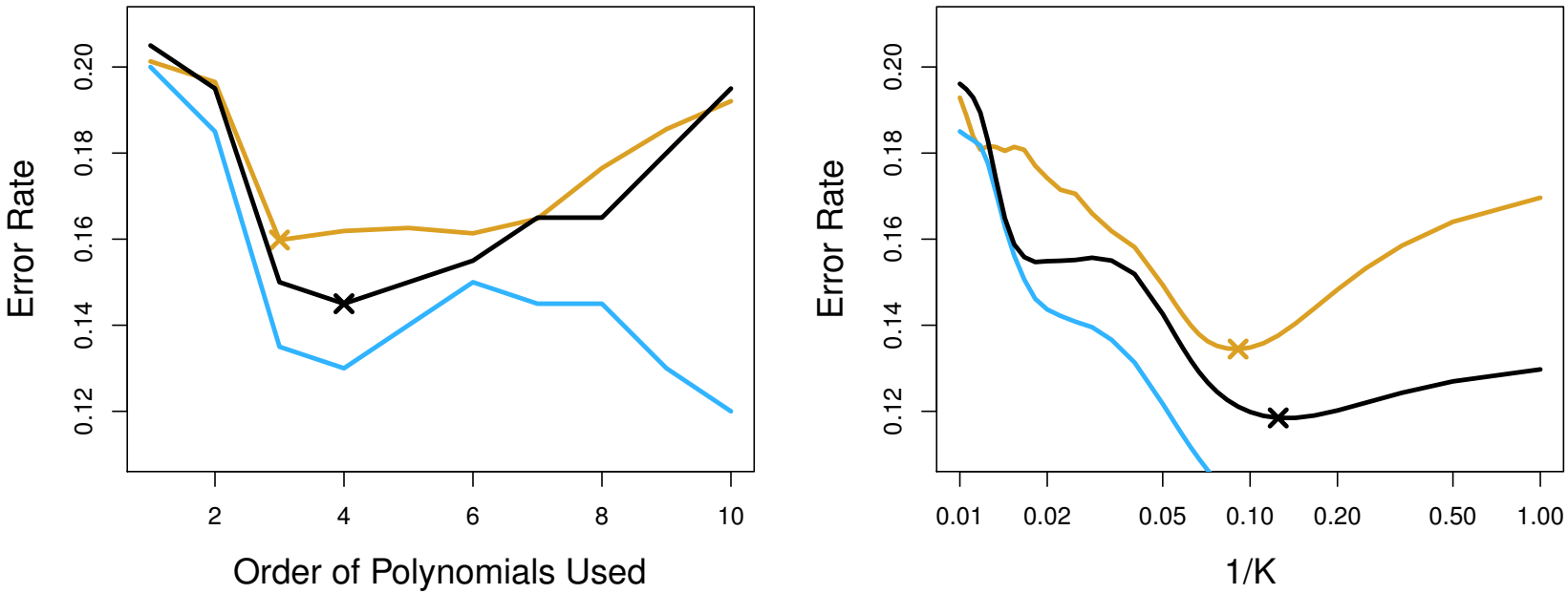


Same problem, same legend. Logistic regression now using cubic and quartic fits.

Bayes error rate: 0.133. (True) test error rates now : 0.160 and 0.162.

- In practice neither Bayes decision boundary, Bayes error rate nor true test error rate available, but CV offers way to choose among previous 4 models.

- In practice neither Bayes decision boundary, Bayes error rate nor true test error rate available, but CV offers way to choose among previous 4 models.



Same problem, same models. Black: 10-fold CV error rates from fitting 10 logistic regression models using polynomial functions of the predictor variables up to degree 10. Brown: true test errors, blue: training set errors. Right: KNN classifier with varying K (now denoting # nearest neighbors).

Resampling Methods

CV in classification setting

Observations:

- Training error decreases (roughly) with model flexibility.
- Test set error displays typical U-shape.
- 10-fold CV estimate provides good approximation of test error rates.
- Minimal for degree 4, matches true minimum well.
- Similar observations for KNN.
- Obvious: training set error not useful for model selection.

Contents

Resampling Methods

The bootstrap

- The **bootstrap** is a widely applicable and powerful statistical tool for quantifying the uncertainty associated with an estimate or statistical learning method.
- Example: linear regression coefficients (although simpler alternatives here).
- Nice introduction: [Efron, 2013][8]

[8] A 250-Year Argument: Belief, Behavior and the Bootstrap. Bull. AMS 50(1) 2013 pp. 129–146.

Resampling Methods

The bootstrap

- The **bootstrap** is a widely applicable and powerful statistical tool for quantifying the uncertainty associated with an estimate or statistical learning method.
- Example: linear regression coefficients (although simpler alternatives here).
- Nice introduction: [Efron, 2013][8]

Source: Things I can't avoid blog

Source: Wikipedia Commons

[8] A 250-Year Argument: Belief, Behavior and the Bootstrap. Bull. AMS 50(1) 2013 pp. 129–146.

- Goal: invest fixed sum of money in portfolio connsisting of 2 financial assets with (random) returns X and Y.

Resampling Methods

The bootstrap: investment (diversification) problem

- Goal: invest fixed sum of money in portfolio connsisting of 2 financial assets with (random) returns X and Y.
- Invest fraction α in X, remaining $1-\alpha$ in Y.

Resampling Methods

The bootstrap: investment (diversification) problem

- Goal: invest fixed sum of money in portfolio connsisting of 2 financial assets with (random) returns X and Y.
- Invest fraction α in X, remaining $1-\alpha$ in Y.
- Choose α to minimize total **risk** (here: variance) of investment, i.e., minimize $\mathbf{Var}(\alpha X + (1-\alpha)Y)$.

- Goal: invest fixed sum of money in portfolio connsisting of 2 financial assets with (random) returns X and Y.
- Invest fraction α in X, remaining $1-\alpha$ in Y.
- Choose α to minimize total **risk** (here: variance) of investment, i.e., minimize $\mathbf{Var}(\alpha X + (1-\alpha)Y)$.
- Can show: risk-minimizing value given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \tag{5.5}$$

where $\sigma_X^2 = \mathbf{Var}\, X$, $\sigma_Y^2 = \mathbf{Var}\, Y$, $\sigma_{XY} = \mathbf{Cov}(X, Y)$.

- Goal: invest fixed sum of money in portfolio connsisting of 2 financial assets with (random) returns X and Y.
- Invest fraction α in X, remaining $1-\alpha$ in Y.
- Choose α to minimize total **risk** (here: variance) of investment, i.e., minimize $\mathbf{Var}(\alpha X + (1-\alpha)Y)$.
- Can show: risk-minimizing value given by

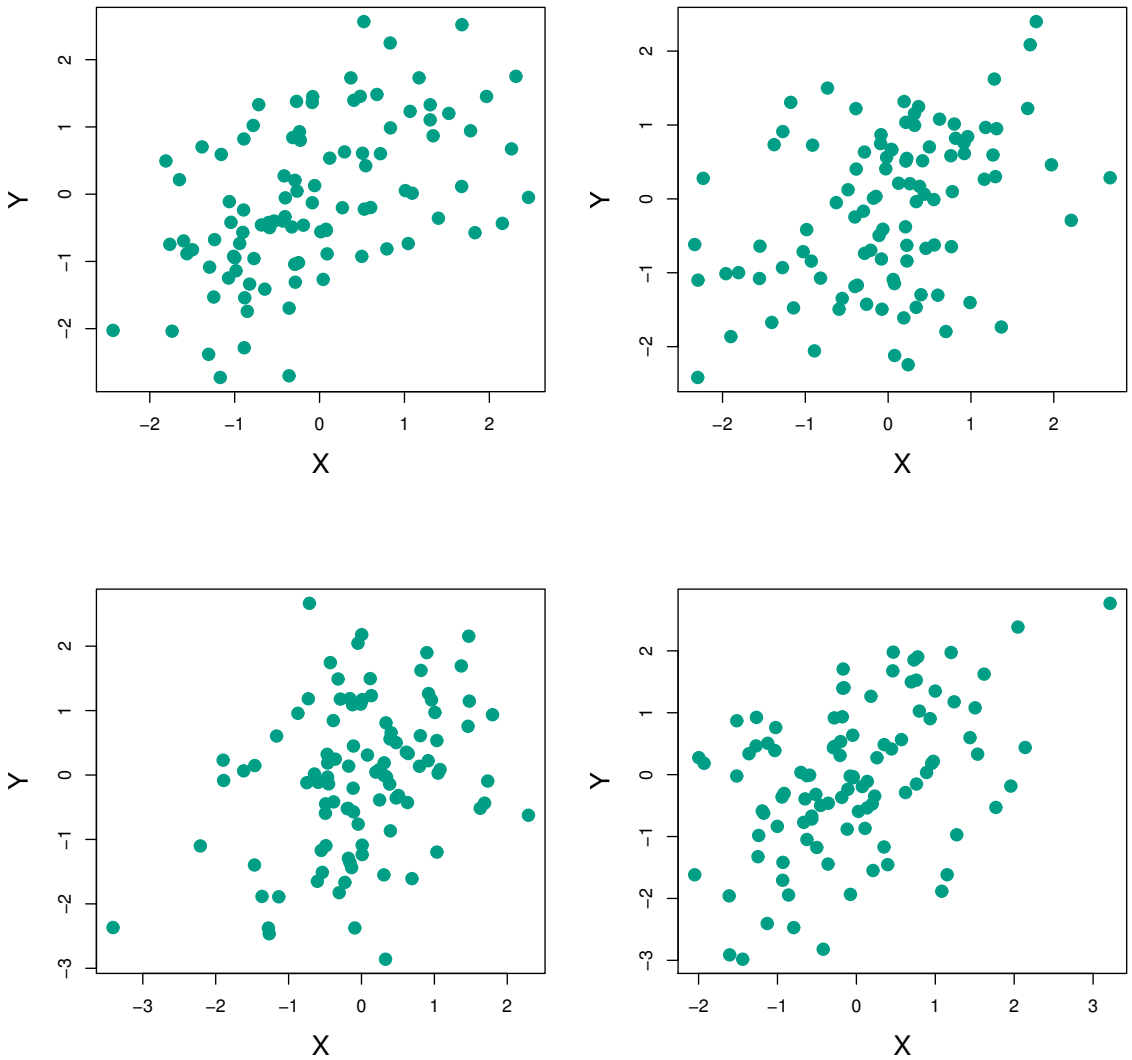
$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \tag{5.5}$$

  where $\sigma_X^2 = \mathbf{Var}\, X$, $\sigma_Y^2 = \mathbf{Var}\, Y$, $\sigma_{XY} = \mathbf{Cov}(X, Y)$.
- These quantities unknown in practice, use instead estimates $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, $\hat{\sigma}_{XY}$ and estimate risk-minimizing ratio as

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\sigma_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}. \tag{5.6}$$

Resampling Methods

The bootstrap: investment (diversification) problem

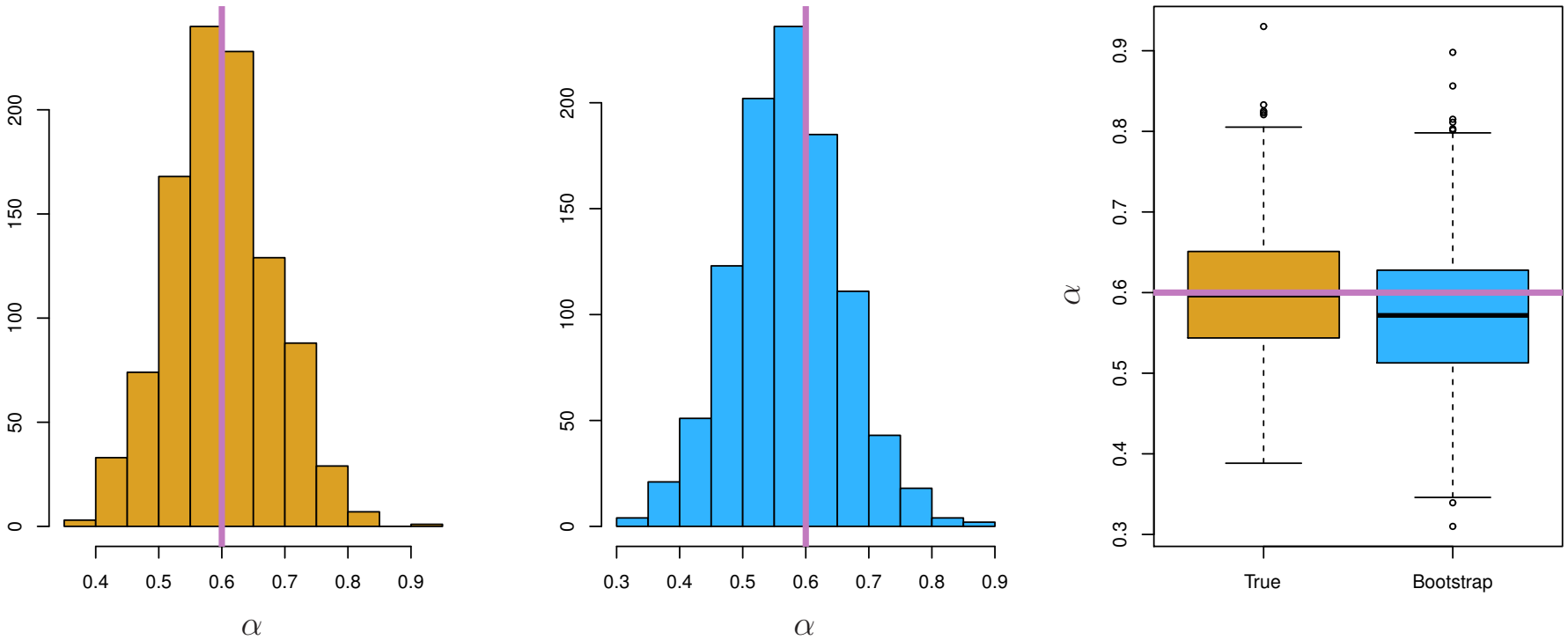


Random sampling: Each panel shows 100 simulated returns X and Y.

Lexicographically, sample variance/covariance estimates result in estimates $\hat{\alpha}$ for α of 0.576, 0.532, 0.657 and 0.651.

Resampling Methods

The bootstrap: investment (diversification) problem



Uncertainty quantification for estimate $\hat{\alpha} \approx \alpha$: 1000 repetitions of simulating 100 (X, Y)-observations and estimating α using (5.6). Left: histogram of $\{\hat{\alpha}_j\}_{j=1}^{1000}$. ($\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5, \Rightarrow \alpha = 0.6$, solid vertical line). Center: bootstrap histogram. Right: boxplots of original data and bootstrap data sets.

Mean over all estimates:

$$\overline{\alpha} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\alpha}_i = 0.5996 \approx 0.6 = \alpha.$$

Sample standard deviation:

$$\sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\alpha}_i - \overline{\alpha})^2} = 0.083,$$

hence $\mathsf{SE}(\hat{\alpha}) \approx 0.083$. We thus expect $\hat{\alpha}$ to deviate from α, on average, by 0.08.

Bootstrap estimate: use only original 100 samples to generate estimate $\hat{\alpha} \approx \alpha$ with a standard error of $\mathsf{SE}(\hat{\alpha}) = 0.087$.

Bootstrap approach:

- In general, can't generate multiple instances of given data.
- Bootstrap: use computer to **emulate** generation of new sample data sets.
- Use these to assess variability of associated estimates.
- Sampling proceeds from *original data set*.
- Sampling proceeds with *replacement*, all components of an observation treated as a unit.
- For $i = 1, \ldots, B$, generate i-th **bootstrap data set** Z^{*i}, each with estimate $\hat{\alpha}^{*i}$.
- Can estimate standard error of these estimates by

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\alpha}^{*i} - \frac{1}{B}\sum_{j=1}^{B}\hat{\alpha}^{*j}\right)^2} \tag{5.7}$$

- Example for data set Z containing $n = 3$ elements:

Resampling Methods

The bootstrap