MAIN CONCEPTS REGRESSION

1

AGENDA

- Estimates of f
- Bias-variance trade-off

ESTIMATES OF f

NOTATION

- *Input variables:* $X = (X_1, X_2, \dots, X_p)$ - independent variables, predictors, features
- *Output variable(s):* Y - response, dependent variables
- We assume some relationship between Y and X in the form

$$Y = f(X) + e, \qquad E[e] = 0,$$

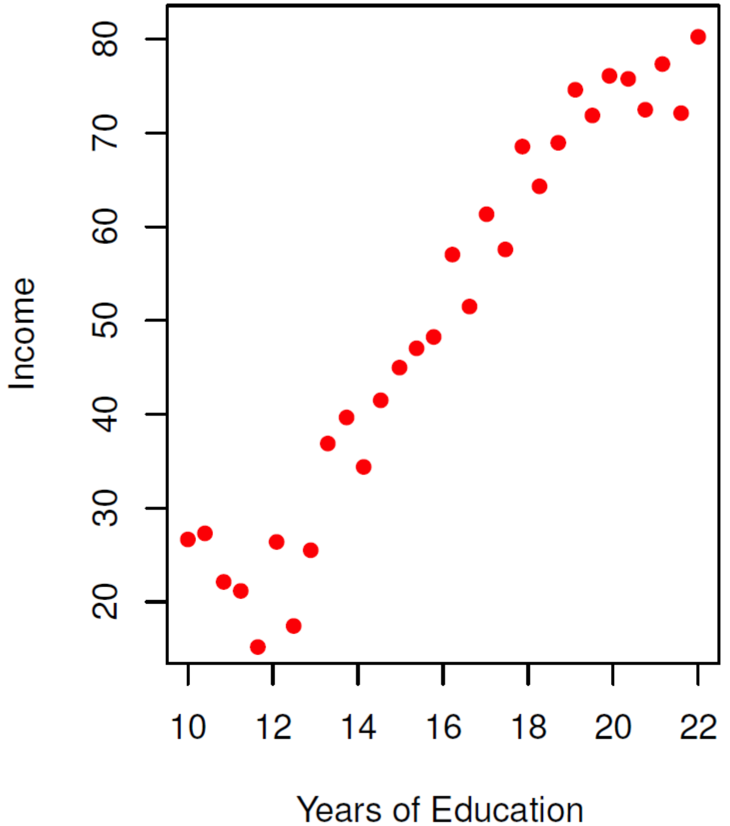
where e is a random error term (stochastic component) , which is independent of X

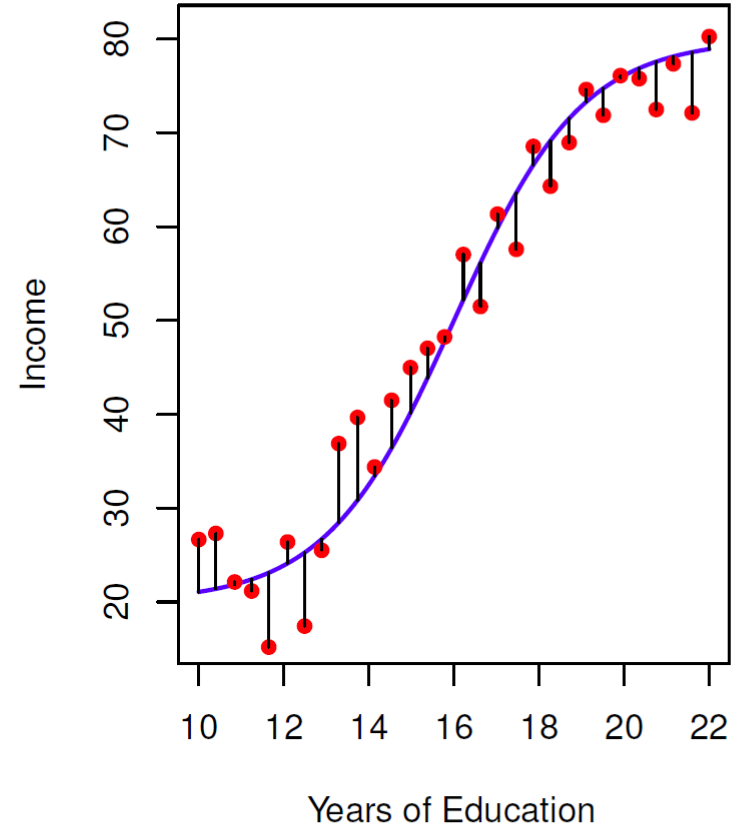
- We can predict Y using

$$\hat{Y} = \hat{f}(X)$$

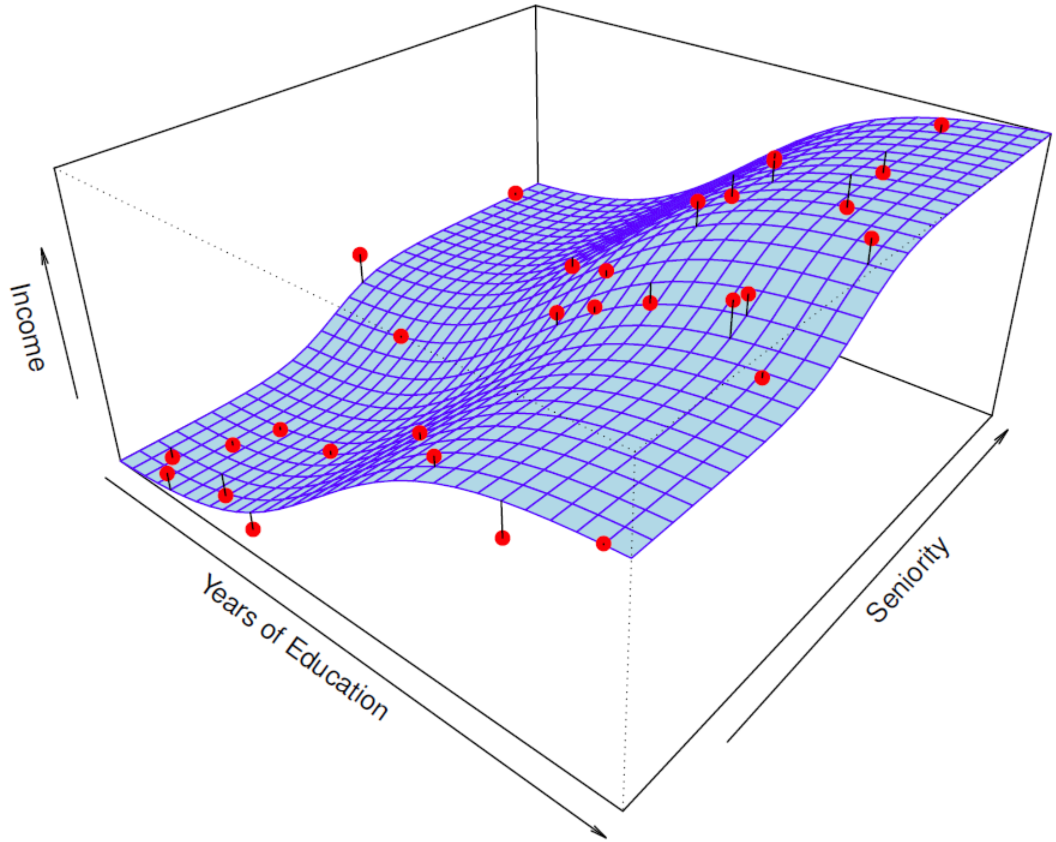
where $\hat{f}$ is estimate of f and $\hat{Y}$ is the prediction of Y

ESTIMATE OF f





ESTIMATE OF f



WHY WE ESTIMATE f?

- Prediction and inference (data understanding)
  - Make predictions of Y at new points
  - Understand which components of X are important in explaining Y
  - Depending on the complexity of f better understand relationship between X and Y (linear or non-linear)

CONDUCTING A DIRECT-MARKETING CAMPAIGN

- Identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual
- Predictors
  - Demographic variables
- Outcome
  - Response to the marketing campaign - Positive or Negative
- The company is not interested in obtaining a deep understanding of the relationships between each predictor and the response
- The company simply wants an accurate model to predict the response using the predictors – **Prediction Problem**

ADVERTISING DATA

- The goal may be answering the questions:
  - Which media contribute to sales?
  - Which media generate the biggest boost in sales?
  - How much increase in sales is associated with a given increase in TV advertising?
- **Inference Problem**

MODELING THE BRAND OF THE PRODUCT

- Model the brand of a product that a customer might purchase based on variables such as price, store location, discount levels, etc.
- How each of the individual variables affects the probability of purchase? What impact will have changing the price of a product on sales?
- **Inference Problem**

REAL ESTATE

- Relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, etc.
- How the individual input variables affect the prices? How much extra will a house be worth if it has a view of the river? – **Inference Problem**
- One may be interested in predicting the value of a home given its characteristics. Is this house under- or over-valued? - **Prediction Problem**

HOW WE ESTIMATE f?

- There is no free lunch in statistics: no one method dominates over all possible data sets
- It is an important task to decide for any given set of data which method produces the best results
- Selecting the best approach can be one of the most challenging parts of statistical learning

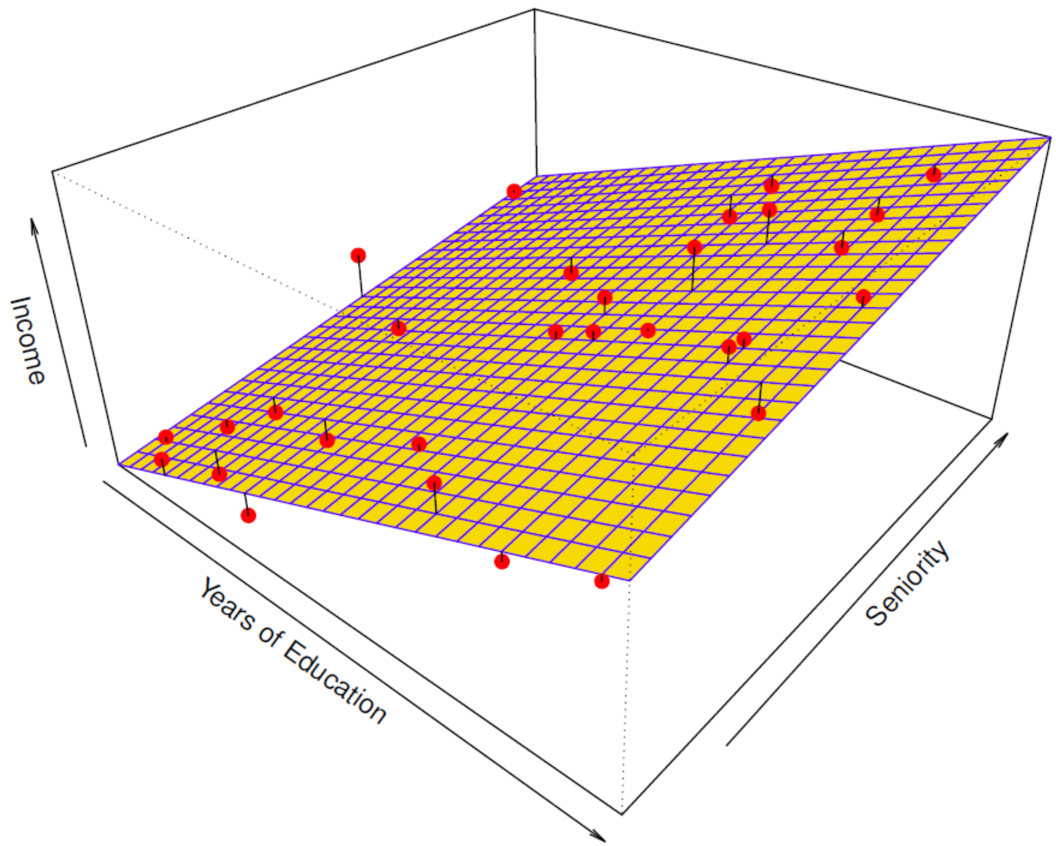
HOW WE ESTIMATE f?

- Method selection alternatives:
  - Regression vs classification
  - Parametric vs non-parametric
  - Quality of fit (data understanding) vs quality of prediction
  - Model flexibility vs model interpretability
  - Model bias vs model variance

PARAMETRIC VS NON-PARAMETRIC

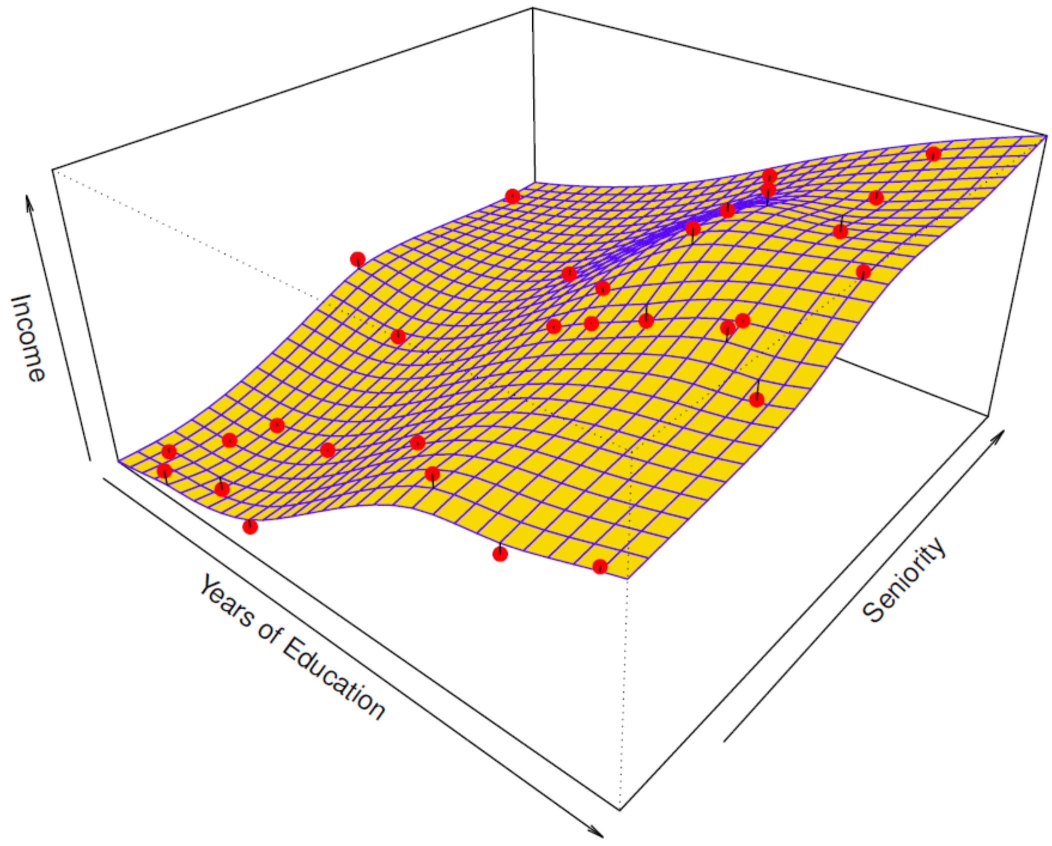
- Parametric method first select a model (linear, quadratic, etc.) and then fit it by training data
- Advantage of parametric models
  - Simplicity
- Disadvantages of parametric models
  - If the chosen model is too far from the true function, then our estimate will be poor
  - We can try more flexible models with greater parameters but it can lead to another problem known as overfitting the data, which essentially means they follow the noise, too closely
  - Non-parametric methods do not make explicit assumptions about the functional form of f

PARAMETRIC VS NON-PARAMETRIC



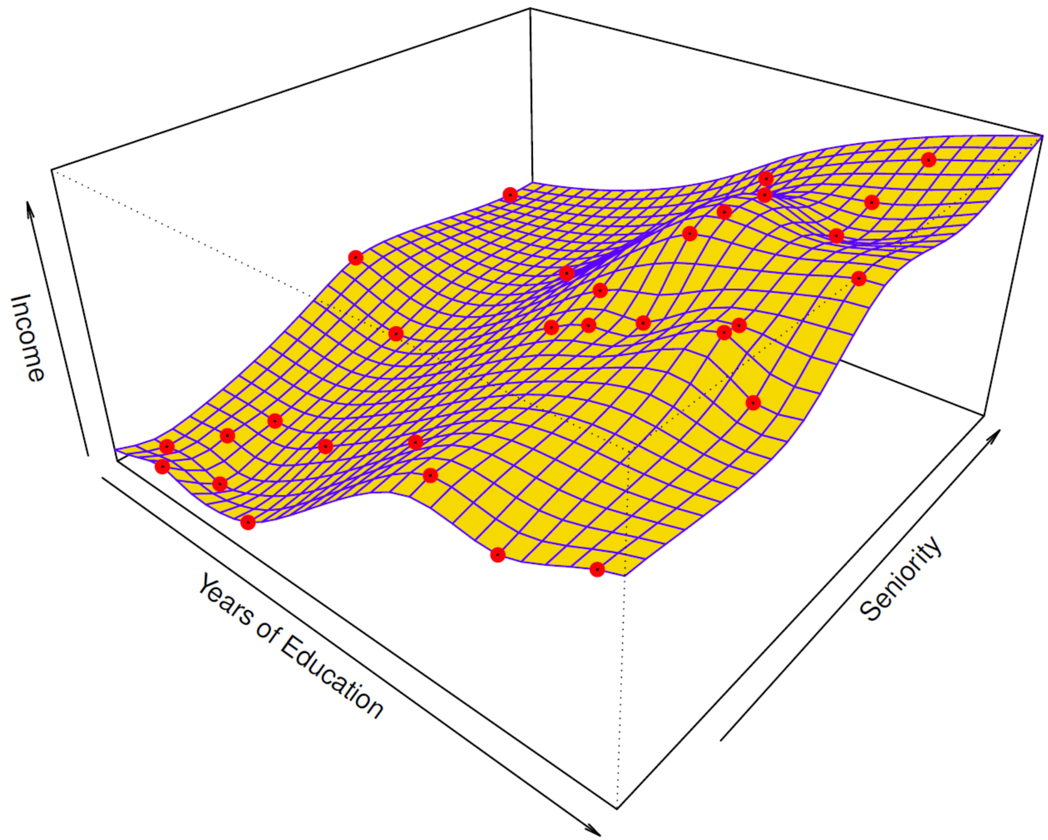
Parametric approach (linear regression) applied to the Income data

PARAMETRIC VS NON-PARAMETRIC



Non-parametric approach: thin-plate spline

PARAMETRIC VS NON-PARAMETRIC



Thin-plate spline application with lower level of smoothness. Perfect fit for the observed data but undesirable variability. More sensitive to noise with worse predictive properties

QUALITY OF FIT VS PREDICTION ACCURACY

- Accuracy of a model

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2 = E\left[\left(Y - \hat{Y}\right)^2\right]$$

- Training data – train MSE (quality of fit)
- Test data, which are previously unseen observations not used to train the statistical learning model – test MSE (quality of prediction)
- We don't care how small is train MSE – Why?
- Can we decrease test MSE by decreasing the train MSE?

INDEPENDENCE

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

- If X and Y are independent

$$Cov(X,Y) = 0$$

$$E[XY] = E[X]E[Y]$$

IRREDUCIBLE AND REDUCIBLE ERRORS

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

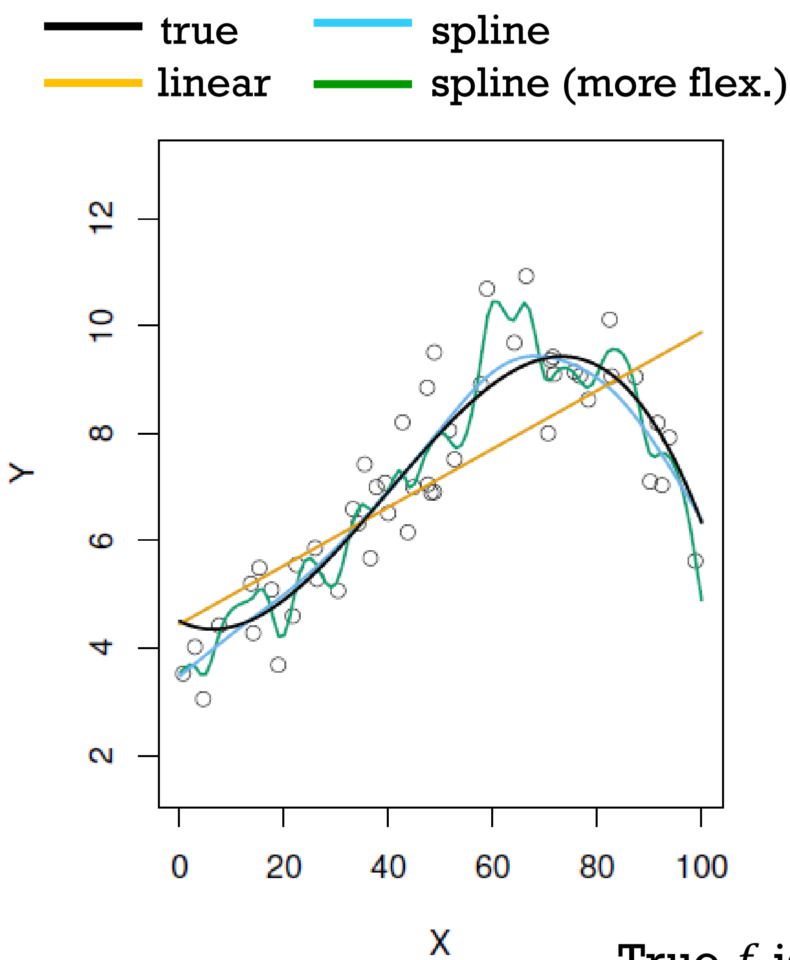
$$E\left[\left(f(X)-\hat{f}(X)\right)^2\right] + E[e^2] + \underbrace{2E\left[e\left(f(X)-\hat{f}(X)\right)\right]}_{0} =$$

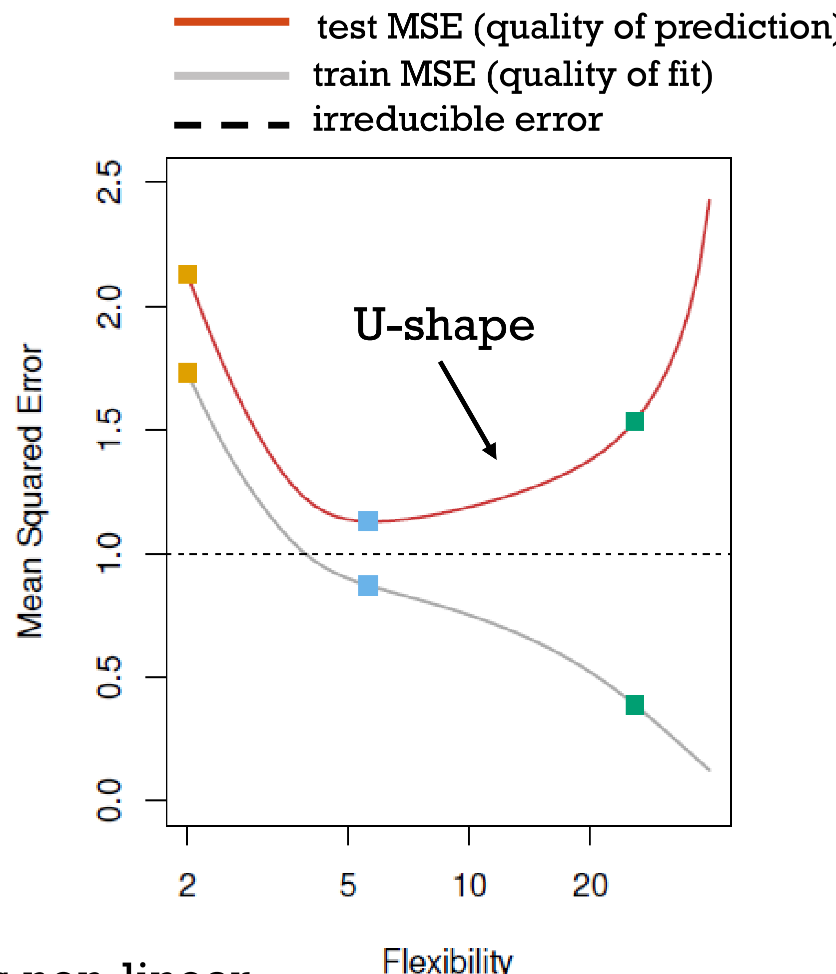
$$= \underbrace{E\left[\left(f(X)-\hat{f}(X)\right)^2\right]}_{Reducible\ Error} \quad + \quad \underbrace{Var[e]}_{Irreducible\ Error}$$

- $X = x_0$

$$MSE = \left(f(x_0)-\hat{f}(x_0)\right)^2 + Var[e]$$

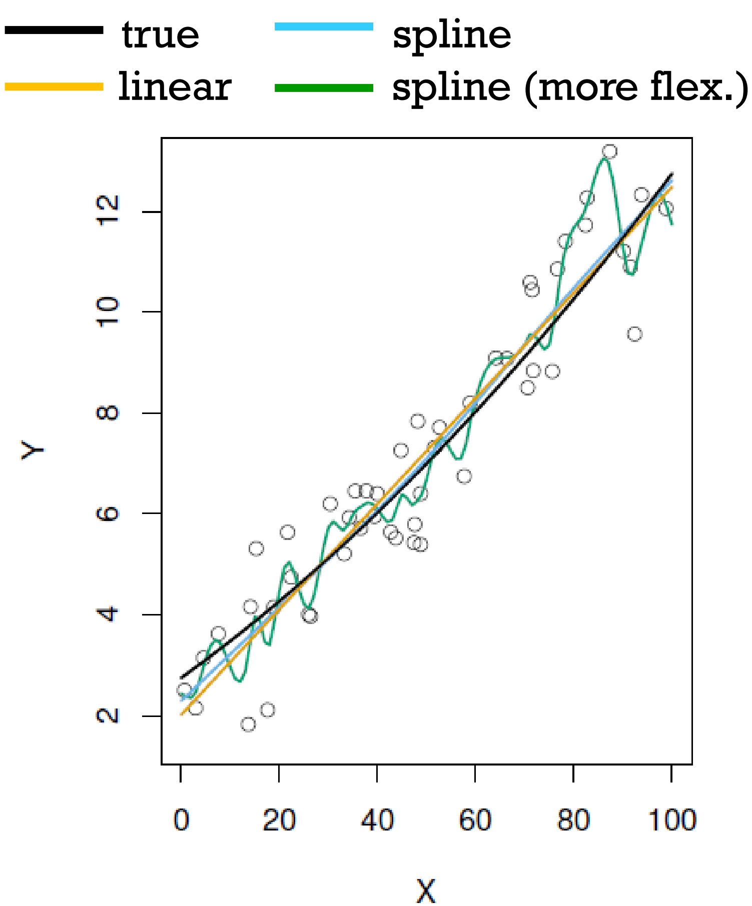
QUALITY OF FIT VS PREDICTION ACCURACY

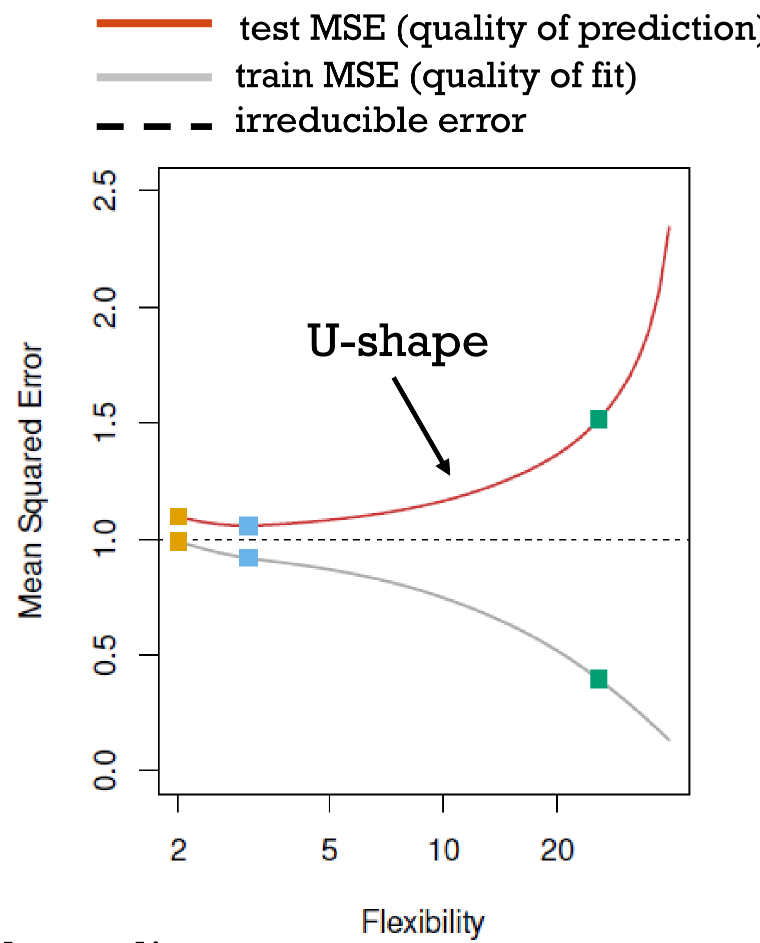




True f is non-linear

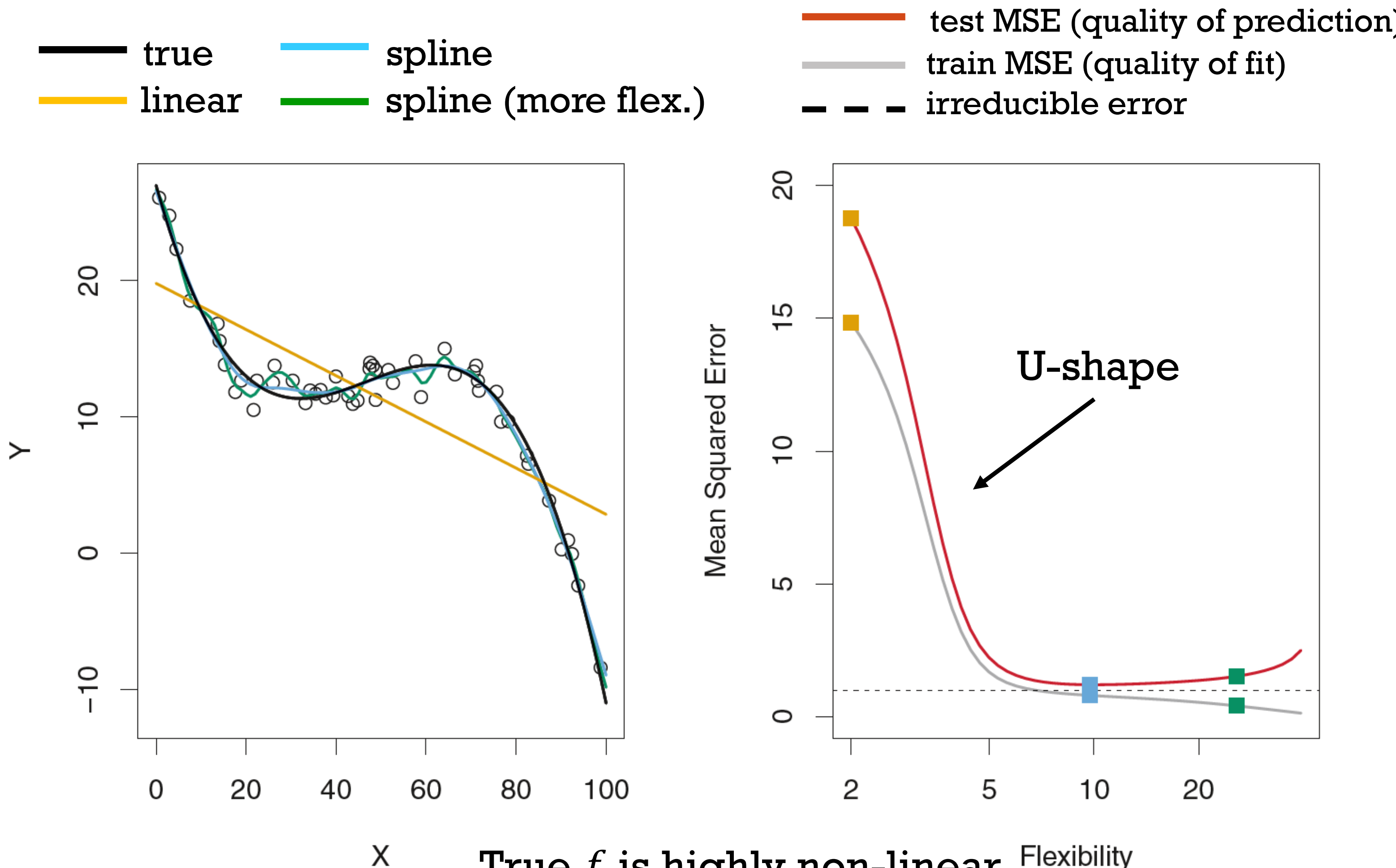
QUALITY OF FIT VS PREDICTION ACCURACY





True f is almost linear

QUALITY OF FIT VS PREDICTION ACCURACY



True f is highly non-linear

QUALITY OF FIT VS PREDICTION ACCURACY

- Test MSE can never lie below $Var(e)$
- As higher is the flexibility as less is the training MSE. Training MSE monotonically decreases
- Test MSE has a U-shape: fundamental property of ML regardless data and model
- When a given method yields a small training MSE but a large test MSE, we are said to be ***overfitting*** the data

FLEXIBILITY VS INTERPRETABILITY

- Linear regression is relatively inflexible approach, as it can generate only linear functions
- Thin plate splines are considerably more flexible as they can generate a much wider class of possible shapes to estimate f

FLEXIBILITY VS INTERPRETABILITY

- There are some reasons why we apply inflexible approaches
  - In general, inflexible methods are less complex
  - Restrictive models are much more interpretable in the sense of statistical inference. In case of flexible methods it is difficult to understand connection between individual predictor and the response
- When inference is the final goal (not prediction accuracy) then inflexible methods have clear advantages
- When prediction is the final goal then flexible (more accurate) methods are preferable. However, for many problems less flexible methods will provide with better accuracy (see bias-variance trade-off problem)

BIAS-VARIANCE TRADE-OFF

VARIANCE AND EXPECTATION

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

- If $X = Y$

$$Cov(X,X) = Var[X] = E[(X - E[X])^2]$$

$$\boldsymbol{E[X^2] = Var[X] + E[X]^2}$$

ERROR DECOMPOSITION

$$Y = f(X) + e$$

$$\hat{Y} = \hat{f}(X)$$

$$MSE = \underbrace{E\left[\left(f(X) - \hat{f}(X)\right)^2\right]}_{Reducible\ Error} \quad + \quad \underbrace{Var[e]}_{Irreducible\ Error}$$

BIAS-VARIANCE-NOISE DECOMPOSITION

- Prediction for $X = x_0$

$$y_0 = f(x_0)\ (deterministic\ \ prediction)$$

- Training set is not fixed

$$X^{(1)}, X^{(2)}, \dots, X^{(m)}, \dots$$

$$\hat{f}_1\left(X^{(1)}\right) = \hat{Y}_1\,, \qquad \hat{f}_2\left(X^{(2)}\right) = \hat{Y}_2, \qquad \dots \qquad \hat{f}_m\left(X^{(m)}\right) = \hat{Y}_m$$

- Prediction for $X = x_0$

$$\hat{y}_0 = \hat{f}(x_0)\ (stochastic\ prediction)$$

$$\hat{f} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m, \dots\}$$

BIAS-VARIANCE-NOISE DECOMPOSITION

$$Reducible\ Error = E\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] =$$

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

$$E\left[\left(E[\hat{f}] - f\ \right)^2\right] + E\left[\left(\hat{f} - E[\hat{f}]\right)^2\right] + \underbrace{2E[(\hat{f} - E[\hat{f}])(E[\hat{f}] - f)]}_{0} =$$

Since $E[f] = f,$

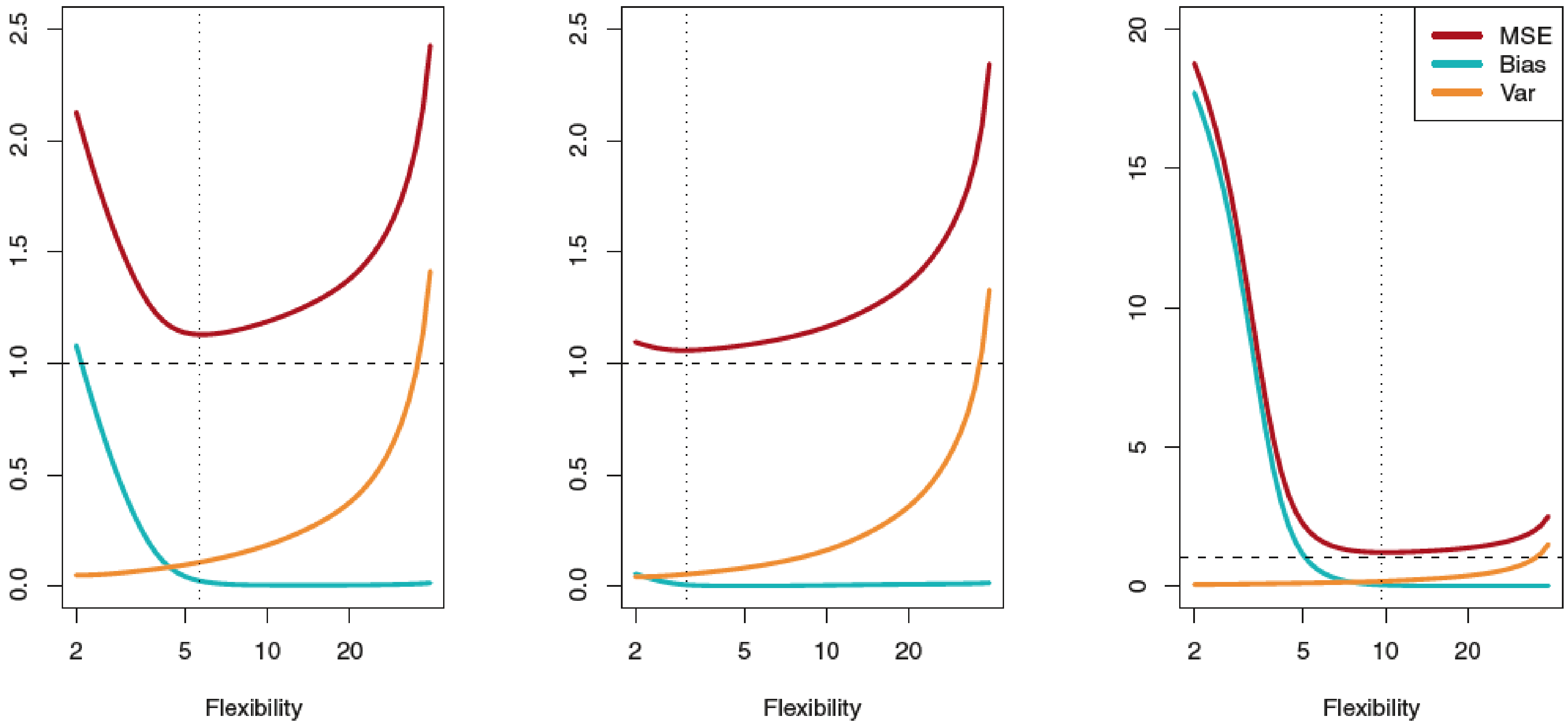
$$\left(E[\hat{f}(x_0)] - f(x_0)\right)^2 + E\left[\left(\hat{f}(x_0) - E[\hat{f}(x_0)]\right)^2\right]$$

$$\boldsymbol{MSE} = \boldsymbol{Bias}[\hat{f}(x_0)]^{\mathbf{2}} + \boldsymbol{Var}[\hat{f}(x_0)] + \boldsymbol{Var}[\boldsymbol{e}]$$

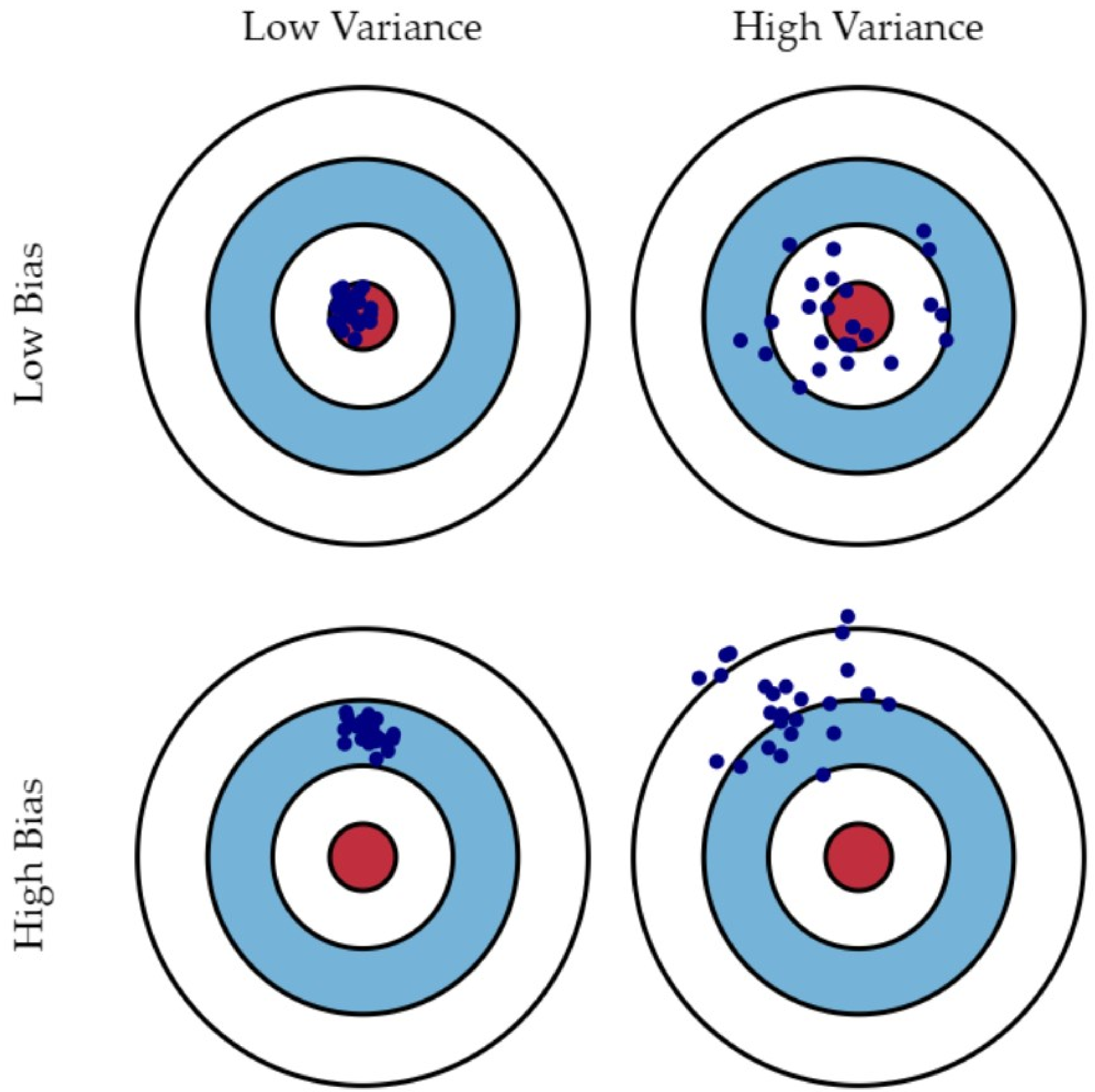
BIAS-VARIANCE TRADE-OFF

- We need to select a statistical learning method that simultaneously achieves *low variance* and *low bias*
- In general, more flexible methods have higher variance
- In general, more flexible methods result in less bias

BIAS-VARIANCE TRADE-OFF



As we use more flexible methods, the variance will increase and the bias will decrease. Bias-variance decomposition explains the U-shape of the test MSE


Low Variance
High Variance
Low Bias
High Bias