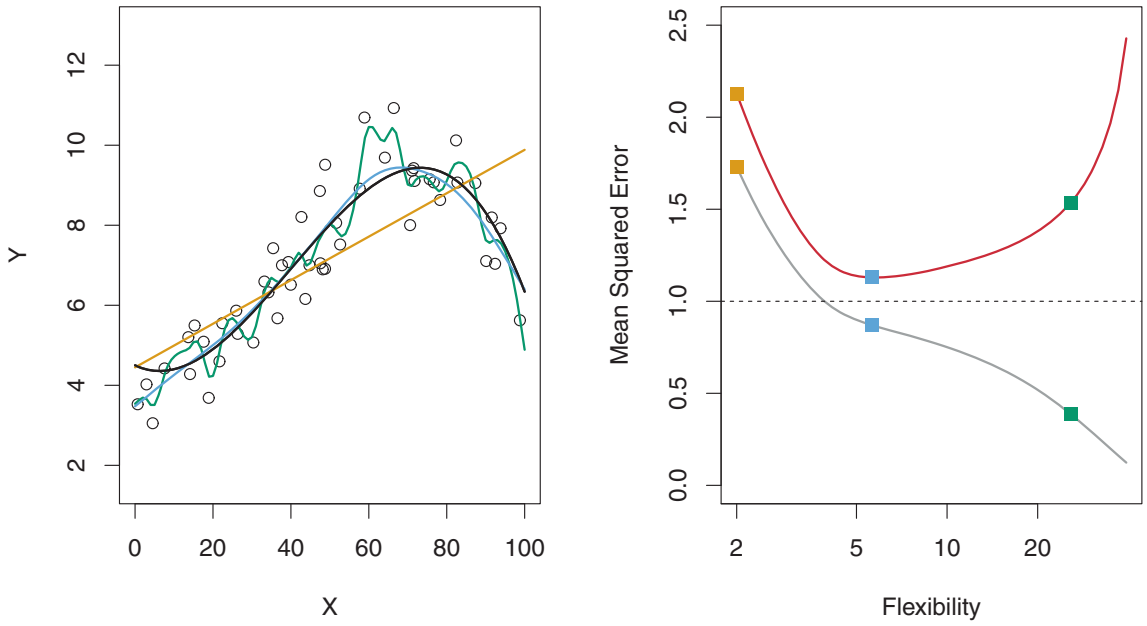
iii. There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.



Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line).

In the right panel, as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE. This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of statistical learning method being used.

when a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data.
overfitting refers to the case in which a less flexible model would have yielded a smaller test MSE.

The Bias-Variance Trade-Off

$\text{Bias}(\hat{f}) = E\hat{f} - f$

The test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ. That is,

show that (*) is true.

$$E\big(y_0 - \hat{f}(x_0)\big)^2 \overset{*}{=} \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

Var(ε) ≤

Here the notation $E\big(y_0 - \hat{f}(x_0)\big)^2$ defines the expected test MSE, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0. The overall expected test MSE can be computed by averaging $E\big(y_0 - \hat{f}(x_0)\big)^2$ over all possible values of x_0 in the test set.

Notes:

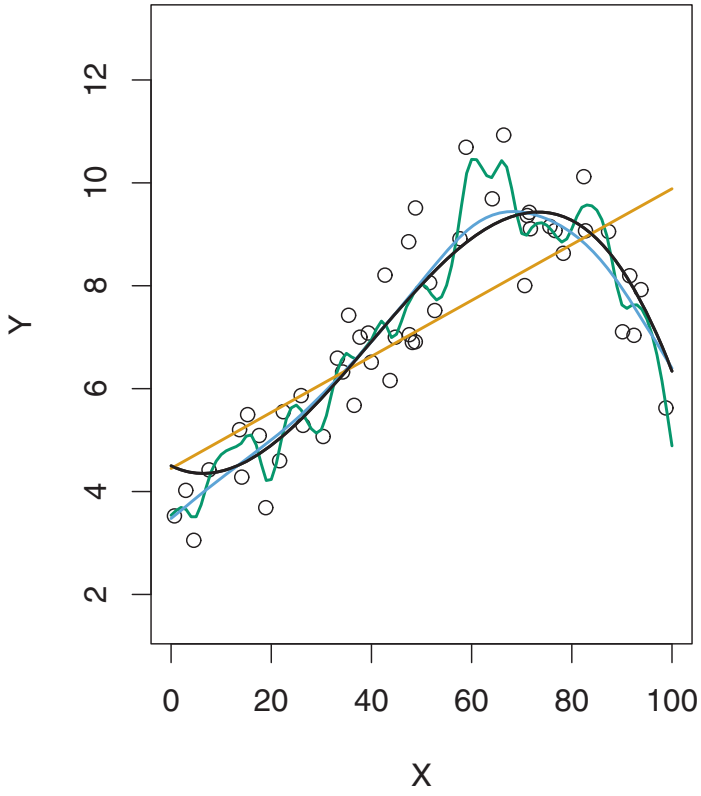
i. In order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias.

ii. Variance (of a statistical learning method) refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.

If a method has high variance then small changes in the training data can result in large changes in $\hat{f}$.

In general, more flexible statistical methods have higher variance.

iii. Bias (of a statistical learning method) refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.



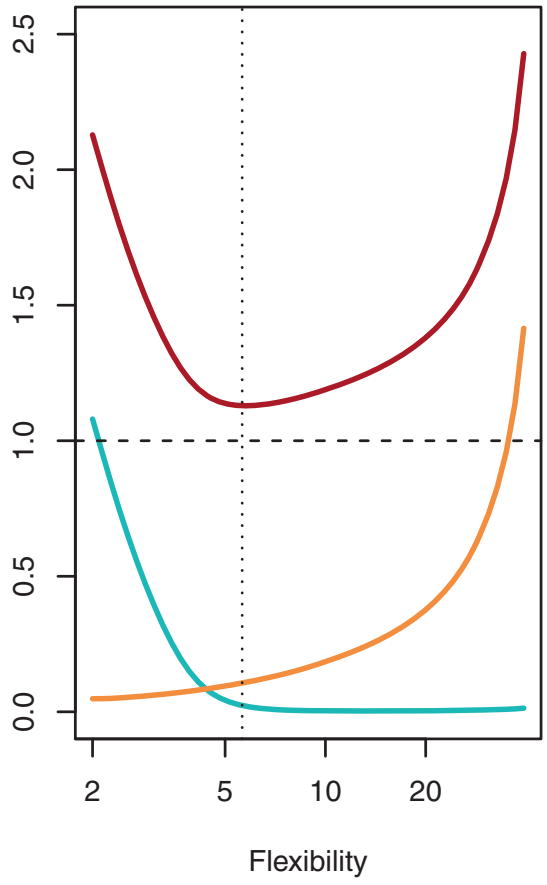
Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).

The flexible green curve is following the observations very closely. It has high variance b/c changing any one of these data points may cause the estimate

f to change considerably. In contrast, the orange line is relatively inflexible and has low variance, b/c moving any single observation will likely cause only small shift in the position of the line.

The true f is substantially non-linear, so linear regression (orange curve) results in high bias in this example.

In general, more flexible methods result in less bias.



Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the previous simulated data set.

Measuring the Quality of Fit in the Classification Setting

Suppose that we seek to estimate a classifier on the basis of training observations $(x_1, y_1), \ldots, (x_n, y_n)$, where now $y_1, \ldots, y_n$ are qualitative. The most common approach for quantifying the accuracy of our estimate is the training error rate, the proportion of mistakes that are made if we apply our estimate to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Here $\hat{y}_i$ is the predicted class label for the ith observation using our estimate.

$I(y_i \neq \hat{y}_i) = \begin{cases} 1 & y_i \neq \hat{y}_i \\ 0 & y_i = \hat{y}_i \end{cases}$

↑ indicator function

If $I(y_i \neq \hat{y}_i) = 0$ then the ith observation was classified correctly by our classification method; otherwise it was misclassified.

The test error rate associated with a set of test observations of the form (x_0, y_0) is given by

$$\text{Ave}(I(y_0 \neq \hat{y}_0))$$

where $\hat{y}_0$ is the predicted class label that results from applying the classifier to the test observation with predictor x_0.

A good classifier is one for which the test error is smallest.

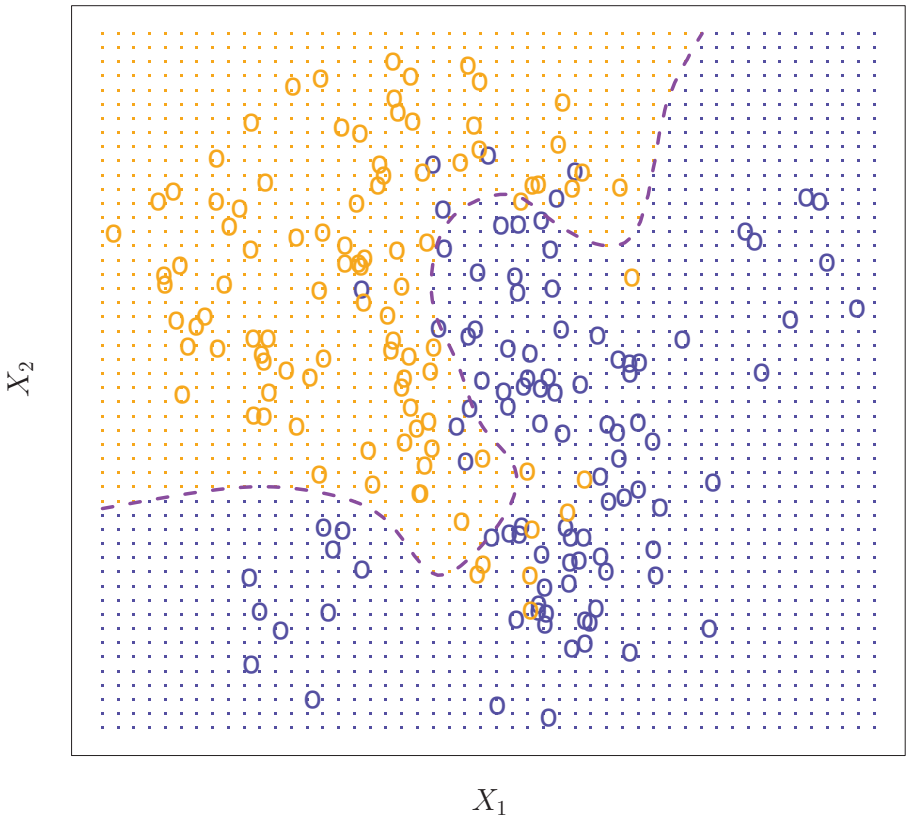
The Bayes Classifier

- It is possible to show that the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values.
- In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which

  $$\Pr(Y = j|X = x_0)$$

  is largest. This very simple classifier is called the *Bayes classifier*.

- In a two-class problem where there are only two possible response values, say class 1 or class 2, the Bayes classifier corresponds to predicting class one if $\Pr(Y = 1|X = x_0) > 0.5$, and class two otherwise.
- The line (curve) formed by the points where the probability is exactly 50% is called the Bayes decision boundary. The Bayes classifier's prediction is determined by the Bayes decision boundary.



A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

- The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate. The overall Bayes error rate is given by

$$1 - E\left(\max_j \Pr(Y = j|X)\right),$$

where the expectation averages the probability over all possible values of X.

The Bayes error rate is analogous to the irreducible error, discussed earlier.