MATH 569 Statistical Learning

Part I: Overview of Supervised Learning

Maggie Cheng

ILLINOIS INSTITUTE OF TECHNOLOGY

Supervised vs. Unsupervised Learning

- Supervised Learning
  - Data include outcome variable (Y) and predictor variable (X): $(x_i, y_i)_{i=1}^{N}$
  - Use the presence of Y to guide the learning process.
  - Two steps:
    - estimate a relationship between Y and X: y=f(x)
    - use f(x) to predict the value of y for a new x
  - Examples:

Supervised vs. Unsupervised Learning

- Unsupervised learning
  - Data only include X; Y is not given
  - Task: not to estimate the relationship between Y and X, or predict Y for a given X; rather to describe how the data are organized or clustered
  - Examples:

Supervised or Unsupervised Learning?

- Gene expression
- Red/blue balls
- Handwritten digits

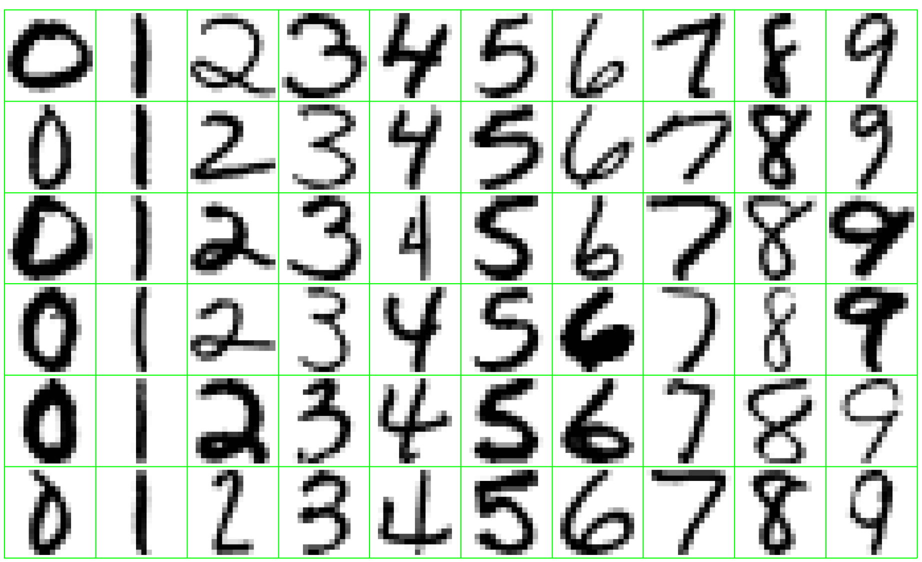
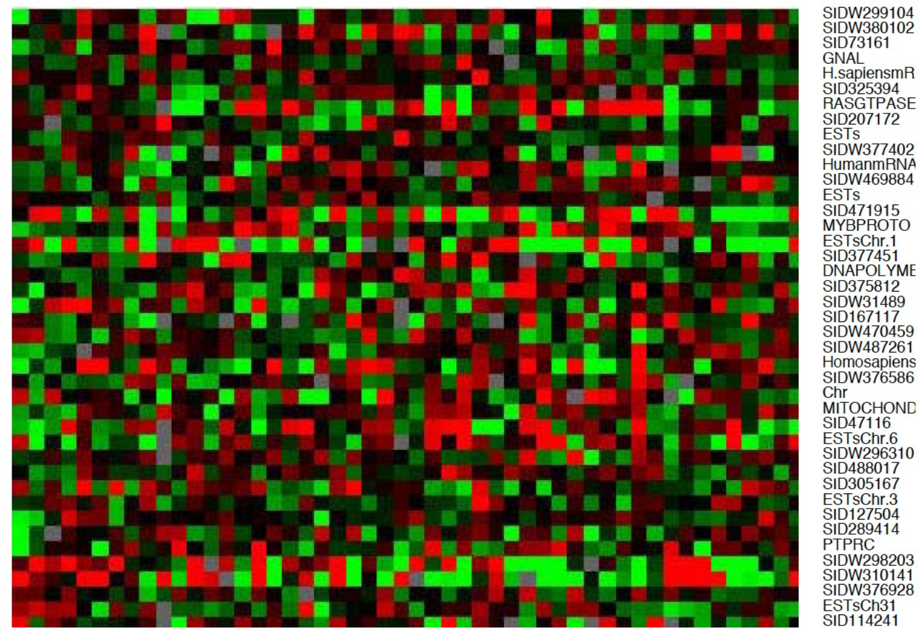


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*



Overview of Supervised Learning

- Supervised Learning:
  - use the inputs to predict the values of the outputs.
- Inputs (X):
  - Predictors
  - Independent variables
  - Features
- Outputs(Y):
  - Responses
  - Dependent variables

Supervised Learning

- Types of variables
  - Quantitative
  - Qualitative, or categorical
  - Ordered categorical
- Tasks
  - Regression: quantitative (continuous) outcome
  - Classification: qualitative (discrete) outcome

Notations

- Upper case: variables
  - X: input, could be a vector. Use X_j to denote the j-th variable
  - Y: output
  - $\hat{Y}$: is the predicted output variable; use $\hat{G}$ for classification
- Lower case: values for the variables

- $\mathbf{X}$: bold upper case, $N \times p$ matrix
- x_i: the p-vector x_i is the ith observation
- $\mathbf{x_j}$: the bold N-vector $\mathbf{x_j}$ consists of all the observations on variable X_j
- All vectors are column vectors

Supervised Learning

- Labelled training data
  - Regression problem:

$$\mathcal{T} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

  - Classification problem:

$$\mathcal{T} = \{(x_1, g_1), \ldots, (x_N, g_N)\}$$

Prediction Rules

- Least squares
- k-nearest neighbors
- …

Fig 2.1, fit by linear regression
Orange: 1
Blue: 0

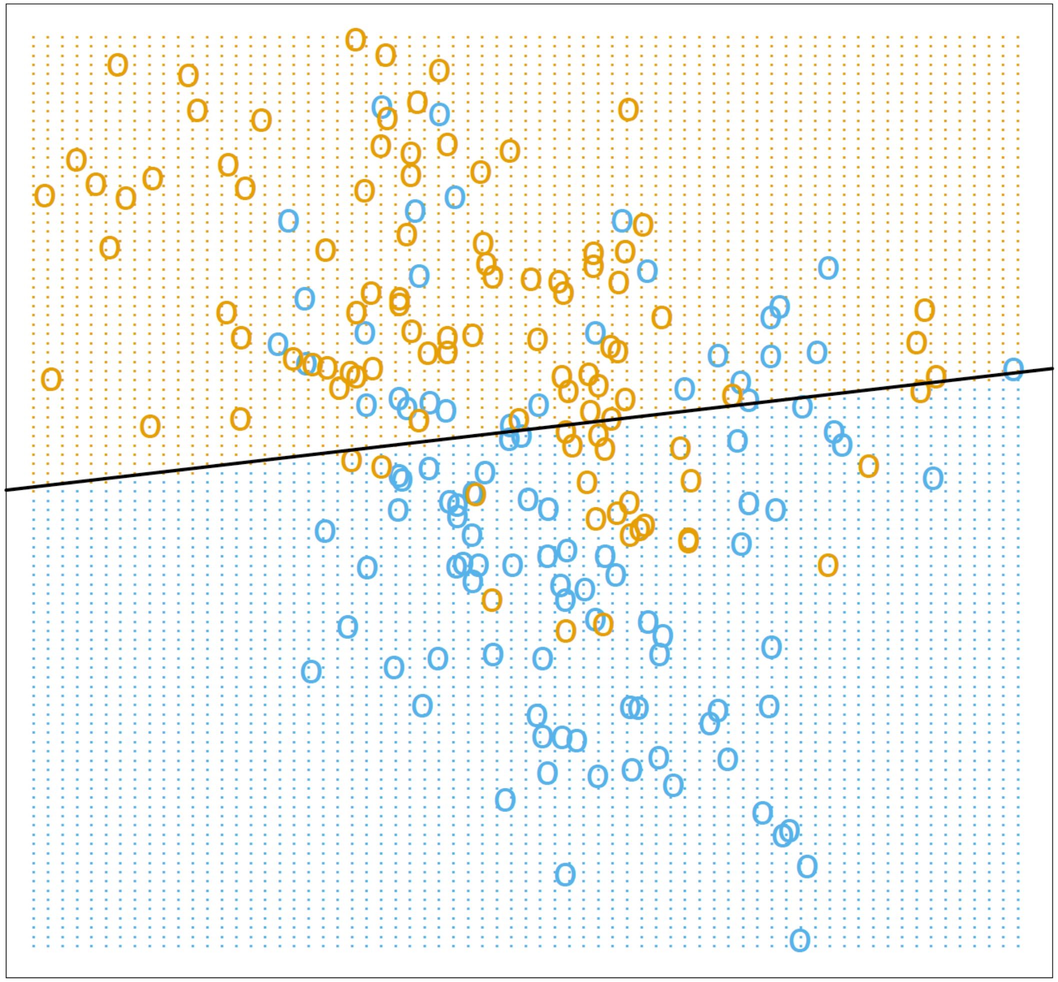


Fig 2.2, fit by K-Nearest Neighbor with k=15

Orange: 1

Blue: 0

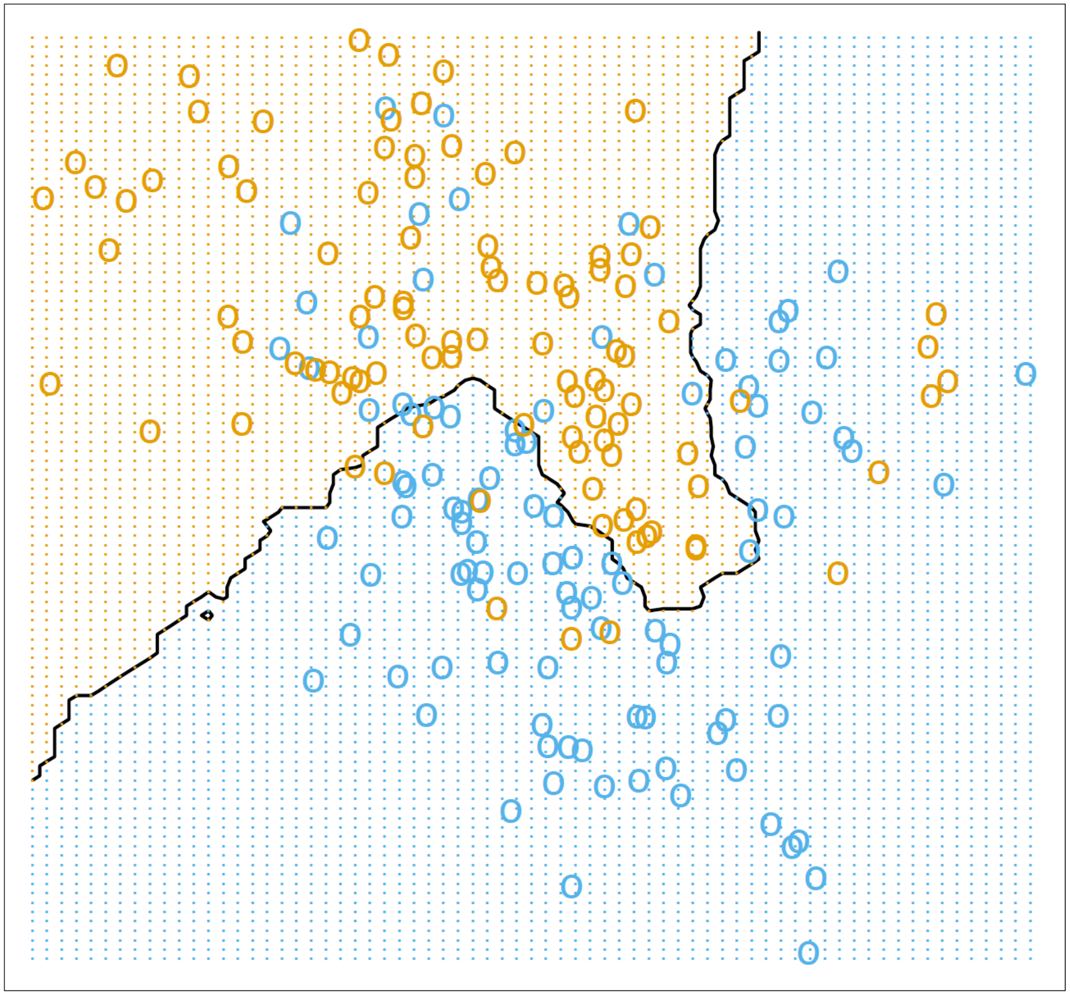


Fig 2.3, fit by K-Nearest Neighbor with k=1
Orange: 1
Blue: 0

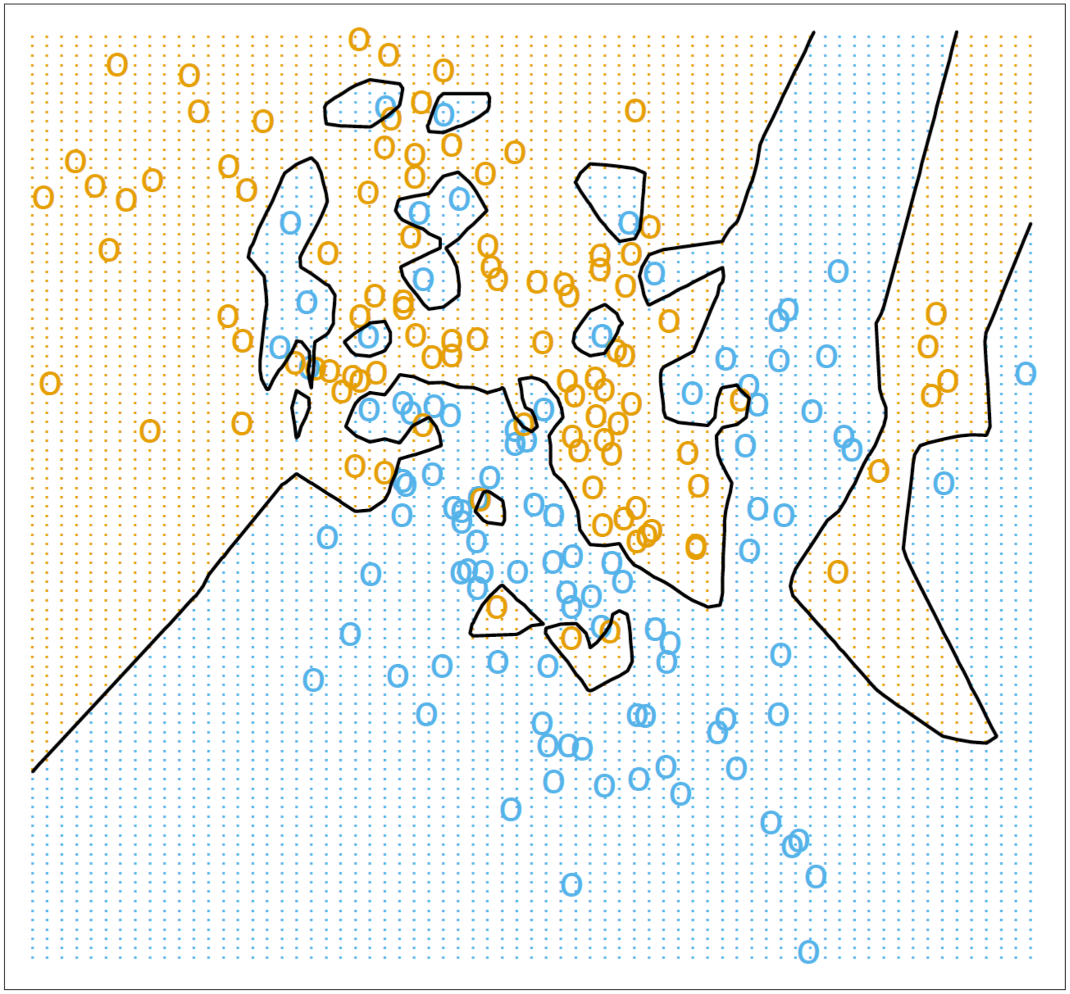


Fig 2.5, fit by Bayes Optimal Classifier

Orange: 1

Blue: 0

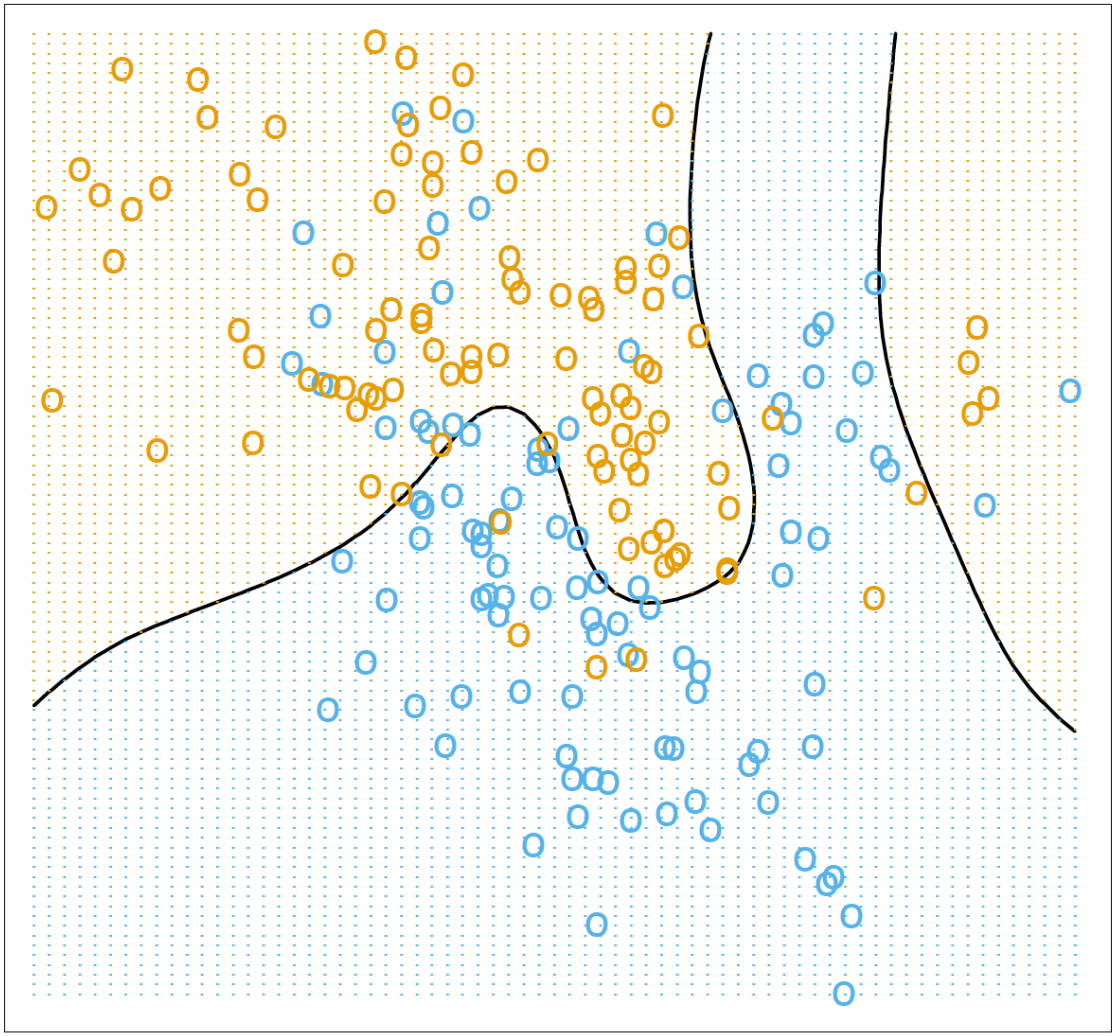


Fig 2.6, the curse of dimensionality

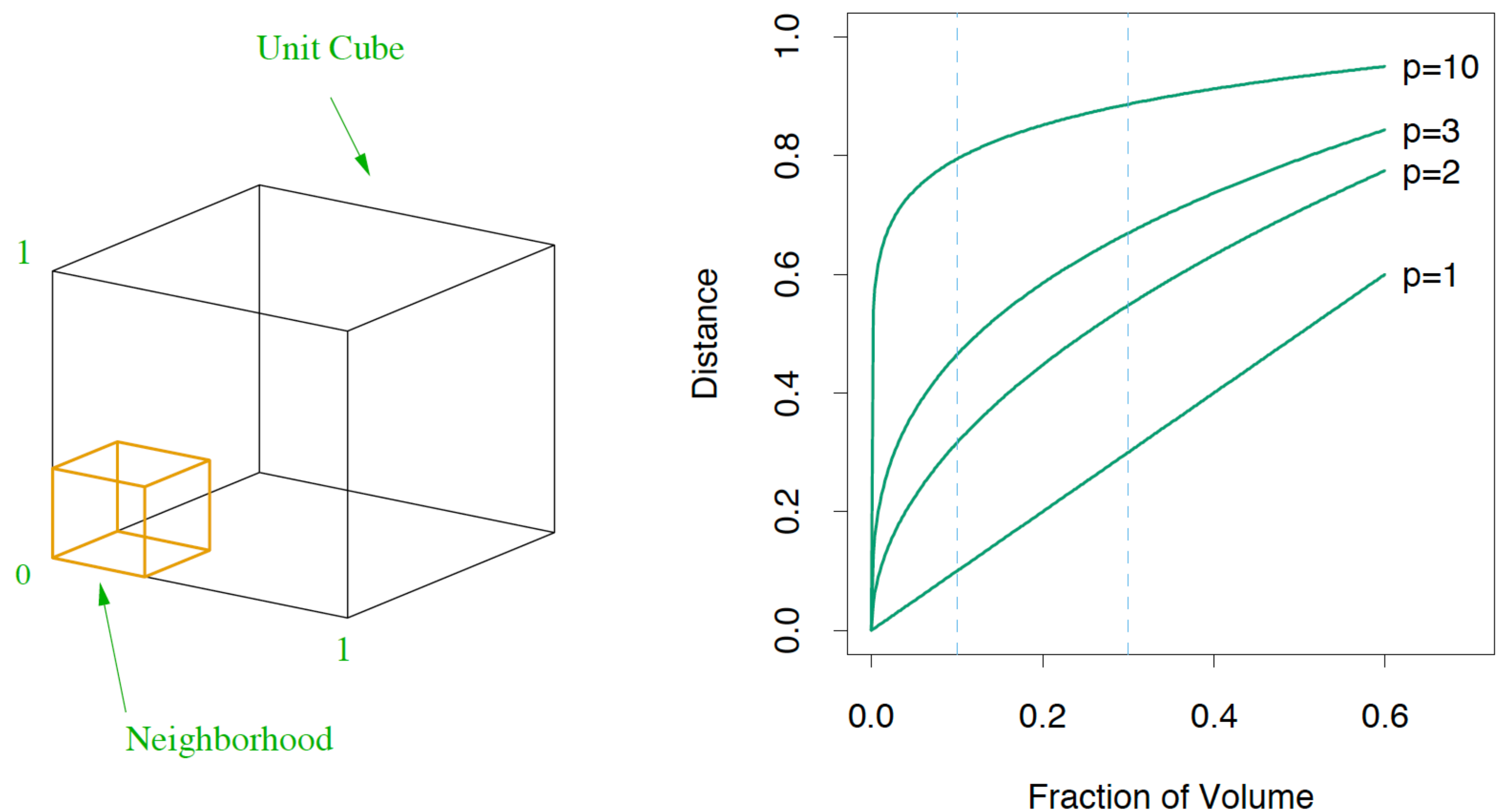


Fig 2.7, the curse of dimensionality

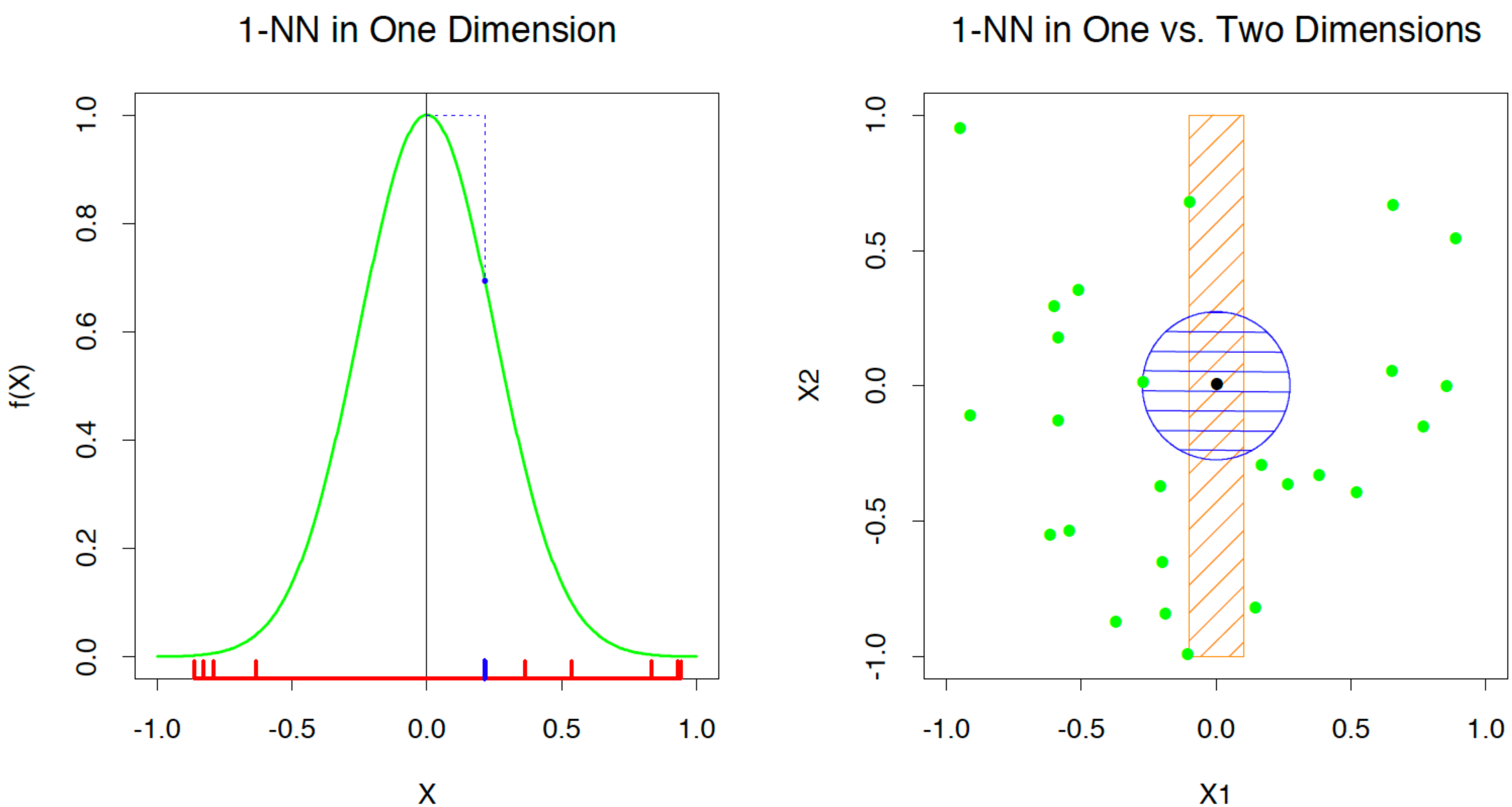


Fig 2.7, the curse of dimensionality

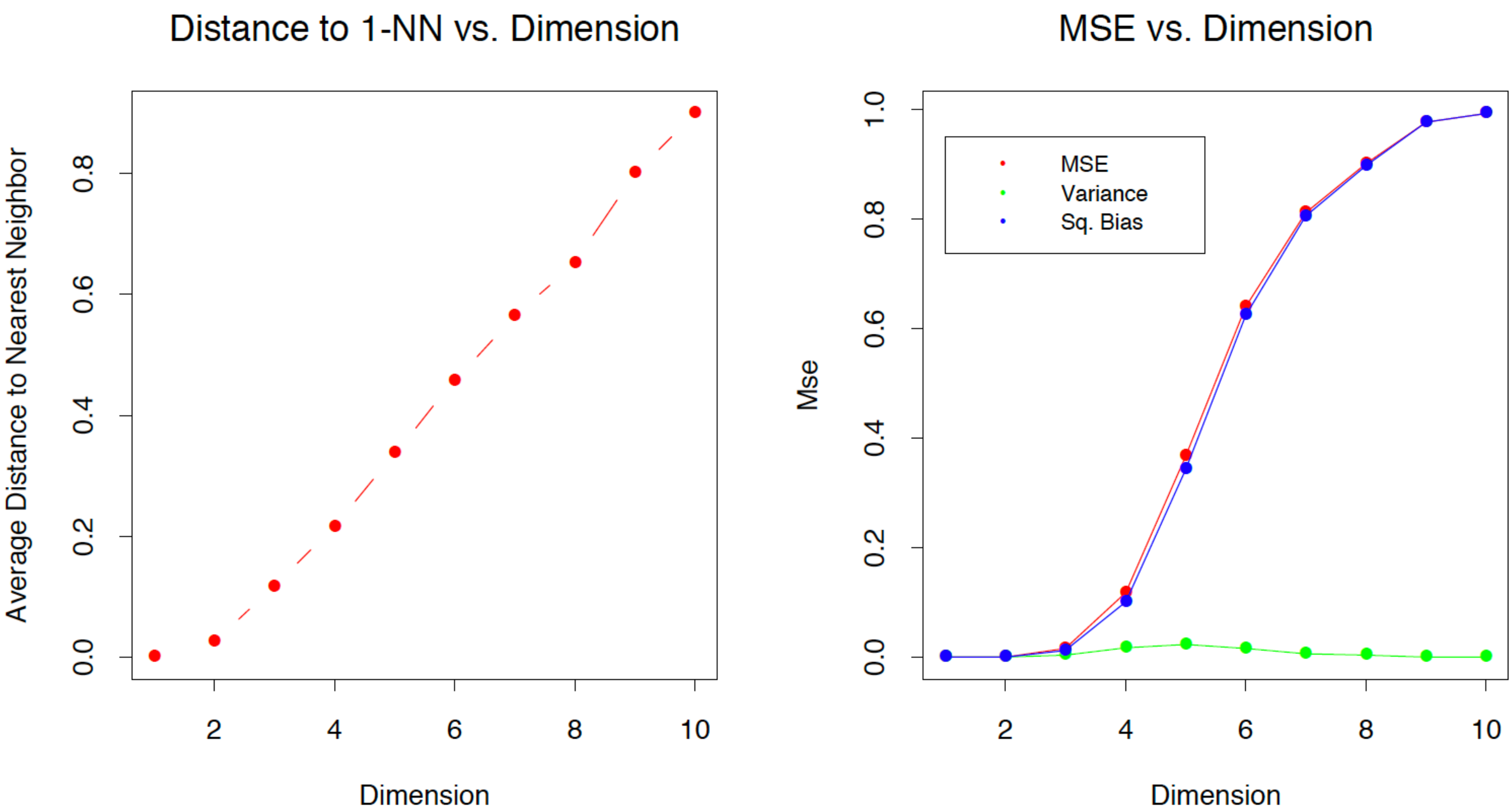


Fig 2.8, still the curse of dimensionality, but variance dominates

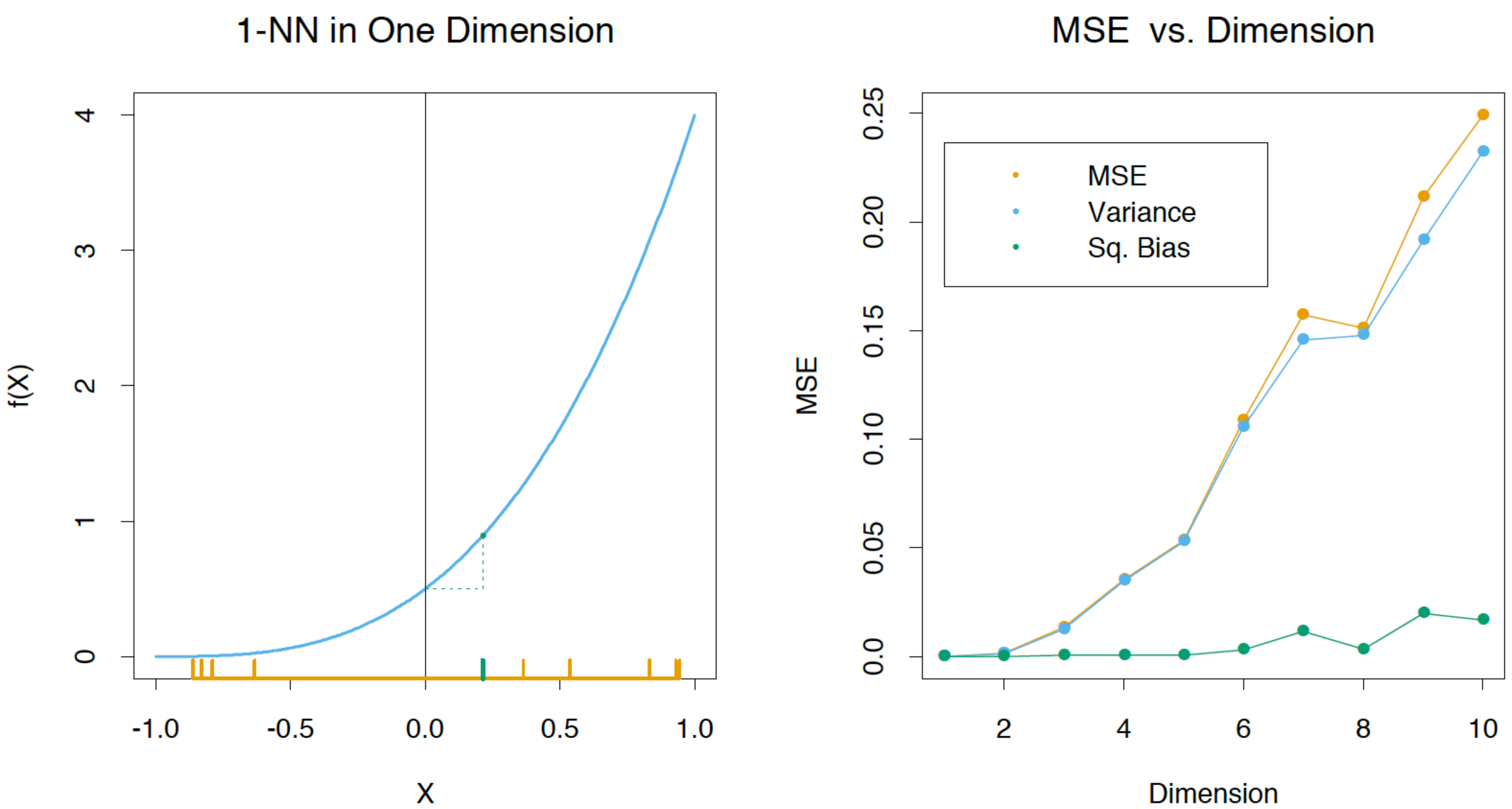


FIGURE 2.8. *A simulation example with the same setup as in Figure 2.7. Here the function is constant in all but one dimension:* $f(X) = \frac{1}{2}(X_1 + 1)^3$. *The variance dominates.*

Fig 2.9, ratio of EPE(1-NN) to EPE(OLS)
OLS is unbiased for the linear case
OLS is biased for the cubic case

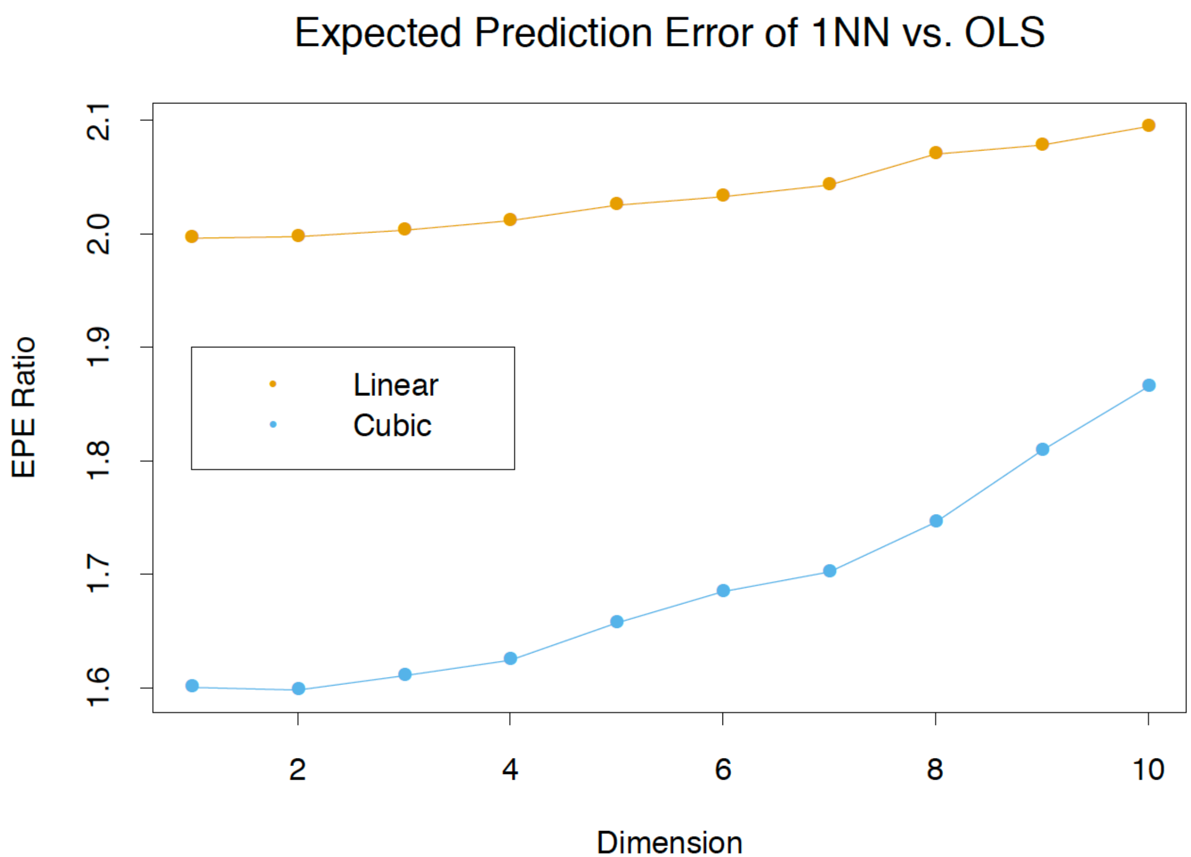


FIGURE 2.9. *The curves show the expected prediction error (at $x_0 = 0$) for 1-nearest neighbor relative to least squares for the model $Y = f(X) + \varepsilon$. For the orange curve, $f(x) = x_1$, while for the blue curve $f(x) = \frac{1}{2}(x_1 + 1)^3$.*

Bias-Variance Tradeoff

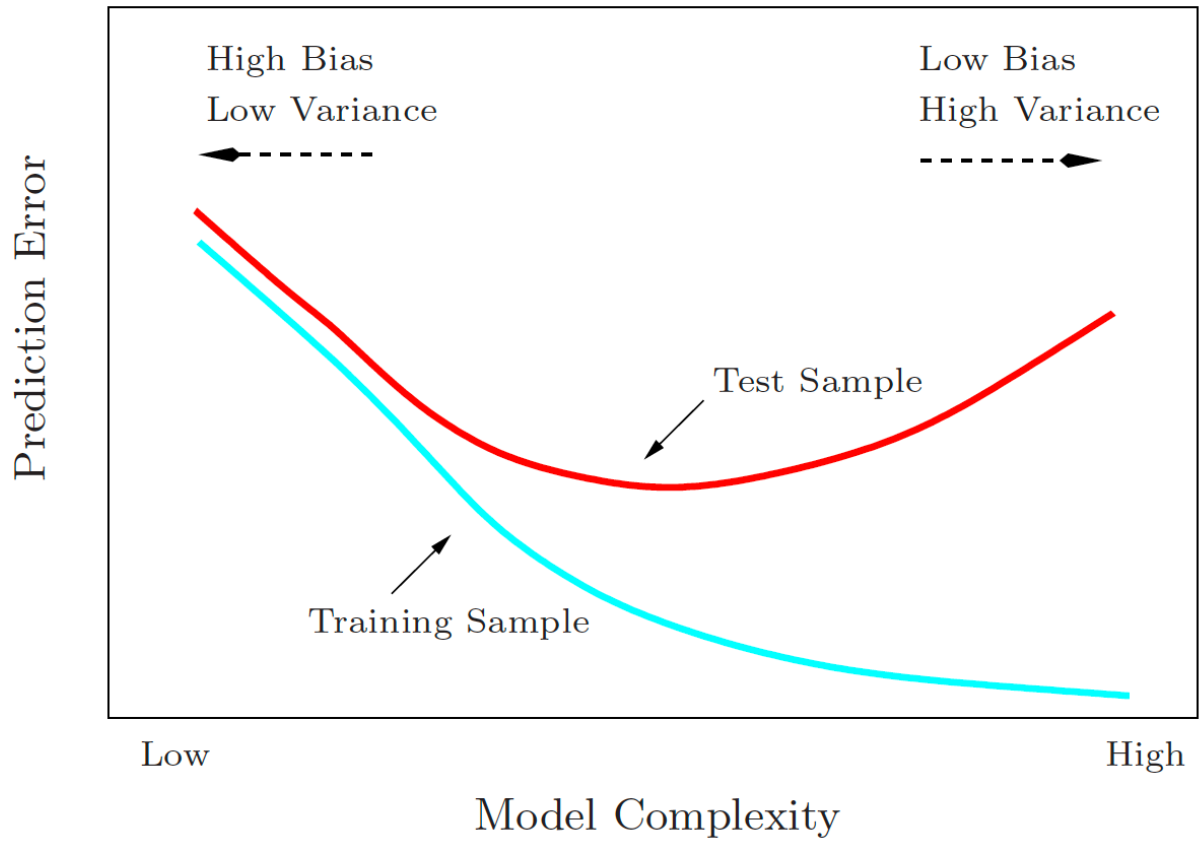
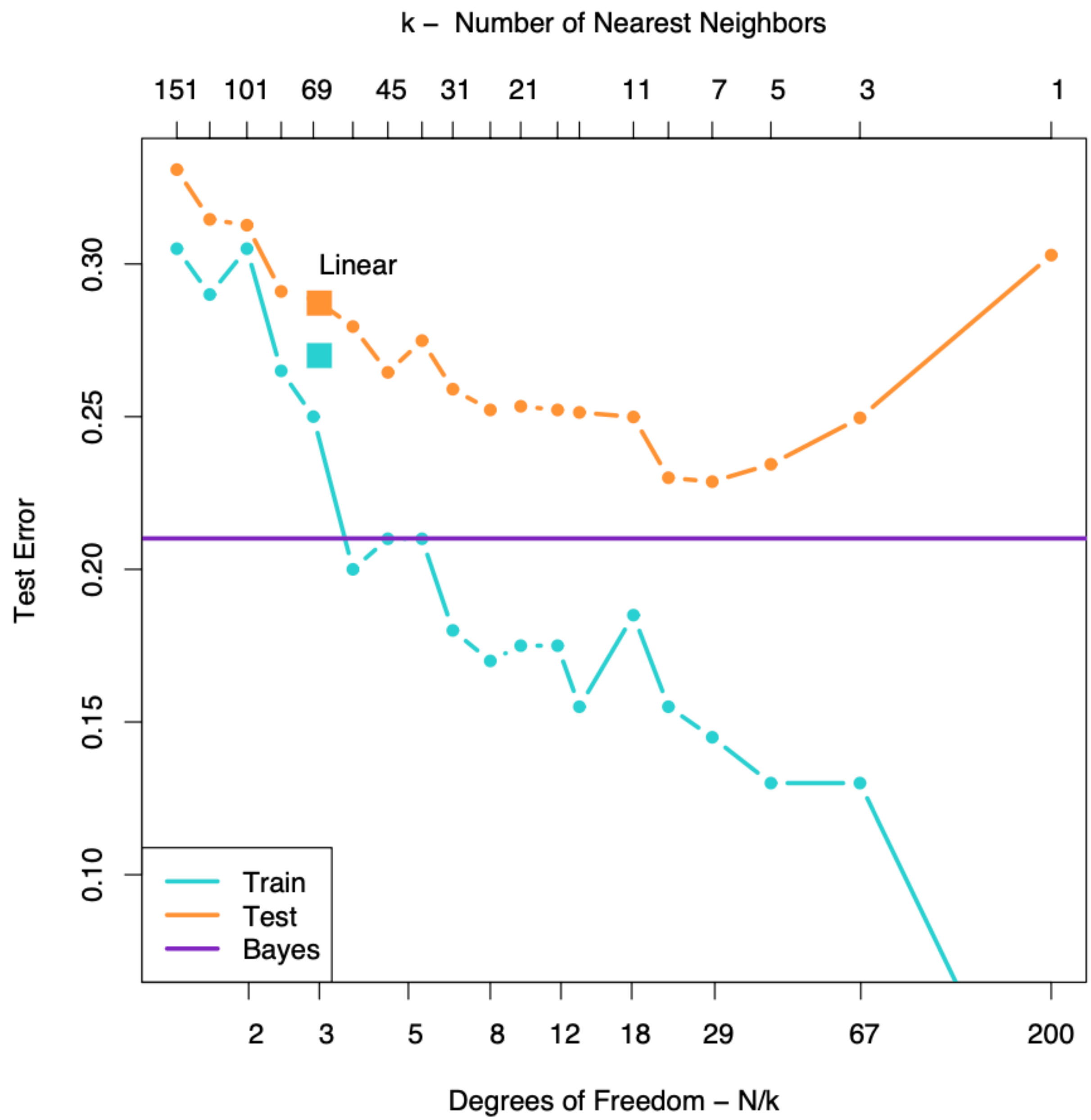


FIGURE 2.11. *Test and training error as a function of model complexity.*



k – Number of Nearest Neighbors
151
101
69
45
31
21
11
7
5
3
1
Linear
Test Error
0.30
0.25
0.20
0.15
0.10
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
Degrees of Freedom – N/k