Bias-Variance Tradeoffs

STAT 37710 / CMSC 35300

Rebecca Willett and Yuxin Chen

Parameter estimation

General problem statement:

We observe

$$z_i \overset{\text{iid}}{\sim} p_\theta, \ \theta \in \Theta$$

and the goal is to determine the θ that produced $\{z_i\}_{i=1}^n$.

Given a collection of observations $z_1, ..., z_n$ and a probability model

$$p(z_1, ..., z_n | \theta)$$

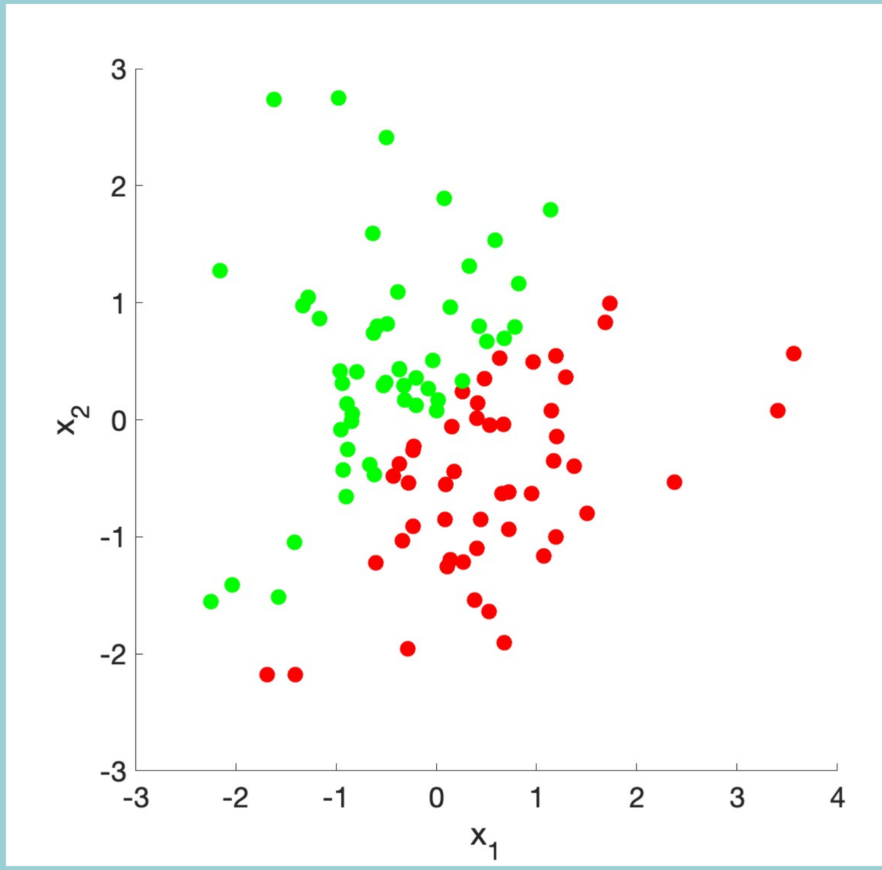
parameterized by the parameter θ, determine the value of θ that best matches the observations.
(We will explore different notions of "best" later.)

Example

$z_i = (x_i, y_i)$, where $x_i \in \mathbb{R}^p$ is a feature vector and $y_i \in \{0,1\}$ is a label

Assume $x_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_p)$ and

$y_i \mid x_i \sim \text{Bernoulli}(p_i)$ where $p_i = \dfrac{1}{1 + \exp(-x_i^\top \theta)}$ (logistic function)



Terminology in Estimation Theory

Consider some estimate $\widehat{\theta}$ of θ, and define

$$\epsilon(\widehat{\theta}) := \widehat{\theta} - \theta$$

Recall that $\widehat{\theta} = \widehat{\theta}(\{z_i\}_{i=1}^n)$ is a function of data $\implies \epsilon(\widehat{\theta})$ is a statistic!

Mean Squared Error:

$$\text{MSE}(\widehat{\theta}) := \mathbb{E}[\epsilon^\top \epsilon] = \mathbb{E}[\sum_{i=1}^{p} (\widehat{\theta}_i - \theta_i)^2]$$

Bias:

$$\text{Bias}(\widehat{\theta}) := \|\mathbb{E}[\widehat{\theta}] - \theta\|$$

Covariance:

$$\text{Cov}(\widehat{\theta}) := \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^\top]$$

Variance:

$$\text{Var}(\widehat{\theta}) := \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^\top (\widehat{\theta} - \mathbb{E}[\widehat{\theta}])] = \mathbb{E}[\|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\|_2^2] = \text{tr}(\text{Cov}(\widehat{\theta}))$$

Bias-variance decomposition

Key fact:

$$\text{MSE}[\widehat{\theta}] = \text{Bias}^2(\widehat{\theta}) + \text{Var}(\widehat{\theta})$$

Bias^2+variance

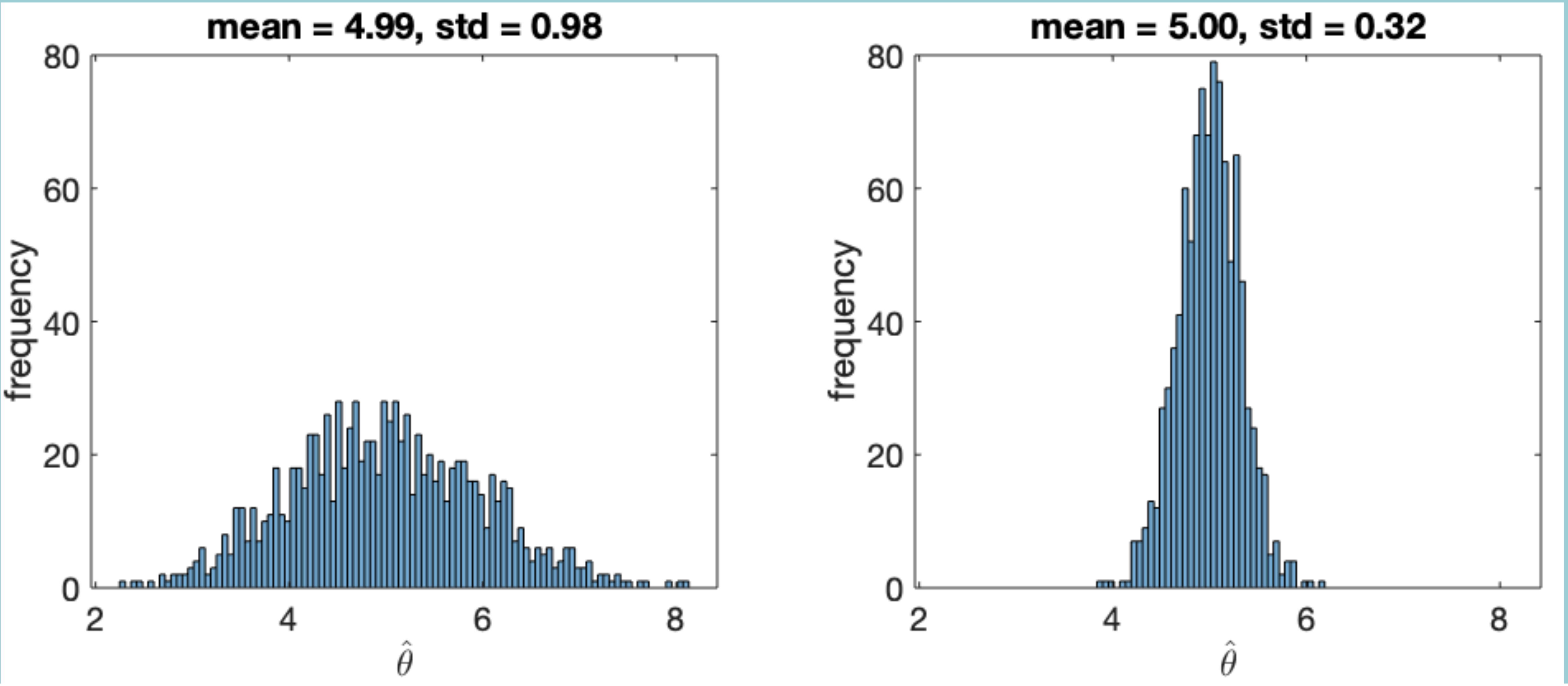
$$\begin{aligned}
&= \|\mathbb{E}[\widehat{\theta}] - \theta\|^2 + \mathbb{E}[\|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\|^2] \\
&= (\mathbb{E}[\widehat{\theta}] - \theta)^\top (\mathbb{E}[\widehat{\theta}] - \theta) + \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^\top (\widehat{\theta} - \mathbb{E}[\widehat{\theta}])] \\
&= \mathbb{E}[\widehat{\theta}]^\top \mathbb{E}[\widehat{\theta}] - 2\theta^\top \mathbb{E}[\widehat{\theta}] + \theta^\top \theta + \mathbb{E}[\widehat{\theta}^\top \widehat{\theta} - 2\widehat{\theta}^\top \mathbb{E}[\widehat{\theta}] + \mathbb{E}[\widehat{\theta}]^\top \mathbb{E}[\widehat{\theta}]] \\
&= \mathbb{E}[\widehat{\theta}]^\top \mathbb{E}[\widehat{\theta}] - 2\theta^\top \mathbb{E}[\widehat{\theta}] + \theta^\top \theta + \mathbb{E}[\widehat{\theta}^\top \widehat{\theta}] - \mathbb{E}[\widehat{\theta}]^\top \mathbb{E}[\widehat{\theta}] \\
&= -2\theta^\top \mathbb{E}[\widehat{\theta}] + \theta^\top \theta + \mathbb{E}[\widehat{\theta}^\top \widehat{\theta}] \\
&= \mathbb{E}[-2\theta^\top \widehat{\theta} + \theta^\top \theta + \widehat{\theta}^\top \widehat{\theta}] \\
&= \mathbb{E}[\|\theta - \widehat{\theta}\|^2] = \text{MSE}[\widehat{\theta}]
\end{aligned}$$

Example 1

$$z_i \sim \mathcal{N}(\theta, 1),\ i = 1, \ldots, n$$

$$\hat{\theta}_1 = z_1 \qquad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} z_i$$

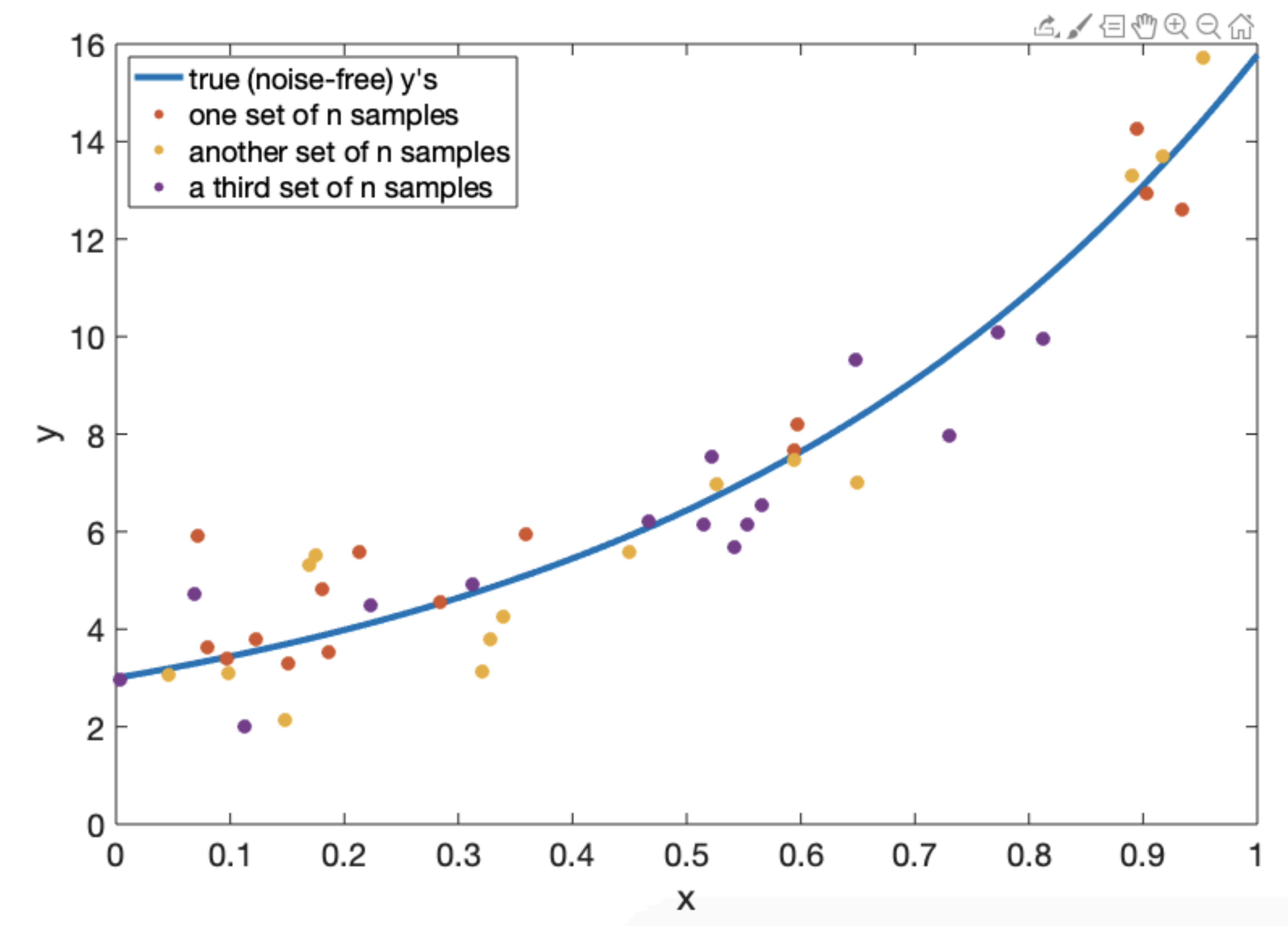
$$\mathbb{E}\hat{\theta}_1 = \theta,\ \mathrm{Var}(\hat{\theta}_1) = 1 \qquad \mathbb{E}\hat{\theta}_2 = \theta,\ \mathrm{Var}(\hat{\theta}_2) = \frac{1}{n}$$



Example 2

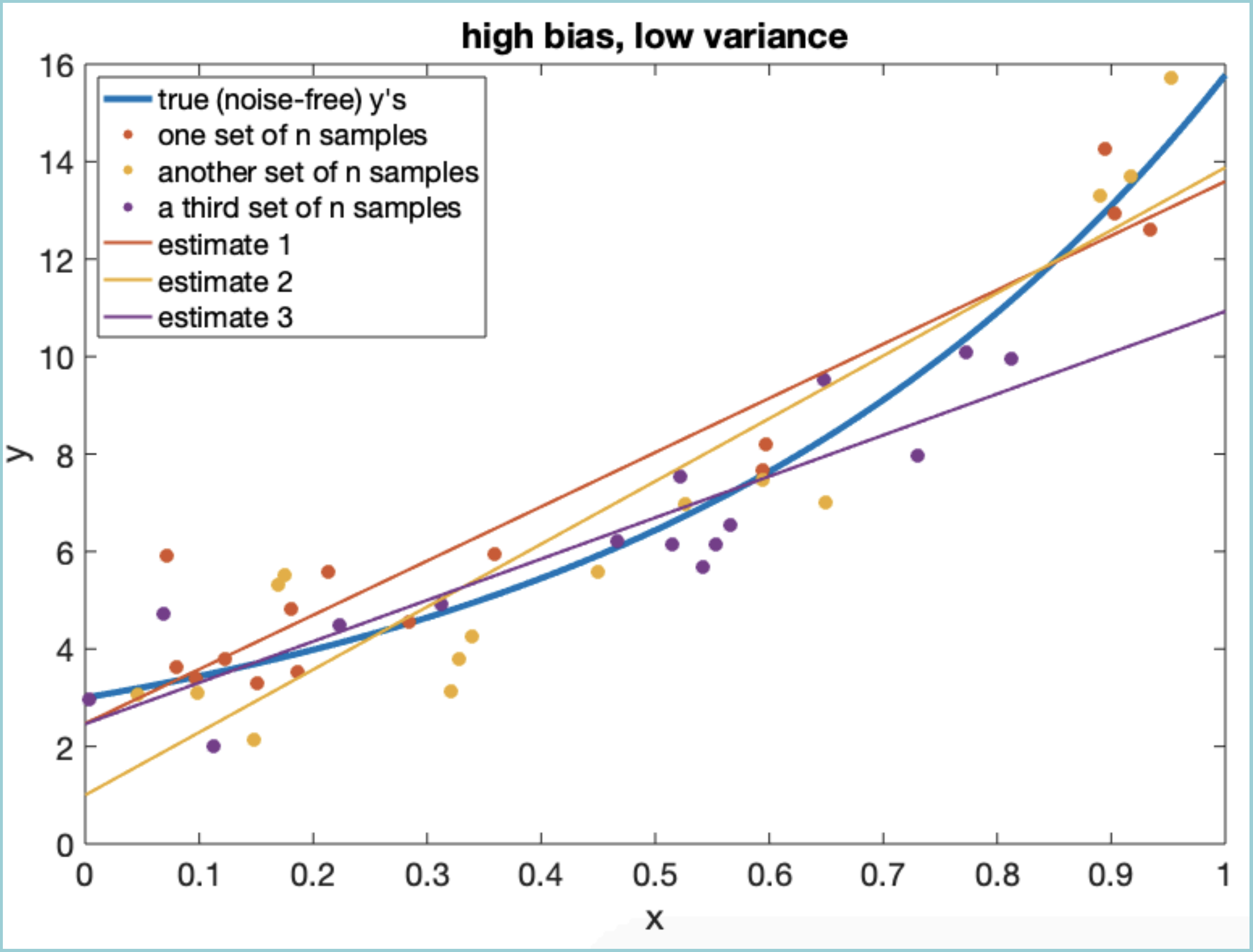
$$z_i = (x_i, y_i),\ i = 1,\ldots,n$$

$$y_i = \theta_1 + \theta_2 \exp\{\theta_3 x_i\} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$$



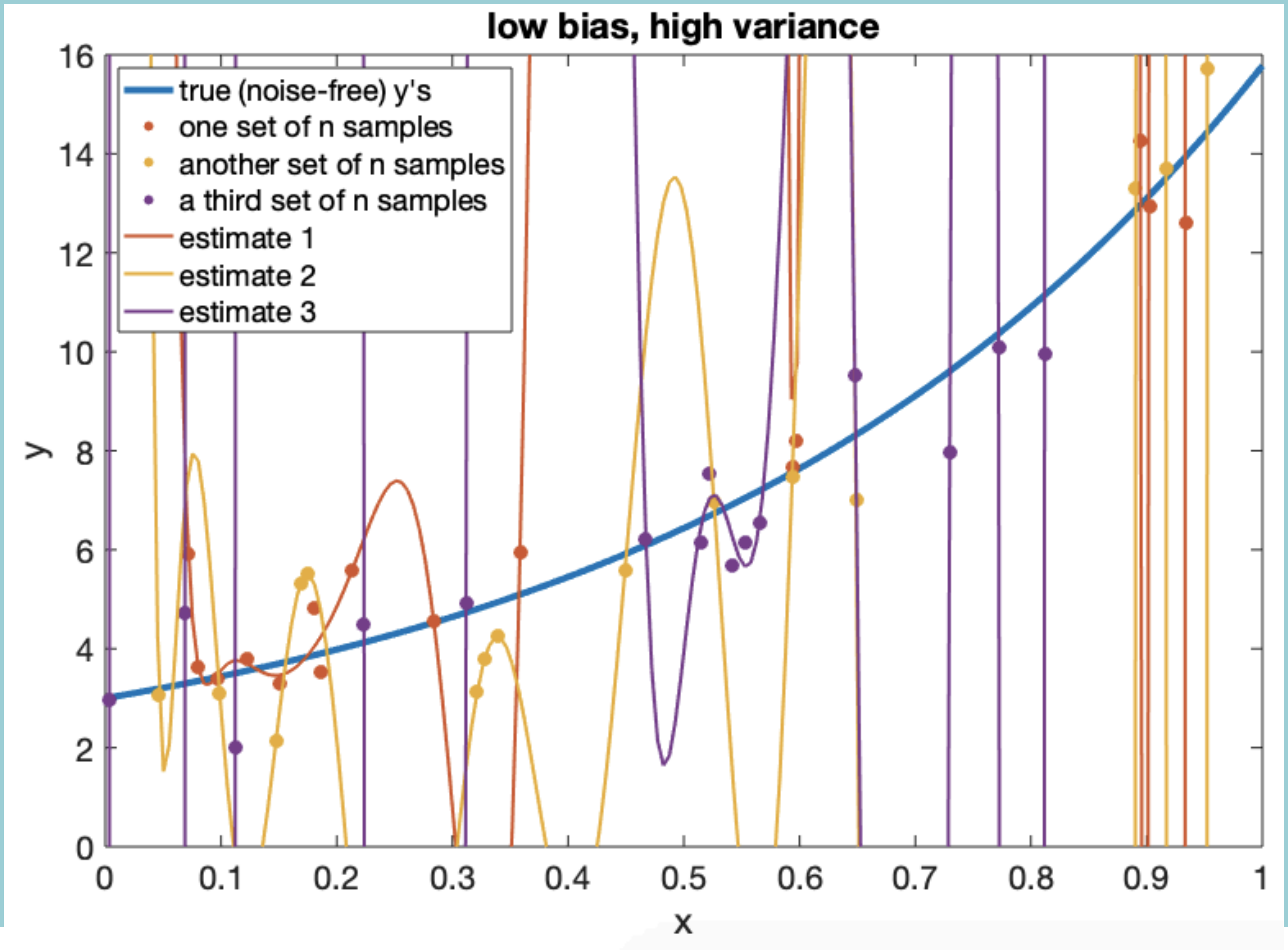
Example 2

Linear estimator



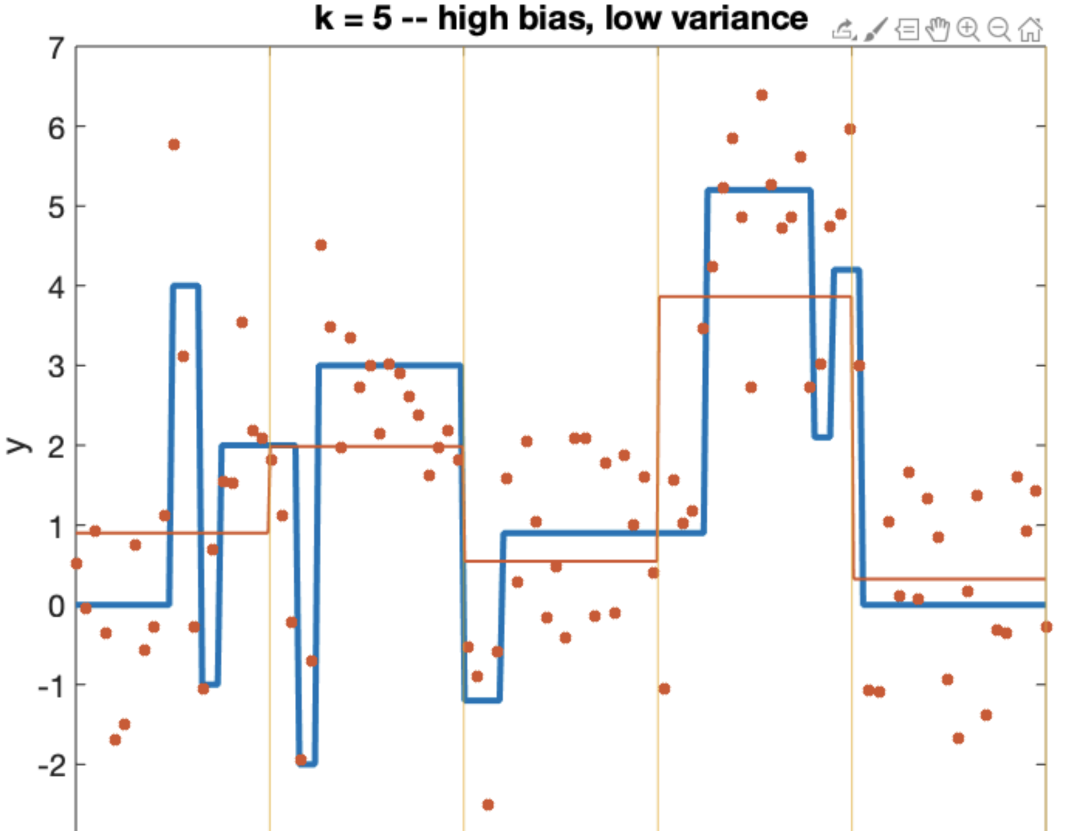
Example 2

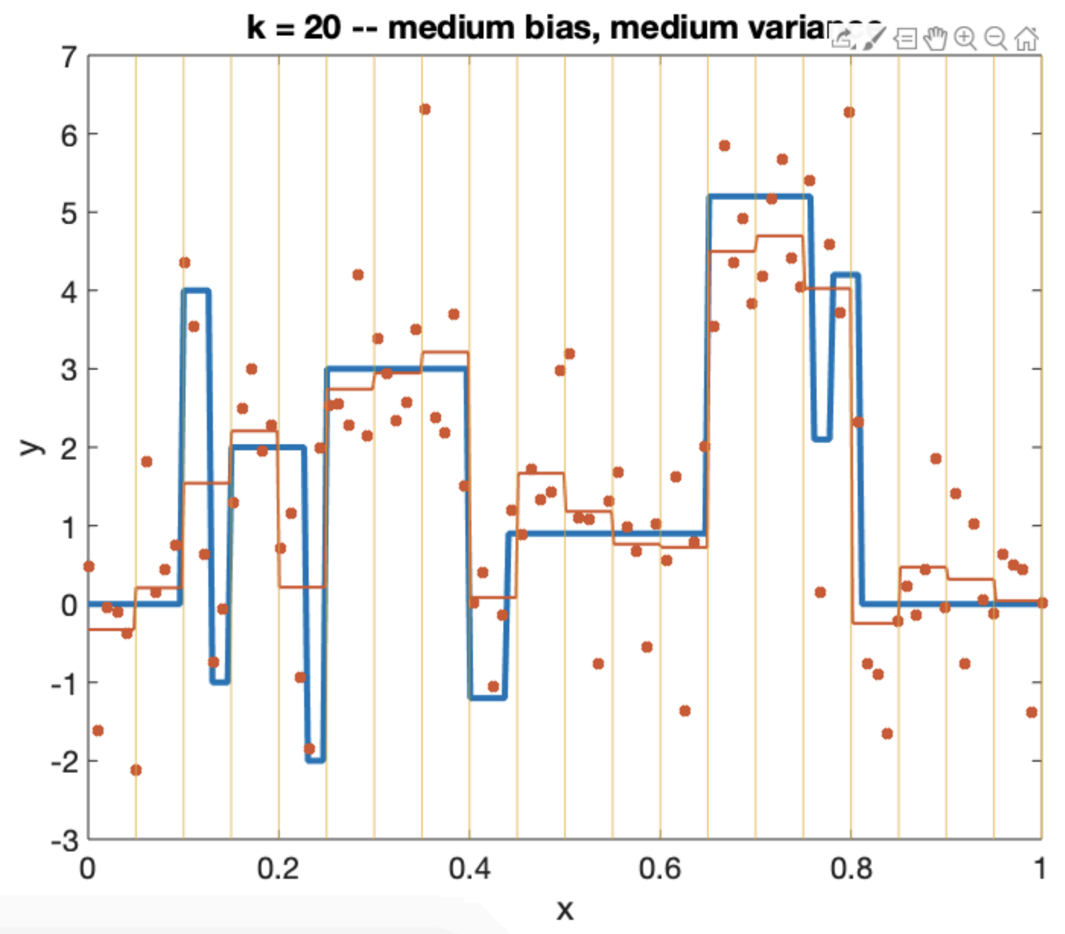
High-degree polynomial estimator

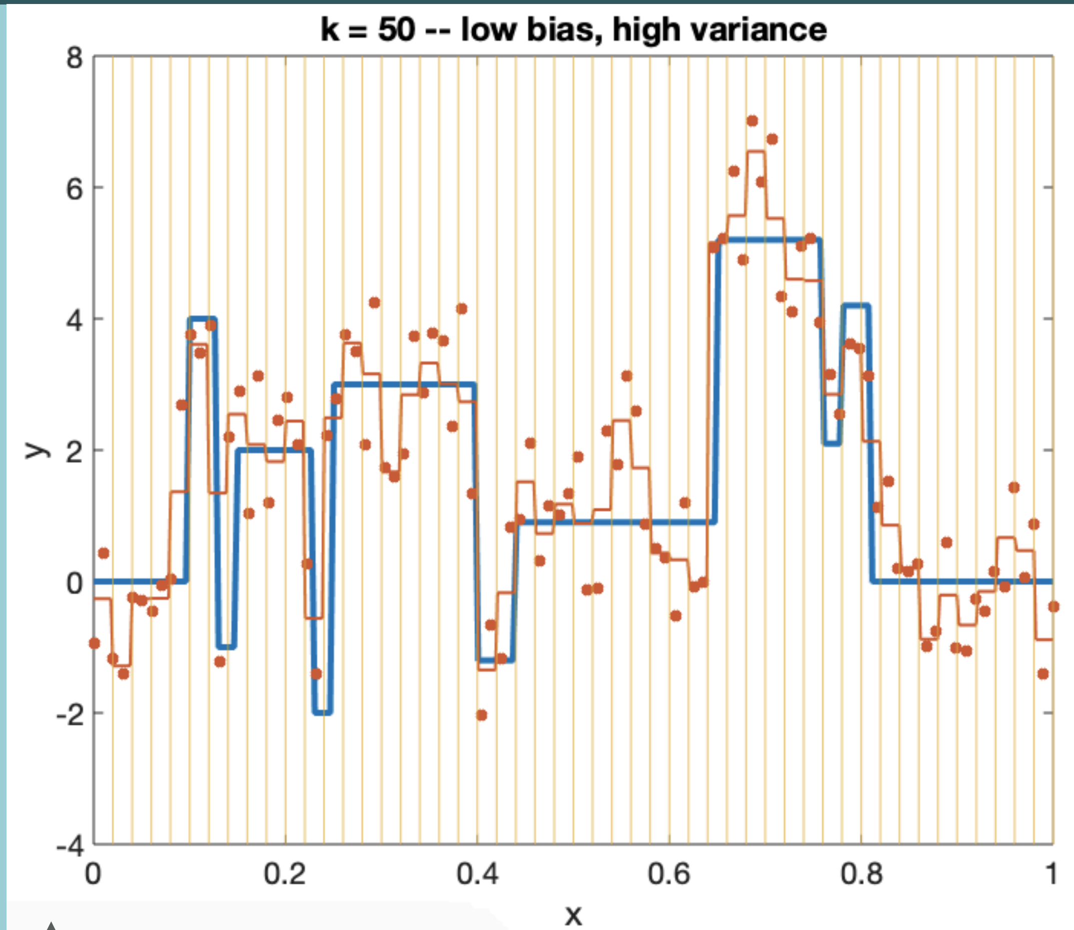


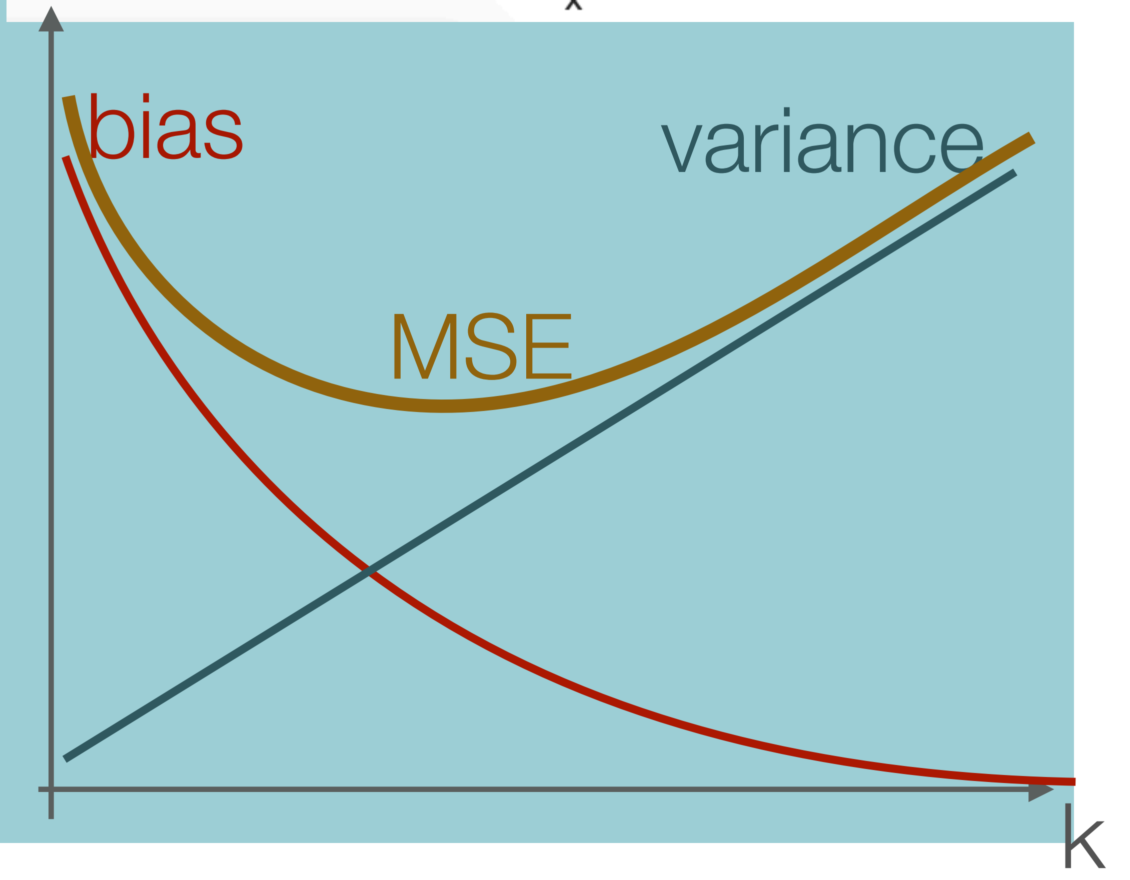
Example 3

Estimator: split horizontal axis into k bins; average samples within each bin. Big bins (small k) average across boundaries in ground truth signal (high bias). Small bins (big k) have few samples per bin (high variance)









Asymptotics

Estimators are often studied as a function of the number of observations:

$$\widehat{\theta}_n := \widehat{\theta}(z_1, \ldots, z_n)$$

$\widehat{\theta}$ is asymptotically unbiased if

$$\lim_{n\to\infty} \mathsf{Bias}(\widehat{\theta}_n) = 0$$

An estimator is consistent if

$$\lim_{n\to\infty} \mathsf{MSE}(\widehat{\theta}_n) = 0$$

A consistent estimator is *at least* asymptotically unbiased. Some estimators are unbiased, but inconsistent.

The latter basically means that our estimation does not improve as the number of data increase. Inconsistent estimators can provide reasonable estimates when we have a small number of data. However, consistent estimators are usually favored in practice.