CH. 4: Parametric Methods

Estimate the parameters of a known model from a given sample.

Two kinds of models are considered:

(1) Deterministic models, e.g., regression functions

Example:

Linear model: $g(x\,|\,\boldsymbol{\theta}) = w_1 x + w_0$

Parameters: $\boldsymbol{\theta} = (w_0, w_1)^T$

Quadratic model: $g(x\,|\,\boldsymbol{\theta}) = w_2 x^2 + w_1 x + w_0$

Parameters: $\boldsymbol{\theta} = (w_0, w_1, w_2)^T$

(2) Nondeterministic models, e.g., probability distribution functions

Example:

Bernoulli probability: $P(x \mid \theta) = p^x (1-p)^{1-x}$

Parameters: $\theta = p$

Gaussian distribution: $p(x \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Parameters: $\boldsymbol{\theta} = (\mu, \sigma)^T$

4.1 Probability Distribution Functions

- Two methods of parameter estimation:

  (a) Maximum likelihood approach

  (b) Bayesian approach

(A) Maximum Likelihood Estimation (MLE)

Let $X = \{\boldsymbol{x}^t\}_{t=1}^{N}$ be the sample drawn from some probability density family, $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Estimate $\boldsymbol{\theta}^*$ that makes X from $p(\boldsymbol{x} \mid \boldsymbol{\theta}^*)$ as likely as possible, i.e., maximizes the likelihood $p(X \mid \boldsymbol{\theta})$ of $\boldsymbol{\theta}$.

- The likelihood $p(X \mid \boldsymbol{\theta})$ of $\boldsymbol{\theta}$ given $X = \{\boldsymbol{x}^t\}_{t=1}^N$ is the probability of obtaining X, which is the product $\prod_{t=1}^N p(\boldsymbol{x}^t \mid \boldsymbol{\theta})$ of the probabilities of obtaining the individual examples, i.e., $p(X \mid \boldsymbol{\theta}) = \prod_{t=1}^N p(\boldsymbol{x}^t \mid \boldsymbol{\theta})$

  Log-likelihood: $L(\boldsymbol{\theta} \mid X) = \log p(X \mid \boldsymbol{\theta})$

  $$= \log \prod_{t=1}^N p(\boldsymbol{x}^t \mid \boldsymbol{\theta}) = \sum_{t=1}^N \log p(\boldsymbol{x}^t \mid \boldsymbol{\theta})$$

- Maximum likelihood estimator: $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid X)$

A.1 Bernoulli Distribution

The random variable X takes values $x \in \{0,1\}$.

Probability function: $P(x \mid \theta) = p^x (1-p)^{1-x}$

p is the only parameter to be estimated, i.e., $\theta = p$.

Given a sample $X = \{x^t\}_{t=1}^N$, the log-likelihood of θ:

$$L(\theta \mid X) = \sum_{t=1}^{N} \log P(x^t \mid \theta) = \sum_{t=1}^{N} \log p^{x^t} (1-p)^{1-x^t}$$

$$= \log p^{x^1} (1-p)^{1-x^1} + \cdots\cdots + \log p^{x^N} (1-p)^{1-x^N}$$

$$= \log \left(p^{x^1} (1-p)^{1-x^1} \, p^{x^2} (1-p)^{1-x^2} \cdots p^{x^N} (1-p)^{1-x^N} \right)$$

$$= \log \left(p^{\sum_t x^t} (1-p)^{N - \sum_t x^t} \right) = \log p^{\sum_t x^t} + \log (1-p)^{N - \sum_t x^t}$$

$$= \sum_t x^t \log p + (N - \sum_t x^t) \log (1-p)$$

The maximum likelihood estimate (MLE) of p:

Let $dL / dp = 0$

$$\frac{dL}{dp} = \frac{d}{dp}\left(\sum_t x^t \log p + \left(N - \sum_t x^t\right)\log(1-p)\right)$$

$$= \frac{1}{p}\sum_t x^t - \frac{1}{1-p}\left(N - \sum_t x^t\right) = \frac{(1-p)\sum_t x^t}{p(1-p)} - \frac{p\left(N - \sum_t x^t\right)}{p(1-p)}$$

$$= 0$$

$$\Leftrightarrow \ (1-p)\sum_t x^t - p\left(N - \sum_t x^t\right) = 0$$

$$\Leftrightarrow \ \sum x^t - p\sum x^t - pN + p\sum_t x^t = 0 \ \Leftrightarrow \ \sum_t x^t - pN = 0$$

$$\boxed{\hat{p} = \sum_{t=1}^{N} x^t / N}$$

A.2 Binomail Distribution

In binomial trials, N identical independent Bernoulli trials (0/1) are conducted.

Random variable X represents the number of 1s.

Probability function: $P(x \mid \theta) = \binom{N}{x} p^x (1-p)^{N-x}$

p is the only parameter to be estimated,

i.e., $\theta = p$.

Given a sample $X = \{x^t\}_{t=1}^{n}$,

derive the maximum likelihood estimate of θ.

The log-likelihood of θ:

$$L(\theta \mid X) = \sum_{t=1}^{n} \log P(x^t \mid \theta) = \sum_{t=1}^{n} \log \binom{N}{x^t} p^{x^t} (1-p)^{N-x^t}$$

$$= \log \binom{N}{x^1} p^{x^1} (1-p)^{N-x^1} + \log \binom{N}{x^2} p^{x^2} (1-p)^{N-x^2}$$

$$+ \cdots + \log \binom{N}{x^n} p^{x^n} (1-p)^{N-x^n}$$

$$= \log \underbrace{\binom{N}{x^1} \cdots \binom{N}{x^N}}_{c} \left(p^{x^1} (1-p)^{N-x^1} \cdots p^{x^n} (1-p)^{N-x^n} \right)$$

$$= \log c\left(p^{x^1}(1-p)^{N-x^1}\cdots p^{x^n}(1-p)^{N-x^n}\right)$$

$$\text{(where } c = \binom{N}{x^1}\cdots\binom{N}{x^n}\text{)}$$

$$= \log c\left(p^{\sum_t x^t}(1-p)^{Nn-\sum_t x^t}\right)$$

$$= \log c + \log p^{\sum_t x^t} + \log(1-p)^{Nn-\sum_t x^t}$$

$$= \log c + \sum_t x^t \log p + (Nn - \sum_t x^t)\log(1-p)$$

The maximum likelihood estimate (MLE) of p:

Let $dL/dp = 0$

$$\frac{dL}{dp} = \frac{d}{dp}\left(\log c + \sum_t x^t \log p + (Nn - \sum_t x^t)\log(1-p)\right)$$

$$= \frac{\sum_t x^t}{p} - \frac{(Nn - \sum_t x^t)}{1-p} = \frac{(1-p)\sum_t x^t}{p(1-p)} - \frac{p(Nn - \sum_t x^t)}{p(1-p)} = 0$$

$$\Leftrightarrow \ (1-p)\sum_t x^t - p(Nn - \sum_t x^t) = 0$$

$$\Leftrightarrow \ \sum_t x^t - p\sum_t x^t - pNn + p\sum_t x^t = 0 \ \Leftrightarrow \ \sum_t x^t - pNn = 0$$

$$\hat{p} = \sum_t x^t / Nn$$

$$\boxed{\hat{p} = \sum_{t=1}^{n} x^t / Nn}$$

A.3 Gaussian (Normal) Distribution

Density function: $p(x\,|\,\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

Parameters to be estimated: $\boldsymbol{\theta} = (\mu, \sigma)^T$

Given a sample: $X = \{x^t\}_{t=1}^N$

Log-likelihood: $L(\boldsymbol{\theta}\,|\,X) = \log \prod_{t=1}^{N} p\left(x^t\,|\,\boldsymbol{\theta}\right)$

$$= \log\left(\prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x^t-\mu)^2}{2\sigma^2}\right]\right)$$

$$= \log\left(\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N \prod_{t=1}^{N}\exp\left[-\frac{(x^t-\mu)^2}{2\sigma^2}\right]\right)$$

$$= \log\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N + \log\left(\prod_{t=1}^{N} \exp\left[-\frac{\left(x^t-\mu\right)^2}{2\sigma^2}\right]\right)$$

$$= N\log\left[\frac{1}{\sqrt{2\pi}\sigma}\right] + \sum_{t=1}^{N} \log\exp\left[-\frac{\left(x^t-\mu\right)^2}{2\sigma^2}\right]$$

$$= -N\log\left[\sqrt{2\pi}\sigma\right] + \sum_{t=1}^{N}\left[-\frac{\left(x^t-\mu\right)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}$$

The maximum likelihood estimate (MLE) of (μ,σ):

Let $\partial L/\partial\mu = 0,\ \partial L/\partial\sigma = 0$

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{-1}{2\sigma^2}\frac{\partial}{\partial\mu}\sum_{t=1}^{N}(x^t-\mu)^2 = \frac{-1}{2\sigma^2}\sum_{t=1}^{N}\frac{\partial}{\partial\mu}(x^t-\mu)^2$$

$$= \frac{-1}{2\sigma^2}\sum_{t=1}^{N}\left(-2(x^t-\mu)\right) = \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu) = 0$$

$$\Rightarrow\ \sum_{t=1}^{N}(x^t-\mu) = 0,\ \sum_{t=1}^{N}x^t - N\mu = 0,\ \boxed{\hat{\mu} = \frac{1}{N}\sum_t x^t}$$

$$\frac{\partial L}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(-\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2} \right)$$

$$= \frac{\partial}{\partial \sigma}\left(-N\log\sigma - \frac{\sum_{t=1}^{N}(x^t-\mu)^2}{2\sigma^2} \right)$$

$$= \frac{-N}{\sigma} - \frac{1}{2}\sum_{t=1}^{N}(x^t-\mu)^2 \frac{\partial}{\partial \sigma}\left(\frac{1}{\sigma^2} \right)$$

$$= \frac{-N}{\sigma} + \frac{1}{\sigma^3}\sum_{t=1}^{N}(x^t-\mu)^2 = 0$$

$$\Rightarrow \; -N + \frac{1}{\sigma^2}\sum_{t=1}^{N}(x^t-\mu)^2 = 0, \quad \boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N}(x^t-\mu)^2}$$

(B) Bayes' Estimation

-- Treat θ as a random variable with prior density $p(\theta)$.

Use Bayes' rule to combine the likelihood $p(X \mid \theta)$

to obtain the posterior $p(\theta \mid X) = p(X \mid \theta)\, p(\theta) / p(X)$.

Bayes' estimator: $\theta_{Bayes} = E[\theta \mid X] = \int \theta\, p(\theta \mid X)\, d\theta$

i.e., the posterior expected value of θ.

ML estimator: $\theta_{ML} = \arg\max_{\theta} p(X \mid \theta)$

MAP estimator: $\theta_{MAP} = \arg\max_{\theta} p(\theta \mid X)$

Example: Given a sample $X = \{x^t\}_{t=1}^N$,

Suppose $x^t \sim N(\theta, \sigma^2)$, $p(x^t) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x^t - \theta)^2}{2\sigma^2}\right]$

$\theta = \mu$ unknown, σ known

Likelihood of θ: $p(X|\theta) = \prod_{t=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x^t - \theta)^2}{2\sigma^2}\right]$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{\sum_t (x^t - \theta)^2}{2\sigma^2}\right]$$

Assume prior: $\theta \sim N(\mu_0, \sigma_0^2)$, μ_0, σ_0: known

$$p(\theta) = \frac{1}{(2\pi)^{1/2}\sigma_0} \exp\left[-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right]$$

Posterior:

$$p(\theta | X) = p(X|\theta)p(\theta)/p(X), \quad p(X) = \sum_\theta p(X|\theta)p(\theta)$$

$$p(\theta | X) = \frac{1}{(2\pi)^{(N+1)/2}\sigma^N\sigma_0}\exp\left[-\left(\frac{\sum_t (x^t-\theta)^2}{2\sigma^2}+\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right)\right] / p(X)$$

$$\theta_{Bay} = E[\theta | X] = \int \theta p(\theta | X) d\theta = \frac{1}{(2\pi)^{(N+1)/2}\sigma^N\sigma_0} \cdot$$

$$\int \theta \exp\left[-\left(\frac{\sum_t (x^t-\theta)^2}{2\sigma^2}+\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right)\right] / p(X) d\theta$$

$$= \frac{N/\sigma^2}{N/\sigma^2+1/\sigma_0^2} m + \frac{1/\sigma_0^2}{N/\sigma^2+1/\sigma_0^2}\mu_0$$

N large $\rightarrow$ θ_{Bay} close to $m = \sum_{t=1}^{N} x^t$;

σ_0^2 small or N small $\rightarrow$ θ_{Bay} close to μ_0.

4.2 Regression Functions

-- Given a sample $(x^t, r^t)_{t=1}^N$, determine the parameters $\boldsymbol{\theta}$ of the function f for $r = f(x)$.

Example:

Linear function $f\left(x|\boldsymbol{\theta} = (w_0, w_1)^T\right) = w_1 x + w_0$

$$E(\boldsymbol{\theta} \mid X) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - f\left(x^t \mid \boldsymbol{\theta}\right)\right]^2 = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - (w_1 x^t + w_0)\right]^2$$

$$\frac{\partial}{\partial w_0} E(w_0, w_1 \mid X) = 0, \quad \sum_t r^t = N w_0 + w_1 \sum_t x^t,$$

$$\frac{\partial}{\partial w_1} E(w_0, w_1 \mid X) = 0, \quad \sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

In vector-matrix form $A\boldsymbol{w} = \boldsymbol{r}$, where

$$A = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}.$$

The solution: $\boldsymbol{w} = A^{-1}\boldsymbol{r}$

Polynomial function

$$f(x|\boldsymbol{\theta} = (w_k, \cdot\cdot, w_0)) = w_k x^k + \cdot\cdot + w_1 x + w_0$$

$$E(\boldsymbol{\theta} \mid X) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - f\left(x^t \mid \boldsymbol{\theta}\right)\right]^2$$

$$= \frac{1}{2}\sum_{t=1}^{N}\left[r^t - (w_k (x^t)^k + \cdot\cdot + w_1 (x^t) + w_0)\right]^2$$

$$\frac{\partial E(\boldsymbol{\theta} \mid X)}{\partial w_0} = 0, \quad \cdots\cdots\cdots\cdots\cdots\cdots, \quad \frac{\partial E(\boldsymbol{\theta} \mid X)}{\partial w_k} = 0$$

$$\sum_t r^t = w_0 \sum_t (x^t)^0 + \cdots\cdots\cdots + w_k \sum_t (x^t)^k,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots,$$

$$\sum_t r^t (x^t)^k = w_0 \sum_t (x^t)^k + \cdots + w_k \sum_t (x^t)^{2k}$$

In vector-matrix form $A\boldsymbol{w} = \boldsymbol{r}$, where

$$A = \begin{bmatrix} \sum_t (x^t)^0 & \sum_t (x^t)^1 & \cdots & \sum_t (x^t)^k \\ \sum_t (x^t)^1 & \sum_t (x^t)^2 & \cdots & \sum_t (x^t)^{k+1} \\ \cdot & \cdot & \cdot & \cdot \\ \sum_t (x^t)^k & \sum_t (x^t)^{k+1} & \cdots & \sum_t (x^t)^{2k} \end{bmatrix}$$

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} \sum_t r^t (x^t)^0 \\ \sum_t r^t (x^t)^1 \\ \vdots \\ \sum_t r^t (x^t)^k \end{bmatrix}$$

Find $\boldsymbol{w}$ that minimizes $e = \|A\boldsymbol{w} - \boldsymbol{r}\|^2$.

Let $\dfrac{de}{d\boldsymbol{w}} = \dfrac{d}{d\boldsymbol{w}} \|A\boldsymbol{w} - \boldsymbol{r}\|^2 = 2A^T (A\boldsymbol{w} - \boldsymbol{r}) = \boldsymbol{0}$

$A^T (A\boldsymbol{w} - \boldsymbol{r}) = \boldsymbol{0}, \;\; A^T A\boldsymbol{w} - A^T \boldsymbol{r} = \boldsymbol{0}, \;\; A^T A\boldsymbol{w} = A^T \boldsymbol{r}$

$\boldsymbol{w} = \left(A^T A\right)^{-1} A^T \boldsymbol{r} = A^+ \boldsymbol{r}$

$A^+ = \left(A^T A\right)^{-1} A^T$: the pseudoinverse of A.

Relate parameter estimations of deterministic and nondeterministic models

Consider equation $r = f(x)$. Introduce uncertainty ε into the equation $r = f(x) + \varepsilon$ and view ε and r as random variables.

Assume $\varepsilon \sim p(e)$, from the probability theorem,

$$r \sim p(r) = p(e)\left|de/dr\right|$$

$$r = f(x) + e, \quad e = r - f(x), \quad \left|\frac{de}{dr}\right| = 1$$

$$p(r) = p(e)\left|de/dr\right| = p(e)$$

Example: Given $r = f(x) + \varepsilon$, let $\varepsilon \sim p(e) = N(0, \sigma^2)$,

i.e., $p(e) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{e^2}{2\sigma^2}\right]$.

From theorem, $r \sim p(r) = p(e)\left|de/dr\right|$.

$$p(r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{e^2}{2\sigma^2}\right]\left|\frac{de}{dr}\right| \qquad \begin{pmatrix} r = f(x) + e \\ e = r - f(x) \\ \left|de/dr\right| = 1 \end{pmatrix}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r - \hat{f}(x|\boldsymbol{\theta})\right]^2}{2\sigma^2}\right] \cdot 1$$

where $\hat{f}(x|\boldsymbol{\theta})$: the estimator of $f(x)$ up to $\boldsymbol{\theta}$.

The log-likelihood of $\boldsymbol{\theta}$: $L(\boldsymbol{\theta} \mid X) = \log \prod_{t=1}^{N} p\left(r^t \mid x^t\right)$

$$L(\boldsymbol{\theta} \mid X) = \sum_{t=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - \hat{f}\left(x^t|\boldsymbol{\theta}\right)\right]^2}{2\sigma^2}\right]$$

$$= \sum_{t=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{t=1}^{N} \log \exp\left[-\frac{\left[r^t - \hat{f}\left(x^t|\boldsymbol{\theta}\right)\right]}{2\sigma^2}\right]^2$$

$$= -N \log\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - \hat{f}\left(x^t|\boldsymbol{\theta}\right)\right]^2$$

$$= -N \log\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} E\left(\boldsymbol{\theta} \mid X\right)$$

$$L(\boldsymbol{\theta} \mid X) = -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} E(\boldsymbol{\theta} \mid X)$$

$\therefore$ Maximizing $L(\boldsymbol{\theta} \mid X)$ $\Leftrightarrow$ Minimizing $E(\boldsymbol{\theta} \mid X)$

(Probabilistic functions) (Regression functions)

4.3 Model Complexity: Bias and Variance

Given M samples $X_i = \{x_i^t, r_i^t\}_{i=1}^{M}$, $t = 1, \cdots, N$ to fit $f(x)$. Let $\hat{f}_i(x)$, $i = 1, \cdots, M$, be the estimates.

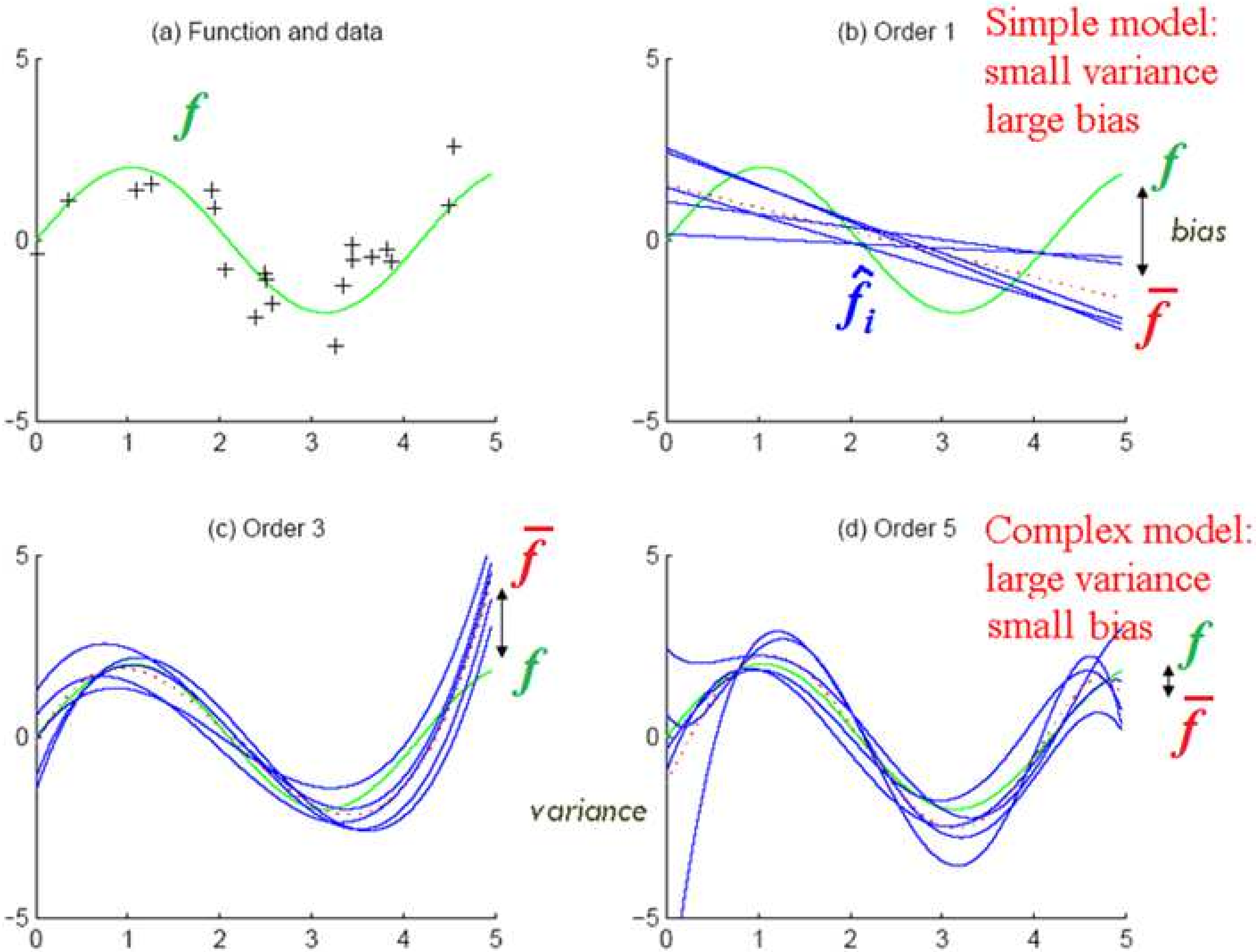
Let $\overline{f}(x) = \frac{1}{M}\sum_i \hat{f}_i(x)$

Bias: $\text{Bias}^2(f) = \frac{1}{N}\sum_t \left[f(x^t) - \overline{f}(x^t)\right]^2$

Variance: $\text{Variance}(f) = \frac{1}{NM}\sum_t \sum_i \left[\hat{f}_i(x^t) - \overline{f}(x^t)\right]^2$

Bias/Variance Dilemma -- Increase model complexity,

⇨ bias decreases (a better fit to data) and variance increases (fit varies more with data)


(a) Function and data
f
(b) Order 1
Simple model:
small variance
large bias
f
bias
$\hat{f}_i$
$\bar{f}$
(c) Order 3
$\bar{f}$
f
variance
(d) Order 5
Complex model:
large variance
small bias
f
$\bar{f}$

Appendix: Derivatives

- $\dfrac{df(x)}{dx}$, e.g., $f(x) = 3x^2 - 4x + 15,\ \dfrac{df(x)}{dx} = 6x - 4$
- $\dfrac{df(\boldsymbol{x})}{d\boldsymbol{x}} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left(\dfrac{\partial f(\boldsymbol{x})}{\partial x_1}\ \dfrac{\partial f(\boldsymbol{x})}{\partial x_2} \cdots \dfrac{\partial f(\boldsymbol{x})}{\partial x_d} \right)^T$,

  where $\boldsymbol{x} = (x_1\ x_2 \cdots x_d)$

  e.g., $f(\boldsymbol{x}) = f(x, y) = 3x - 4xy + 15y - 6$,

$$\frac{df(\boldsymbol{x})}{d\boldsymbol{x}} = \left(\frac{\partial f(x, y)}{\partial x}\ \frac{\partial f(x, y)}{\partial y} \right)^T = \begin{pmatrix} 3 - 4y \\ -4x + 15 \end{pmatrix}$$

- $$\frac{d\boldsymbol{f}(x)}{dx}=\frac{d(f_1(x)\ \cdots\ f_n(x))^T}{dx}=\left(\frac{df_1(x)}{dx}\ \cdots\ \frac{df_n(x)}{dx}\right)^T$$

e.g., $\boldsymbol{f}(x)=\left(f_1(x)\ f_2(x)\right)^T=\left((3x-4)\ (x^2+5x-6)\right)^T$,

$$\frac{d\boldsymbol{f}(x)}{dx}=\frac{d\left(f_1(x)\ f_2(x)\right)^T}{dx}=\frac{d\left((3x-4)\ (x^2+5x-6)\right)^T}{dx}$$

$$=\left(\frac{d(3x-4)}{dx}\quad\frac{d(x^2+5x-6)}{dx}\right)^T$$

$$=\left(3\ \ 2x+5\right)^T$$

- $$\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \frac{d(f_1(\boldsymbol{x}) \ \cdots f_n(\boldsymbol{x}))^T}{d\boldsymbol{x}} = \left(\frac{df_1(\boldsymbol{x})}{d\boldsymbol{x}} \ \cdots \ \frac{df_n(\boldsymbol{x})}{d\boldsymbol{x}} \right)^T$$

$$= \left(\nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}) \ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}) \ \cdots\cdot \ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}) \right)^T$$

$$= \begin{pmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdot\cdot & \frac{\partial f_1(\boldsymbol{x})}{\partial x_d} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdot\cdot & \frac{\partial f_2(\boldsymbol{x})}{\partial x_d} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \frac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_n(\boldsymbol{x})}{\partial x_2} & \cdot\cdot & \frac{\partial f_n(\boldsymbol{x})}{\partial x_d} \end{pmatrix} = J(\boldsymbol{f})$$

Jacobian matrix of $\boldsymbol{f}$

- $$\frac{dF(\boldsymbol{x})}{d\boldsymbol{x}} = \frac{d}{d\boldsymbol{x}}\begin{pmatrix} \boldsymbol{f}_{11} & \boldsymbol{f}_{12} & \cdot\cdot & \boldsymbol{f}_{1m} \\ \boldsymbol{f}_{21} & \boldsymbol{f}_{22} & \cdot\cdot & \boldsymbol{f}_{2m} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \boldsymbol{f}_{n1} & \boldsymbol{f}_{n2} & \cdot\cdot & \boldsymbol{f}_{nm} \end{pmatrix} = \begin{pmatrix} \frac{d\boldsymbol{f}_{11}}{d\boldsymbol{x}} & \cdot\cdot & \frac{d\boldsymbol{f}_{1m}}{d\boldsymbol{x}} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \frac{d\boldsymbol{f}_{n1}}{d\boldsymbol{x}} & \cdot\cdot & \frac{d\boldsymbol{f}_{nm}}{d\boldsymbol{x}} \end{pmatrix}$$

$$= \begin{pmatrix} J_{11} & J_{12} & \cdot\cdot & J_{1m} \\ J_{21} & J_{22} & \cdot\cdot & J_{2m} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ J_{n1} & J_{n2} & \cdot\cdot & J_{nm} \end{pmatrix}, \text{ where } J_{ij} = J(\boldsymbol{f}_{ij}(\boldsymbol{x}))$$

Hessian matrix of F

- $\dfrac{df(X)}{dX}$, where $X = \begin{pmatrix} x_{11} & \cdot\cdot & x_{1n} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ x_{m1} & \cdot\cdot & x_{mn} \end{pmatrix}$
- $\dfrac{d\boldsymbol{f}(X)}{dX}$
- $\dfrac{dF(X)}{dX}$