Modeling Inference and Averaging

Bootstrap / ML / Bayesian

19.02.15 임소현

The Bootstrap
and
Maximum Likelihood Methods

1

2

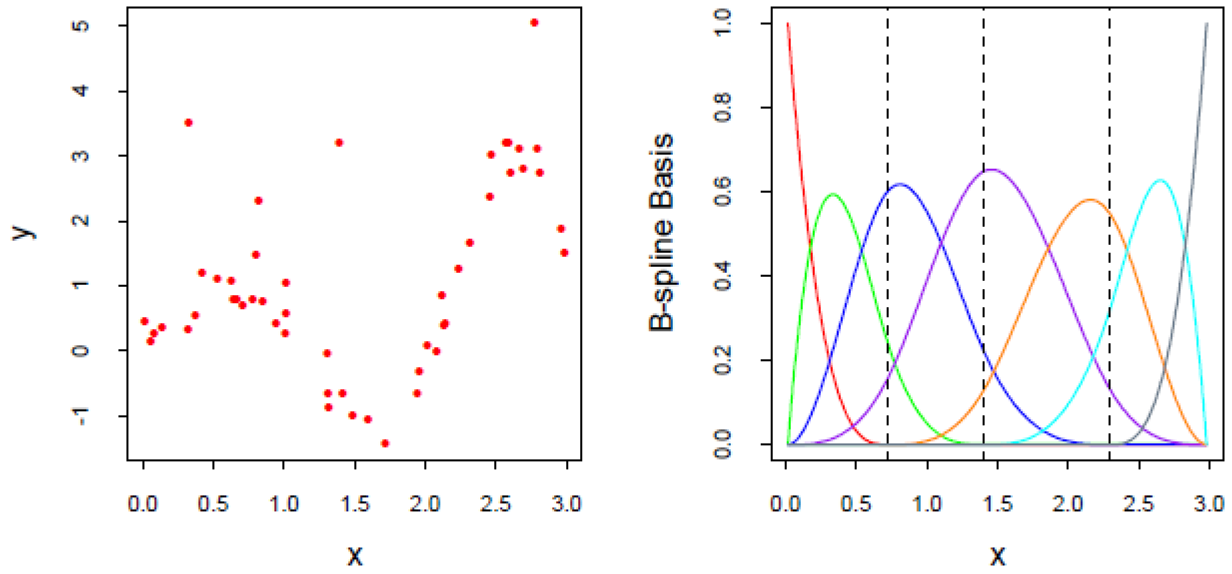
Bayesian Methods

Relationship Between
The Bootstrap
And
Bayesian Inference

3

01
The Bootstrap and Maximum Likelihood Methods

1.1 / A Smoothing Example



- N = 50 data points shown in the left panel
- A cubic spline to the data, with three knots placed at the quartiles of the X values
- a linear expansion of B-spline basis functions

$$\mu(x) \;=\; \sum_{j=1}^{7} \beta_j h_j(x).$$

- hj(x), j = 1, 2, . . . , 7 are the seven functions shown in the right panel

1.1 / A Smoothing Example

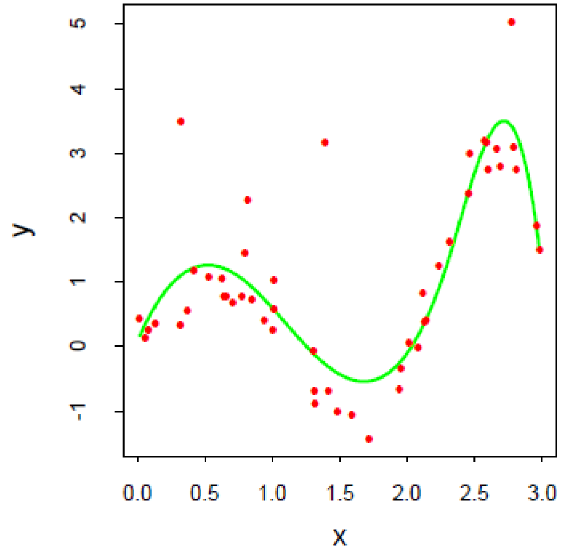
- Let H be the N ×7 matrix with ijth element hj(xi)

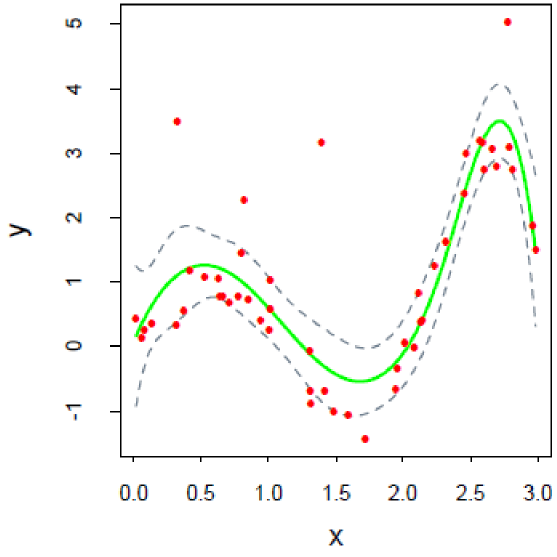
- Estimate of β, obtained by minimizing the squared error $\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}.$
- The estimated covariance matrix

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (\mathbf{H}^T\mathbf{H})^{-1}\hat{\sigma}^2$$

- Letting $h(x)^T = (h_1(x), h_2(x), \ldots, h_7(x)),$

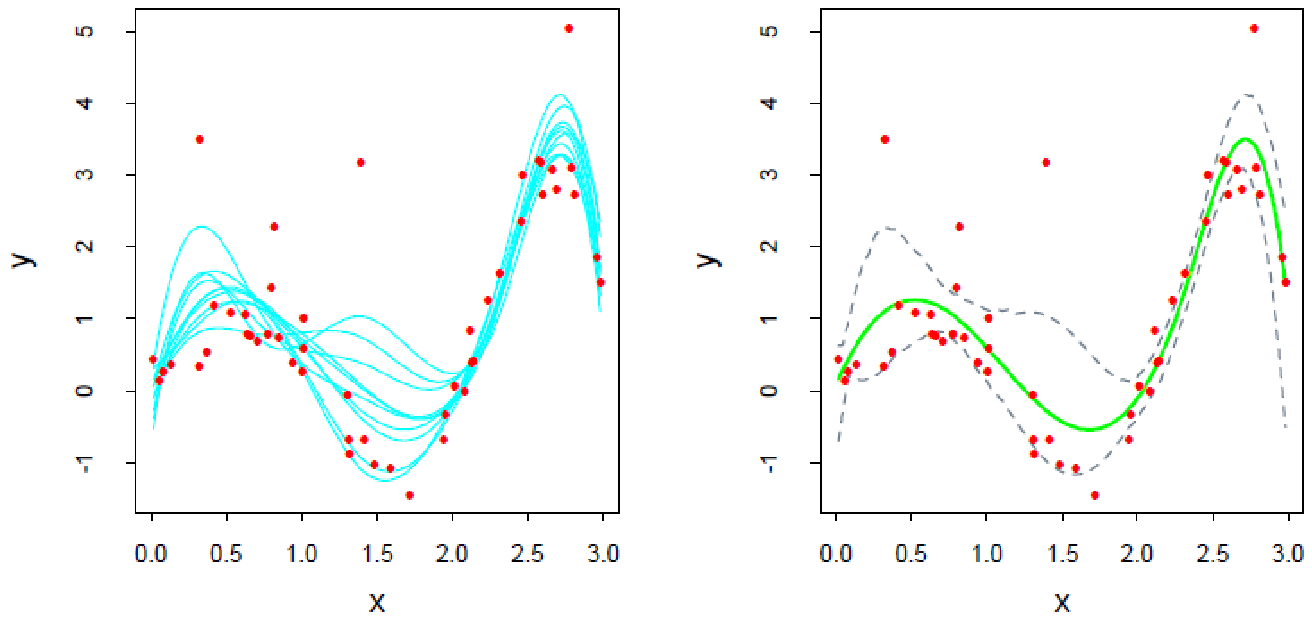
$$\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x) \qquad \widehat{\mathrm{se}}[\hat{\mu}(x)] = [h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)]^{\frac{1}{2}}\hat{\sigma}.$$





1.1 / A Smoothing Example

- draw B=200 datasets each of size N=50 with replacement from our training data
- to each bootstrap dataset Z∗ , fit a cubic spline ˆμ∗(x)
- the fits from ten such samples are shown in the bottom left panel



1.1 / A Smoothing Example

- assume that the model errors are Gaussian,

$$\begin{aligned} Y &= \mu(X) + \varepsilon; \;\; \varepsilon \sim N(0, \sigma^2), \\ \mu(x) &= \sum_{j=1}^{7} \beta_j h_j(x). \end{aligned}$$

- sample with replacement from the training data, is called the *nonparametric bootstrap*

- consider a variation of the bootstrap, called the *parametric bootstrap*
- simulate new responses by adding Gaussian noise to the predicted values -> repeated B times

$$y_i^* = \hat{\mu}(x_i) + \varepsilon_i^*; \quad \varepsilon_i^* \sim N(0, \hat{\sigma}^2); \;\; i = 1, 2, \ldots, N$$

- A function estimated from a bootstrap sample y∗ is given by $\hat{\mu}^*(x) = h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}^*$

$$\hat{\mu}^*(x) \sim N(\hat{\mu}(x), h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)\hat{\sigma}^2)$$

-> the mean of this distribution is the least squares estimate

1.2 / Maximum Likelihood Inference

- $z_i \sim g_\theta(z)$, θ : one or more unknown parameters of Z, called a *parametric model* for Z
- if Z has a normal distribution

$$\theta = (\mu, \sigma^2). \qquad g_\theta(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(z-\mu)^2/\sigma^2}$$

- Maximum likelihood is based on the *likelihood function*,

$$L(\theta; \mathbf{Z}) = \prod_{i=1}^{N} g_\theta(z_i).$$

- ***log-likelihood***

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \ell(\theta; z_i) = \sum_{i=1}^{N} \log g_\theta(z_i).$$

- maximum likelihood chooses the value θ = $\hat{\theta}$ to maximize $\ell(\theta; \mathbf{Z})$

1.2 / Maximum Likelihood Inference

- Let $\theta = (\beta, \sigma^2)$
- the log-likelihood is

$$\ell(\theta) = -\frac{N}{2}\log \sigma^2 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - h(x_i)^T\beta)^2.$$

- By setting $\partial\ell/\partial\beta = 0$ and $\partial\ell/\partial\sigma^2 = 0$, ,

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y},$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum(y_i - \hat{\mu}(x_i))^2$$

- The information matrix

$$\mathbf{I}(\beta) = (\mathbf{H}^T\mathbf{H})/\sigma^2$$

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (\mathbf{H}^T\mathbf{H})^{-1}\hat{\sigma}^2.$$

→ agrees with the least squares estimate

02
Bayesian Methods

02/ Bayesian Methods

- we specify a sampling model Pr(Z|θ), and a prior distribution for the parameters Pr(θ)

- compute the posterior distribution,

$$\Pr(\theta|\mathbf{Z}) = \frac{\Pr(\mathbf{Z}|\theta)\cdot\Pr(\theta)}{\int \Pr(\mathbf{Z}|\theta)\cdot\Pr(\theta)d\theta}$$

- The posterior distribution also provides the basis for predicting the values of a future observation, via the ***predictive distribution***

$$\Pr(z^{\text{new}}|\mathbf{Z}) = \int \Pr(z^{\text{new}}|\theta)\cdot\Pr(\theta|\mathbf{Z})d\theta.$$

- unlike the predictive distribution, this does not account for the uncertainty in estimating θ

02/ Bayesian Methods

- The parametric model and σ^2 is known

$$
\begin{aligned}
Y &= \mu(X) + \varepsilon; \ \ \varepsilon \sim N(0, \sigma^2), \\
\mu(x) &= \sum_{j=1}^{7} \beta_j h_j(x).
\end{aligned}
$$

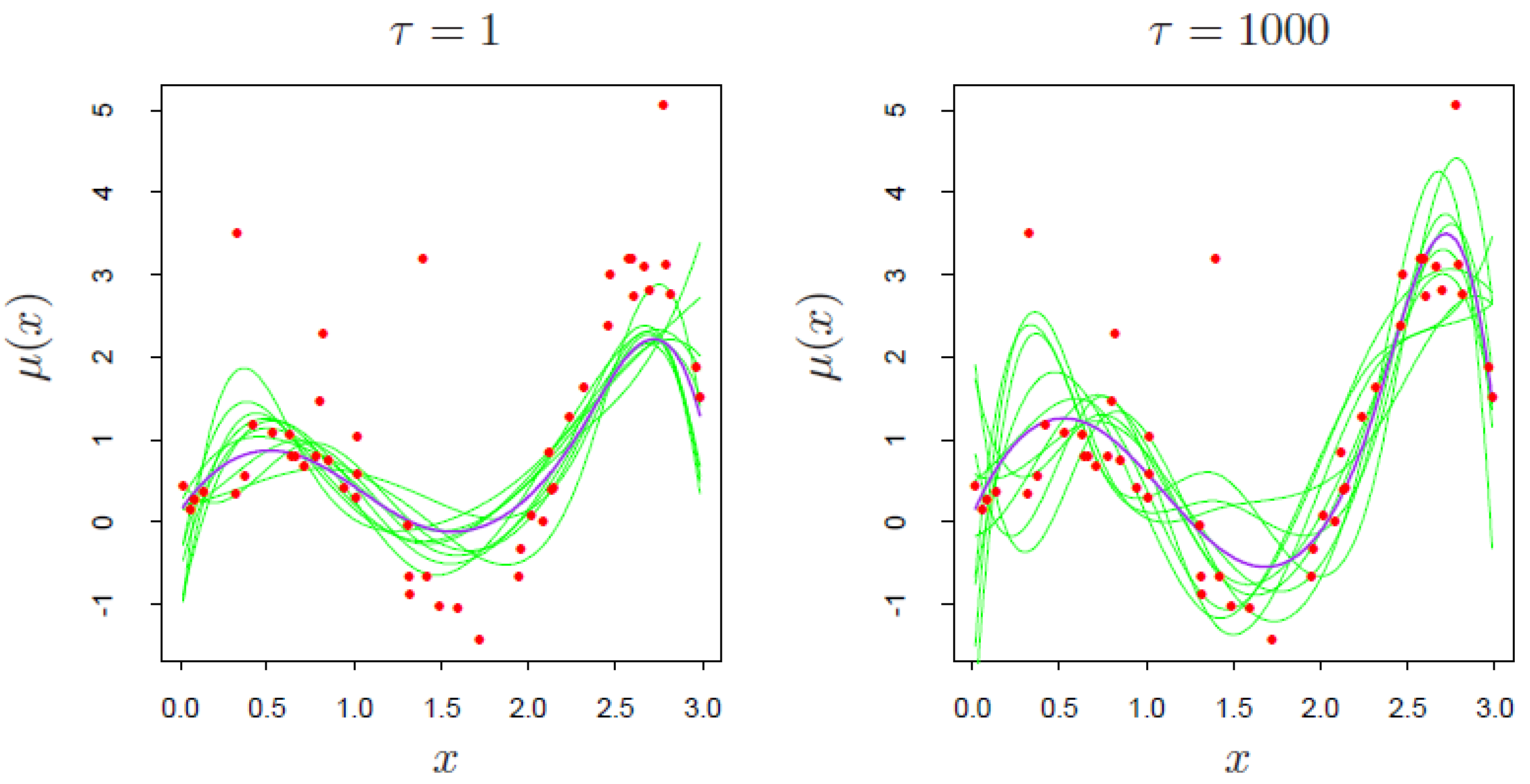
- a prior for the coefficients β, and this defines a prior for μ(x).

$$\beta \sim N(0, \tau\mathbf{\Sigma})$$

- The posterior distribution for β is also Gaussian, with mean and covariance

$$
\begin{aligned}
\mathrm{E}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\mathbf{H}^T\mathbf{y}, \\
\mathrm{cov}(\beta|\mathbf{Z}) &= \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\mathbf{\Sigma}^{-1}\right)^{-1}\sigma^2,
\end{aligned}
$$

02/ Bayesian Methods



- as $\tau \rightarrow \infty$, the posterior distribution and the bootstrap distribution coincide
- $\tau = 1$, the posterior curves $\mu(x)$ are smoother than the bootstrap curves, because we have imposed more prior weight on smoothness.

03
Relationship
Between
The Bootstrap
And
Bayesian Inference

03/ Relationship Between the Bootstrap and Bayesian Inference

- a single observation z , $z \sim N(\theta, 1)$.
- Prior distribution $\theta \sim N(0, \tau)$

$$\theta | z \sim N\left(\frac{z}{1+1/\tau}, \frac{1}{1+1/\tau}\right)$$

- as $\tau \to \infty$,

$$\theta | z \sim N(z, 1)$$

- same as a parametric bootstrap distribution in which we generate bootstrap values $z*$ from the maximum likelihood estimate of the sampling density $N(z, 1)$

1. The choice of noninformative prior for θ.
2. The dependence of the log-likelihood $\ell(\theta; \mathbf{Z})$ on the data $\mathbf{Z}$ only through the maximum likelihood estimate $\hat{\theta}$. Hence we can write the log-likelihood as $\ell(\theta; \hat{\theta})$.
3. The symmetry of the log-likelihood in θ and $\hat{\theta}$, that is, $\ell(\theta; \hat{\theta}) = \ell(\hat{\theta}; \theta) + \text{constant}$.

03/ Relationship Between the Bootstrap and Bayesian Inference

- Let wj be the probability that a sample point falls in category j
- $\widehat{wj}$ the observed proportion in category j (with L categories)

- a prior distribution for *w* a symmetric Dirichlet distribution with parameter *a*

$$w \sim \mathrm{Di}_L(a1)$$

- the posterior density of *w*

$$w \sim \mathrm{Di}_L(a1 + N\hat{w}) \xrightarrow{a \to 0} w \sim \mathrm{Di}_L(N\hat{w})$$

- the bootstrap distributioncan be expressed as sampling the category proportions from a multinomial distribution.

$$N\hat{w}^* \sim \mathrm{Mult}(N, \hat{w})$$

same mean and nearly the same covariance matrix
Hence, **the bootstrap distribution** will closely approximate **the posterior distribution**

THANK YOU