Model Inference and Averaging

8.1 - 8.4

Adrian Płoszczyca

Plan prezentacji

- wprowadzenie
- związek metody bootstrap z funkcją największej wiarogodności
- bootstrap parametryczny
- związek podejścia Bayesowskiego z funkcją największej wiarogodności

Rozważany przypadek

Rozważamy następujące dane:

$$\mathbf{Z} = \{z_1, z_2, \ldots, z_N\}$$

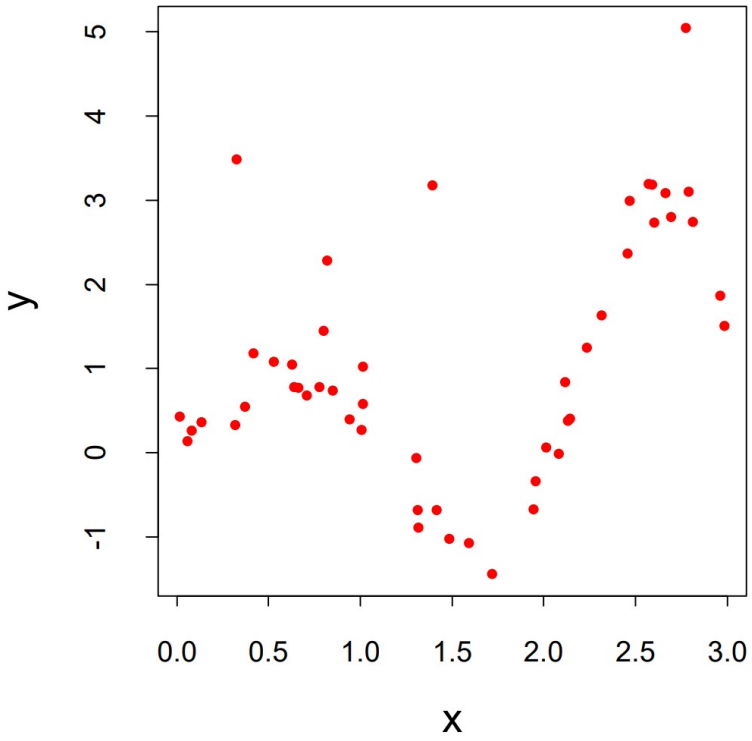
$$z_i = (x_i, y_i), \quad i = 1, 2, \ldots, N.$$

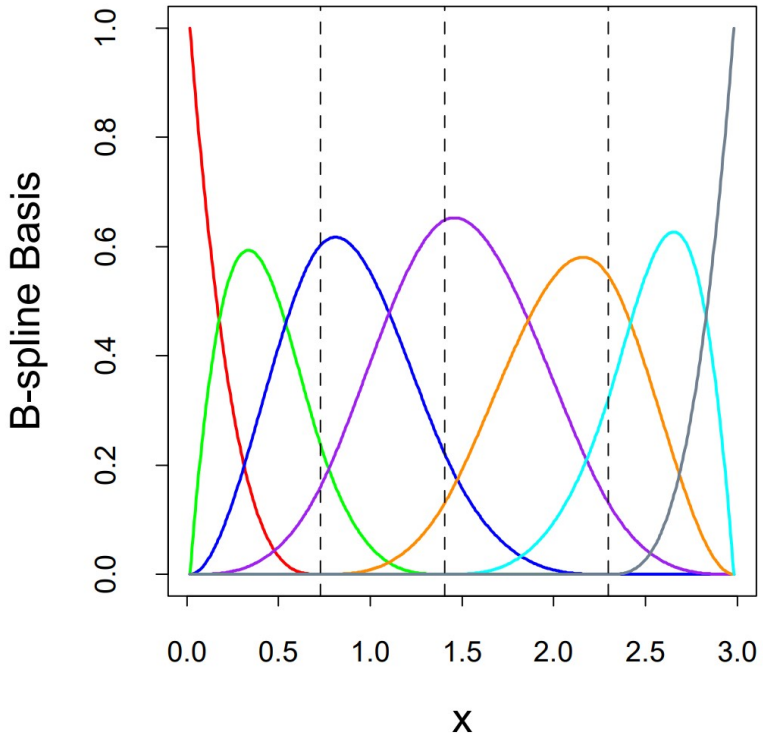
Użyjemy B-splines:

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

$$j = 1, 2, \ldots, 7$$

B-spline

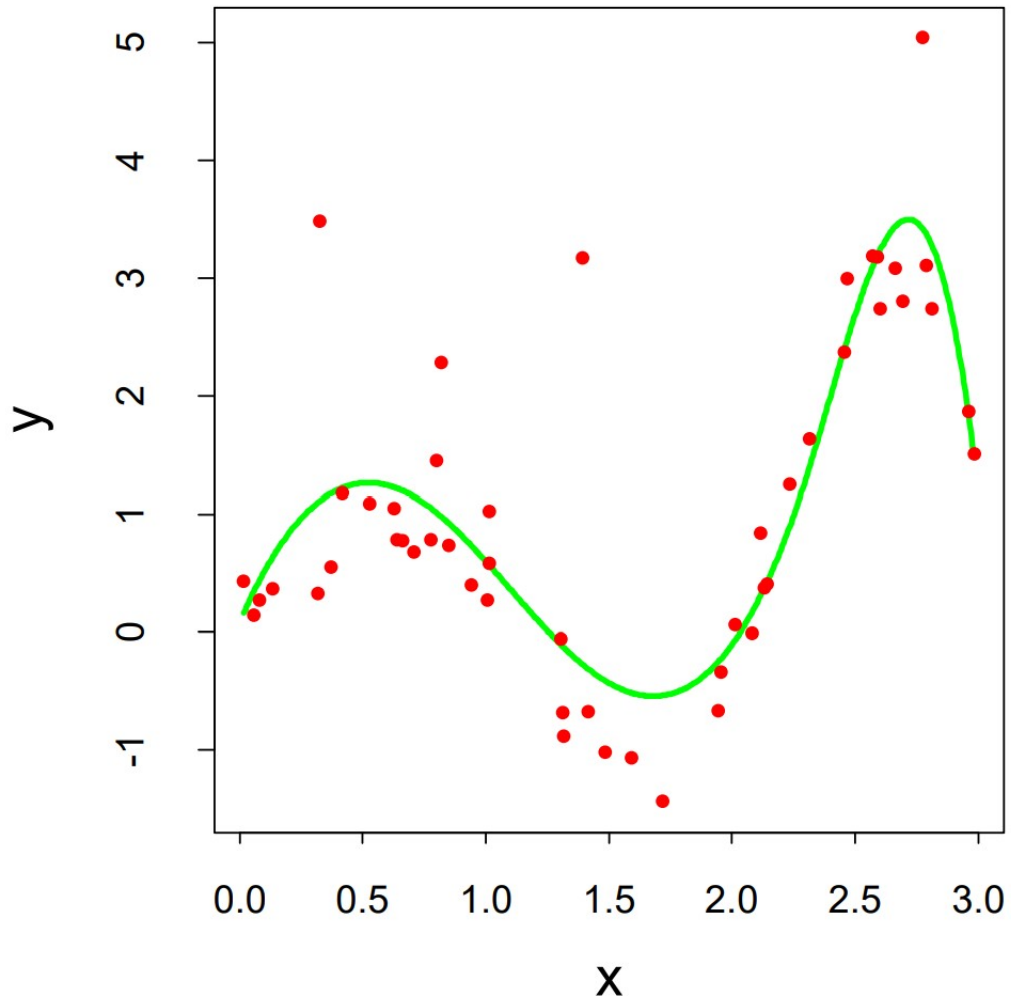

y
x
-1
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0


B-spline Basis
x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Estymator β

- Niech $\mathbf{H}$ będzie macierzą wymiaru $N \times 7$
- $\mathbf{H}_{ij} = h_j(x_i)$
- estymator β jest postaci

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



$$\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x)$$

Estymator wariancji i $\widehat{\text{se}}$

Estymator wariancji:

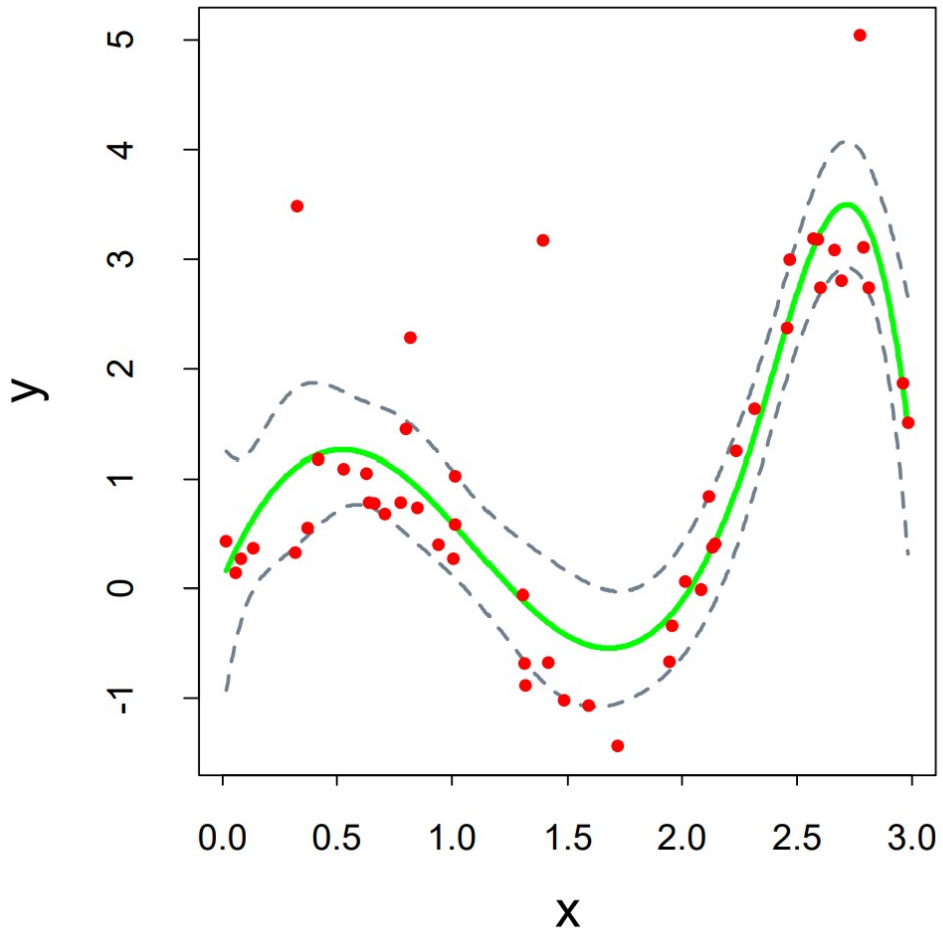
$$\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{H}^T\mathbf{H})^{-1}\hat{\sigma}^2$$

$$\hat{\sigma}^2 = \sum_{i=1}^{N}(y_i - \hat{\mu}(x_i))^2/N$$

Jeżeli zapiszemy $h(x)$ jako $h(x)^T = (h_1(x), h_2(x), \ldots, h_7(x))$, to $\widehat{\text{se}}[\hat{\mu}(x)]$ dla $\hat{\mu}(x) = h(x)^T\hat{\beta}$ jest postaci

$$\widehat{\text{se}}[\hat{\mu}(x)] = [h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}h(x)]^{1/2}\hat{\sigma}$$

Przedział ufności

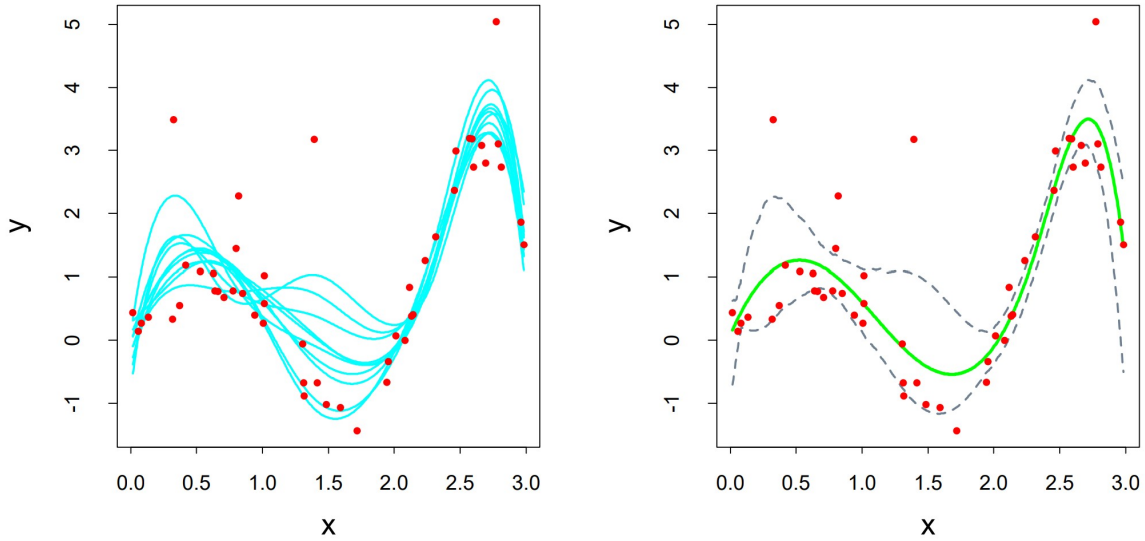


$$\hat{\mu}(x) \pm 1.96 \cdot \widehat{\mathsf{se}}[\hat{\mu}(x)]$$

Zastosowanie metody bootstrap

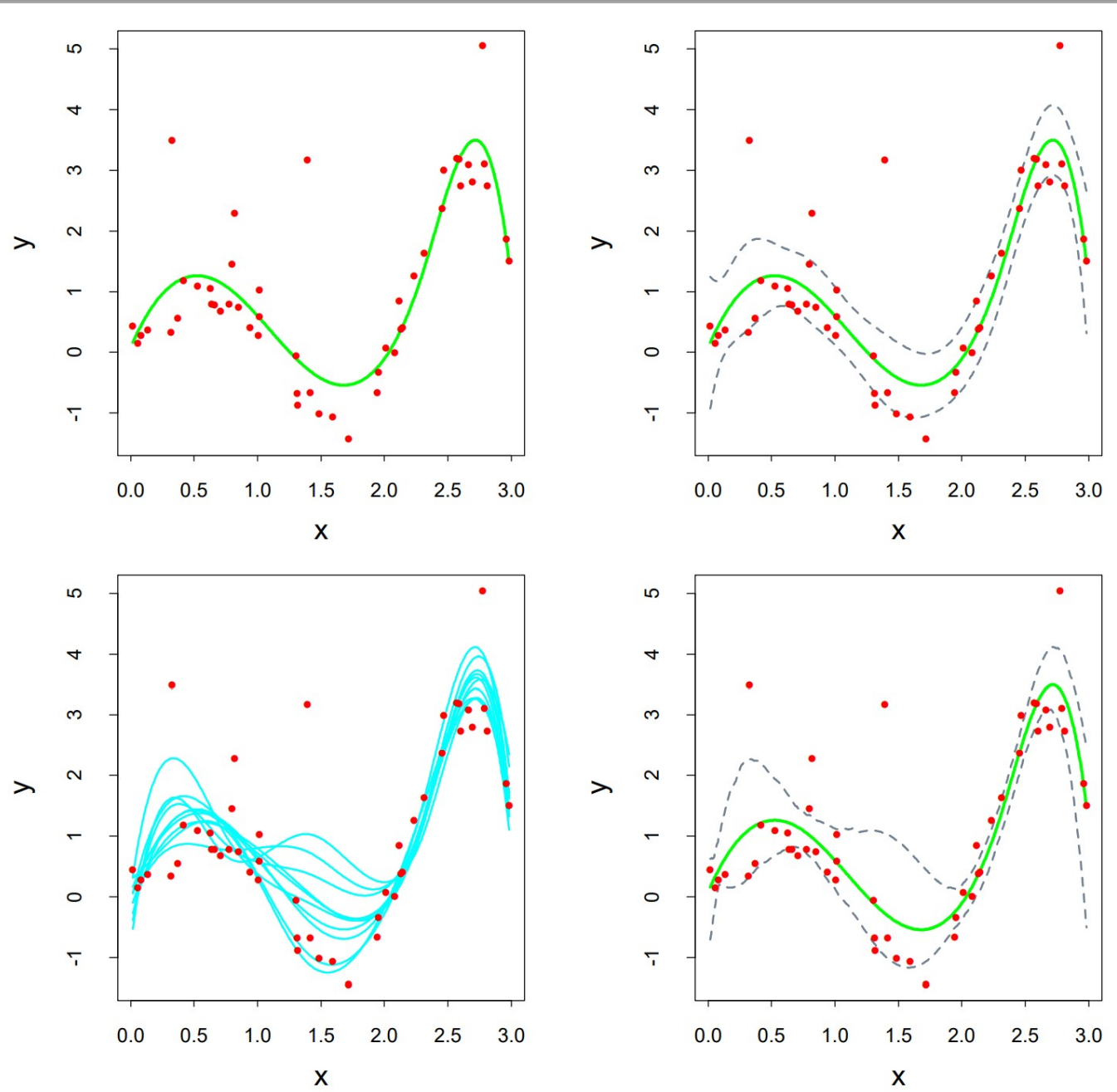
- losujemy z powtórzeniami B zbiorów rozmiaru N ze zbioru treningowego
- losujemy punkty $z_i = (x_i, y_i)$
- w wyniku dopasowania na bazie $\mathbf{Z}^*$ otrzymujemy $\hat{\mu}^*(x)$

Zastosowanie metody bootstrap



Bootstrapowe przedziały ufności obliczamy w oparciu o kwantyle z wygenerowanych danych.

Bootsrtap vs metoda najmniejszych kwadratów



Porównanie metody bootstrap i metody najmniejszych kwadratów

Model

$$Y = \mu(X) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$$\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x)$$

Bootstrap

- nieparametryczny
- parametryczny

Bootstrap parametryczny

Losujemy ϵ_i^* z rozkładu $\epsilon_i^* \sim N(0, \hat{\sigma}^2)$ i otrzymujemy

$$y_i^* = \hat{\mu}(x_i) + \epsilon_i^*,$$

$$i = 1, 2, \ldots, N.$$

Nowe zbiory są postaci

$$(x_1, y_1^*), \ldots, (x_N, y_N^*).$$

Porównanie metody bootsrtap i metody najmniejszych kwadratów

Gdy liczba prób bootstrapowych dąży do nieskończoności, bootstrapowe przedziały ufności zgadzają się z przedziałami otrzymanymi przy użyciu metody najmniejszych kwadratów.

Postać funkcji wstymowanej przy użyciu bootstrapu parametrycznego

$$\hat{\mu}^*(x) = h(x)^T(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}^*$$

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

Zgodność bootstrapu z funkcją największej wiarogodności

Rozkład z:

$$z_i \sim g_\theta(z)$$

Dla z pochodzącego z rozkładu normalnego:

$$\theta = (\mu, \sigma^2)$$

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

Funkcja największej wiarogodności

Estymator największej wiarogodności jest oparty na funkcji największej wiarogotności

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

Metoda największej wiarogodności wybiera taką wartość θ, która maksymalizuje logarytm wiarogodności

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

Rozkład estymatora największej wiarogodności

Macierz informacji

$$\mathbf{I}(\theta) = -\sum_{i=1}^{N} \frac{\partial^2 l(\theta, z_i)}{\partial\theta\partial\theta^T}$$

Macierz informacji Fishera

$$\mathbf{i}(\theta) = \mathbb{E}_\theta[\mathbf{I}(\theta)]$$

Rozkład próbkowy estymatora największej wiarogodności

$$\hat{\theta} \to N(\theta_0, \mathbf{i}(\theta_0)^{-1})$$

Przybliżenie próbkowego rozkładu estymatora θ

Możemy przybliżyć próbkowy rozkład estymatora θ rozkładem normalnym

$$N(\hat{\theta}, \mathbf{i}(\hat{\theta})^{-1}) \quad \text{lub} \quad N(\hat{\theta}, \mathbf{I}(\hat{\theta})^{-1})$$

Estymatory błędów standardowych:

$$\sqrt{\mathbf{i}(\hat{\theta})^{-1}_{jj}} \quad \text{lub} \quad \sqrt{\mathbf{I}(\hat{\theta})^{-1}_{jj}}$$

Przedziały ufności:

$$\hat{\theta}_j - z^{(1-\alpha)}\sqrt{\mathbf{i}(\hat{\theta})^{-1}_{jj}} \quad \text{lub} \quad \hat{\theta}_j - z^{(1-\alpha)}\sqrt{\mathbf{I}(\hat{\theta})^{-1}_{jj}}$$

PU na bazie funkcji wiarogodności

Przybliżenie rozkładem chi-kwadrat:

$$2[l(\hat{\theta}) - l(\theta_0)] \sim \chi^2_p$$

Przedział ufności:

$$2[l(\hat{\theta}) - l(\theta_0)] \leq \chi_p^{2(1-2\alpha)}$$

Funkcja wiarogodności a metoda najmniejszych kwadratów

Rozważamy parametr $\theta = (\beta, \sigma^2)$ oraz logarytm wiarogodności

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

Estymatory największej wiarogodności:

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

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

Macierz informacji:

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

Funkcja wiarogodności a bootstrap

- bootstrap pozwala na osiąganie wyników zgodnych z podejściem opartym na funkcji największej wiarogodności
- przewagą metody bootstrap jest to, że możemy jej użyć gdy nie dysponujemy wzorami analitycznymi

Podejście Bayesowskie

- określamy model zadany jako

$$\mathbb{P}(\mathbf{Z}|\theta)$$

- rozkład a priori

$$\mathbb{P}(\theta)$$

- rozkład a posteriori

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

Predykcja

- podejście Bayesowskie

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

- metoda największej wiarogodności

$$\mathbb{P}(z^{\text{new}}|\hat{\theta})$$

Zastosowanie podejścia Bayesowskiego

Mamy taki model jak wcześniej:

$$Y = \mu(X) + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

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

Zakładamy, że znamy σ^2 oraz, że wartości $x_1, ..., x_N$ są ustalone.

Rozkład β

Rozkład a priori:

$$\beta \sim N(0, \tau\Sigma)$$

Rozkład a posteriori:

$$\mathbb{E}(\beta|\mathbf{Z}) = \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma\right)^{-1} \mathbf{H}^T\mathbf{y}$$

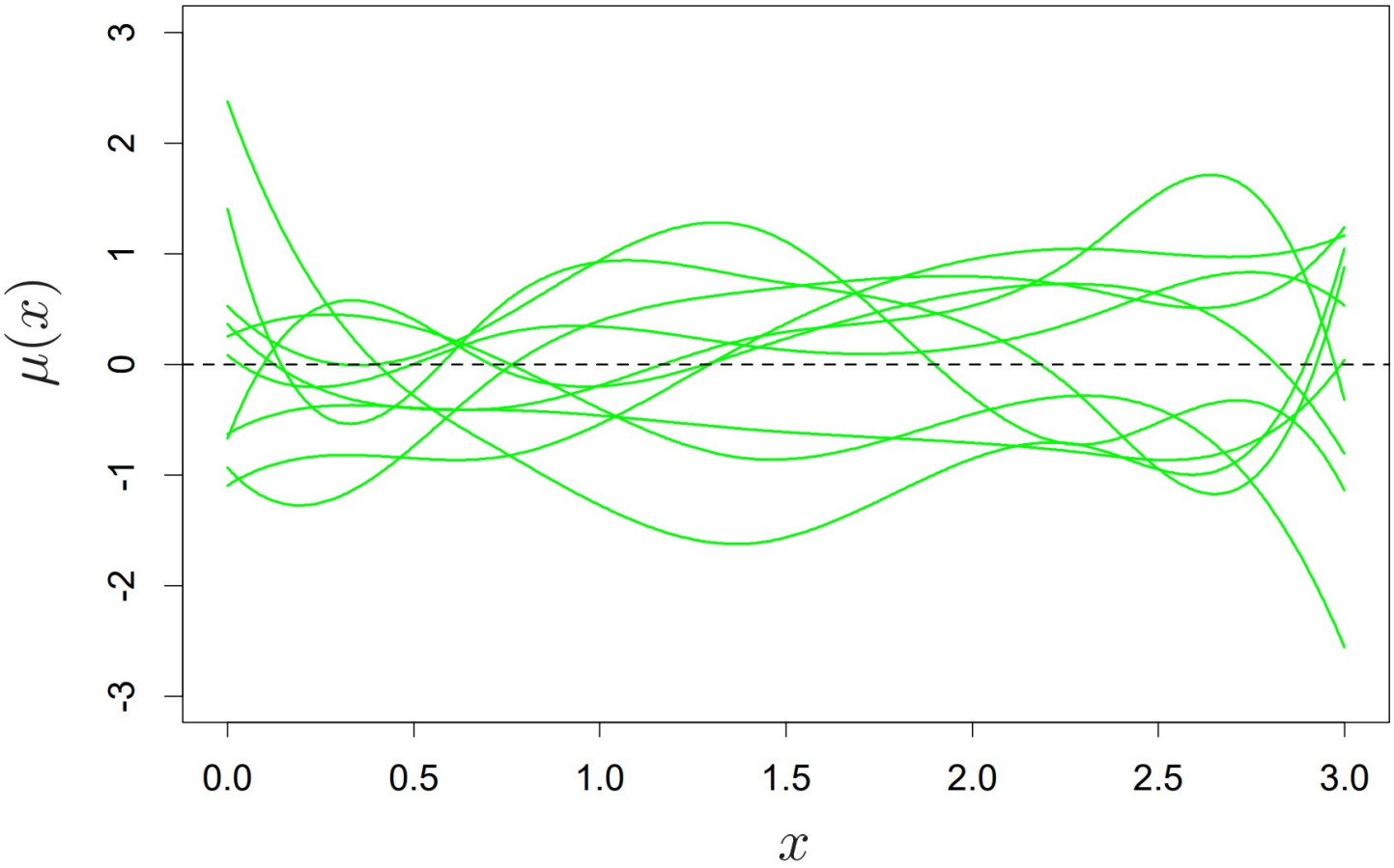
$$\text{cov}(\beta|\mathbf{Z}) = \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma\right)^{-1} \sigma^2$$

Rozkład a posteriori $\hat{\mu}(x)$

$$\mathbb{E}(\mu(x)|\mathbf{Z}) = h(x)^T \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma \right)^{-1} \mathbf{H}^T\mathbf{y}$$

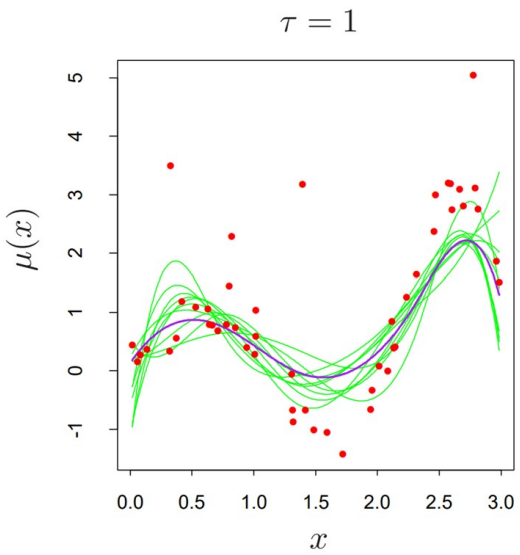
$$\text{cov}[\mu(x), \mu(x')|\mathbf{Z}) = h(x)^T \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma \right)^{-1} h(x')\sigma^2$$

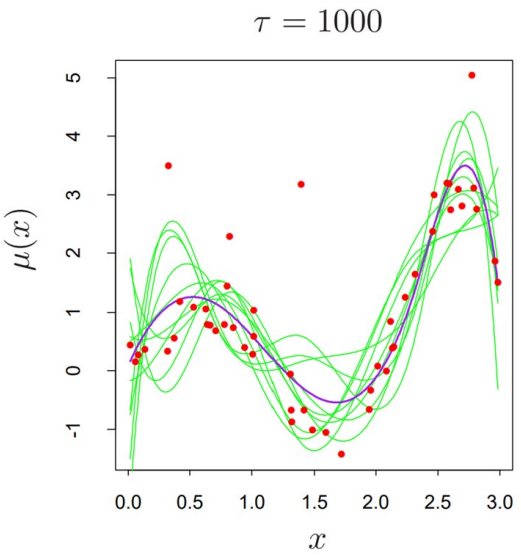
Wybieramy a priori $\Sigma = \mathbf{I}$.

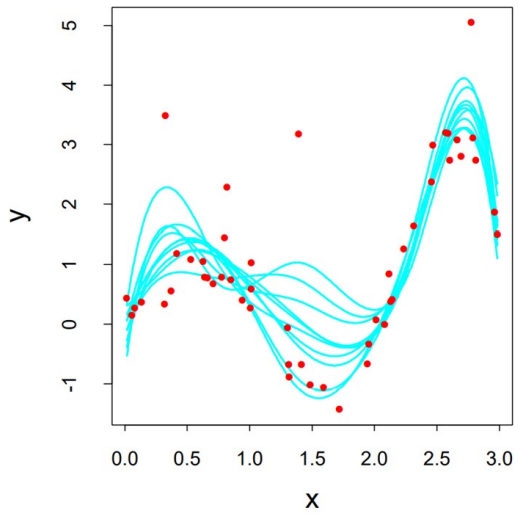

$\mu(x)$
3
2
1
0
-1
-2
-3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x

Próby z rozkładu a posteriori

$$\mu'(x) = \sum_{j=1}^{7} \beta'_j h_j(x)$$







Podejście Bayesowskie a funkcja wiarogodności

- Dla $\tau \to \infty$ otrzymujemy nieinformacyjny rozkład a priori.
- $\mathbb{P}(\theta|\mathbf{Z}) \propto \mathbb{P}(\mathbf{Z}|\theta)\mathbb{P}(\theta)$
- Typowo w podejściu Bayesowskim zakłada się dla rozkładu a priori σ, że

$$g(\sigma) \propto 1/\sigma.$$

Związek pomiędzy bootstrapem a wnioskowaniem Bayesowskim

Rozważamy $z \sim N(\theta, 1)$.

Określamy rozkład a priori $\theta \sim N(0, \tau)$.

Otrzymujemy rozkład a posteriori postaci

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

Dla parametrycznego bootstrapu mamy rozkład $N(z, 1)$.

Związek pomiędzy bootstrapem a wnioskowaniem Bayesowskim

1. Wybór nieinformacyjnego rozkładu a priori dla θ
2. Możemy zapisać $l(\theta, \mathbf{Z})$ jako $l(\theta, \hat{\theta})$
3. $l(\theta, \hat{\theta}) = l(\hat{\theta}, \theta) + const$

Związek pomiędzy bootstrapem a wnioskowaniem Bayesowskim

- mamy L kategorii
- w_j - prawdopodobieństwo, że obserwacja jest z j-tej kategorii

$$w = (w_1, w_2, \ldots, w_L)$$

- $\hat{w}_j$ - obserwowana proporcja obserwacji w j-tej kategorii

$$\hat{w} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_L)$$

- $S(\hat{w})$ - estymator
- rozkład a priori dla w:

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

Związek pomiędzy bootstrapem a wnioskowaniem Bayesowskim

Wtedy:

Funkcja prawdopodobieństwa jest proporcjonalna do $\prod_{l=1}^{L} w_l^{a-1}$.

Rozkład a posteriori dla w

$$w \sim \mathrm{Di}_L(a + N\hat{w}).$$

Przy $a \to 0$ mamy

$$w \sim \mathrm{Di}_L(N\hat{w}).$$

Związek pomiędzy bootstrapem a wnioskowaniem Bayesowskim

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

gdzie $\text{Mult}(N, \hat{w})$ oznacza rozkład wielomianowy o funkcji prawdopodobieństwa postaci

$$\binom{N}{N\hat{w}_1^*, \ldots, N\hat{w}_L^*} \prod \hat{w}_l^{N\hat{w}_l^*}.$$

Bootstrapowy rozkład $S(\hat{w}^*)$ przybliża rozkład a posteriori $S(w)$.

KONIEC

Dziękuję za uwagę