# Estymacja gęstości prawdopodobieństwa metodą selekcji modelu

M. Wojtyś

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Wisła, 7 grudnia 2009

Próba losowa z rozkładu prawdopodobieństwa o nieznanej gęstości $f$:

$$X_1, X_2, \ldots, X_n \sim f$$

Niech

$$\mathcal{M} = \{f_\theta : \ \theta \in \mathbb{R}^m\}$$

będzie $m$-wymiarowym modelem parametrycznym, gdzie funkcje $f_\theta$ to gęstości p-stwa dane w pewnej ustalonej postaci.

Próba losowa z rozkładu prawdopodobieństwa o nieznanej gęstości $f$:

$$X_1, X_2, \ldots, X_n \sim f$$

Niech

$$\mathcal{M} = \{f_\theta : \ \theta \in \mathbb{R}^m\}$$

będzie $m$-wymiarowym modelem parametrycznym, gdzie funkcje $f_\theta$ to gęstości p-stwa dane w pewnej ustalonej postaci.

**Przykład 1.** Liniowe kombinacje funkcji ortogonalnych:

$$f_\theta(x) = \sum_{j=1}^{m} \theta_j b_j(x),$$

gdzie $b_1, b_2, \ldots, b_m$ jest układem ortogonalnym w zadanej przestrzeni funkcyjnej.

**Przykład 2.** Rodzina wykładnicza rozkładów:

$$f_\theta(x) = c(\theta) \exp\{\sum_{j=1}^{m} \theta_j b_j(x)\},$$

gdzie $c(\theta)$ jest stałą normującą.

**Przykład 1.** Liniowe kombinacje funkcji ortogonalnych:

$$f_\theta(x) = \sum_{j=1}^{m} \theta_j b_j(x),$$

gdzie $b_1, b_2, \ldots, b_m$ jest układem ortogonalnym w zadanej przestrzeni funkcyjnej.

**Przykład 2.** Rodzina wykładnicza rozkładów:

$$f_\theta(x) = c(\theta) \exp\{\sum_{j=1}^{m} \theta_j b_j(x)\},$$

gdzie $c(\theta)$ jest stałą normującą.

# Wstęp

Dla $\gamma \in 2^{\{1,\ldots,m\}}$ zdefiniujmy model $\mathcal{M}_\gamma \subset \mathcal{M} = \{f_\theta : \theta \in \mathbb{R}^m\}$ taki, że

$$\mathcal{M}_\gamma = \{f_\theta(\cdot) \in \mathcal{M} : \theta_j = 0 \text{ dla } j \notin \gamma\}.$$

Oczywiście $\mathcal{M}_{\{1,\ldots,m\}} = \mathcal{M}$.

Problem selekcji modelu polega na wybraniu na podstawie próby losowej jednego z $2^m$ powyższych modeli w celu estymacji $f$.

Dla $\gamma \in 2^{\{1,\ldots,m\}}$ zdefiniujmy model $\mathcal{M}_\gamma \subset \mathcal{M} = \{f_\theta : \theta \in \mathbb{R}^m\}$ taki, że

$$\mathcal{M}_\gamma = \{f_\theta(\cdot) \in \mathcal{M} : \theta_j = 0 \text{ dla } j \notin \gamma\}.$$

Oczywiście $\mathcal{M}_{\{1,\ldots,m\}} = \mathcal{M}$.

Problem selekcji modelu polega na wybraniu na podstawie próby losowej jednego z $2^m$ powyższych modeli w celu estymacji $f$.

# Wstęp

Metody selekcji modelu mają za zadanie wybrać model, w którym możliwe jest uzyskanie estymatora, który jednocześnie:

- jest dopasowany do próby losowej $X_1, \ldots, X_n$,
- odzwierciedla ogólne własności rozkładu $f$.

Szukamy zatem najprostszego modelu, w obrębie którego możliwa jest dobra aproksymacja $f$.

**Przykład**
Jeśli $f \in L^2([a,b])$, to można ją rozwinąć w szereg Fouriera w zadanej bazie funkcji:

$$f(x) = \sum_{j=1}^{\infty} \theta_j b_j(x).$$

Wtedy wybór modelu polega na znalezieniu skończonego podzbioru $\gamma \subset \mathbb{N}$ dającego dobry estymator postaci

$$\hat{f}(x) = \sum_{j \in \gamma} \hat{\theta}_j b_j(x)$$

w modelu $\mathcal{M}_\gamma$.

# Przykłady metod selekcji

**AIC**: bazuje na odległości Kullbacka-Leiblera $\mathbb{E}_f\left(\log \frac{f}{\hat{f}}\right)$ jako funkcji straty.

$$\hat{\gamma} = \underset{\gamma \in 2^{\{1,\ldots,m\}}}{\arg\max}\ AIC(\mathcal{M}_\gamma),$$

$$\text{gdzie } AIC(\mathcal{M}_\gamma) = \log \prod_{i=1}^{n} \hat{f}_\gamma(X_i) - |\gamma|,$$

gdzie $\hat{f}_\gamma$ to estymator największej wiarogodności dla $f$ w modelu $\mathcal{M}_\gamma$.

**BIC**:

$$\hat{\gamma} = \underset{\gamma \in 2^{\{1,\ldots,m\}}}{\arg\max}\ BIC(\mathcal{M}_\gamma),$$

$$\text{gdzie } BIC(\mathcal{M}_\gamma) = \log \prod_{i=1}^{n} \hat{f}_\gamma(X_i) - |\gamma|\frac{\log n}{2}.$$

# Przykłady metod selekcji

**AIC**: bazuje na odległości Kullbacka-Leiblera $\mathbb{E}_f\left(\log \frac{f}{\hat{f}}\right)$ jako funkcji straty.

$$\hat{\gamma} = \underset{\gamma \in 2^{\{1,\ldots,m\}}}{\arg\max}\ AIC(\mathcal{M}_\gamma),$$

$$\text{gdzie } AIC(\mathcal{M}_\gamma) = \log \prod_{i=1}^{n} \hat{f}_\gamma(X_i) - |\gamma|,$$

gdzie $\hat{f}_\gamma$ to estymator największej wiarogodności dla $f$ w modelu $\mathcal{M}_\gamma$.

**BIC**:

$$\hat{\gamma} = \underset{\gamma \in 2^{\{1,\ldots,m\}}}{\arg\max}\ BIC(\mathcal{M}_\gamma),$$

$$\text{gdzie } BIC(\mathcal{M}_\gamma) = \log \prod_{i=1}^{n} \hat{f}_\gamma(X_i) - |\gamma| \frac{\log n}{2}.$$

# Progowanie

Niech $\hat{\theta}$ będzie estymatorem parametru $\theta \in \mathbb{R}^m$ w modelu $\mathcal{M}$ opartym o próbę $X_1, \ldots, X_n$.
Dla zadanego $\varepsilon_n > 0$ oraz stałych $C_j > 0$ zdefiniujmy zbiór

$$\Gamma_n = \{j : |\hat{\theta}_j| > C_j \varepsilon_n, 1 \leqslant j \leqslant m\}.$$

Wówczas jeśli $f$ jest postaci $f_{\theta^0}$ dla pewnego $\theta^0 \in \mathbb{R}^\infty$, to $\Gamma_n$ przybliża zbiór

$$\gamma := \{j : |\theta_j^0| > 0\},$$

a tym samym przybliża najmniejszy model, do którego należy nieznane $f_{\theta^0}$.

# Zgodność reguły wyboru modelu

## Twierdzenie

Załóżmy, że

$$f(x) = c(\theta)\exp\{\sum_{j=1}^{k}\theta_j b_j(x)\},$$

gdzie $k < \infty$ lub $k = \infty$. Jeśli:

- $\max_{1\leqslant j\leqslant m}||b_j||_\infty = O(m^\omega)$ dla pewnego $\omega \geqslant 0$,
- $||\sum_{j=m+1}^{\infty}\theta_j b_j||_\infty = O(m^{-r})$ gdy $m \to \infty$, gdzie $r > \frac{1}{2} + \omega$,
- $n/m_n^{2+4\omega} \to \infty$ oraz $n\varepsilon_n^2/m_n^{1+2\omega} \to \infty$ gdy $n \to \infty$,
- $\liminf_{n\to\infty} m_n \geqslant k$, $\varepsilon_n \to 0$, $C_j \leqslant C_{j+1}$ dla $j \in \mathbb{N}$,

to

$$\lim_{n\to\infty} P(\Gamma_n = \gamma) = 1, \text{ gdy } k < \infty.$$

Jeśli dodatkowo $|\theta_j| > (C_1 + C_j)\varepsilon_n$ dla $j \in \gamma_m := \gamma \cap \{1, \ldots, m\}$, to

$$\lim_{n\to\infty} P(\Gamma_n = \gamma_m) = 1, \text{ gdy } k = \infty.$$

# Zgodność estymatora postselekcyjnego

## Twierdzenie

Załóżmy, że

$$f(x) = c(\theta)\exp\{\sum_{j=1}^{k}\theta_j b_j(x)\},$$

gdzie $k < \infty$ lub $k = \infty$. Jeśli:

- $\max_{j=1,\ldots,m}||b_j||_\infty = O(m^\omega)$ dla pewnego $\omega \geqslant 0$,
- $||\sum_{j=m+1}^{\infty}\theta_j b_j||_\infty = O(m^{-r})$ gdy $m \to \infty$, gdzie $r > \frac{1}{2} + \omega$,
- $n/m_n^{2+4\omega} \to \infty$ i $m_n \to \infty$ gdy $n \to \infty$, $\varepsilon_n \to 0$, $C_j \leqslant C_{j+1}$ dla $j \in \mathbb{N}$,

to

$$KL(f||\hat{f}_n) \xrightarrow{\mathcal{P}} 0, \quad \text{gdy} \quad n \to \infty,$$

gdzie $\hat{f}_n$ jest estymatorem największej wiarogodności w modelu wykładniczym wybranym metodą progowania.

# Zgodność

Przy podobnych założeniach można wykazać, że jeśli

$$f(x) = \sum_{j=1}^{k} \theta_j b_j(x),$$

gdzie $k < \infty$ lub $k = \infty$, to

$$||f - \hat{f}||_{L^2} \xrightarrow{\mathcal{P}} 0, \quad \text{gdy} \quad n \to \infty.$$

Ponadto $P(\Gamma_n = \gamma) \to 1$, jeśli tylko $\lim\limits_{n\to\infty} m_n \geqslant k$.

# Wybór progu

Próg $C_j\varepsilon_n$ w definicji:

$$\Gamma_n = \{j : |\hat{\theta}_j| > C_j\varepsilon_n, 1 \leqslant j \leqslant m\}.$$

jest deterministyczny. Możemy jednak rozważać progi zależne od danych. Na przykład:

$$C_j := ||\hat{\theta}||_m C_j$$

lub

$$C_j := R_j = \text{ ranga } |\hat{\theta}_j| \text{ wśród } \{|\hat{\theta}_1|, \ldots, |\hat{\theta}_m|\}$$

przy uporządkowaniu malejącym (tzn. element największy ma rangę równą 1).

Dla tak zdefiniowanych reguł selekcji tezy twierdzeń o zgodności są prawdziwe przy podobnych założeniach.

# Przykłady obliczeniowe - trafność wyboru modelu

| $\theta$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ |
|---|---|---|---|---|
| $(0, 0.1, 0, 0, 0)$ | 0.0002 | 0.202 | 0.085 | 0.136 |
| $(0, 0.2, 0, 0, 0)$ | 0.009 | 0.375 | 0.351 | 0.345 |
| $(0, 0.4, 0, 0, 0)$ | 0.361 | 0.674 | 0.935 | 0.823 |
| $(0, 0.6, 0, 0, 0)$ | 0.931 | 0.846 | 0.999 | 0.980 |
| $(0, 0.8, 0, 0, 0)$ | 0.970 | 0.936 | 1.000 | 0.998 |
| $(0, 1.0, 0, 0, 0)$ | 0.956 | 0.968 | 0.998 | 1.000 |

$$\Gamma_1 = \{j : |\hat{\theta}_j| > j\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_2 = \{j : |\hat{\theta}_j| > j||\hat{\theta}||_m\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_3 = \{j : |\hat{\theta}_j| > R_j\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_4 = \{j : |\hat{\theta}_j| > R_j||\hat{\theta}||_m\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$

$$\varepsilon_n = \sqrt{\log n/n},\ n = 100,\ m = 5.$$

# Przykłady obliczeniowe - trafność wyboru modelu

| $\theta$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ |
|---|---|---|---|---|
| $(0, 0, 0.1, 0, 0)$ | 0.000 | 0.115 | 0.129 | 0.152 |
| $(0, 0, 0.2, 0, 0)$ | 0.000 | 0.261 | 0.429 | 0.355 |
| $(0, 0, 0.4, 0, 0)$ | 0.019 | 0.553 | 0.944 | 0.781 |
| $(0, 0, 0.6, 0, 0)$ | 0.340 | 0.704 | 0.985 | 0.954 |
| $(0, 0, 0.8, 0, 0)$ | 0.744 | 0.775 | 0.970 | 0.988 |
| $(0, 0, 1.0, 0, 0)$ | 0.758 | 0.810 | 0.951 | 0.995 |

$$\Gamma_1 = \{j : |\hat{\theta}_j| > j\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_2 = \{j : |\hat{\theta}_j| > j||\hat{\theta}||_m\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_3 = \{j : |\hat{\theta}_j| > R_j\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$
$$\Gamma_4 = \{j : |\hat{\theta}_j| > R_j||\hat{\theta}||_m\varepsilon_n,\ 1 \leqslant j \leqslant m\},$$

$$\varepsilon_n = \sqrt{\log n/n},\ n = 100,\ m = 5.$$

# Przykłady obliczeniowe - $MISE(\hat{f})$

| $\theta$ | $S$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ |
|---|---|---|---|---|---|
| $(0, 0.1, 0, 0, 0)$ | 0.025 | 0.013 | 0.028 | 0.020 | 0.046 |
| $(0, 0.2, 0, 0, 0)$ | 0.044 | 0.047 | 0.031 | 0.041 | 0.063 |
| $(0, 0.4, 0, 0, 0)$ | 0.035 | 0.133 | 0.025 | 0.025 | 0.052 |
| $(0, 0.6, 0, 0, 0)$ | 0.039 | 0.038 | 0.024 | 0.015 | 0.024 |
| $(0, 0.8, 0, 0, 0)$ | 0.046 | 0.019 | 0.022 | 0.017 | 0.018 |
| $(0, 1.0, 0, 0, 0)$ | 0.054 | 0.022 | 0.022 | 0.021 | 0.018 |

$S$ - reguła Schwarza,

$$\Gamma_1 = \{j : |\hat{\theta}_j| > j\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_2 = \{j : |\hat{\theta}_j| > j||\hat{\theta}||_m\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_3 = \{j : |\hat{\theta}_j| > R_j\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_4 = \{j : |\hat{\theta}_j| > R_j||\hat{\theta}||_m\varepsilon_n, 1 \leqslant j \leqslant m\},$$

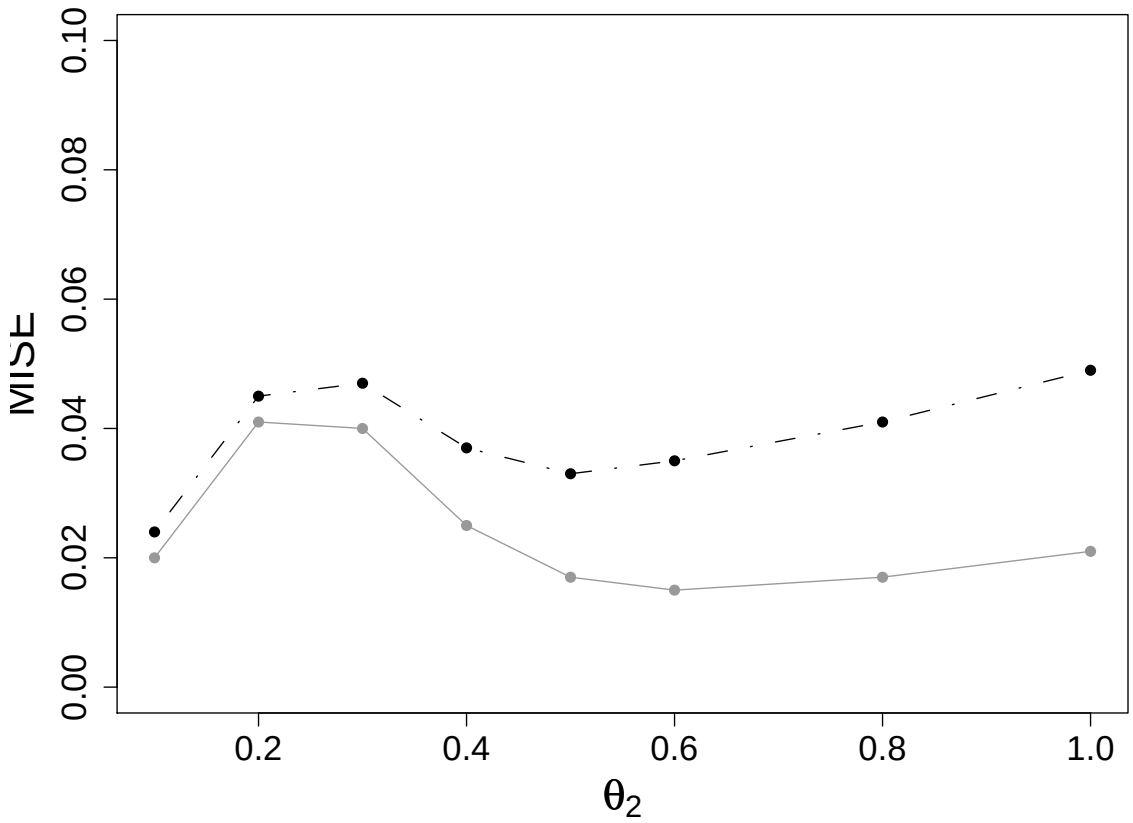
# Przykłady obliczeniowe - $MISE(\hat{f})$

| $\theta$ | $S$ | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ |
|---|---|---|---|---|---|
| $(0, 0, 0.1, 0, 0)$ | 0.024 | 0.013 | 0.031 | 0.020 | 0.038 |
| $(0, 0, 0.2, 0, 0)$ | 0.052 | 0.043 | 0.038 | 0.035 | 0.037 |
| $(0, 0, 0.4, 0, 0)$ | 0.056 | 0.163 | 0.029 | 0.020 | 0.026 |
| $(0, 0, 0.6, 0, 0)$ | 0.044 | 0.233 | 0.027 | 0.019 | 0.021 |
| $(0, 0, 0.8, 0, 0)$ | 0.053 | 0.081 | 0.030 | 0.030 | 0.024 |
| $(0, 0, 1.0, 0, 0)$ | 0.063 | 0.042 | 0.037 | 0.045 | 0.033 |

$S$ - reguła Schwarza,

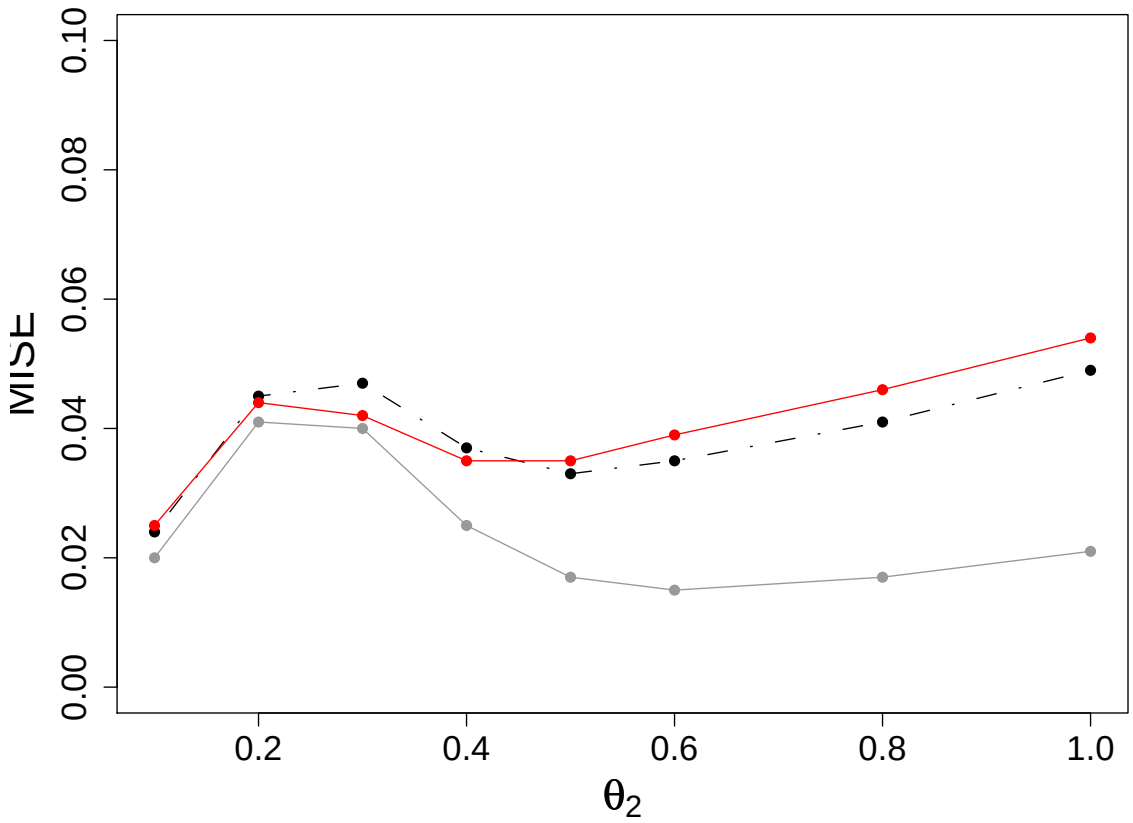
$$\Gamma_1 = \{j : |\hat{\theta}_j| > j\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_2 = \{j : |\hat{\theta}_j| > j||\hat{\theta}||_m\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_3 = \{j : |\hat{\theta}_j| > R_j\varepsilon_n, 1 \leqslant j \leqslant m\},$$
$$\Gamma_4 = \{j : |\hat{\theta}_j| > R_j||\hat{\theta}||_m\varepsilon_n, 1 \leqslant j \leqslant m\},$$

# Przykłady obliczeniowe - $MISE(\hat{f})$




linia szara - $MISE(\hat{f})$ dla reguły $\Gamma_3$,
linia przerywana - $MISE(\hat{f})$ dla reguły $\hat{k} := \max \Gamma_3$.

# Przykłady obliczeniowe - $MISE(\hat{f})$




linia szara - $MISE(\hat{f})$ dla reguły $\Gamma_3$,
linia przerywana - $MISE(\hat{f})$ dla reguły $\hat{k} := \max \Gamma_3$,
linia czerwona - $MISE(\hat{f})$ dla reguły Schwarza.