Selekcja modelu liniowego i predykcja metodami losowych podprzestrzeni

Paweł Teisseyre

Instytut Podstaw Informatyki, Polska Akademia Nauk

Plan prezentacji

1. Wysoko-wymiarowy model regresji liniowej.
2. Dwustopniowe procedury wyboru modelu.
3. Metoda Losowych Podprzestrzeni (RSM).
4. Metoda RSM + kryteria informacyjne.

Motywacja- modele

1. Regresja liniowa to najpopularniejszy model w sytuacji, gdy zmienna odpowiedzi jest ilościowa.
2. Wybrane obszary zastosowań:
   - bioinformatyka (dane mikro-macierzowe, QTL, GWAS, QSAR),
   - finanse,
   - nauki społeczne i ekonomiczne (modelowanie wskaźników makro i mikro-ekonomicznych),
   - analiza danych tekstowych (przewidywanie cech osób na podstawie wypowiedzi),
   - i wiele innych...

Motywacja- wybór modelu

- Dlaczego selekcja modelu (u nas: pewnego podzbioru zmiennych objaśniających) jest ważna?
  - odkrycie nieznanej zależności funkcyjnej na podstawie dostępnych danych,
  - wybór modelu o dobrych własnościach predykcyjnych,
  - ocena istotności zmiennych objaśniających.

Model regresji liniowej

- Obiekty opisane parą $(\mathbf{x}, y)$, gdzie:
  - $y \in R$ - zmienna odpowiedzi,
  - $\mathbf{x} \in R^p$ - wektor atrybutów.
- W modelu liniowym zakładamy, że:

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

gdzie:

- $\boldsymbol{\beta} \in R^p$ jest wektorem parametrów,
- ε błędem losowym o rozkładzie $N(0, \sigma^2)$.

Uwaga:
Dopuszczamy sytuację: $p \geqslant n$.

Selekcja zmiennych

- Minimalny model prawdziwy: $t := \{k : \beta_k \neq 0\}$, t.j.
  - **dla regresji liniowej:** minimalny model taki, że $\mathbf{E}(y|\mathbf{x}) = \mathbf{x}_t'\beta_t$,

  gdzie: dolny indeks t oznacza wybór współrzędnych odpowiadających modelowi t.
- **Cel:** Identyfikacja zbioru t na podstawie niezależnych obserwacji $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$.

Procedury dwustopniowe wyboru modelu

1. Zmienne $\{1, \ldots, p\}$ są porządkowane wg pewnej miary istotności:

$$W_{i_1} \geqslant W_{i_2} \geqslant \ldots \geqslant W_{i_p}.$$

2. Wybieramy model z zagnieżdżonej rodziny:

$$\mathcal{M}_{\text{nested}} := \{\{0\}, \{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, \ldots, i_p\}\}$$

Uwaga:

- W drugim kroku sprawdzamy $p + 1$ modeli zamiast 2^p (przy pełnym przeszukiwaniu).

Procedura Zhenga i Loha dla modelu liniowego

1. Dopasuj model liniowy zawierający wszystkie zmienne $1, \ldots, p$.
2. Zmienne $\{1, \ldots, p\}$ są porządkowane wg kwadratu statystyki T:

$$T_{i_1}^2 \geqslant T_{i_2}^2 \geqslant \ldots \geqslant T_{i_p}^2.$$

3. Wybieramy model z zagnieżdżonej rodziny:

$$\mathcal{M}_{\text{nested}} := \{\{0\}, \{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, \ldots, i_p\}\}.$$

Uwagi:

- Użycie w drugim kroku Bayesowskiego kryterium wybory zmiennych (BIC) prowadzi do zgodnej procedury selekcji.
- Procedura nie może być zastosowana gdy $p \geqslant n$.

Metoda RSM dla klasyfikacji

- Metoda zaproponowana w pracy:
  T. K. Ho, *The Random Subspace Method for Constructing Decision Forests*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 8, 1998.
- Budowa komitetu klasyfikatorów na bazie losowo wybranych podzbiorów atrybutów.
- Efektywne narzędzie w przypadku dużego wymiaru przestrzeni cech.

Metoda RSM dla modelu liniowego

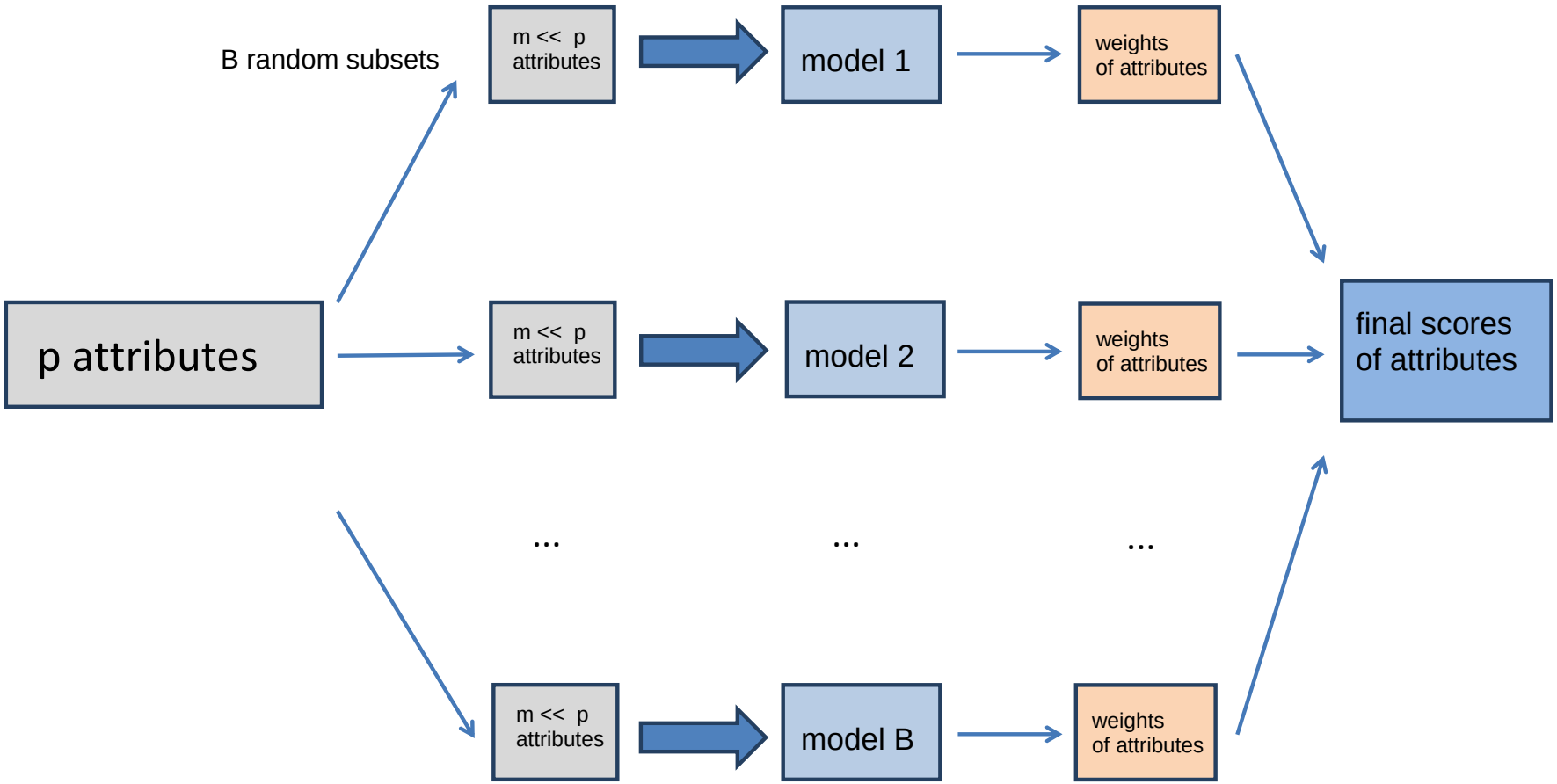
Algorytm RSM

1. Wejście: Dane $(\mathbf{Y}, \mathbf{X})$, liczba symulacji B, wielkość podprzestrzeni $|m| < \min(p, n)$.
2. Powtarzaj procedurę dla $k = 1, \ldots, B$ z $C_{i,0} = 0$ dla każdego i.
   - Wylosuj zbiór zmiennych $m^* = \{i_1^*, \ldots, i_{|m|}^*\}$ z przestrzeni cech.
   - Dopasuj model $y \sim \mathbf{x}_{m^*}$ i oblicz wagi $w_n(i, m^*) \geqslant 0$ dla zmiennych $i \in m^*$. Ustaw $w_n(i, m^*) = 0$ jeżeli $i \notin m^*$.
   - $C_{i,k} = C_{i,k-1} + I\{i \in m^*\}$.
3. Dla wszystkich zmiennych i oblicz końcowe wagi:

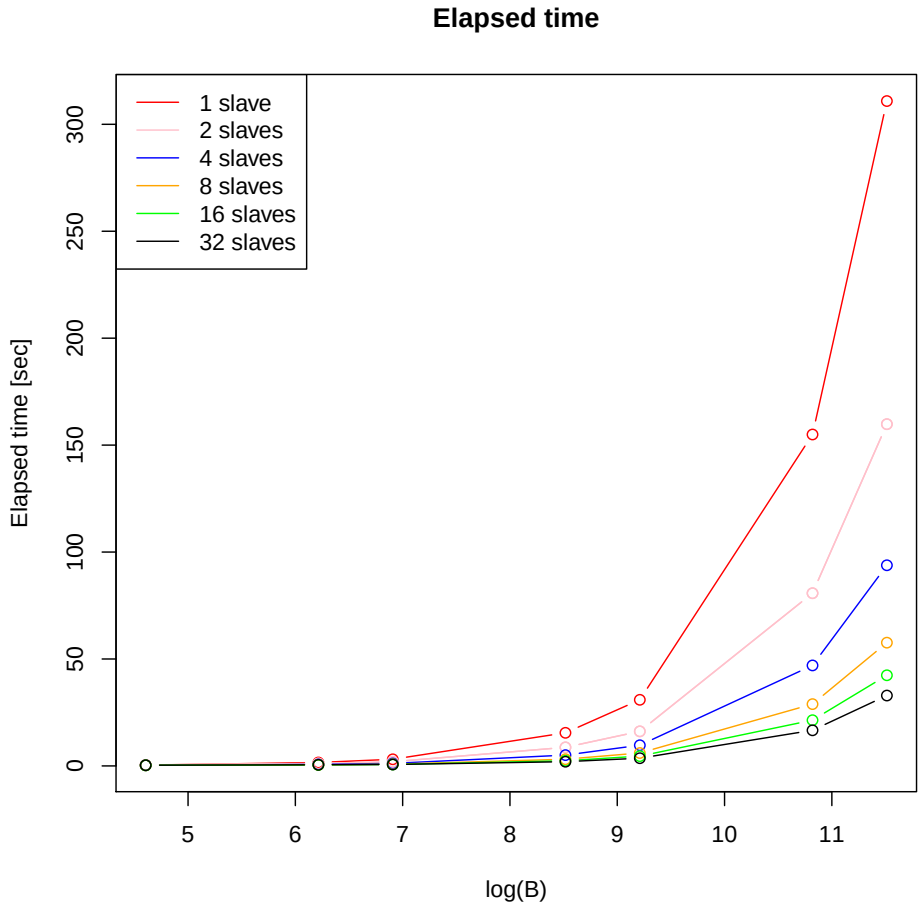
$$W_i^* = \frac{1}{C_{i,B}} \sum_{m^* : i \in m^*} w_n(i, m^*).$$

4. Posortuj zmienne wg końcowych wag W_i^*: $W_{i_1}^* \geqslant W_{i_2}^* \ldots \geqslant W_{i_p}^*$.
5. Wyjście: uporządkowana lista zmiennych $\{i_1, \ldots, i_p\}$.

Metoda RSM dla modelu liniowego



Czas obliczeń dla $p = 1000$, $n = 100$, $|m| = 50$.



Rysunek : Maszyna:2x Intel(R) Xeon(R) CPU E5-2630L @ 2.00GHz (6 cores, 12 threads) - 24 logical cores in total, 64 GB RAM

Metoda RSM- wybór wag $w_n(i, m)$

Wybór wag:

$$w_n(i, m) := T^2_{i,m},$$

gdzie $T_{i,m}$ oznacza statystykę T dla zmiennej i, obliczoną na podstawie dowolnego podmodelu m.

- Zauważmy, że:

$$\frac{T^2_{i,m}}{n - |m|} = \underbrace{(R^2_m - R^2_{m\setminus\{i\}})}_{\text{istotność zm. i}} \cdot \underbrace{\frac{1}{1 - R^2_m}}_{\text{dopasowanie modelu m}},$$

gdzie R^2_m jest współczynnikiem determinacji dla modelu m.

Asymptotyczna postać wag końcowych W_i^*

- Można pokazać asymptotyczną równoważność:

$$W_i^* \approx \frac{1}{|\mathcal{M}_{i,|m|}|} \sum_{m \in \mathcal{M}_{i,|m|}} \frac{MSEP(m \setminus \{i\}) - MSEP(m)}{MSEP(m)},$$

- $|\mathcal{M}_{i,|m|}|$ to liczba modeli o liczności $|m|$ które zawierają zmienną i,
- Błąd predykcji dla modelu m:

$$MSEP(m) := \lim_{n \to \infty} n^{-1} \mathbf{E}[||Y^* - \mathbf{X}_m \hat{\beta}_m||^2 | \mathbf{X}],$$

gdzie $Y^* = \mathbf{X}\beta + \varepsilon^*$, ε^* niezależna kopia ε.

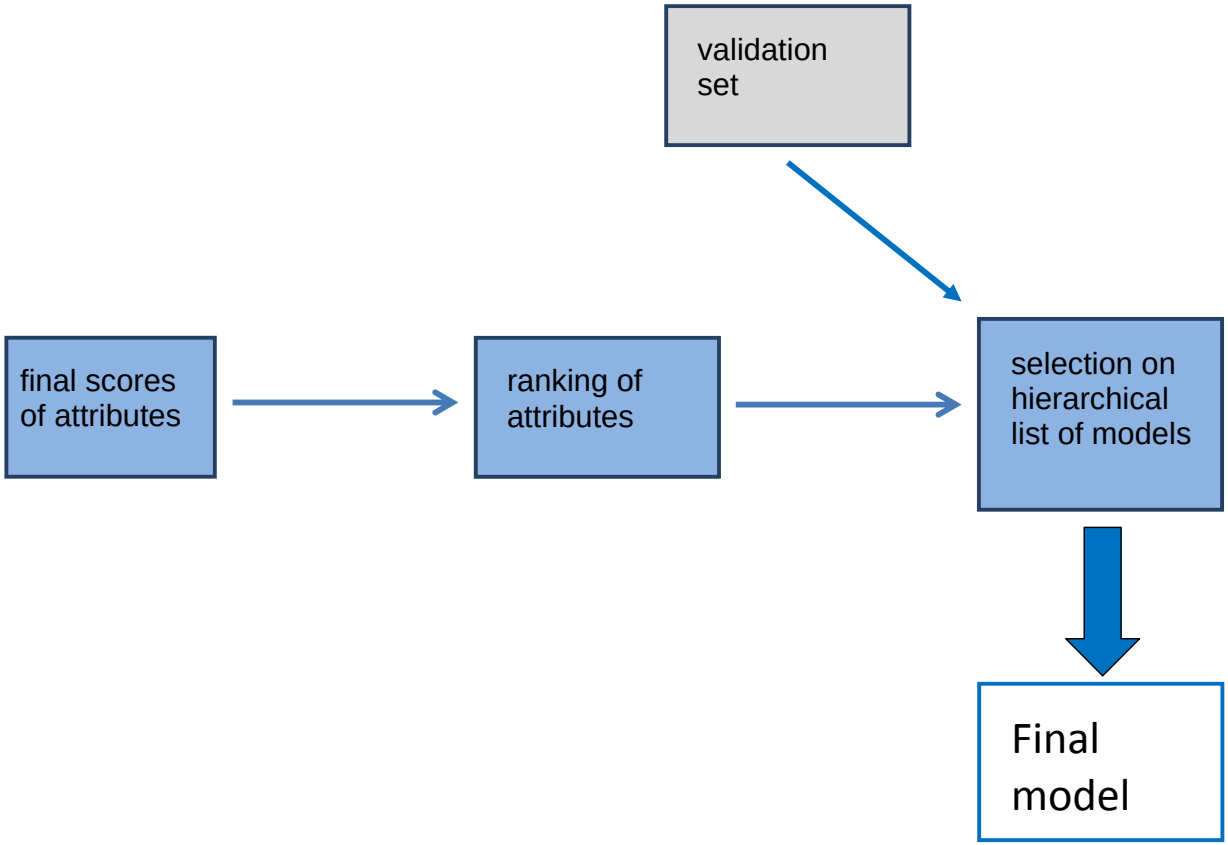
Procedura wyboru modelu:

1. Dane $(\mathbf{Y}, \mathbf{X})$ dzielone na część treningową: $(\mathbf{Y}^t, \mathbf{X}^t)$ oraz walidacyjną $(\mathbf{Y}^v, \mathbf{X}^v)$.
2. Procedura RSM jest realizowana na części treningowej. Zmienne są porządkowane wg. wag końcowych:

$$W^*_{i_1} \geqslant \ldots, \geqslant W^*_{i_p}.$$

3. Z zagnieżdżonej listy modeli $\{\{0\}, \{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, \ldots, i_{\min(n,p)-1}\}\}$ wybieramy model m_{opt} dla którego błąd na próbie walidacyjnej $n^{-1}||\mathbf{Y}^v - \mathbf{X}^v\hat{\beta}_{m_{\text{opt}}}||^2$ jest najmniejszy.
(tutaj: $\hat{\beta}_{m_{\text{opt}}}$- estymator ML oparty na modelu m_{opt}, obliczony na próbie $(\mathbf{Y}^t, \mathbf{X}^t)$).

Procedura wyboru modelu



Kryteria Informacyjne

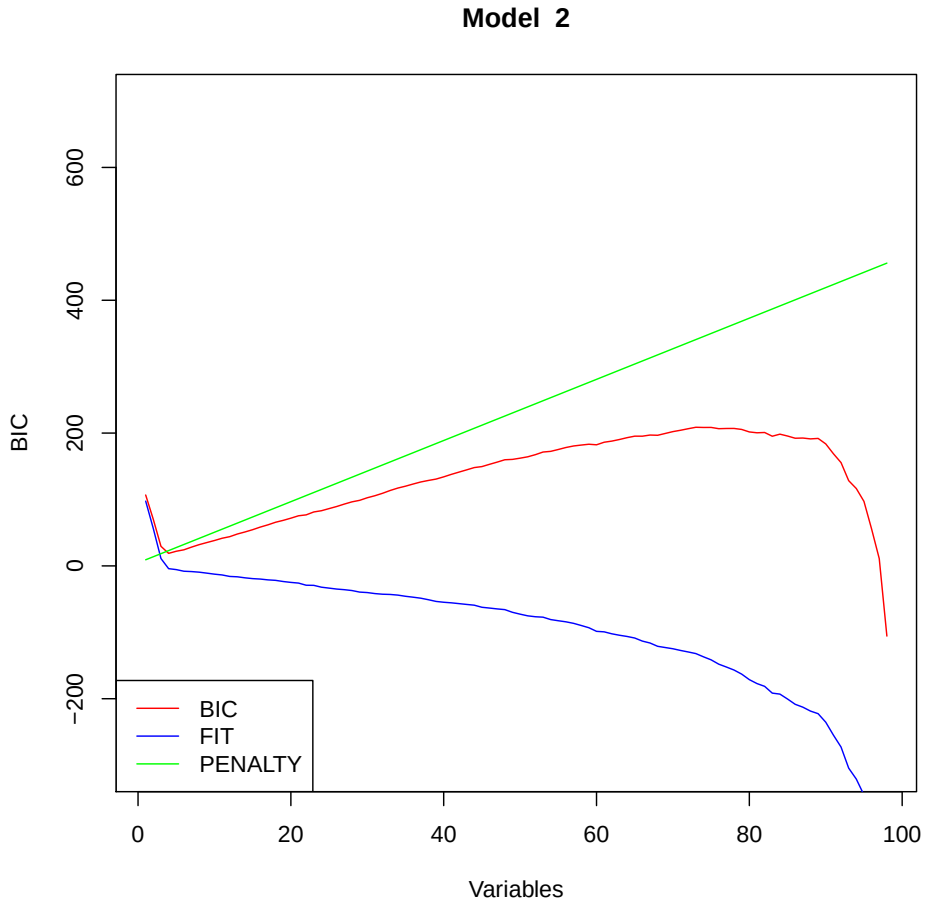
- **Wada procedury opisanej powyżej:** konieczność wydzielenia próby walidacyjnej (duży problem w sytuacji małej liczby obserwacji).
- Kryterium Bayesowskie:

$$BIC(m) = \underbrace{-2l(\hat{\beta}_m)}_{\text{dopasowanie modelu}} + \underbrace{log(n)|m|}_{\text{kara za liczbę parametrów}} \rightarrow \min,$$

gdzie: $l(\cdot)$ to funkcja log-wiarogodności, $|m|$ to liczba parametrów w modelu m.

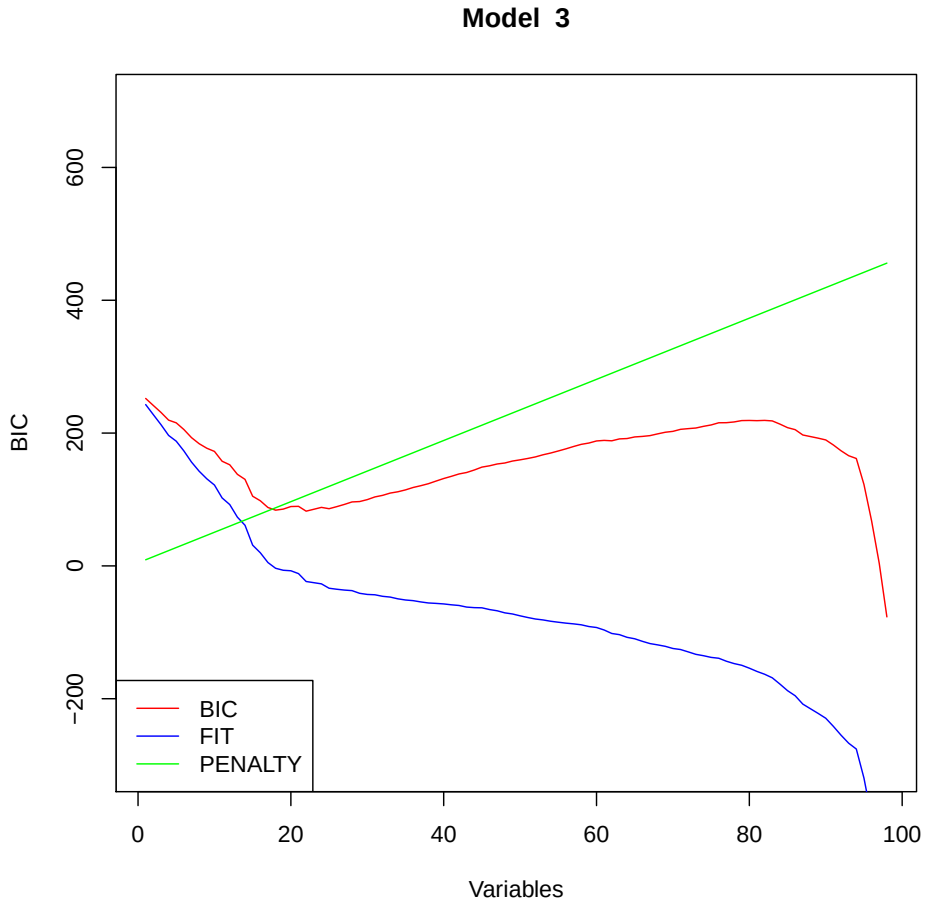
- **Procedura oparta na BIC:** z zagnieżdżonej rodziny $\{\{0\}, \{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, \ldots, i_{\min(n,p)-1}\}\}$ wyznaczonej na podstawie metody RSM wybieramy model które minimalizuje BIC.

Kryteria Informacyjne- problem



Rysunek : Problem: BIC działa niepoprawnie gdy liczba zmiennych jest duża w porównaniu z n (model prawdziwy t zawiera 3 zmienne).

Kryteria Informacyjne- problem



Rysunek : Problem: BIC działa niepoprawnie gdy liczba zmiennych jest duża w porównaniu z n (model prawdziwy t zawiera 10 zmiennych).

Wyniki symulacji- metody

- Metoda lasso.
- Metoda RSM + BIC.
- Metoda WRSM + BIC.
- Metoda Univariate + BIC.
- Metoda CAR + BIC [$\mathrm{CAR} = corr(y, P^{-1/2} X_{\mathrm{std}})$, P- correlation matrix of attributes].

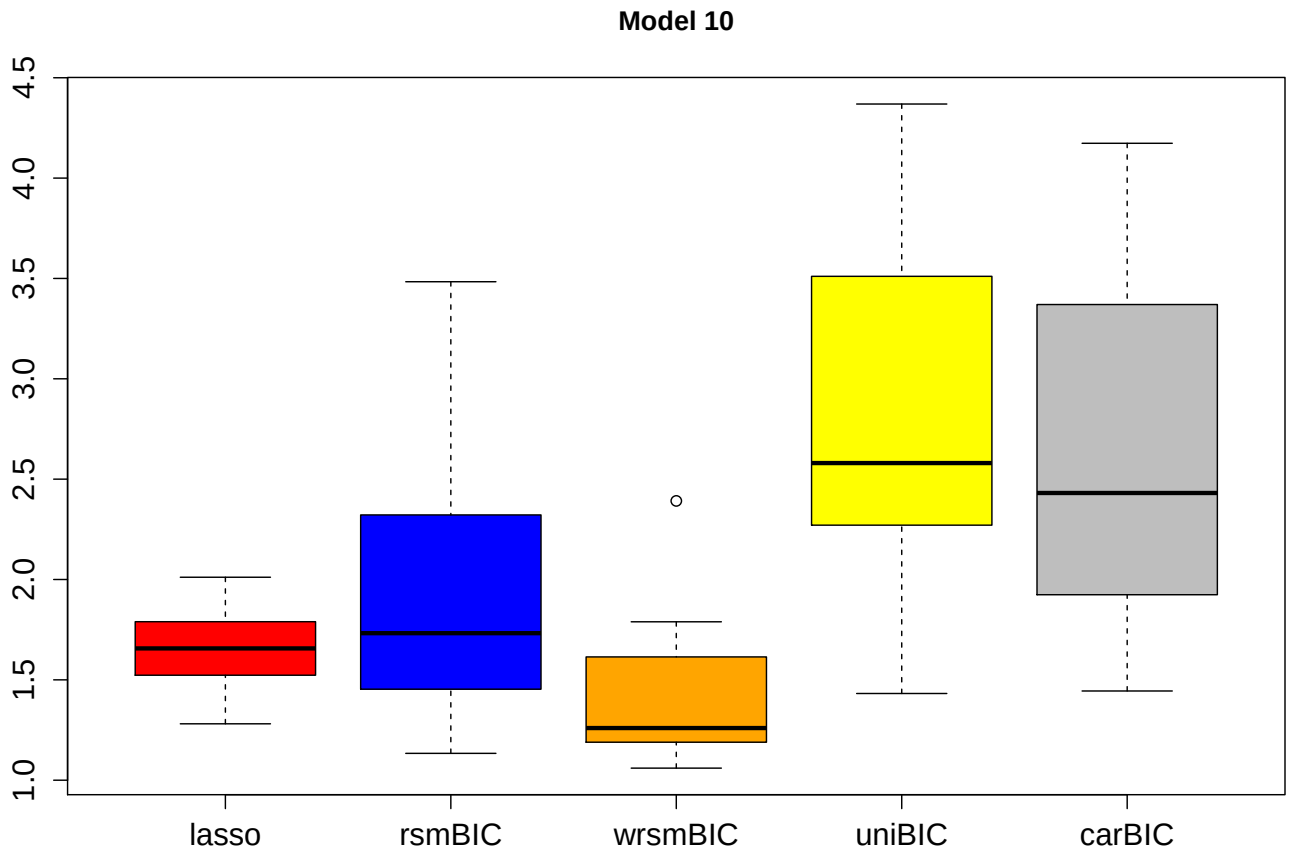
Punkt odcięcia:

- Sztywny punkt odcięcia: $(n-1)/2$.
- 5% spermutowanych kopii originalnych zmiennych ma większą korelacje z y niż zmienne oryginalne.

Wyniki symulacji- miary oceny

- (CS): pstwo wyboru modelu t: $P(\hat{t} = t)$,
- (PSR): $\mathbf{E}(|\hat{t} \cap t|/|t|)$,
- (FDR): $\mathbf{E}(|\hat{t} \setminus t|/|\hat{t}|)$,
- (PE): Błąd predykcji na niezależnym zbiorze testowym.
- (CO): pstwo poprawnego uporządkowania w pierwszym kroku procedury dwustopniowej: $P[\max_{i \notin t} T^2_{i,f} < \min_{i \in t} T^2_{i,f}]$.

Wyniki symulacji



Rysunek : Błędy predykcji dla wybranego modelu (model prawdziwy t zawiera 50 zmiennych).

Wyniki symulacji

Model	1 miejsce	2 miejsce	3 miejsce	4 miejsce	5 miejsce
1	lasso	carBIC	uniBIC	rsmBIC	wrsmBIC
2	rsmBIC	uniBIC	carBIC	lasso	wrsmBIC
3	wrsmBIC	rsmBIC	carBIC	uniBIC	lasso
4	rsmBIC	carBIC	uniBIC	wrsmBIC	lasso
5	wrsmBIC	rsmBIC	lasso	carBIC	uniBIC
6	wrsmBIC	lasso	rsmBIC	uniBIC	carBIC
7	wrsmBIC	rsmBIC	lasso	carBIC	uniBIC
8	carBIC	uniBIC	rsmBIC	wrsmBIC	lasso
9	wrsmBIC	rsmBIC	lasso	carBIC	uniBIC
10	wrsmBIC	lasso	rsmBIC	carBIC	uniBIC

Tabela : Ranking badanych metod ze względu na błąd predykcji.

Wyniki symulacji- PSR

Model	$\lvert t \rvert$	lasso	rsmBIC	wrsmBIC	uniBIC	carBIC	Max. PSR
1	1	0.000	0.367	0.433	0.467	0.467	UNI, CAR
2	3	1.000	1.000	1.000	1.000	1.000	wszystkie
3	10	1.000	1.000	1.000	1.000	1.000	wszystkie
4	5	1.000	1.000	1.000	1.000	1.000	wszystkie
5	15	0.996	0.838	0.973	0.816	0.829	lasso
6	15	0.998	0.769	0.940	0.731	0.733	lasso
7	20	1.000	0.982	0.995	0.963	0.967	lasso
8	8	0.854	0.817	0.888	0.829	0.833	WRSM
9	50	0.995	0.922	0.979	0.845	0.870	lasso
10	50	1.000	0.960	0.991	0.893	0.908	lasso

Tabela : Wskaźniki PSR.

Wyniki symulacji- FDR

Model	lasso	rsmBIC	wrsmBIC	uniBIC	carBIC	Min. FDR
1	1.000	0.954	0.980	0.926	0.931	UNI
2	0.124	0.021	0.608	0.033	0.025	RSM
3	0.410	0.290	0.074	0.384	0.358	WRSM
4	0.329	0.069	0.454	0.123	0.109	RSM
5	0.216	0.179	0.199	0.203	0.220	RSM
6	0.297	0.260	0.156	0.231	0.191	WRSM
7	0.271	0.217	0.018	0.312	0.260	WRSM
8	0.111	0.074	0.467	0.050	0.059	WRSM
9	0.419	0.208	0.100	0.233	0.198	WRSM
10	0.427	0.327	0.097	0.302	0.275	WRSM

Tabela : Wskaźniki FDR.

RSM- wnioski

- WRSM zazwyczaj działa lepiej niż konkurencyjne metody (biorąc pod uwagę PE).
- FDR jest zazwyczaj mniejsze dla RSM niż dla metody lasso oraz metody univariate.
- Stosując metodę RSM otrzymujemy mniej złożone modele (jest to potwierdzone przez eksperymenty na zbiorach rzeczywistych).
- Zastosowanie wersji ważonej (WRSM) pozwala zmniejszyć liczbę symulacji i w ten sposób zredukować koszt obliczeniowy.

RSM- plany

- Zastosowanie metody RSM dla innych modeli (n.p. modelu logistycznego).
- Połączenie metody RSM i metod wyboru zmiennych wykorzystujących kryteria informacyjne (zastosowanie innych kryteriów informacyjnych, modyfikacja metody znajdującej punkt odcięcia).
- Nowe warianty metody WRSM (n.p. użycie wag końcowych RSM jako wag zmiennych w WRSM).
- Dopracowanie pakietu zawierającego implementacje równoległą.

Literatura

1. J. Mielniczuk, P. Teisseyre *Using Random Subspace Method for Prediction and Variable Importance Assessment in Linear Regression*, Computational Statistics and Data Analysis, http://www.sciencedirect.com/science/article/pii/S0167947312003477.
2. T. K. Ho, *The Random Subspace Method for constructing decision forests*, IEEE Trans. Pattern Anal. Machine Intell., Vol. 20, No. 8, pages 832–844, 1998.
3. L. Breiman, *Random forests*, Machine Learning, Vol. 45, No. 1, pages 5–32, 2001.
4. C. Lai, M. J. T. Reinders, L. Wessels, *Random Subspace Method for multivariate feature selection*, Pattern Recognition Letters, Vol. 27, pages 1067-1076, 2006.
5. M. Draminski et. al. *Monte carlo feature selection for supervised classification*, BIOINFORMATICS, 24(1):110-117, 2008.

Dziękuje za uwagę!