Kryteria informacyjne

ALEKSANDER MILACH

27 kwietnia 2021

Prosty model linowy

Rozważmy model liniowy

$$Y = X\beta + \epsilon,$$

gdzie Y to wartości zmiennej objaśnianej, macierz X jest znane, ϵ to nieznany wektor szumu o rozkładzie $N(0, \sigma^2)$, a β to nieznany wektor parametrów, który chcemy estymować. Pierwszym standardowym podejściem, jest metoda najmniejszych kwadratów, która w tym przypadku daje ten sam wynik, co estymacja metodą największej wiarogodności. Otrzymujemy wówczas wzór

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y.$$

Ten znany estymator ma dobre własności, w szczególności jest on nieobciążony i ma rozkład normalny $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.

Problemy przy rosnącej liczbie zmiennych

Zauważmy najpierw, że gdy p zbliża się do n, rośnie wariancja estymatora. Za jej wartość będą odpowiadały wartości na przekątnej macierzy $(X'X)^{-1}$, zatem w przypadku, gdy wartości na przekątnej macierzy $X'X$ będa bliskie zeru, wariancja estymatora może bardzo wzrosnąć. Skutkować to będzie spadkiem mocy testu na istotność współczynnika dla danej zmiennej, nawet do ustalonej wartości prawdopodobieństwa błędu pierwszego rodzaju - nasz model będzie wykrywał bardzo mało zmiennych i wiele z nich będzie fałszywymi odkryciami.

Przykład 1

Klasyczne kryteria informacyjne

Pewnym rozwiązaniem powyższych problemów są kryteria informacyjne. Stosujemy je w następujący sposób. Dla ustalonego zbioru modeli obliczamy wartość kryterium, poczym wybieramy model, dla którego wartość kryterium jest najmniejsza. Dwa najstarsze i najbardziej znane to AIC i BIC.

$$AIC(\hat{\beta}) = -2\log(L(\hat{\beta})) + 2k$$

$$BIC(\hat{\beta}) = -2\log(L(\hat{\beta})) + k\log(n),$$

gdzie k to ilość niezerowych współrzędnych wektora współczynników $\hat{\beta}$. Pierwszy składnik odpowiada za jakość estymacji, natomiast drugi ma za zadanie ograniczać ilość zmiennych w modelu.

Przykład 2

Problem ilości porównywanych modeli

W powyższym przykładzie porównaliśmy ze sobą 5 narzuconych modeli przy pomocy AIC. W szczególności nie mamy podstaw do uważania, że wśród nich jest optymalny model względem naszego kryterium. Aby znaleźć najlepszy model powinniśmy sprawdzić wszystkie możliwe modele. Jest ich 2^p, ponieważ każdą z p zmiennych możemy uznać za istotną bądź nie.
Ilość modeli bardzo szybko rośnie wraz z p i dla $p > 50$ obliczenie wartośći kryterium dla każdego z nich jest właściwie niemożliwe, przez czas wykonania odpowiednich obliczeń.

Problem ilości porównywanych modeli c.d.

W praktyce stosuje się metody, które z dużym prawdopodobieństwem znajdują modele optymalne, bądź niewiele różniące się od optymalnych.

- procedura forward - startujemy z pustego modelu i dodajemy zmienne pojedyńczo w taki sposób, aby w każdym kroku maksymalnie poprawiać wartość kryterium
- procedura backward - startujemy z modelu ze wszystkimi zmiennymi i usuwamy zmienne w taki sposób, aby w każdym kroku maksymalnie poprawiać wartość kryterium
- procedura stepwise - na zmianę dodajemy i usuwamy zmienne zgodnie z powyższą regułą

Wady AIC i BIC

Przypuśćmy, że macierz planu X jest ortogonalna i przeskalowana tak, że $X^T X = I_{p \times p}$. Pomimo tego, że szanse, że w rzeczywistych danych do tego dojdzie są bardzo niewielkie, na tym prostym przykładzie najłatwiej pokazać wady AIC i BIC, które występują w ogólniejszych przypadkach. Dzięki orgotonalności zmienne objaśniające są od siebie niezależne, a wzór na wartość współczynnika odpowiadającego zmiennej obliczonego przy pomocy metody najmniejszych kwadratów sprowadza się do $\hat{\beta}_j = X_j^T Y$, gdzie X_j jest kolumną macierzy X. Wówczas powyższe estymatory mają rozkład normalny $N(\beta_j, \sigma^2)$, a hipotezy $\beta_j = 0$ przeciwko $\beta_j \neq 0$ testujemy przy pomocy statystyki $Z_j = \sqrt{n}\hat{\beta}_j/\sigma$, która przy hipotezie ma rozkład standardowy normalny.

Wady AIC i BIC c.d.

W tej sytuacji składnik log-wiarogodności ma postać:

$$-2\log(L(\hat{\beta})) = c + ||y - \sum_{j=1}^{p} X_j\hat{\beta}_j||^2/\sigma^2 = c + (Y - X\hat{\beta})'(Y - X\hat{\beta})/\sigma^2 =$$

$$= c + (Y'Y + \hat{\beta}'X'X\hat{\beta} - 2Y'X\hat{\beta})/\sigma^2 = c + (Y'Y + \hat{\beta}'\hat{\beta} - 2\hat{\beta}'\hat{\beta})/\sigma^2 =$$

$$= c + (Y'Y - \hat{\beta}'\hat{\beta})/\sigma^2 = c + (Y'Y - ||\hat{\beta}||^2)/\sigma^2$$

Wady AIC i BIC c.d.

Zatem dodanie do modelu zmiennej X_j zmniejsza składnik log-wiarogodności o $\hat{\beta}_j^2/\sigma^2$, a zwiększa karę o 2. Korzystając z rozkładu estymatora mamy, że wartość kryterium AIC zmaleje, gdy wartość bezwzględna estymatora będzie większa niż $\sqrt{2}\sigma$. Oznacza to, że prawdopodobieństwo błędu pierwszego rodzaju wynosi $2(1 - \Phi(\sqrt{2})) = 0.157$. Dodanie zmiennej do modelu odbywa się niezależnie od tego, ile zmiennych już zostało do niego wybranych, a ilość fałszywych odkryć rośnie liniowo z p.

Wady AIC i BIC c.d.

Podobnie przy ortogonalności macierzy planu zachowuje się BIC. Z analogicznych rozważań otrzymamy, że wartość kryterium zmaleje, gdy wartość bezwzględna estymatora będzie większa niż $\sqrt{\log(n)}\,\sigma$. Prowadzi to do prawdopodobieństwa błędu pierwszego rodzaju na poziomie $2(1-\Phi(\sqrt{\log(n)}))$. Przykładowo dla $n = 100$ jest to 0.032.
Mimo tego kryterium BIC ma pewną przewagę, jest zgodne, przy $n \to \infty$ prawdopodobieństwo błędu maleje do zera.

RIC

W związku w wadami standardowych metod zaszła potrzeba znalezienia nowych kryteriów o użytecznych własnościach w przypadkach, gdy $n \sim p$. Pierwszym rozwiązaniem było kryterium RIC postaci

$$RIC(\hat{\beta}) = -2\log(L(\hat{\beta})) + 2k\log(p)$$

Sprawdzając, podobnie jak dla powyższych kryteriów, prawdopodobieństwo błędu pierwszego rodzaju w sytuacji ortogonalnej dochodzimy do zależności $|\hat{\beta}_j| > \sigma\sqrt{2\log p}$.
Dla rosnącego p wartość bezwzględna estymatora musi być coraz większa, aby zmienna została włączona do modelu, równoważnie prawdopodobieństwo błędu pierwszego rodzaju maleje.
Konsekwentnie, frakcja fałszywych odkryć maleje do 0, jednak bardzo wolno ze względu na pojawiający się czynnik $\log p$.

RIC c.d.

W praktyce dla $p = 10$ family wise error rate (FWER), czyli prawdopodobieństwo uzyskania niezerowej liczby fałszywych odkryć wynosi 0,35, a dla $p = 1000$ nadal około 0,2.

mBIC

Użyteczna w praktyce okazała się dopiero następująca modyfikacja BIC:

$$mBIC(\hat{\beta}) = -2\log(L(\hat{\beta})) + k\log n + 2k\log(p/E)$$

mBIC łączy kary BIC i RIC. Dzięki temu dla dużych wartości n czynnik z kryterium BIC będzie odpowiadał za minimalizowanie prawdopodobieństwa błędu pierwszego rodzaju, zaś dla dużego p będzie to kara z kryterium RIC.

mBIC c.d.

Stała E jest oczekiwaną liczbą istotnych zmiennych w naszym modelu, gdy nie zakładamy żadnej konkretnej ilości istotnych zmiennych należy przyjąć $E = 4$. Już dla tego ogólnego przypadku kryterium osiąga lepsze wyniki w symulacjach.
Dla $n = 150$ już dla $p \geq 10$ FWER jest kontrolowane na poziomie 0,1, dla $p = 1000$ otrzymujemy ograniczenie na poziomie 0,065, natomiast dla $n = 500$ FWER jest mniejsze niż 0,05 dla $p \geq 10$ i mniejsze niż 0,035 dla $p = 1000$.

mBIC2

Powyższe kryterium można zmodyfikować, aby osiągało minimalne wartości frakcji fałszywych odkryć (FDR), nie zaś FWER jak powyżej. Kryterium realizującym tę własność jest mBIC2.

$$mBIC2(\hat{\beta}) = -2\log(L(\hat{\beta})) + k\log n + 2k\log(p/E) - 2\log k!$$

Zauważmy, że w tym przypadku zmniejszamy karę w stosunku do mBIC, właśnie ze względu na osiągnięcie jak lepszej kontroli FDR. Działanie to znajduje uzasadnienie teoretyczne w związku z korektą Benjaminiego - Hochberga na wielokrotne testowanie. Kryterium mBIC osiąga kontrolę FDR na poziomie zależnym od n - $\mathrm{FDR}_n \sim (n\log n)^{-1/2}$. Zatem dla $n = 100$ spodziewamy się kontroli FDR na poziomie około 0,1, dla $n = 1000$ na poziomie 0,031.

Przykład 3

Wyniki przykładu 3

	AIC	BIC	mBIC	mBIC2
Moc	1	1	0,25	1
FDR	0,64	0,23	0	0,2

Symulacje dla mBIC2

We wszystkich symulacjach rozważamy model regresji liniowej dla $n = p = 500$, błąd ϵ pochodzi z rozkładu normalnego $N(0, I)$, k - liczba niezerowych współczynników wektora β przyjmuje wartości $k \in \{10, 20, 40, 60, 80, 100\}$. Wyróżniamy dwa poziomy siły sygnału. Słaby sygnał utożsamiamy z wektorem β postaci:

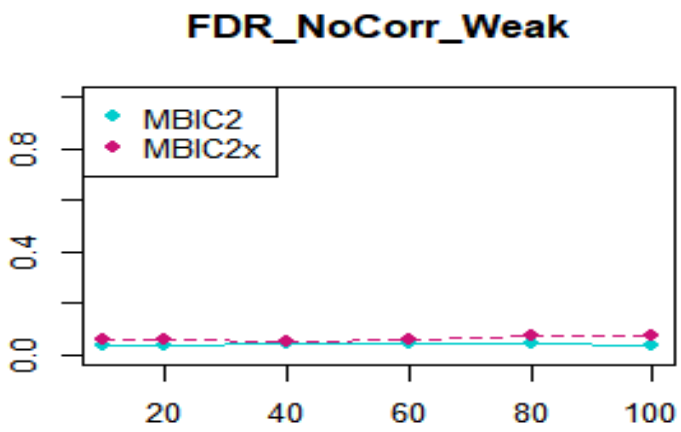
$$\beta_1 = \beta_2 = ... = \beta_k = 1.3\sqrt{2 \log p}$$

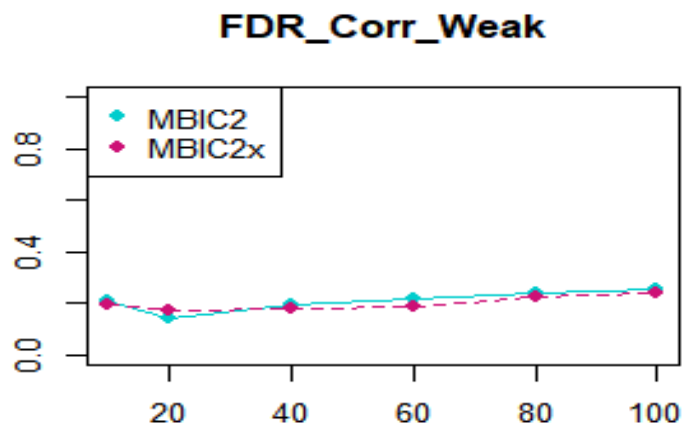
natomiast silny sygnał z

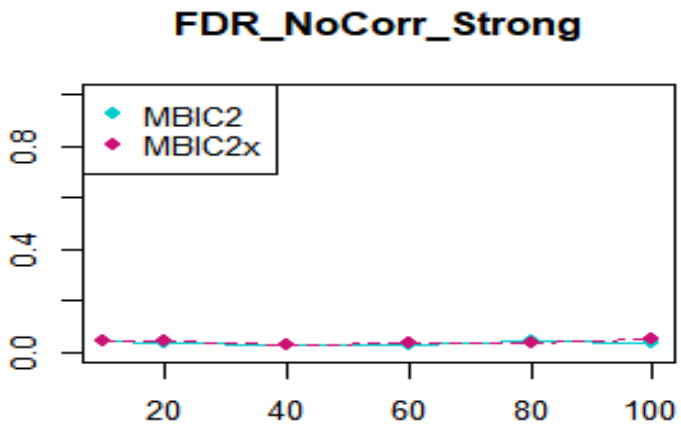
$$\beta_1 = \beta_2 = ... = \beta_k = 2\sqrt{2 \log p}.$$

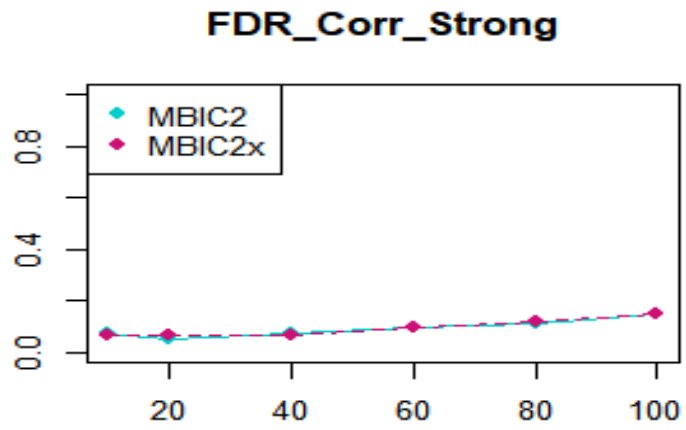
Rozważamy dwa przypadki: niezależnych zmiennych objaśniających, gdzie $\Sigma = I$ i skorelowanych zmiennych objaśniających, gdzie $\Sigma_{i,j} = 1$, gdy $i = j$ i $\Sigma_{i,j} = 0,5$, gdy $i \neq j$.

Symulacje dla mBIC2 c.d.

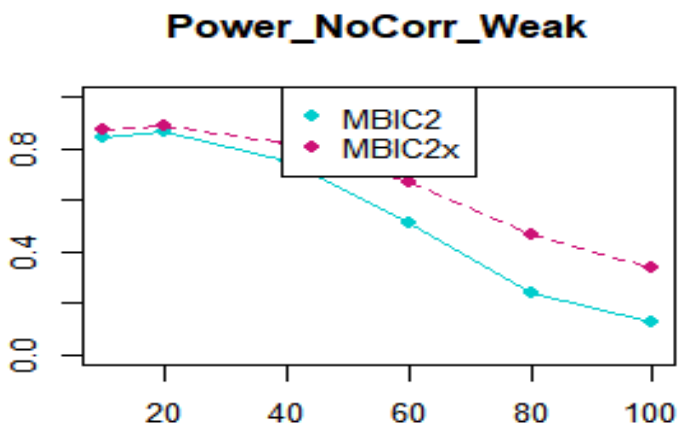


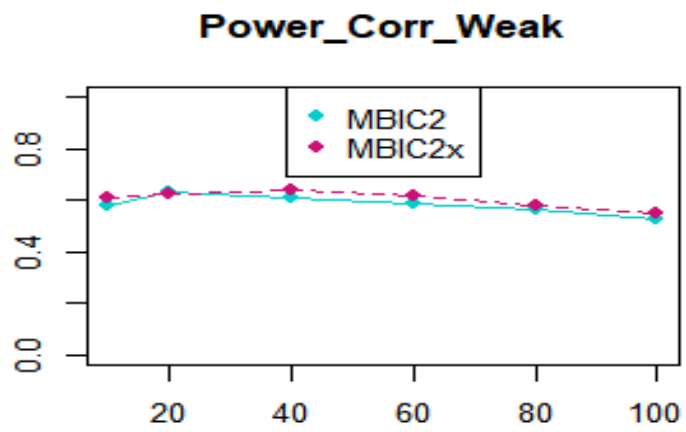


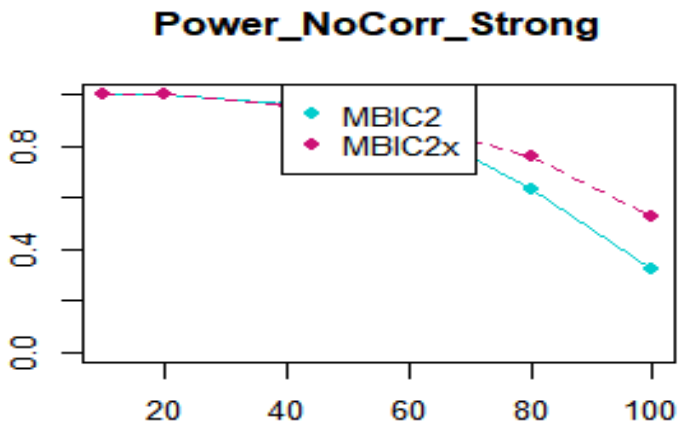


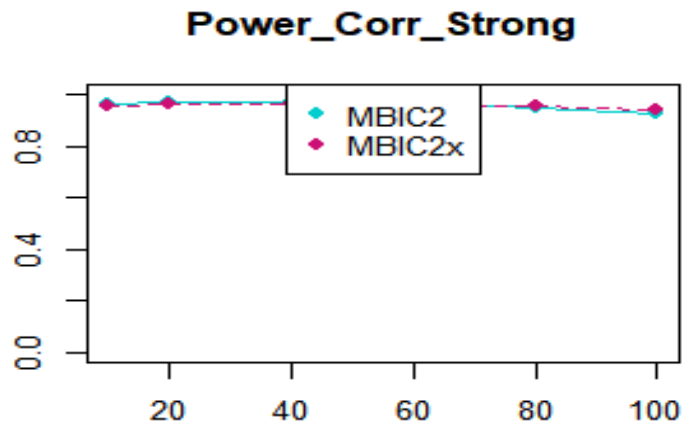


Symulacje dla mBIC2 c.d.

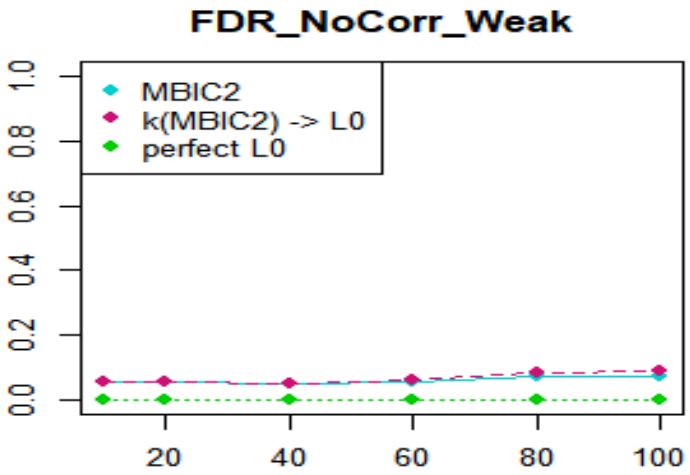


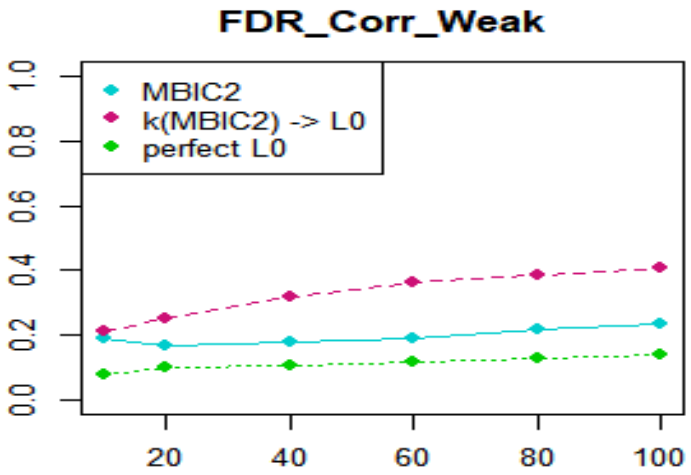


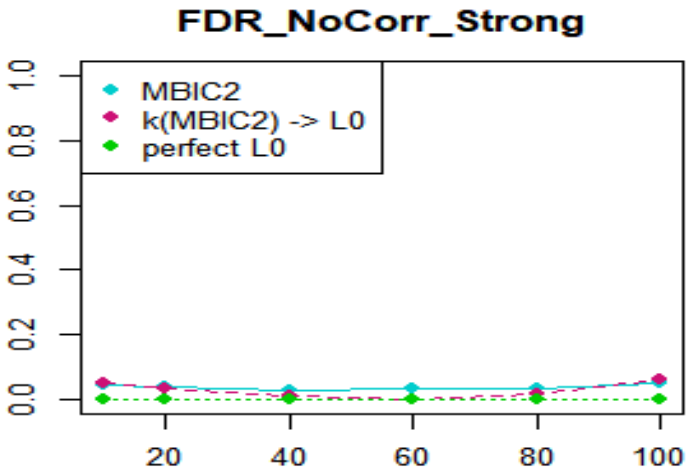


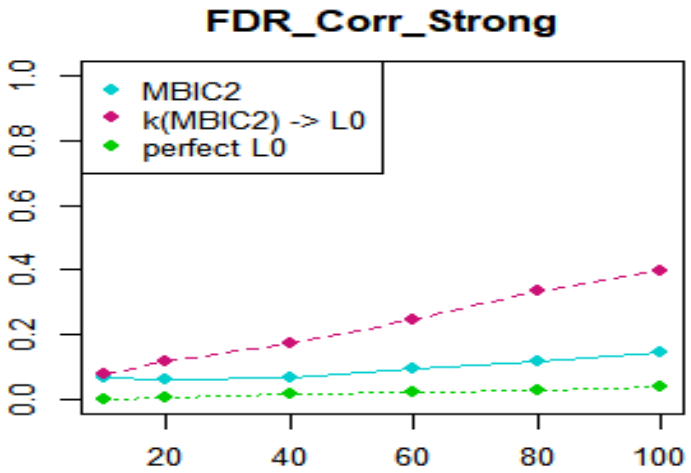


Inne metody

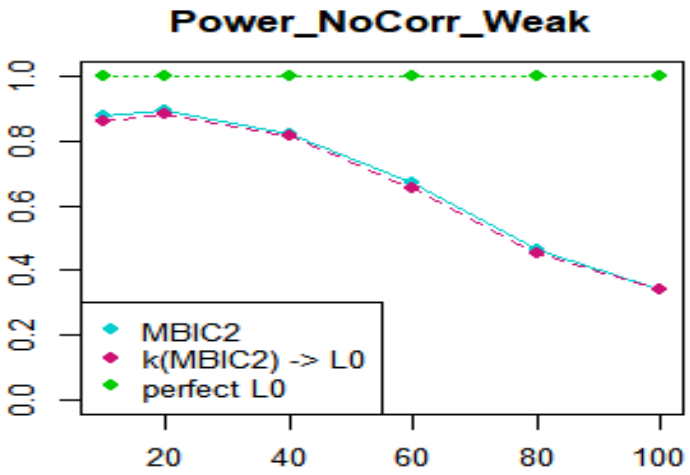

FDR_NoCorr_Weak
MBIC2
k(MBIC2) -> L0
perfect L0
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

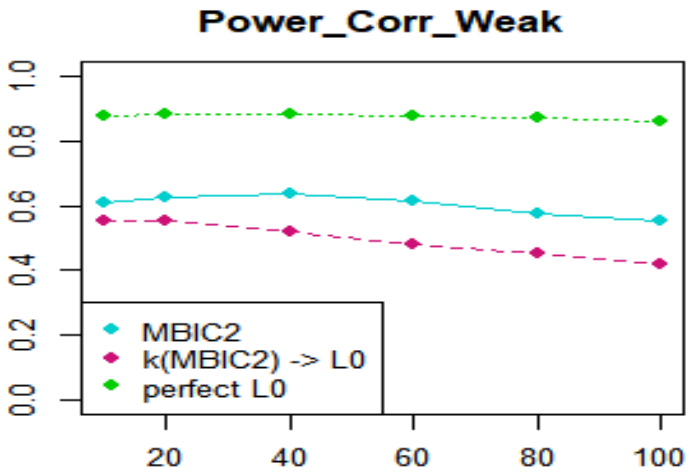

FDR_Corr_Weak
MBIC2
k(MBIC2) -> L0
perfect L0
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

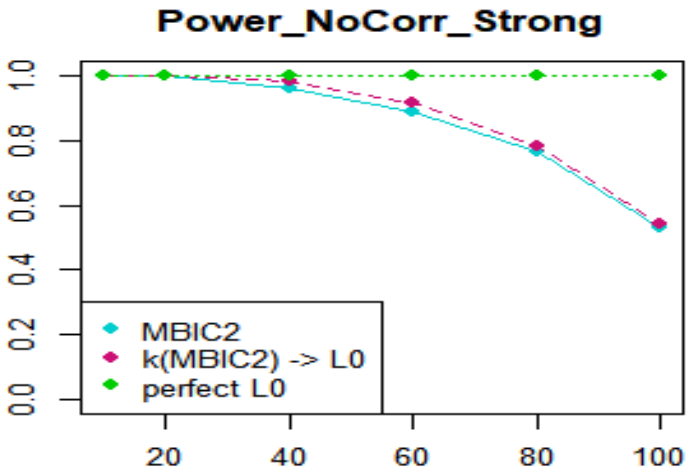

FDR_NoCorr_Strong
MBIC2
k(MBIC2) -> L0
perfect L0
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

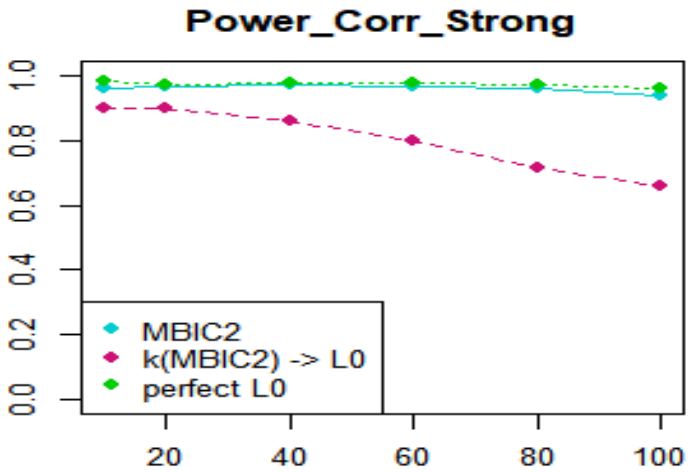

FDR_Corr_Strong
MBIC2
k(MBIC2) -> L0
perfect L0
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

Inne metody c.d.









Jak to otrzymałem?

CDN.