Metody regulacyjne w regresji wielorakiej - regresja grzbietowa i LASSO

Sylwia Patrijas

2025-05-23

Celem niniejszego raportu jest analiza i porównanie różnych metod estymacji współczynników regresji w wielowymiarowym modelu liniowym, w którym liczba zmiennych predykcyjnych jest równa liczbie obserwacji ($n = p = 500$). Zastosowane zostaną klasyczne i nowoczesne techniki regresyjne, takie jak regresja grzbietowa, regresja LASSO oraz adaptacyjne LASSO.

W analizie zostanie wykorzystany model liniowy postaci $Y = X\beta + \epsilon$, gdzie: X_{nxp} - macierz planu, $n = p = 500$, $\epsilon \sim N(0, I)$. Rozważone zostanie sześć przypadków, w zależności od wektora współczynników regresji, który wyraża się następującymi wzorami:

- $\beta_1 = ... = \beta_k = 4$, $\beta_{k+1} = ... = \beta_p = 0$,
- $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$, $\beta_{k+1} = ... = \beta_p = 0$

dla $k \in \{5, 20, 100\}$.

PRZYPADEK MACIERZY ORTONORMALNEJ

Jako pierwszy zostanie rozważony przypadek **ortonormalnej macierzy planu** X, to znaczy takiej, że $X^T X = I$. Wygeneruję taką macierz, korzystając z polecenia *randortho*, a następnie sprawdzę poprawność wygenerowanej macierzy, wyliczając $X^T X$.

Maksymalna różnica między elementami na przekątnej iloczynu macierzy $X^T X$ a liczbą 1 to wartość **bardzo bliska zera**. Podobnie, maksymalna różnica między elementami poza przekątną tego iloczynu a liczbą 0 jest **bardzo zbliżona do zera**. Zatem, potwierdza to **poprawność wygenerowania macierzy ortonormalnej**, gdyż niewielkie różnice pojawiają się z powodu błędów numerycznych.

OBCIĄŻENIE, WARIANCJA I MSE DLA ESTYMATORA W REGRESJI GRZBIETOWEJ

Na początku dla każdego $i \in \{1, ..., p\}$ wyliczę teoretycznie obciążenie, wariancję i błąd średniokwadratowy dla estymatora $\hat{\beta}_i^{RR}$ uzyskanego za pomocą regresji grzbietowej z parametrem wygładzającym γ.

Regresja grzbietowa polega na szukaniu estymatora $\hat{\beta}_{RR}$ w następującej postaci: $\hat{\beta}_{RR} = argmin_{b \in \mathbb{R}^p} L(b)$, gdzie $L(b) = ||Y - Xb||^2 + \gamma||b||^2$, a $\gamma > 0$ - parametr. Wynika z tego, iż owy estymator regresji grzbietowej wyraża się następującym wzorem:

$$\hat{\beta}_{RR} = (X'X + \gamma I)^{-1} X'Y.$$

W naszym przypadku, $X'X = I$ oraz $\sigma^2 = 1$. Z tego wynika, iż estymator regresji grzbietowej wygląda następująco: $\hat{\beta}_{RR} = (I + \gamma I)^{-1} X'Y = \frac{1}{1+\gamma} X'Y$ i (na podstawie tego, iż jest on przekształceniem liniowym wektora $Y \sim N(X\beta, \sigma^2 I)$) pochodzi on z rozkładu normalnego $N(\frac{1}{1+\gamma} X'X\beta, \sigma^2 \frac{1}{(1+\gamma)^2} X'X) = N(\frac{1}{1+\gamma}\beta, \frac{\sigma^2}{(1+\gamma)^2} I) = N(\frac{1}{1+\gamma}\beta, \frac{1}{(1+\gamma)^2} I)$. Stąd mamy, że:

$$\hat{\beta}_i^{RR} \sim N(\frac{1}{1+\gamma}\beta_i, \frac{1}{(1+\gamma)^2}).$$

Na tej podstawie możemy wyliczyć następujące wielkości:

- **obciążenie estymatora** $\hat{\beta}_i^{RR}$: $Bias(\hat{\beta}_i^{RR}) = E(\hat{\beta}_i^{RR}) - \beta_i = \frac{1}{1+\gamma}\beta_i - \beta_i = \beta_i \cdot (\frac{1}{1+\gamma} - 1) = -\frac{\gamma}{1+\gamma}\beta_i$,
- **wariancja estymatora** $\hat{\beta}_i^{RR}$: $Var(\hat{\beta}_i^{RR}) = \frac{1}{(1+\gamma)^2}$,
- **błąd średniokwadratowy estymatora** $\hat{\beta}_i^{RR}$: $MSE(\hat{\beta}_i^{RR}) = [Bias(\hat{\beta}_i^{RR})]^2 + Var(\hat{\beta}_i^{RR}) = [-\frac{\gamma}{1+\gamma} \cdot \beta_i]^2 + \frac{1}{(1+\gamma)^2} = \frac{\gamma^2}{(1+\gamma)^2} \cdot \beta_i^2 + \frac{1}{(1+\gamma)^2} = \frac{\gamma^2\beta_i^2+1}{(1+\gamma)^2}$.

OPTYMALNY PARAMETR γ DLA REGRESJI GRZBIETOWEJ

W tej części, dla każdego z sześciu omawianych przypadków, wyznaczę teoretycznie parametr γ, który umożliwia osiągnięcie minimalnej wartości błędu średniokwadratowego $MSE = E||\hat{\beta} - \beta||^2$.

Na podstawie wyliczeń z poprzedniej części możemy wywnioskować, że:

$$MSE(\hat{\beta}^{RR}) = \sum_{i=1}^{p} MSE(\hat{\beta}_i^{RR}) = \sum_{i=1}^{p} \frac{\gamma^2\beta_i^2 + 1}{(1+\gamma)^2} = \frac{p + \gamma^2||\beta||^2}{(1+\gamma)^2} = f(\gamma).$$

Następnie, wybieramy taki parametr γ, że:

$$\gamma_{opt} = argmin_\gamma f(\gamma) = argmin_\gamma \frac{p + \gamma^2||\beta||^2}{(1+\gamma)^2}.$$

Aby to zrobić, wyliczę pochodną funkcji $f(\gamma)$ względem γ, a następnie przyrównam ją do 0:

$$\frac{df}{d\gamma} = \frac{2\gamma||\beta||^2 \cdot (1+\gamma)^2 - (p + \gamma^2||\beta||^2) \cdot 2 \cdot (1+\gamma)}{(1+\gamma)^4} = \frac{2\gamma||\beta||^2 \cdot (1+\gamma) - (p + \gamma^2||\beta||^2) \cdot 2}{(1+\gamma)^3} =$$

$$= \frac{2\gamma||\beta||^2 + 2\gamma^2||\beta||^2 - 2p - 2\gamma^2||\beta||^2}{(1+\gamma)^3} = \frac{2\gamma||\beta||^2 - 2p}{(1+\gamma)^3},$$

$$\frac{df}{d\gamma} = 0 \Leftrightarrow 2\gamma||\beta||^2 - 2p = 0 \Leftrightarrow \gamma||\beta||^2 = p \Leftrightarrow \gamma = \frac{p}{||\beta||^2}.$$

Wynika z tego, że:

$$\gamma_{opt} = \frac{p}{||\beta||^2}.$$

Znając już teoretyczny wzór, wyznaczę te wartości dla wszystkich sześciu omawianych przypadków i przedstawię uzyskane wyniki w tabeli, gdzie beta1 oznacza wektor $\beta_1 = ... = \beta_k = 4$, $\beta_{k+1} = ... = \beta_p = 0$, a beta2 - $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$, $\beta_{k+1} = ... = \beta_p = 0$.

Table 1: Optymalne wartości γ

	k = 5	k = 20	k = 100
beta1	6.25	1.56	0.31
beta2	6.25	6.25	6.25

OPTYMALNY PARAMETR λ DLA LASSO

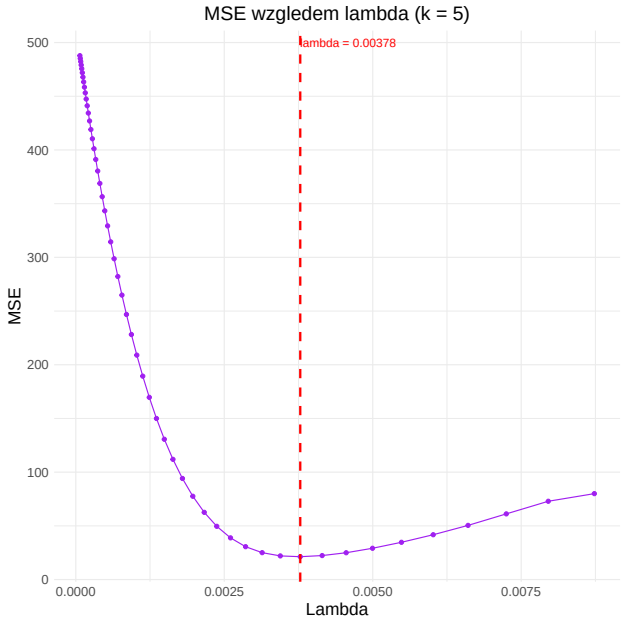
W tej części, dla każdego z sześciu omawianych przypadków, wyznaczę parametr λ dla LASSO, który umożliwia osiągnięcie minimalnej wartości błędu średniokwadratowego $MSE = E||\hat{\beta} - \beta||^2$.

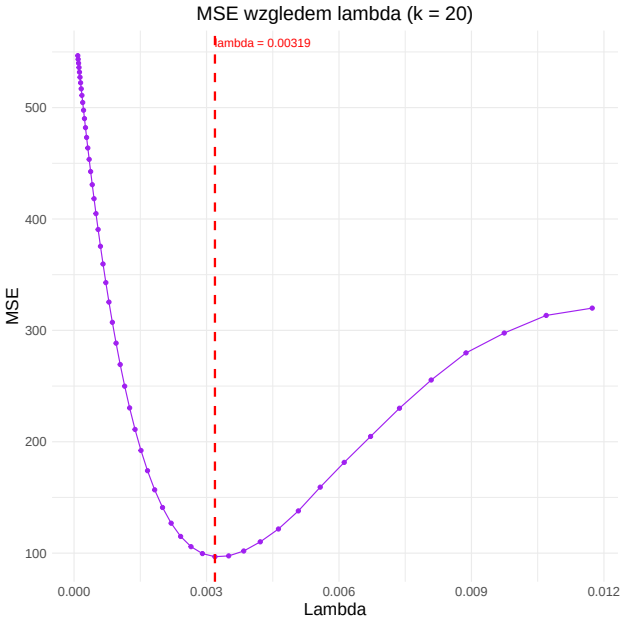
LASSO polega na szukaniu estymatora $\hat{\beta}_L$ w następującej postaci: $\hat{\beta}_L = argmin_{b \in \mathbb{R}^p} L(b)$, gdzie $L(b) = \frac{1}{2}||Y - Xb||_2^2 + \lambda||b||_1$, a $\lambda > 0$ - parametr.

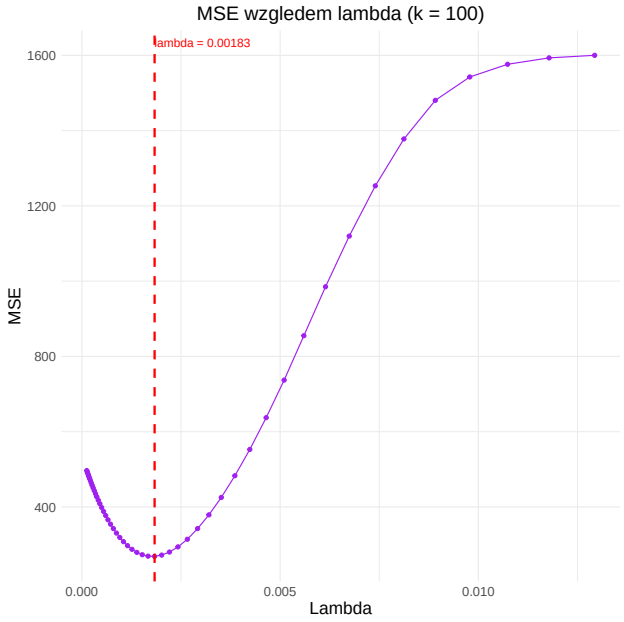
W przeciwieństwie do regresji grzbietowej, gdzie możliwe było analityczne wyznaczenie parametru γ minimalizującego wartość średniokwadratowego błędu estymacji, w przypadku regresji LASSO nie istnieje ogólna postać zamknięta pozwalająca na bezpośrednie wyznaczenie optymalnego parametru regularyzacyjnego λ. Z tego powodu jego doboru dokonam w oparciu o symulacje komputerowe. Dla każdej z analizowanych konfiguracji dopasuję model LASSO za pomocą funkcji *glmnet*, a następnie posłużę się siatką wartości λ automatycznie wygenerowaną przez zastosowaną funkcję. Dla każdej wartości λ z tego zbioru obliczę błąd średniokwadratowy względem prawdziwego wektora β i za optymalną wybiorę tę wartość, która minimalizuje obliczone MSE.

Ukażę wykresy przedstawiające obliczone wartości błędu średniokwadratowego dla zastosowanego zbioru wartości λ z zaznaczeniem tej wartości, która została uznana za optymalną.

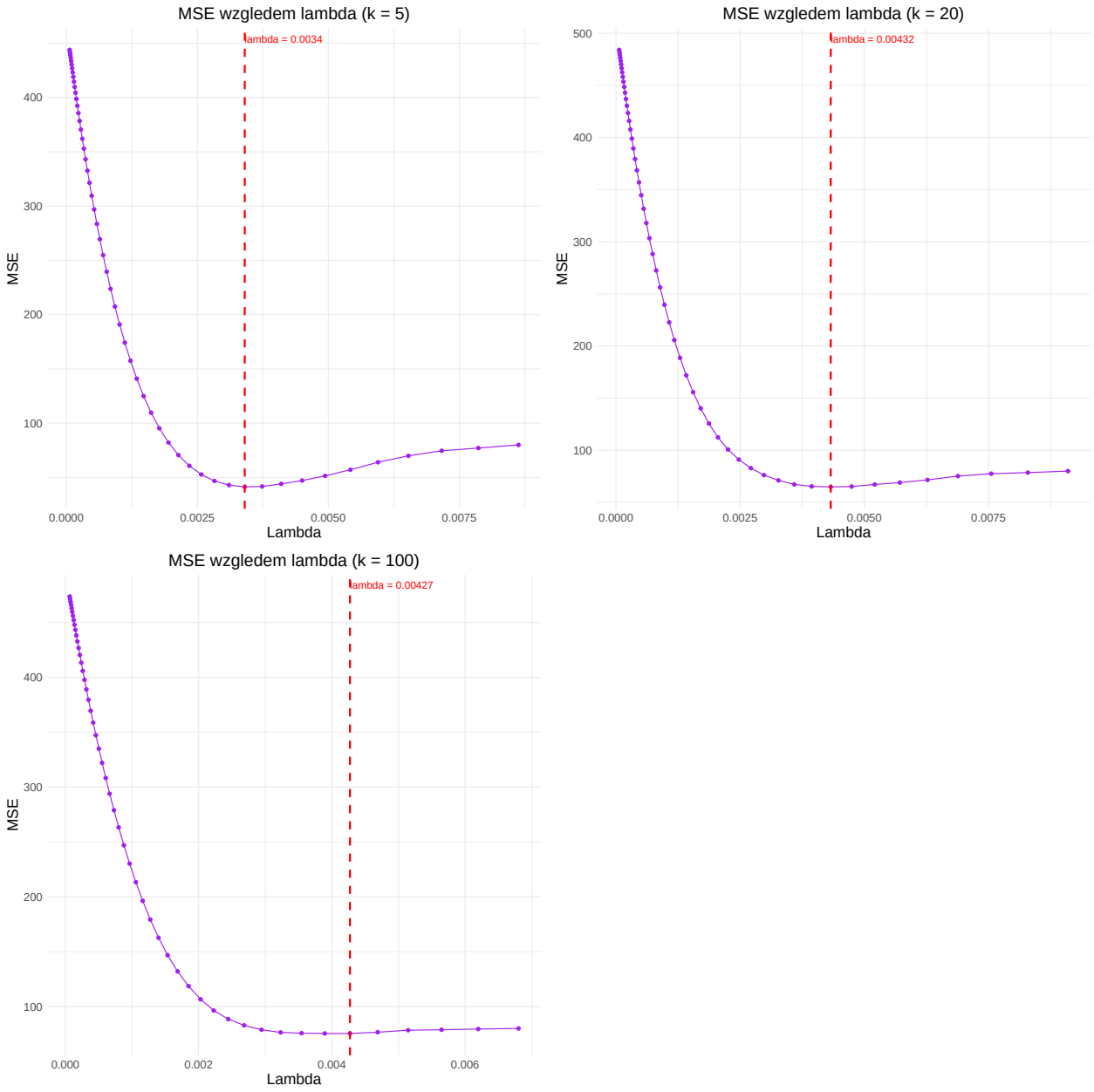
Dla wektora $\beta_1 = ... = \beta_k = 4$**,** $\beta_{k+1} = ... = \beta_p = 0$**:**







Dla wektora $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$**,** $\beta_{k+1} = ... = \beta_p = 0$**:**



Widzimy, iż, faktycznie, przedstawiona funkcja MSE osiąga najmniejszą wartość na wykresie w wybranym optymalnym punkcie λ.

Następnie, przedstawię również tabelę uzyskanych wyników, gdzie, ponownie, beta1 oznacza wektor $\beta_1 = ... = \beta_k = 4$, $\beta_{k+1} = ... = \beta_p = 0$, a beta2 - $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$, $\beta_{k+1} = ... = \beta_p = 0$.

Table 2: Optymalne wartości λ

	k = 5	k = 20	k = 100
beta1	0.0038	0.0032	0.0018
beta2	0.0034	0.0043	0.0043

PORÓWNANIE BŁĘDU ŚREDNIOKWADRATOWEGO MSE

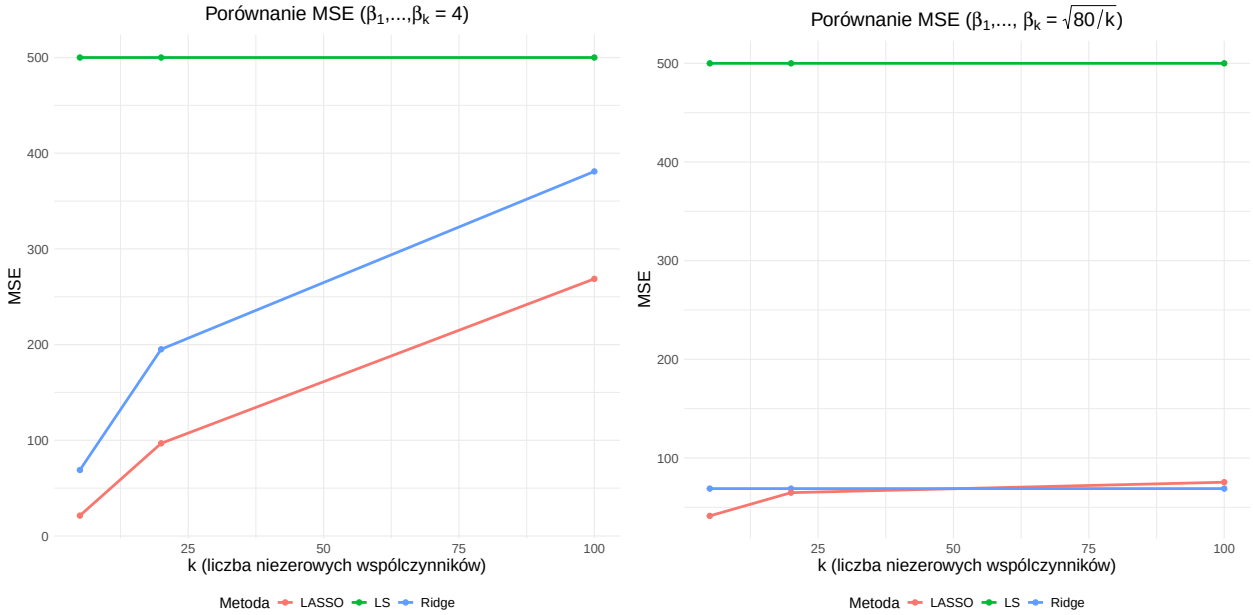
Porównam teraz w każdym z omawianych przypadków wartości błędu średniokwadratowego MSE uzyskane za pomocą metody najmniejszych kwadratów z optymalnymi wartościami MSE, które można uzyskać za pomocą regresji grzbietowej i LASSO.

Wiemy, iż estymator uzyskany **metodą najmniejszych kwadratów** wygląda w tym przypadku następująco: $\hat{\beta}^{LS} = (X'X)^{-1}X'Y = X'Y$ oraz (ponieważ $Y \sim N(X\beta, \sigma^2 I)$) rozkład $\hat{\beta}^{LS}$ to $N(\beta, \sigma^2 I) = N(\beta, I)$. Mając już takie informacje, możemy wyznaczyć następujący wzór na błąd średniokwadratowy:

$$MSE(\hat{\beta}^{LS}) = E||\hat{\beta}^{LS} - \beta||^2 = E\sum_{i=1}^{p}(\hat{\beta}_i^{LS} - \beta_i)^2 = \sum_{i=1}^{p}E(\hat{\beta}_i^{LS} - \beta_i)^2 = \sum_{i=1}^{p}Var(\hat{\beta}_i^{LS}) = \sum_{i=1}^{p}\sigma^2 = p \cdot \sigma^2 = p.$$

Następnie, wartość MSE dla **regresji grzbietowej** obliczę, korzystając ze wzoru wyprowadzonego we wcześniejszej części: $MSE(\hat{\beta}^{RR}) = \frac{\gamma^2||\beta||^2+p}{(1+\gamma)^2}$. Z kolei dla **LASSO** wykorzystam wartość MSE uzyskaną w poprzedniej części dla optymalnej wartości λ.

Przedstawię teraz dwa wykresy ukazujące uzyskane wartości MSE w zależności od postaci wektora β oraz od liczby istotnych współczynników i zastosowanej metody.



Na pierwszym wykresie widzimy, iż w przypadku, gdy $\beta_1 = ... = \beta_k = 4$ **LASSO wypada najlepiej** - osiąga najniższe wartości MSE, szczególnie przy mniejszej liczbie istotnych zmiennych. Jest tak dlatego, że metoda ta zeruje nieistotne współczynniki, co w przypadku rzadkich wektorów β prowadzi do mniejszych błędów estymacji. **Regresja klasyczna ma bardzo wysoki i stały błąd średniokwadratowy** - nie radzi sobie ona w sytuacji, gdy liczba zmiennych p jest zbliżona do liczby obserwacji n. Z kolei **regresja grzbietowa** jest lepsza niż klasyczna regresja, ale gorsza od LASSO pod względem wartości MSE - regularizacja pomaga, ale nie eliminuje wpływu nieistotnych zmiennych tak skutecznie jak LASSO. Regresja grzbietowa nie zeruje współczynników - tylko je tłumi, dlatego też nie usuwa ona szumu w pełni. Możemy również zauważyć, że wraz ze wzrostem liczby niezerowych współczynników k rośnie również MSE dla LASSO i dla regresji grzbietowej - wówczas model staje się mniej rzadki i LASSO ma więcej istotnych zmiennych do wykrycia, więc łatwiej o błędy eliminacji, a regresja grzbietowa ma więcej współczynników do oszacowania, w wyniku czego suma kar za ich wielkość rośnie.

Patrząc na drugi wykres, widzimy, iż w przypadku, gdy $\beta_1 = ... = \beta_k = \sqrt{80/k}$ **regresja klasyczna** nadal ma najwyższe, stałe wartości błędu średniokwadratowego. W tej sytuacji zawsze $||\beta||^2 = k \cdot (\sqrt{\frac{80}{k}})^2 = 80$,

więc współczynniki są mniejsze dla większego k, ale nie są aż tak trudne do wykrycia, gdyż suma ich wpływów jest ograniczona. Dlatego też dla **LASSO** oraz dla **regresji grzbietowej** wartości MSE są znacznie niższe niż w poprzedniej sytuacji. Widzimy także, iż początkowo LASSO ma najniższe wartości błędu średniokwadratowego, ale dla $k = 100$ to regresja grzbietowa okazuje się lepsza.

ŚREDNIA LICZBA FAŁSZYWYCH ODKRYĆ DLA LASSO

W tej części raportu wyznaczę teoretycznie **średnią liczbę fałszywych odkryć** dla LASSO z optymalną wartością parametru λ.

Załóżmy, że badamy następujące hipotezy: $H_{0i} : \beta_i = 0$, $H_{1i} : \beta_i \neq 0$. Wówczas, przy prawdziwości hipotezy zerowej, estymator wyznaczony metodą najmniejszych kwadratów $\hat{\beta}_i^{LS} \sim N(0, 1)$. Zatem, prawdopodobieństwo fałszywego odkrycia mogę wyliczyć następująco: $P(FD) = P(\text{LASSO wybiera} X^i | \beta_i = 0) = P(|\hat{\beta}_i^{LS}| > \lambda | \beta_i = 0) = 2 \cdot (1 - \Phi(\lambda))$. Z tego wynika, iż średnia liczba fałszywych odkryć to otrzymany właśnie wynik przemnożony przez liczbę prawdziwych hipotez zerowych:

$$E(PD) = (p - k) \cdot 2 \cdot (1 - \Phi(\lambda)),$$

gdzie: p - liczba zmiennych, k - liczba istotnych zmiennych.

Wykorzystam owy wzór i przedstawię tabelę ukazującą uzyskane średnie liczby fałszywych odkryć w rozważanych przypadkach, gdzie, ponownie, beta1 oznacza wektor $\beta_1 = ... = \beta_k = 4$, $\beta_{k+1} = ... = \beta_p = 0$, a beta2 - $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$, $\beta_{k+1} = ... = \beta_p = 0$.

Table 3: Średnia liczba fałszywych odkryć

	k = 5	k = 20	k = 100
beta1	493.507	478.778	399.415
beta2	493.656	478.344	398.637

Widzimy, iż uzyskane wyniki dla obu postaci wektora β są do siebie bardzo zbliżone. Ponadto, w każdym przypadku średnia liczba fałszywie wybranych zmiennych przez LASSO jest bardzo bliska liczbie wszystkich zmiennych nieistotnych. Zatem, teoretycznie, LASSO z optymalnie dobranym parametrem λ nie radzi sobie dobrze z identyfikacją nieistotnych zmiennych.

MOC IDENTYFIKACJI ISTOTNYCH ZMIENNYCH DLA LASSO

Następnie, w sposób teoretyczny wyznaczę **moc identyfikacji istotnych zmiennych dla LASSO** z optymalną wartością parametru λ.

Ponownie, załóżmy, że badamy następujące hipotezy: $H_{0i} : \beta_i = 0$, $H_{1i} : \beta_i \neq 0$. Wówczas, przy prawdziwości hipotezy alternatywnej, estymator wyznaczony metodą najmniejszych kwadratów $\hat{\beta}_i^{LS} \sim N(\beta_i, 1)$. Wówczas, moc takiego pojedynczego testu wyznacza się następująco: $P(\text{LASSO wybiera} X^i | \beta_i \neq 0) = P(|\hat{\beta}_i^{LS}| > \lambda | \beta_i \neq 0) = P(\hat{\beta}_i^{LS} > \lambda | \beta_i \neq 0) + P(\hat{\beta}_i^{LS} < -\lambda | \beta_i \neq 0) = P(\hat{\beta}_i^{LS} - \beta_i > \lambda - \beta_i | \beta_i \neq 0) + P(\hat{\beta}_i^{LS} - \beta_i < -\lambda - \beta_i | \beta_i \neq 0) = 1 - \Phi(\lambda - \beta_i) + \Phi(-\lambda - \beta_i)$. Mając już ten wynik, możemy wyznaczyć moc identyfikacji w następujący sposób:

$$\text{moc} = \frac{1}{k} \cdot \sum_{i:\beta_i \neq 0} (1 - \Phi(\lambda - \beta_i) + \Phi(-\lambda - \beta_i)) = \frac{k \cdot (1 - \Phi(\lambda - \beta_i) + \Phi(-\lambda - \beta_i))}{k} = 1 - \Phi(\lambda - \beta_i) + \Phi(-\lambda - \beta_i).$$

Wykorzystam owy wzór i przedstawię tabelę ukazującą uzyskane moce identyfikacji istotnych zmiennych w rozważanych przypadkach, gdzie, ponownie, beta1 oznacza wektor $\beta_1 = ... = \beta_k = 4$, $\beta_{k+1} = ... = \beta_p = 0$, a beta2 - $\beta_1 = ... = \beta_k = \sqrt{\frac{80}{k}}$, $\beta_{k+1} = ... = \beta_p = 0$.

Table 4: Moc identyfikacji istotnych zmiennych

	k = 5	k = 20	k = 100
beta1	1	1.0000	1.0000
beta2	1	0.9995	0.9977

Widzimy, iż, teoretycznie, LASSO bardzo dobrze identyfikuje zmienne istotne – w przypadku $\beta_1 = ... = \beta_k = 4$ moc wynosi w przybliżeniu 1 dla każdego k, co oznacza, że praktycznie wszystkie zmienne istotne zostaną poprawnie wybrane. Spadek mocy w drugim przypadku (zwłaszcza przy $k = 100$) wynika z tego, że wartości współczynników są wtedy mniejsze – sygnał jest słabszy i nieco trudniej odróżnić zmienne istotne od szumu. Zatem, teoretycznie, LASSO z optymalnie dobranym parametrem λ bardzo skutecznie identyfikuje zmienne istotne, jednak jednocześnie źle radzi sobie z eliminowaniem zmiennych nieistotnych, co może prowadzić do słabej interpretowalności modelu.

PRZYPADEK MACIERZY O WYRAZACH Z ROZKŁADU NORMALNEGO

W drugiej części owego raportu rozważę sytuację takiej macierzy planu, której elementy są niezależnymi zmiennymi losowymi z rozkładu normalnego $N(0, \sigma = \frac{1}{\sqrt{n}})$.

REGRESJA GRZBIETOWA, LASSO ORAZ ADAPTACYJNE LASSO

Na początku, wyestymuję wektor β na cztery następujące sposoby:

- za pomocą **regresji grzbietowej** z parametrem wybranym za pomocą walidacji krzyżowej - zrobię to, używając polecenia *cv.glmnet* z parametrem *alpha = 0*,
- za pomocą **LASSO** z parametrem wybranym za pomocą walidacji krzyżowej - zrobię to, używając polecenia *cv.glmnet* z parametrem *alpha = 1*,
- za pomocą **adaptacyjnego LASSO** z parametrem wybranym za pomocą walidacji krzyżowej z wagami $1/|\hat{\beta}_i|$, gdzie $\hat{\beta}_i$ zostały uzyskane w pierwszym punkcie (za pomocą regresji grzbietowej) - zrobię to w następujących krokach:

1) obliczę wagi $w_i = \frac{1}{|\hat{\beta}_i|}$, wykorzystując współczynniki regresji uzyskane za pomocą regresji grzbietowej,
2) przekształcę macierz planu X w następujący sposób: $X_{temp} = X \cdot diag(1/w_i)$,
3) na przekształconych danych X_{temp} zastosuję regresję LASSO z wyborem parametru λ za pomocą walidacji krzyżowej (używając polecenia *cv.glmnet*),
4) po uzyskaniu estymatora $\hat{\beta}_{temp}$ z regresji LASSO na przeskalowanych danych, przekształcę go z powrotem do oryginalnej skali: $\hat{\beta}_{aL} = \frac{\hat{\beta}_{temp}}{w_i}$,

- za pomocą **adaptacyjnego LASSO** z parametrem wybranym za pomocą walidacji krzyżowej z wagami wybranymi w następujących krokach:

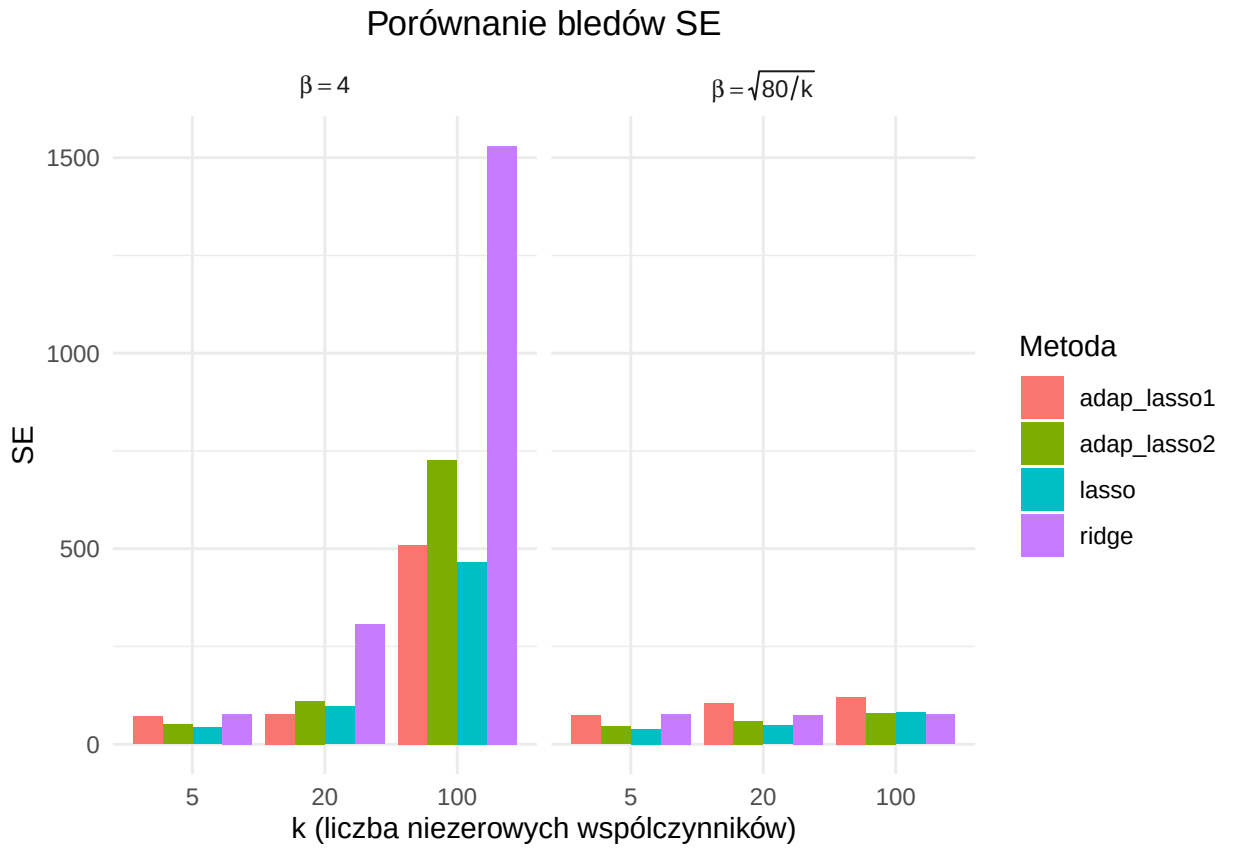
1) używając estymatora wektora β uzyskanego w punkcie drugim (za pomocą LASSO), obliczę estymator σ jako: $\hat{\sigma} = \sqrt{\frac{RSS}{n-k}}$, gdzie RSS wyznaczę w oparciu o LASSO z kros-walidacją, k to liczba zmiennych wybranych przez LASSO, a $n = 500$,
2) obliczę wagi w następujący sposób: $w_i = \frac{\hat{\sigma}}{|\hat{\beta}_i|}$, wykorzystując współczynniki regresji uzyskane za pomocą LASSO, przy czym zmienne, dla których $\hat{\beta}_i = 0$ zostaną pominięte - ich współczynniki w ostatecznym estymatorze przyjmują wartość 0,
3) wyznaczę parametr wygładzający $\lambda = \frac{\hat{\sigma}}{n} \cdot \Phi^{-1}(1 - \frac{0.2}{2p})$, gdzie $n = p = 500$,
4) dla zmiennych wybranych przez klasyczne LASSO utworzę nową macierz X_{small}, a następnie przekształcę ją przez przemnożenie przez macierz diagonalną z $1/w_i$,
5) na przeskalowanej w ten sposób macierzy X_{small} zastosuję regresję LASSO z wcześniej wyznaczonym parametrem λ,

6) uzyskany wektor przeskaluję w następujący sposób, by otrzymać ostateczny estymator: $\hat{\beta}_{aL} = \frac{\hat{\beta}_{temp}}{w_i}$ (zmiennym nieuwzględnionym zostanie przypisana wartość 0).

Wyznaczę zatem owe cztery estymatory wektora β dla wszystkich sześciu analizowanych przypadków.

PORÓWNANIE SE

Porównam teraz na wykresach $SE = ||\hat{\beta} - \beta||^2$ dla czterech otrzymanych właśnie estymatorów.



Bardzo wyraźnie widzimy, iż w przypadku $k = 5$ uzyskane wartości SE dla obu postaci wektora β są bardzo do siebie zbliżone, gdyż wówczas dla drugiego wektora również $\sqrt{80/k} = \sqrt{80/5} = \sqrt{16} = 4$, więc sytuacje te są analogiczne. Jednak dla większych wartości k postać wektora β z $\beta_i = \sqrt{80/k}$ daje znacznie niższe wyniki między innymi dlatego, że ma on mniejszą długość - w wyniku tego SE jest mniejsze, bo różnice są liczone względem mniejszych wartości. Możemy także zauważyć, iż dla wektora β z $\beta_i = 4$ wraz ze wzrostem liczby niezerowych współczynników k rosną również otrzymane wartości SE dla wszystkich metod - im więcej niezerowych współczynników, tym więcej możliwości popełnienia błędu. Ponadto, możemy zauważyć, że najwyższe wartości SE uzyskiwane są dla regresji grzbietowej - metoda ta nie odrzuca zmiennych nieistotnych, tylko przypisuje im małe wartości, w wyniku czego dodaje dużo błędów w estymacji zerowych współczynników, które sumują się w wysokie wartości SE. Dokładniejsze wnioski będzie można wyciągnąć w dalszej części raportu, gdy doświadczenie to zostanie powtórzone 100 razy.

TECHNIKA KNOCKOFFÓW

W tej części, zastosuję również **technikę knockoffów dla LASSO i regresji grzbietowej** tak, aby kontrolować FDR na poziomie 0.2. Następnie, porównam FDR i moc tych procedur z FDR i mocą dla metod stosowanych powyżej.

Procedura knockoffów opiera się na założeniu, iż fałszywe odkrycia mogą być spowodowane tym, że

nieistotne zmienne są mocno skorelowane z istotnymi zmiennymi. Z tego powodu, aby określić, które estymatory są przypadkowo duże ze względu na korelację z odpowiednimi istotnymi zmiennymi, procedura ta polega na stworzeniu zmiennych kontrolnych, które zachowują taką samą strukturę korelacji jak zmienne w oryginalnej bazie danych.

Procedura ta ma następujące kroki:

1) stworzę kopie zmiennych oryginalnych - macierz X_{copy} o tej samej strukturze, co oryginalna macierz planu X, a następnie złączę te dwie macierze w jedną dużą macierz X_{large},
2) stworzę regresję grzbietową/regresję LASSO przy użyciu polecenia *cv.glmnet* na macierzy X_{large} - w ten sposób otrzymam wektor estymatorów $\hat{\beta}$ dla wszystkich $2p$ zmiennych,
3) dla każdej zmiennej oryginalnej obliczę statystykę w_j w następujący sposób: $w_j = |\hat{\beta}_j| - |\hat{\beta}_j^{knockoff}|$,
4) wyznaczę próg $\hat{t}$ według następującego wzoru:

$$\hat{t} = min\{t > 0 : \frac{1 + \#\{j : w_j \leq -t\}}{\#\{j : w_j \geq t\} \bigvee 1} \leq \alpha\},$$

gdzie, w naszym przypadku, $\alpha = 0.2$ - zrobię to w następujących krokach:

- zdefiniuję funkcję logiczną, która sprawdza, czy dla danego progu t spełniony jest powyższy warunek,
- będę iterować po możliwych progach $t = |w_j|$ - wybór progu spośród tych wartości gwarantuje, że rozpatrzone zostaną wszystkie możliwe zmiany w zbiorze wybranych zmiennych - i wybiorę najmniejszy $\hat{t}$ spośród tych, dla których warunek przedstawiony wyżej jest spełniony,

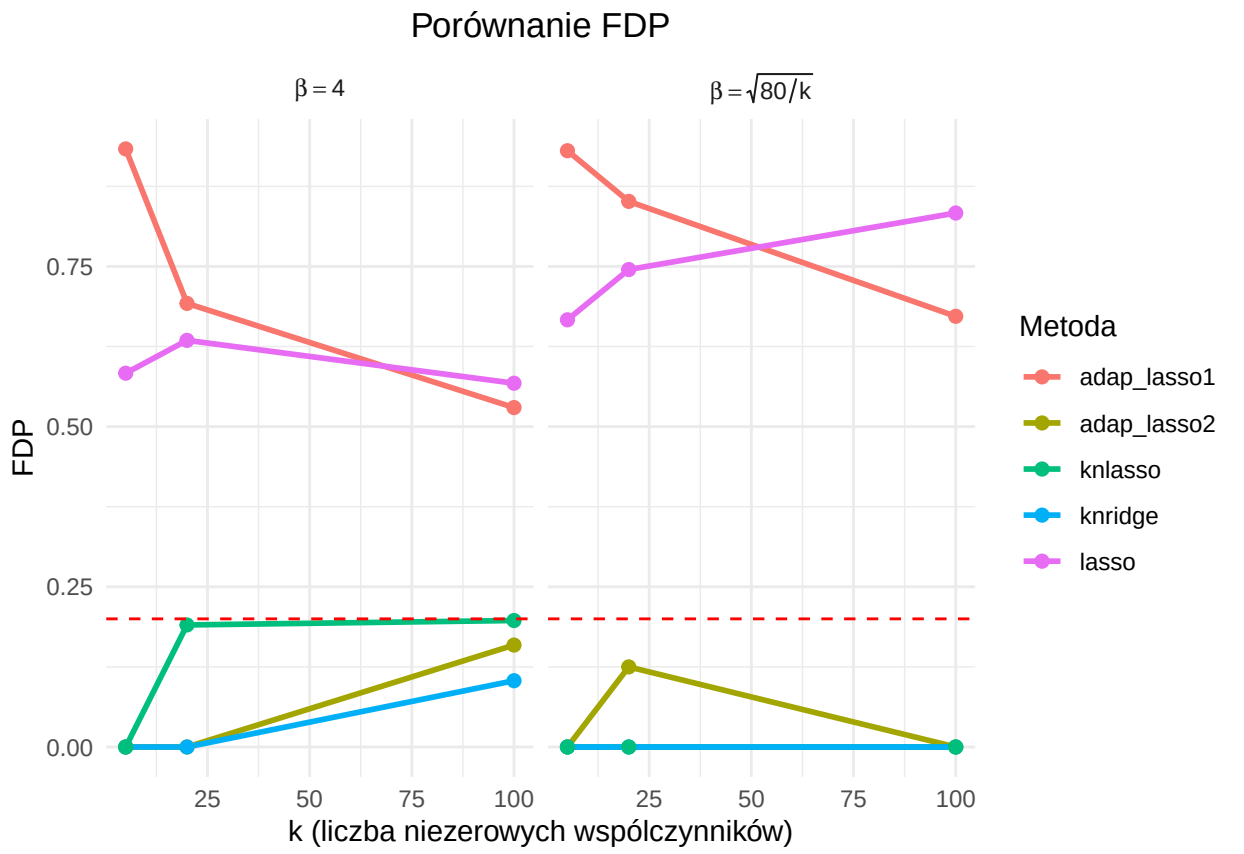
5) na podstawie wyznaczonego progu selekcji zachowam tylko te współczynniki estymatora regresji grzbietowej/regresji LASSO, dla których odpowiadająca im statystyka knockoff w_j jest nie mniejsza niż próg $\hat{t}$ - współczynniki pozostałych zmiennych zostają wyzerowane.

PORÓWNANIE FDP

Porównam teraz **FDP** - proporcję fałszywych odkryć dla wszystkich zastosowanych metod we wszystkich sześciu omawianych przypadkach.

FDP obliczę, dzieląc liczbę rzeczywiście nieistotnych zmiennych, które zostały wybrane przez daną metodę przez maksimum z liczby 1 oraz liczby wszystkich zmiennych wybranych przez daną metodę.

Przedstawię teraz wykresy ukazujące uzyskane wyniki.



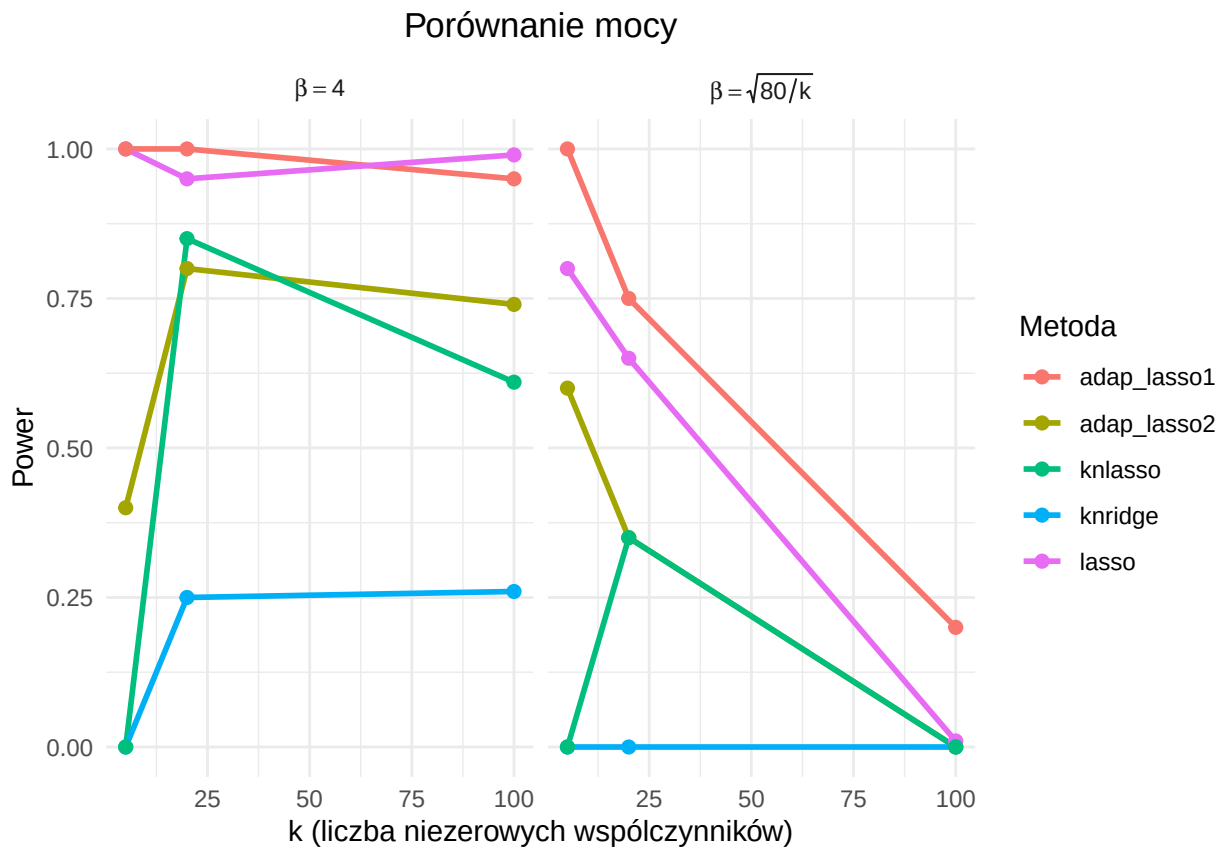
Widzimy przede wszystkim, iż obie zastosowane **procedury knockoffów** utrzymują FDP poniżej poziomu 0.2, niezależnie od liczby niezerowych współczynników k oraz rodzaju wektora β. Widzimy także, iż w obu sytuacjach to **LASSO** oraz **pierwsze adaptacyjne LASSO** uzyskują najwyższe wartości FDP. LASSO nie posiada żadnego mechanizmu kontroli FDR, z kolei wagi zastosowane w pierwszym adaptacyjnym LASSO mogą być niestabilne, a użyta walidacja krzyżowa może prowadzić do niedokładnej selekcji. Ponadto, możemy zauważyć, iż wartości FDP dla **adaptacyjnego LASSO 2** są bardzo niskie - może to być spowodowane tym, iż metoda ta usuwa zmienne odrzucone przez LASSO, a pozostałe penalizuje z użyciem wag uwzględniających poziom szumu i siłę sygnału. Dodatkowo, wybór parametru λ oparty jest tam na kwantylu rozkładu normalnego, co także pomaga kontrolować liczbę fałszywych odkryć. Widzimy też, że przedstawione wyniki dla obu postaci wektora β są do siebie zbliżone, przy czym metody LASSO oraz pierwsze adaptacyjne LASSO mają nieco wyższe FDP w drugim scenariuszu. Dokładniejsze wnioski będzie można wyciągnąć w dalszej części raportu, gdzie owe doświadczenie zostanie powtórzone 100 razy.

PORÓWNANIE MOCY

Porównam teraz **moc** dla wszystkich zastosowanych metod we wszystkich sześciu omawianych przypadkach.

Moc obliczę, dzieląc liczbę rzeczywiście istotnych zmiennych, które zostały wybrane przez daną metodę przez liczbę wszystkich rzeczywiście istotnych zmiennych.

Przedstawię teraz wykresy ukazujące uzyskane wyniki.



Widzimy, iż dla wektora β z $\beta_i = 4$ metody **LASSO** oraz **pierwsze adaptacyjne LASSO** uzyskują bardzo wysoką moc, ponieważ synał jest silny i łatwy do wykrycia. Obie te metody mają tendencję do wybierania wielu predyktorów, co zwiększa szansę wybrania również tych istotnych. W tym przypadku widzimy też, że moc dla **adaptacyjnego LASSO 2** oraz dla **metody knockoffów dla LASSO** jest dość wysoka, a najniższe wartości uzyskane są dla **metody knockoffów dla regresji grzbietowej**. Niskie wartości mocy dla metod knockoffów wynikają z tego, iż są one zaprojektowane w taki sposób, by kontrolować poziom fałszywych odkryć, co często skutkuje właśnie niższą mocą. Widzimy też, że dla drugiej postaci wektora β moce dla wszystkich metod są niższe, co może wynikać z tego, iż sygnał słabnie przy większej liczbie istotnych zmiennych, co prowadzi do trudności z wyborem zmiennych rzeczywiście istotnych. Dokładniejsze wnioski będzie można wyciągnąć w dalszej części raportu, gdzie owe doświadczenie zostanie powtórzone 100 razy.

POWTÓRZENIE DOŚWIADCZENIA 100 RAZY

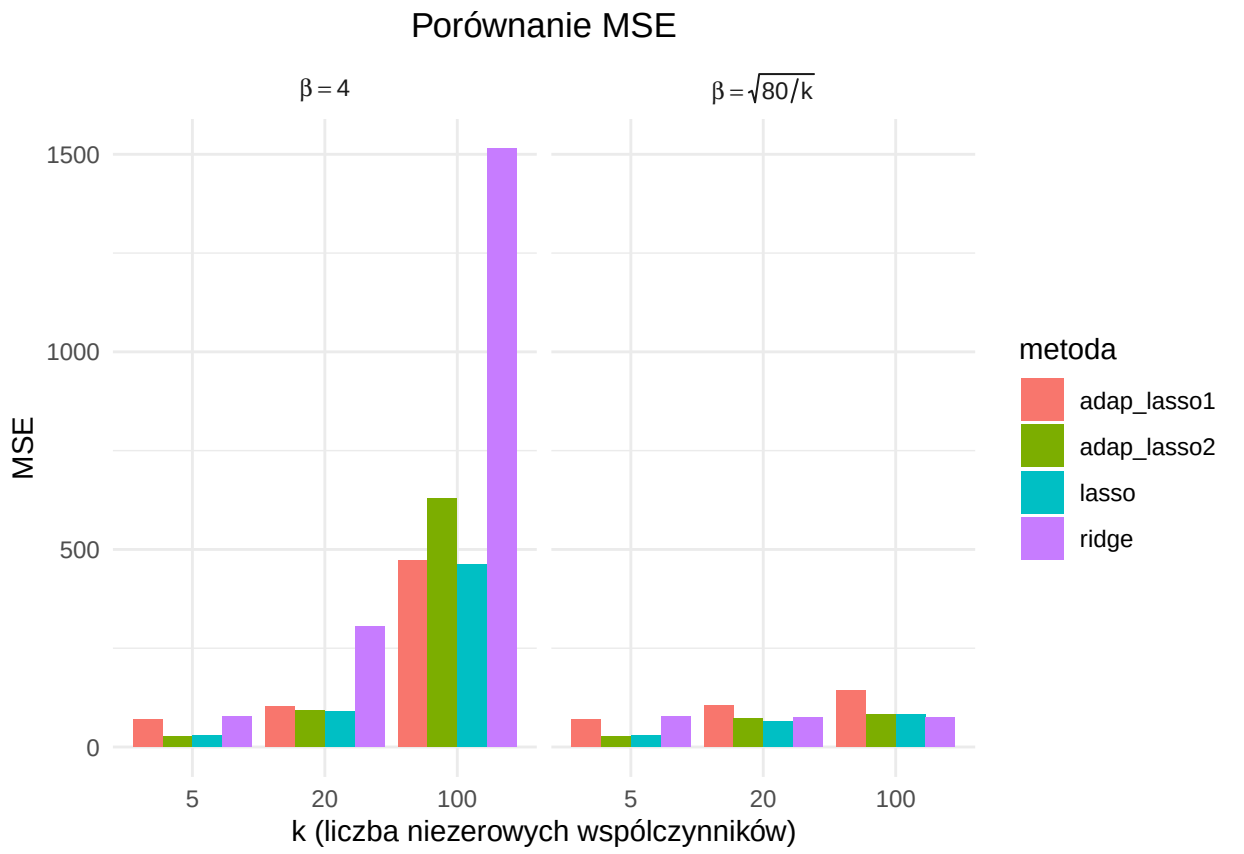
Powtórzę teraz wszystkie powyższe punkty 100 razy, by porównać MSE, FDR oraz moc dla analizowanych metod.

PORÓWNANIE MSE

Porównam teraz **MSE** - błąd średniokwadratowy dla wszystkich zastosowanych metod we wszystkich sześciu omawianych przypadkach.

MSE obliczę, wyliczając średnią z usykanych wartości SE, przy powtórzeniu doświadczenia 100 razy.

Przedstawię teraz wykresy ukazujące uzyskane wyniki.



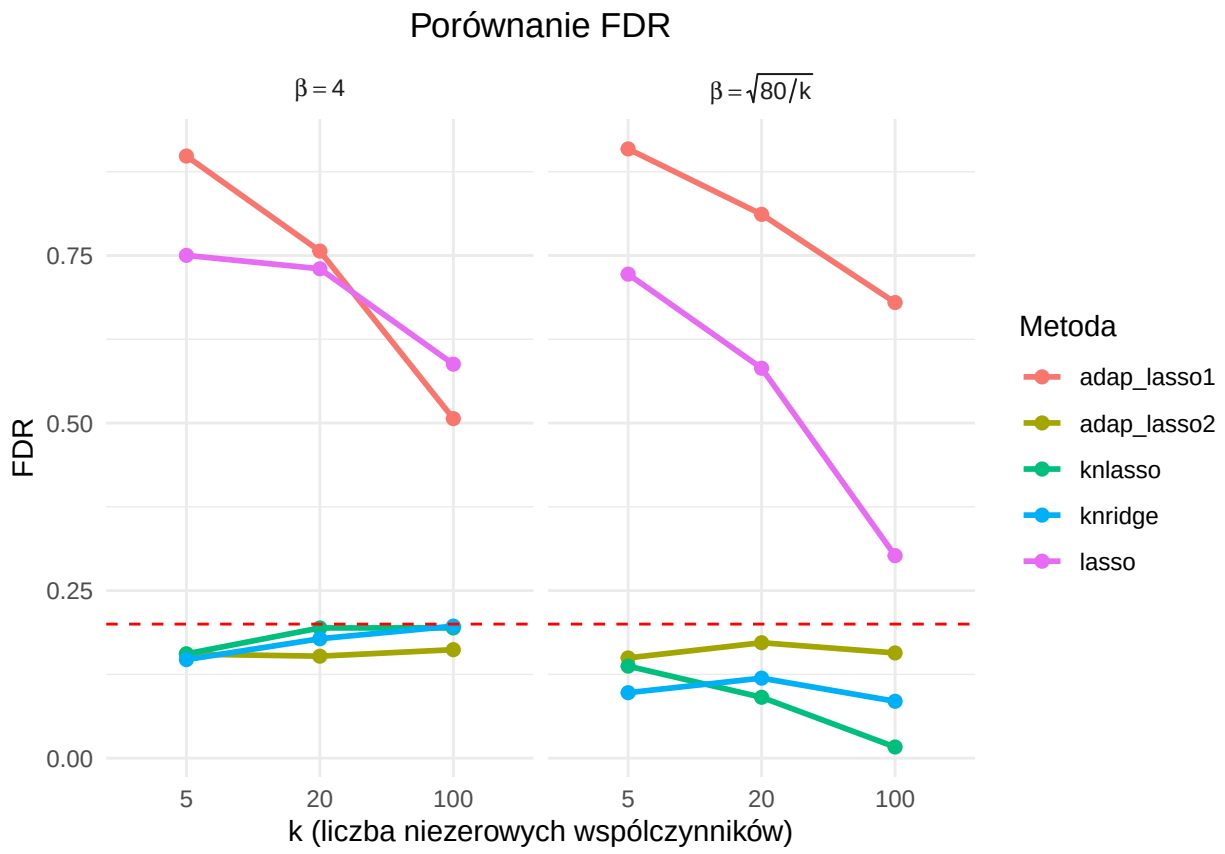
Widzimy ponownie, iż w przypadku $k = 5$ uzyskane wartości MSE dla obu postaci wektora β są bardzo do siebie zbliżone, gdyż wówczas analizowane sytuacje są analogiczne. Jednak w sytuacji wektora β z $\beta_i = \sqrt{80/k}$ dla większych wartości k uzyskane zostały znacznie niższe wyniki - między innymi dlatego, iż wektor ten ma mniejszą długość, w wyniku czego MSE jest mniejsze, bo różnice liczone są względem mniejszych wartości. Widzimy też, że dla tej postaci wektora β uzyskane wartości są bardzo do siebie zbliżone, niezależnie od liczby k. Z kolei dla pierwszej postaci wektora β, gdzie $\beta_i = 4$, wartości MSE rosną wraz ze wzrostem liczby niezerowych współczynników k - gdy liczba niezerowych współczynników rośnie, model musi oszacować więcej istotnych zmiennych (większa złożoność modelu). Bardzo wyraźnie widzimy też, że najwyższe wartości błędu średniokwadratowego zauważalne są dla **regresji grzbietowej** - metoda ta nie zeruje współczynników, a rozkłada wagę na wszystkie zmienne - również te nieistotne, co powoduje wzrost błędu. Z kolei wyniki dla pozostałych metod są do siebie zbliżone.

PORÓWNANIE FDR

Porównam teraz **FDR** - współczynnik fałszywych odkryć dla wszystkich zastosowanych metod we wszystkich sześciu omawianych przypadkach.

FDR obliczę, wyliczając średnią z usykanych wartości FDP, przy powtórzeniu doświadczenia 100 razy.

Przedstawię teraz wykresy ukazujące uzyskane wyniki.



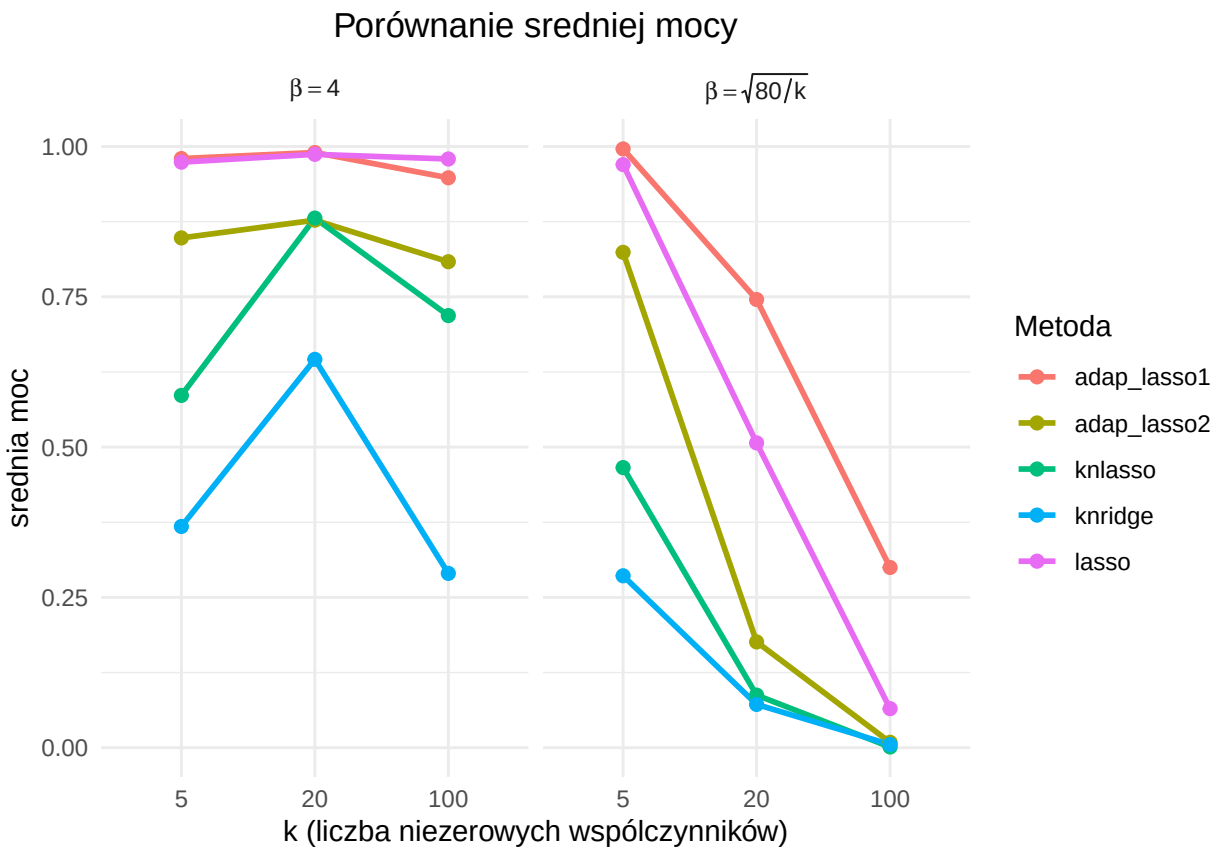
Widzimy, iż dla obu postaci wektora β najwyższe wartości FDR uzyskiwane są dla metod **LASSO** oraz **adaptacyjne LASSO 1**. Oznacza to, że metody te wybierają wiele fałszywych zmiennych, co może wynikać z tego, iż ich parametryzacja oparta na walidacji krzyżowej nie wystarcza do kontrolowania błędnych odkryć. Widzimy też, że FDR maleje tutaj wraz z rosnącym k, gdyż sygnał staje się mniej rzadki i może to ułatwiać odróżnienie zmiennych istotnych od nieistotnych. Dalej, możemy zauważyć, że **metody knockoffów** skutecznie kontrolują FDR poniżej poziomu 0.2, zgodnie z założeniem tej techniki. Są one skonstruowane właśnie w taki sposób, by kontrolować FDR na zadanym poziomie. Z kolei metoda **adaptacyjne LASSO 2** również osiąga niskie wyniki - metoda ta lepiej dopasowuje parametr λ, uwzględniając zarówno strukturę szumu, jak i estymację wag, co zmniejsza liczbę fałszywych odkryć. Ponadto, widzimy też, iż w większości sytuacjach wartości FDR maleją wraz ze wzrostem liczby k - wówczas względna liczba możliwych fałszywych odkryć spada. Z kolei analizując obie postacie wektora β, widzimy, iż trend nie jest jednoznaczny - na przykład dla wektora β z $\beta_i = \sqrt{80/k}$ metoda LASSO oraz techniki knockoffów uzyskały niższe FDR, a metody adaptacyjnego LASSO - wyższe.

PORÓWNANIE ŚREDNIEJ MOCY

Porównam teraz **średnią moc** dla wszystkich zastosowanych metod we wszystkich sześciu omawianych przypadkach.

Średnią moc obliczę, wyliczając średnią z usykanych wartości mocy, przy powtórzeniu doświadczenia 100 razy.

Przedstawię teraz wykresy ukazujące uzyskane wyniki.



Widzimy, iż w obu przypadkach postaci wektora β najwyższą moc uzyskały metody **LASSO** oraz **adaptacyjne LASSO 1** - są one bardzo skuteczne w wykrywaniu istotnych zmiennych, jednak kosztem również wielu odkryć fałszywych. Dalej, **adaptacyjne LASSO 2** także uzyskuje dość wysoką moc, jednak niższą niż dwie poprzednie metody. Natomiast najniższe wyniki w obu przypadkach uzyskały **techniki knockoffów** - nie są one skuteczne w wykrywaniu zmiennych istotnych, ponieważ zostały zaprojektowane do kontroli FDR, co często wiąże się właśnie z niższą mocą. Możemy również zauważyć, że w przypadku wektora β z $\beta_i = \sqrt{80/k}$ osiągnięte moce są znacznie niższe dla wszystkich metod, przy czym maleją one wraz ze wzrostem liczby k - trudniej wykryć istotne zmienne, gdy ich wpływ jest coraz słabszy. Wartości współczynników są tutaj znacznie mniejsze, co oznacza słabszy sygnał, który trudniej odróżnić od szumu.

PODSUMOWANIE

W raporcie tym zostały porównane **metody regulacyjne** w regresji wielorakiej - ich wady oraz zalety w oparciu o ich MSE, FDR i średnią moc. Metoda **LASSO** wykazała się bardzo wysoką mocą, jednak kosztem również wysokiego współczynnika FDR. **Regresja grzbietowa** uzyskała najwyższe błędy MSE, zwłaszcza przy większej liczbie istotnych zmiennych, gdyż metoda ta nie wykonuje selekcji zmiennych. Metoda **adaptacyjne LASSO 1**, podobnie jak LASSO, osiągnęła wysoką moc, jednak kosztem również wielu odkryć fałszywych. Z kolei metoda **adaptacyjne LASSO 2** okazała się najlepszym kompromisem między mocą a kontrolą FDR - wykazała się dość wysoką mocą oraz niskim FDR. Natomiast **metody knockoffów** wyróżniły się tym, iż kontrolują one wartości FDR na zadanym poziomie (0.2), jednak ich wadą była niska moc, zwłaszcza w przypadku słabych efektów i dużej liczby istotnych zmiennych.