Knockoff - wprowadzenie, motywacja, podstawowe pojęcia, możliwe zastosowania i ich rozszerzenia

1 Wprowadzenie

W wielu dziedzinach nauki widzimy wszechstronne zastosowania matematyki, które pomagają nam wyjaśnić i opisać dane. Kiedyś ktoś powiedział, że matematyka jest językiem opisującym świat i to dzięki niej wiemy aż tak dużo. Weźmy pod uwagę medycynę, która nieustannie się rozwija. To właśnie tam obserwujemy zmienną odpowiedzi wraz z duża liczbą potencjalnych zmiennych objaśniających. Naszym zadaniem, w tak postawionym problemie, jest znalezienie zmiennych, które są naprawdę powiązane z naszą zmienną odpowiedzi przy jednoczesnej kontroli pewnego czynnika. Tym czynnikiem jest frakcja fałszywych odkryć, czyli oczekiwany odsetek fałszywych odkryć wśród wszystkich odkryć. Aby zapewnić naukowca, że większość okryć jest rzeczywiście prawdziwa i możliwa do odtworzenie, dany czynnik powinien być dość mały. Z reguły standardowa kontrola FDR na poziomie 0.05 lub 0.1 jest zadowalająca. Relatywnie nową procedurą selekcji zmiennych kontrolującą frakcje fałszywych odkryć w statystycznym modelu liniowym jest filtr knockoff - przy założeniu, że ilość obserwacji jest co najmniej tyle co zmiennych. Dana metoda daje dokładną kontrolę FDR w skończonej próbie, bez względu na macierz planu czy zmienne towarzyszące oraz bez względu na wahania współczynników oraz nie wymaga znajomości szumu. Odwołując się do nazwy, metoda polega na wytwarzani zmiennych knockoff, które są mało restrykcyjne, czyli nie wymagają dużych założeń aby poprawnie działać oraz tanie - ich konstrukcja nie wymaga nowych danych i są tak skonstruowane aby naśladować strukturę korelacyjną znalezioną w istniejący zmiennych. Ta własność pozwala nam na dokładną kontrolę FDR, wykraczając poza standardowe metody dzięki metodą opartym na permutacjach. Dzięki bardzo ogólnej formie i elastyczności metoda może współpracować z szeroką klasą statystyk testowych. W literaturze możemy spotkać się z wykorzystaniem metody regularyzacyjnej LASSO do wyliczania statystyk. LASSO dzięki swojej skłonności do rzadkiej regresji pokazuje (w analizie empirycznej), że otrzymana metoda ma znacznie wyższą moc, niż istniejące reguły wyboru, gdy odsetek zmiennych zerowych jest wysoki.

Jest wiele spekulacji na temat skąd narodziła się ta procedura, ale najbardziej wiarygodną wydaje się, że jej początek możemy znaleźć w medycynie. W takim razie, gdzie możemy zaobserwować kopię bardzo zbliżonych danych, ale nie wynikającej z klonowania danych komórek? Odpowiedz jest dość prosta, bo chodzi tutaj o bliźnięta jednojajowe, które w swojej strukturze niosą bardzo podobną strukturę DNA. Pewien naukowiec zauważył, że jedno z bliźniąt jest chore na pewną chorobę, a drugie jest zdrowe. I stąd właśnie mogła narodzić się idea stworzenia tej metody.

2 Wstęp

Jednym z najważniejszych tematów obecnych badań statystyki teoretycznej jest zrozumienie właściwości wnioskowania o skończonej próbie procedur, które wybierają i dopasowują model do naszych danych. Dana metoda dotyczy właśnie tego projektu i skupia się na dokładności doboru zmiennych w klasycznym modelu liniowych regresji liniowej.

Załóżmy hipotetycznie, ze zaobserwowaliśmy interesującą nas zmienną odpowiedzi y i wiele potencjalnych objaśniających zmiennych X_j na n jednostkach obserwacyjnych. Dodatkowo wprowadźmy, że nasze obserwacje są zgodne z klasycznym modelem regresji liniowej

$$y = \mathbf{X}\beta + \epsilon,$$

gdzie jak zwykle $y \in \mathrm{R}^p$ jest wektorem odpowiedz, $\mathbf{X} \in \mathrm{R}^{n\times p}$ jest znaną macierzą planu zawierającą wartości naszych predykotrów, $\beta \in \mathrm{R}^p$ jest nieznanym wektorem współczynników oraz $\epsilon \sim N(0, \sigma^2\mathbf{I})$ to szum Gaussa. Ze względu, że interesuje nas tylko prawidłowe wnioskowanie na podstawie skończenie wielu próbek, to ograniczamy się do przypadku, gdzie ilość obserwacji jest większa bądź równa ilości

potencjalnych zmiennych objaśniających ($n \geqslant p$), ponieważ w przeciwnym razie nasz model nie byłby nawet identyfikowalny. Obecnie w nowoczesnych warunkach jest tak, że wśród wielu, które zostały zarejestrowane, zazwyczaj jest tylko kilka istotnych zmiennych. Przykładem może być dziedzina medycyny, w której zazwyczaj oczekujemy, że tylko kilka genów jest powiązanych z fenotypem, który nas interesuje. W przypadku omawianego w danym raporcie modelu liniowego oznaczałoby to, ze tylko kilka składowych parametru ma być niezerowa. W świeci statystyki czy układów dynamicznych mamy ogromny wachlarz metod i procedur, które strategicznie dopasują się do naszych danych, ale dla większości z nich nie mam zagwarantowanej kontroli oczekiwanej liczy proporcji fałszywych odkryć. Natomiast proponowana kontrola współczynnika frakcji fałszywych odkryć wśród wszystkich wybranych zmiennych, czyli wszystkich zmiennych uwzględnionych w naszym modelu jest zachowana.

Nieformalnie współczynnik frakcji fałszywych odkryć jest zdefiniowany jako oczekiwana proporcja błędnie wybranych zmiennych fałszywych. Fałszywe odkrycia to zmienne, które nie pojawią nie pojawiają się w prawdziwym modelu. Natomiast formalnie FDR procedury selekcji zwracającej podzbiór $\mathcal{S}$ zmiennych jest zdefiniowany jako

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\{j\,:\,\beta_j = 0,\, j \in \hat{\mathcal{S}}\}}{\max\{\#\{j \in \hat{\mathcal{S}}\},\, 1\}}\right] \tag{1}$$

Powiemy, ze reguła kontroluje FDR na poziomie q, jeśli gwarantuje się, że jego FDR będzie co najwyżej q niezależnie od wartości współczynników. Wyobraźmy sobie, że mamy procedurę, dzięki której dokonano właśnie 100 odkryć. Zakładając, że nasza procedura kontroluje FDR na poziomie 10%, oznaczać to będzie, że możemy oczekiwać, że co najwyżej 10 z tych odkryć będzie fałszywa, co za tym idzie co najmniej 90 zmiennych będzie prawdziwych. Rozpatrując dane pochodzące (lub będące wynikiem) z eksperymentu naukowego, spodziewalibyśmy się, że większość zmiennych wybranych w procedurze knockoff odpowiada rzeczywistym efektom, które można odtworzyć w wyniku eksperymentów uzupełniających.

W języku testowania hipotez interesują nas hipotezy $\mathrm{H_j}$; $\beta_j \neq 0$ i chcemy znaleźć procedurę wielokrotnych porównań, która mogłaby w stanie odrzucić indywidualne hipotezy, kontrolując FDR. Znacznie częściej używana jest terminologia z testowania hipotez, dlatego wprowadzimy inne słownictwo, które będzie przydatne w dalszej części prezentacji. Możemy powiedzieć, że H_j została odrzucona, co będzie oznaczało, że cecha j została wybrana lub możemy powiedzieć, że dane dostarczają dowodów przeciwko H_j, co oznacza, że zmienna j-ta prawdopodobnie należy do modelu.

3 Filtr knockoff

W tej części przedstawimy procedurę kontrolną FDR, która gwarantuje działanie w ramach dowolnej oraz stałej macierzy planu $\mathbf{X}$, o ile oczywiście $n \geqslant p$, a odpowiedz jest zgodna z liniowym modelem Gaussa. Ważną cechą tej procedury jest to, że nie wymaga ona znajomości poziomu szumu. Nie zakłada też żadnej wiedzy na temat ilości zmiennych w modelu, która może być dowolna. Poniżej przedstawiamy trzy etapy konstruowania knockoff.

Krok 1. Konstruowanie knockoff

Dla każdej cechy $\mathbf{X}_j$ w modelu (tzn. dla każdej j-tej kolumny macierzy $\mathbf{X}$) konstruujemy cechę knockoff $\tilde{X}$. Celem zmiennych knockoff jest naśladowanie struktury korelacyjnej oryginalnych cech w bardzo specyficzny sposób, który pozwoli na kontrole FDR.

W szczególności aby wyliczyć knockoff musimy obliczyć macierz Gram dla oryginalnych cech i po dokonaniu normalizacji tej macierz, tak aby $\Sigma_{jj} = 1$ dla każdego $j = 1, \ldots, p$. Zapewnimy, że te funkcje są zgodnie na postawie poniższych wzorów

$$\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} = \Sigma \qquad \mathbf{X}^{\mathrm{T}}\tilde{\mathbf{X}} = \Sigma - \mathrm{diag}\{s\}$$

gdzie s jest p-wymiarowym nieujemnym wektorem. Innymi słowy, możemy powiedzieć, że $\tilde{\mathbf{X}}$ ma tę samą strukturę kowariancyjną co oryginalna macierz planu z dodatkowym czynnikiem, że kowariancja między zmienna oryginalną a jej kopią knockoff jest możliwie najmniejsza. Aby upewnić się, że nasza metoda ma dobrą moc statystyczną do wykrywania sygnałów, to powinniśmy wybierać takie wielkości s tak duże, jak to możliwe, żeby zmienna $\mathbf{X}_j$ nie była zbyt podobna do jej knockoff $\mathbf{X}_j$.

Strategią konstrukcji knockoff jest taki wybór parametrów s, dla których $\mathrm{diag\{s\}} \succcurlyeq \Sigma$ i procedura konstrukcji knockoff wygląda następująco

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \Sigma^{-1}\mathrm{diag}\{s\}) + \tilde{\mathbf{U}}\mathbf{C}$$

gdzie $\tilde{\mathbf{U}}$ jest macierzą rozmiaru $n \times p$ ortogonalną do macierzy planu $\mathbf{X}$, oraz $\mathbf{C}^{\mathrm{T}}\mathbf{C} = 2\mathrm{diag}\{s\} - \mathrm{diag}\{s\}\Sigma^{-1}\mathrm{diag}\{s\}$ odpowiada dekompozycji Choleskiego.

Krok 2. Obliczanie statystyki dla każdej pary zmiennych oryginalnych i knockoff

Chcemy teraz wprowadzić statystykę W_j dla każdego $j = 1, \ldots, p$, co pomoże nam oddzielić te zmienne, które są w modelu, od tych które nie są. Te W_j są tak konstruowane, że duże wartości dodatnie są dowodem przeciwko hipotezie zerowej H_j. W naszym przypadku będzie to model LASSO, który wykorzystuje regresję z czynnikiem karu w 'L-One', która przyjmuje rzadkie oszacowanie współczynników, podane przez formułę LASSO:

$$\hat{\beta}(\lambda) = \frac{1}{2}||Y - \mathbf{X}\beta||_2^2 + ||\beta||_1$$

W przypadku rzadkich modeli liniowych wiemy, że LASSO jest asymptotycznie dokładny zarówno dla wyboru zmiennych, jak i dla oszacowania współczynnika lub sygnału. Nawet w przypadku nieasymptotycznym będziemy patrzyli na dane rozwiązanie, ponieważ wtedy nasza metoda wrzuci zasadniczo znaczące zmienne i dodatkowo będziemy mieli w zbiorze kilka fałszywych zmiennych. Przyjmując Z_j jako punkt na ścieżce LASSO, w którym X_j pierwszy raz wchodzi do modelu

$$Z_j = \sup\left\{\lambda : \hat{\beta}_j \neq 0\right\}.$$

Mamy wtedy nadzieję, że Z_j jest duże dla większości sygnałów i małe dla większości zmiennych zerowych. Jednak, aby móc to określić ilościowo i wybrać odpowiedni próg do wyboru zmiennych, musimy użyć zmiennych knockoff do skalibrowania naszego progu. Mając to na uwadze, zamiast tego obliczamy statystyki pogrupowane w pary tzn. $(X_j;\, \tilde{X}_j)$ dla każdego $j = 1, \ldots, p$ i używamy innej statystyki

$$W_j = \max\{Z_j,\, \tilde{Z}_j\} \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j \\ -1, & Z_j < \tilde{Z}_j \end{cases}$$

(dodatkowo możemy ustalić 0 w przypadku, gdy $Z_j = \tilde{Z}_j$). Duża dodatnia wartość W_j wskazuje, że zmienna X_j wchodzi do modelu LASSO wcześnie (przy pewnej dużej wartości) i robi to przed swoją kopią knockoff $\tilde{X}_j$. Jest to więc wskazówka, że zmienna ta jest prawdziwym sygnałem i należy do modelu. Możemy również rozważyć inne alternatywy dla konstruowania W_j, np. na przykład, zamiast rejestrować wprowadzanie zmiennych do modelu LASSO, możemy rozważyć metody wyboru w przód i zarejestrować kolejność, w jakiej zmienne są dodawane do modelu. W późniejszej części wskażemy kilka innych możliwości definiowania statystyki W.

Krok 3. Obliczanie progu dla statystyk zależny od danych.

Chcemy wybrać takie zmienne, aby W_j było duże i dodatnie tzn. takie $W_j > t$ dla pewnego $t > 0$. Ustalamy dowolny poziom q, który odpowiada za kontrole FDR. Dla takiego problemu zdefiniujemy próg τ następującej postaci

$$\tau = \min\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leqslant -t\}}{\max\{\#\{j : W_j \geqslant t\}, 1\}} \leqslant q\right\}$$

gdzie zbiór $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\}$ jest zbiorem unikatowych wartości niezerowych osiąganych przez W_j. Zobaczymy, że ułamek pojawiający się powyżej jest oszacowaniem odsetka fałszywych odkryć, gdybyśmy wybrali wszystkie cechy $W_j \geqslant t$. Z tego powodu często będziemy nazywać ten ułamek 'szacunkową wartością FDP'.

Na rysunku przedstawiono reprezentacje danego korku, na którym wykreślono punkty $(Z_j, \tilde{Z}_j)$ dla każdej cechy j. Czarne kropki oznaczają cechy zerowe, a natomiast czerwone punkty oznaczają sygnał rzeczywisty. Przypomnijmy, że W_j jest dodatnia, jeśli oryginalna zmienna zostanie wybrana przed jej knockoffem ($Z_j > \tilde{Z}_j$) w przeciwnym razie jest ujemna ($Z_j < \tilde{Z}_j$). Dlatego cecha j, której punkt leży poniżej przerywanej linii ukośnej na rysunku, ma wtedy dodatnią wartość W_j, podczas gdy punktom powyżej przekątnej przypisuje się ujemne W_j.

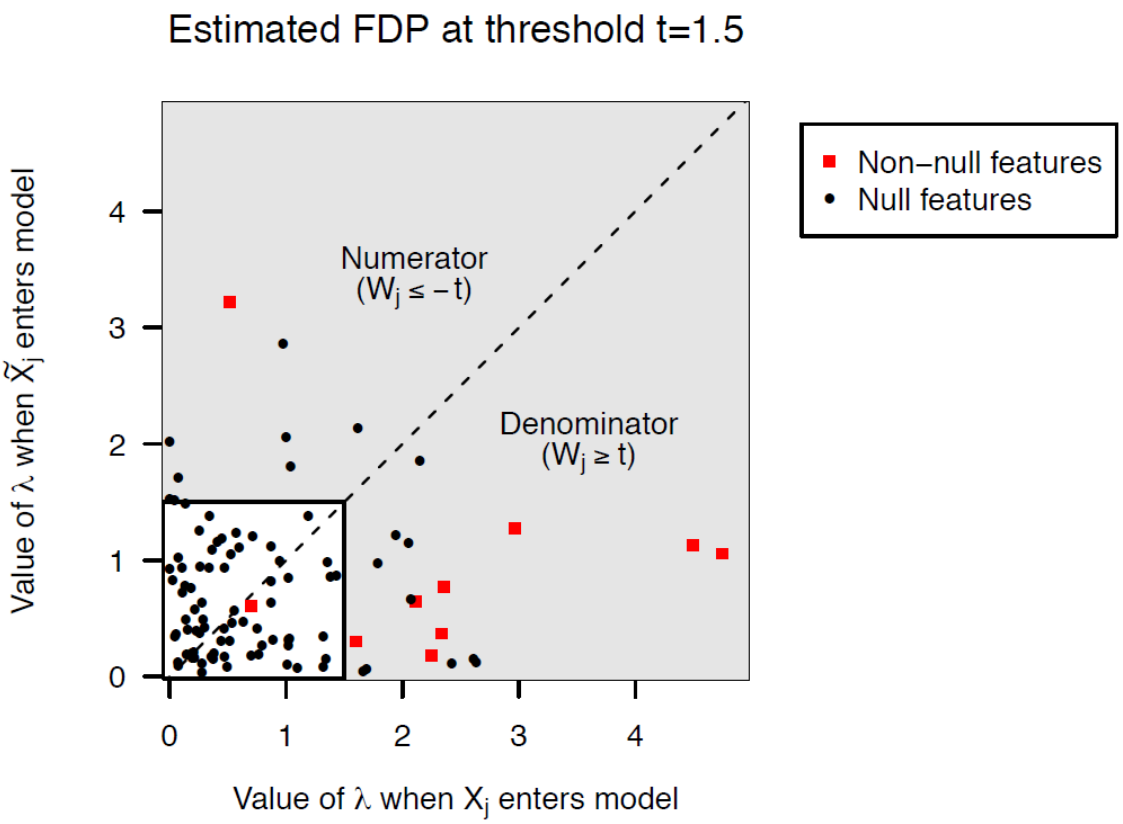
Przejdźmy teraz do formalnego zdefiniowania pierwszej wersji knockoff

Definicja 1 (**Knockoff**). *Skonstruuj $\tilde{X}$ jak w powyższych powyższej procedurze i oblicz statystyki W_j spełniające właściwości wystarczalności i antysymetrii. Następnie wyliczamy zbiór potencjalnych fałszywych hipotez zerowych*

$$\hat{S} = \{j : W_j \geqslant T\}$$

gdzie T jest progiem dla procedury ściśle związanym z danymi.

Jedną z najważniejszych cech tej procedury jest kontrola FDR na poziomi q. Mówi o tym następujące twierdzenie



Rysunek 1: Przykład wykorzystania knockoffów w przypadku macierzy gaussowskiej oraz gdzie sygnał jest dość słaby.

Twierdzenie 1. *Dla dowolnego $q \in [0,1]$ metoda knockoff spełnia wymagania*

$$\mathrm{E}\left[\frac{\#\{j:\ \beta_j = 0 \ \text{ oraz } \ j \in \hat{S}\}}{\#\{j:, j \in \hat{S}\} + q^{-1}}\right] \leqslant q$$

gdzie przyjmujemy szum Gaussa z modelu liniowego, dla którego $\boldsymbol{X}$ oraz $\tilde{\boldsymbol{X}}$ są ustalone.

Istnieje również bardziej konserwatywna procedura, która jest zdefiniowana w następujący sposób

Definicja 2 (**Knockoff+**). *Wybieramy model zgodnie z **Definicją 1** ale z innym progiem T*

$$T = \min\left\{t \in \mathcal{W}:\ \frac{1 + \#\{j:\ W_j \leqslant -t\}}{\max\{\#\{j:\ W_j \geqslant t\}, 1\}} \leqslant q\right\}$$

Wyjaśniliśmy, dlaczego duża dodatnia wartość W_j niesie pewne dowody przeciwko hipotezie zerowej $H_j:\ \beta_j = 0$ a teraz podajemy krótką intuicję, w jaki sposób nasz konkretny wybór progu umożliwia kontrolę FDR (lub zmodyfikowanego FDR). Sposób, w jaki skonstruowano statystyki W, sugeruje, że znaki W_j są i.i.d. dla 'hipotez zerowych'; to znaczy dla tych j jest takich, że $\beta_j = 0$. Dlatego dla dowolnego progu t zachodzi następująca równość rozkładów

$$\#\{j:\ \beta_j = 0 \ \text{ oraz } \ W_j \geqslant t\} \stackrel{d}{=} \#\{j:\ \beta_j = 0 \ \text{ oraz } \ W_j \leqslant -t\}$$

Powracając do naszego rysunku, który przedstawia nam przykład zagadnienia knockoff, możemy zauważyć, że punkty 'zerowe' są rozmieszczone w przybliżeniu symetrycznie na przekątnej.

Wszystko jest idealnie ale nadal nie mamy estymatora proporcji fałszywych odkryć, która jest istotna w przypadku wyliczania i kontroli FDR. Ale to możemy w dość łatwy sposób przedstawić, ponieważ FDP (proporcja fałszywych odkryć), to po prostu ilość fałszywych odkryć podzielona przez ilość wszystkich odkryć, co można zapisać w następującej postaci

$$\frac{\#\{j:\ \beta_j = 0 \ \text{ oraz } \ W_j \geqslant t\}}{\max\{\#\{j:\ W_j \geqslant t\}, 1\}} \approx \frac{\#\{j:\ \beta_j = 0 \ \text{ oraz } \ W_j \leqslant -t\}}{\max\{\#\{j:\ W_j \geqslant t\}, 1\}} \leqslant \frac{\#\{j:\ W_j \leqslant -t\}}{\max\{\#\{j:\ W_j \geqslant t\}, 1\}} = \widehat{\mathrm{FDP}}(t)$$

gdzie $\widehat{\text{FDP}}(t)$ jest estymatorem FDP(t). Procedurę knockoff można interpretować jako znalezienie progu przez $T = \{t \in \mathcal{W} : \widehat{\text{FDP}}(t) \leqslant q\}$.

4 Konstruowanie próbek knockoff

Zaczniemy od konstruowania knockoff $\tilde{X}$ i macierzy $\Sigma = X^{\text{T}}X$. Abyśmy mogli zastosować tę metodę musimy założyć, że $n \geqslant p$.

4.1 Naturalny przypadek $n \geqslant 2p$

Zarys danego przypadku został już przedstawiony wcześniej, ale teraz zajmiemy się bardziej szczegółowo tym przypadkiem. Macierz $\tilde{X}$ powinna spełniać następującą strukturę kowariancji

$$[X\ \tilde{X}]^{\text{T}}[X\ \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix} = \mathbf{G}$$

gdzie s jest p-wymiarowym wektorem. Warunkiem koniecznym i wystarczającym, aby $\tilde{X}$ istniał, jest to, aby macierz G była pół-dodatnio określona. Używając reguły Schur możemy zredukować trochę problem do następującej postaci

$$A = \Sigma - (\Sigma - \text{diag}\{s\})\Sigma^{-1}(\Sigma - \text{diag}\{s\}) = 2\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\text{diag}\{s\}$$

macierz A powinna być również pół-dodatnio określona. Zapisując macierz A w postaci macierzowej mamy

$$\begin{bmatrix} \Sigma & \text{diag}\{s\} \\ \text{diag}\{s\} & 2\text{diag}\{s\} \end{bmatrix} \succcurlyeq 0 \Leftrightarrow \begin{array}{l} \text{diag}\{s\} \succcurlyeq 0 \\ 2\Sigma - \text{diag}\{s\} \succcurlyeq 0 \end{array}$$

Na tej postawie możemy już skonstruować kopie knockoff, która jest następującej postaci

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \Sigma^{-1}\text{diag}\{s\}) + \tilde{\mathbf{U}}\mathbf{C}$$

gdzie macierz $\tilde{\mathbf{U}}$ jest macierzą wymiaru $n \times p$ ortogonalną do kolumn macierzy planu $\mathbf{X}$ (czyli $\tilde{\mathbf{U}}\mathbf{X} = 0$), dopóki macierz A jest dodatnio określona to możemy dokonać dekompozycji Choleskiego ($A = C^{\text{T}}C$), gdzie $\mathbf{C}$ jest wymiaru $p \times p$.

Teraz, gdy rozumiemy warunek s niezbędny do zaistnienia cech knockoff z pożądaną strukturą korelacji, pozostaje przedyskutować, który z nich powinniśmy skonstruować, to znaczy określić wybór s. Metodologia będzie użyteczna tylko wtedy, gdy zmienne, które naprawdę należą do modelu, są wybierane przed ich efektami negatywnymi, ponieważ w przeciwnym razie nie mielibyśmy żadnej mocy. Wyobraź sobie, że zmienna X_j jest w prawdziwym modelu. Następnie chcielibyśmy, aby X_j wchodziło przed $\tilde{X}_j$. Aby tak się stało, potrzebujemy, aby korelacja między $\tilde{X}_j$ a rzeczywistym sygnałem była mała, aby $\tilde{X}_j$ nie wchodziło wcześnie do modelu LASSO. Innymi słowy, chcielibyśmy, aby X_j i $\tilde{X}_j$ były względem siebie jak najbardziej ortogonalne. W sytuacji, w której cechy są znormalizowane $\Sigma_{jj} = 1$ dla wszystkich $j = 1, \ldots, p$, chcemy aby $\tilde{X}_j^{\text{T}}X_j = 1 - s_j$ było jak najbliższe zeru. Poniżej rozważmy dwa szczególne przypadki:

- *Equi-correlated knockoffs*: W tym przypadku $s_j = \min\{2 \cdot \lambda_{\min}(\Sigma), 1\}$ dla każdego $j = 1, \ldots, p$, więc wszystkie korelacje przyjmują tę samą wartość

$$< X_j; \tilde{X}_j >= 1 - \min\{2 \cdot \lambda_{\min}(\Sigma), 1\}$$

- *SDP knockoff*: Inną możliwością jest wybranie odchyleń, tak aby średnia korelacja między oryginalną zmienną a jej odchyleniem była minimalna. Odbywa się to poprzez rozwiązanie problemu wypukłości

$$\begin{array}{ll} \text{minimize} & \sum_j |1 - s_j| \\ \text{subject to} & s_j \geqslant 0 \\ & 2\,\Sigma \succcurlyeq \text{diag}\{s\} \end{array}$$

4.2 Przypadek gdy $p \leqslant n < p$

Kiedy $n < 2p$, nie możemy już znaleźć podprzestrzeni wymiaru p, która jest ortogonalna do X, więc nie możemy skonstruować $\tilde{\mathbf{U}}$ jak powyżej. Możemy jednak nadal używać filtru knockoff, o ile poziom szumu jest znany lub można go oszacować - na przykład w modelu szumu Gaussa (1.1) możemy wykorzystać fakt, że rezydualna suma kwadratów z pełnego model ma rozkład $||y - X\hat{\beta}||_2^2 \sim \chi^2_{n-p}$, gdzie $\hat{\beta}$ jest estymatorem najmniejszych kwadratów. Teraz niech $\hat{\sigma}^2$ będzie estymatorem σ^2 oraz nasz nowy wektor $y' \sim N(0, \hat{\sigma}^2\mathbf{I})$. Jeśli $n - p$ jest duże, wówczas $\hat{\sigma}$ będzie niezwykle dokładnym oszacowaniem i możemy postępować tak, jakby σ i $\hat{\sigma}$ były równe. Następnie zwiększamy wektor odpowiedzi y nowym wektorem y' długości $(n - p)$ i powiększamy macierz projektu X o $n - p$ wierszy zer. Tym samym

$$\begin{bmatrix} y \\ y' \end{bmatrix} \sim N\left(\begin{bmatrix} X \\ 0 \end{bmatrix}\beta,\ \sigma^2\mathbf{I}\right)$$

Mamy teraz model liniowy ze zmiennymi p i obserwacjami $2p$, więc możemy zastosować filtr knockoff do tych danych powiększonych wierszami, używając metody opisanej dla ustawienia $n \geqslant 2p$.

5 Własności statystyki W

1. Mówi się, że statystyka W jest zgodna z właściwością wystarczalności, jeśli W zależy tylko od macierzy Grama i od wewnętrznych iloczynów odpowiedzi cechy; to znaczy możemy pisać

$$W = f\left([X\ \tilde{X}]^{\mathrm{T}}[X\ \tilde{X}],\ [X\ \tilde{X}]^{\mathrm{T}}y\right)$$

2. Mówi się, że statystyka W jest zgodna z właściwością antysymetrii, jeśli zamiana X_j i $\tilde{X}_j$ ma wpływ na zmianę znaku W_j - to znaczy dla dowolnego S

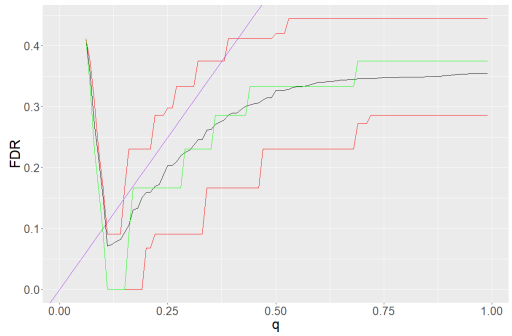
$$W_j\left([X\ \tilde{X}]_{\mathrm{swap}(S)}, y\right) = W_j\left([X\ \tilde{X}], y\right) \cdot \begin{cases} +1 & j \in S \\ -1 & j \notin S \end{cases}$$

6 Możliwe zastosowania

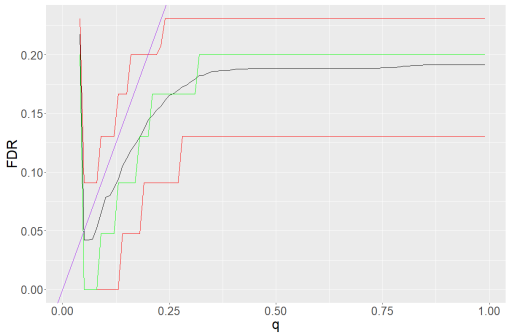
Przedstawimy tutaj kilka możliwych zastosowań knockoffów:

1. Gaussowskie modele - zwykła regresja liniowa, gdzie zakładamy związek liniowy pomiędzy y a macierzą planu X
2. Modele Markowa
3. Ukryte modele Markowa

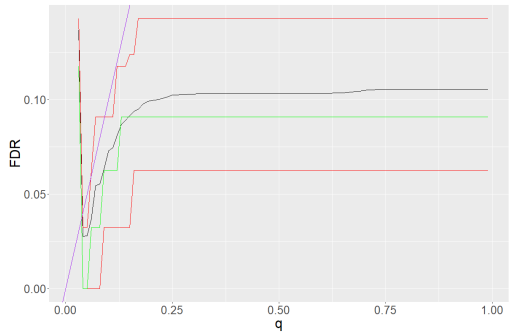
7 Symulacje numeryczne



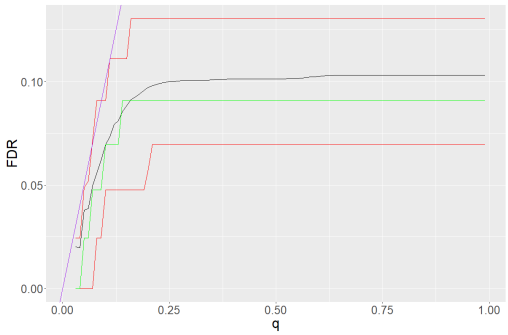
(a) $k = 10$



(b) $k = 20$

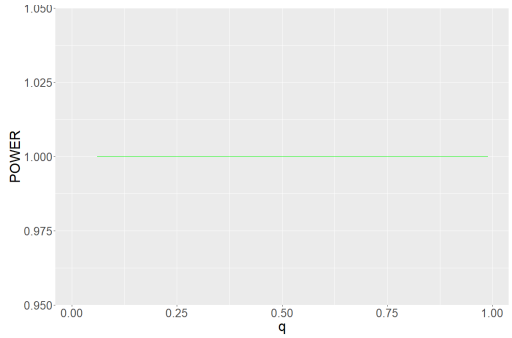


(c) $k = 30$

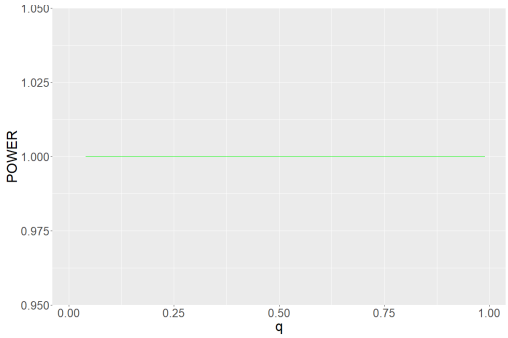


(d) $k = 40$

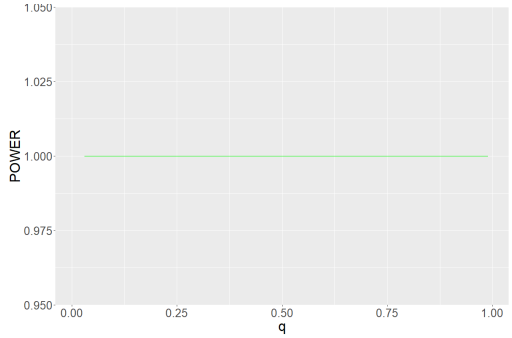
Rysunek 2: FDR



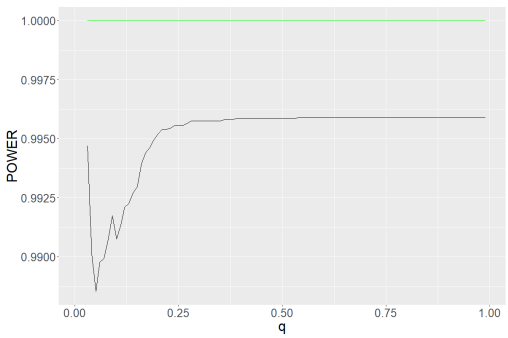
(a) $k = 10$



(b) $k = 20$



(c) $k = 30$



(d) $k = 40$

Rysunek 3: Moc

Literatura

[1] R. Foygel Barber E. J. Candès *Controlling the false discovery rate via knockoffs* The Annals of Statistics. 2015. link

[2] M. Sesia C. Sabatti E. J. Candès *Gene hunting with knockoffs for hidden Markov models (with discussion)* Biometrika. 2019 link

[3] `https://github.com/zhimeir/adaptiveKnockoffs/`

[4] `https://github.com/zhimeir/derandomKnock`