Porównanie błędu predykcji dla różnych metod estymacji współczynników w modelu liniowym, scenariusz p bliskie lub większe od n

Przemyslaw.Biecek@gmail.com, MIM Uniwersytet Warszawski

Plan prezentacji:

1. Motywacja;
2. Błąd predykcji a różne metody estymacji współczynników w modelu liniowym;
3. Co nam daje wyjście poza liniowy predyktor;
4. Komitet predyktorów i bootstrapowy estymator błędu predykcji;
5. Podsumowanie.

Motywacja

Przeprowadzono badania mikromacierzowe na szeroką skalę.
Rozważmy badanie, którego celem jest umożliwienie predykcji wartości hodowlanej osobników, np. mleczności bydła.
Dla 2000 osobników wyznaczono ich wartość hodowlaną (zmienna objaśniana Y) oraz zebrano informację o genotypach w $p = 1800$ pozycjach (zmienne objaśniające X).

Zakładamy, że zależność pomiędzy zmiennymi objaśniającymi a zmienną objaśnianą można opisać modelem liniowym, tz.

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

gdzie $\varepsilon \sim N(0, \sigma^2)$. Nie znamy ani β, ani σ^2.

Cel: Na podstawie zebranych danych poproszono nas o zbudowanie „możliwie dobrego" predyktora $\hat{f}(x, X, Y)$ wartości hodowlanej.

Podejście I - estymujemy wszystkie efekty w modelu

Zbudujmy predyktor w oparciu o model liniowy tz.

$$\hat{f}(x, X, Y) = X\hat{\beta}.$$

Współczynniki $\hat{\beta}$ wyznaczmy korzystając z estymatorów BLUE dla modelu liniowego, czyli

$$\hat{\beta}^{lin} = (X^T X)^- X^T Y,$$

gdzie A^- oznacza uogólnioną odwrotność Moore-Penrose'a macierzy A.
Jak wiadomo tak wyznaczone współczynniki minimalizują błąd dopasowania

$$RSS = (Y - X\hat{\beta}^{lin})^T (Y - X\hat{\beta}^{lin}).$$

Nas jednak interesuje błąd predykcji!
Można go różnie definiować. My przyjmiemy następującą definicję

$$PE(X, Y, \hat{f}(.)) = \sum_{x_i \in X} (x_i\beta - \hat{f}(x_i, X, Y))^2 = ||X\beta - X\hat{\beta}^{lin}||^2.$$

Podejście I - estymujemy wszystkie efekty w modelu

Symulacyjnie sprawdzimy jak dla rozważanych parametrów (p=1800, n=2000) zachowuje się błąd predykcji dla predyktora opartego o model liniowy $\hat{f}^{lin}(x, X, Y) = X\hat{\beta}^{lin}$.

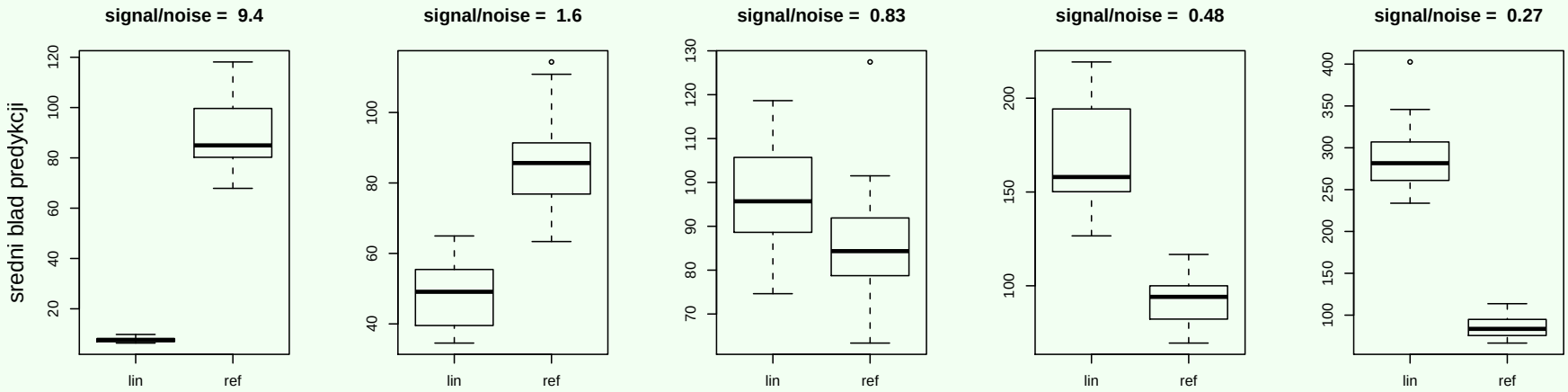
Do porównania, jako referencyjny predyktor zastosujemy predyktor będący zwykłą średnią arytmetyczną (tzw. model zerowy) $\hat{f}^{ref}(x, X, Y) = \bar{X}$.

Schemat symulacji:
Wykonując $N = 1000$ powtórzeń losujemy macierze X (kolumny losujemy niezależnie) i badamy rozkład rzeczywistego błędu predykcji PE w zależności od stosunku szumu do sygnału.

Szumem określamy $E[\varepsilon^T\varepsilon]$ a sygnałem $E[(X\beta)^T(X\beta)]$ (macierz X jest generowana losowo, β jest nieznane ustalone). Współczynnik sygnału do szumu wyznaczany jest jako

$$s/n = \frac{E[(X\beta)^T(X\beta)]}{E[\varepsilon^T\varepsilon]}.$$

Podejście I - estymujemy wszystkie efekty w modelu, wyniki



Jak widzimy im większy udział szumu tym gorsze właściwości predykcyjne predyktora liniowego (to zachowanie intuicyjne).

Jednocześnie z zaskoczeniem obserwujemy, że dla dużego szumu błąd predykcji dla modelu liniowego przewyższa błąd predykcji dla naiwnego predyktora - średniej!!!

Czyżby dla niskiego sygnału najlepszym predyktorem była zwykła średnia z obserwacji?

Podejście II - estymujemy efekty w „optymalnym" modelu

Estymacja wszystkich efektów w tak dużym modelu oczwiście niesie ze sobą problemy. Jak pamiętamy

$$\hat{\beta}^{lin} \sim \mathcal{N}(\beta, (X^T X)^{-} \sigma^2).$$

Więc dla macierzy $X^T X$ bliskiej osobliwej wariancje $\hat{\beta}_i^{lin}$ są bardzo duże. Stąd prawdopodobnie problemy, które zaobserwowaliśmy.
Dlatego też często do predykcji wykorzystuje się model tylko z wybranymi kolumnami X odpowiadającymi najbardziej istotnie różnym od zera wartościom $\hat{\beta}$ (w rzeczywistości wszystkie β są różne od 0).

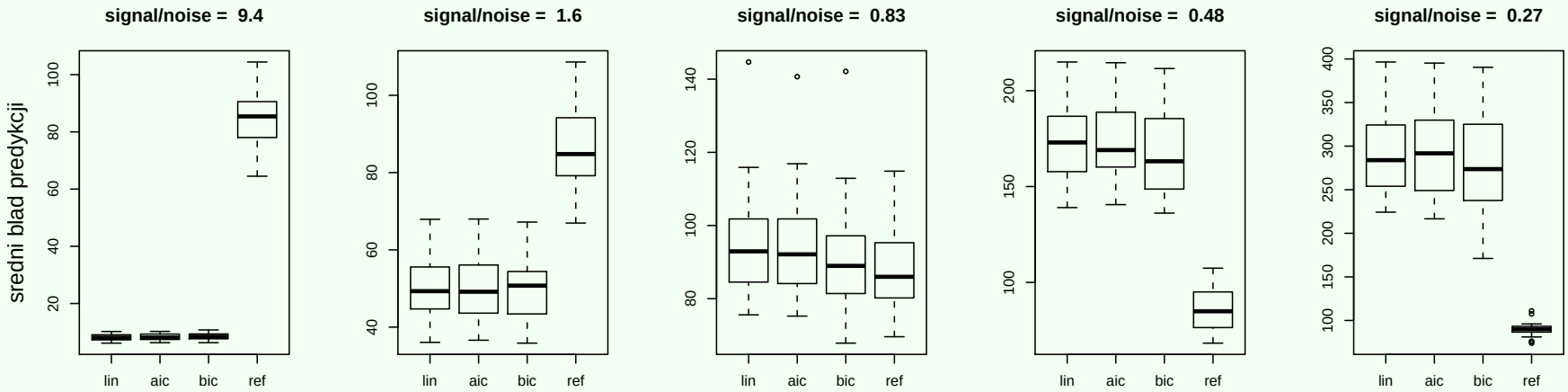
Zbudujemy predyktor oparty na modelu liniowym, ale z podzbiorem tych kolumn macierzy X, które wybierze kryterium AIC lub BIC.

Predyktorem opartym o kryterium AIC jest

$$\hat{f}^{AIC}(x, X, Y) = X\hat{\beta}^{AIC},$$

gdzie $\hat{\beta}^{AIC}$ są współczynnikami wyestymowanymi w modelu wybranym przez kryterium AIC. Podobnie konstruujemy predyktor $\hat{f}^{AIC}(x, X, Y)$.
Kryterium AIC jest uznawane za kryterium o dobrych właściwościach predykcyjnych!!!

Podejście II - estymujemy efekty w „optymalnym” modelu, wyniki



W rozpatrywanym scenariuszu, błąd predykcji przy modelu wybranym przez AIC lub BIC jest podobny do błędu predykcji pełnego modelu.

Dla niskiego sygnału błąd predyktora opartego o kryterium AIC czy BIC również przekracza błąd predykcji dla predyktora równemu średniej!

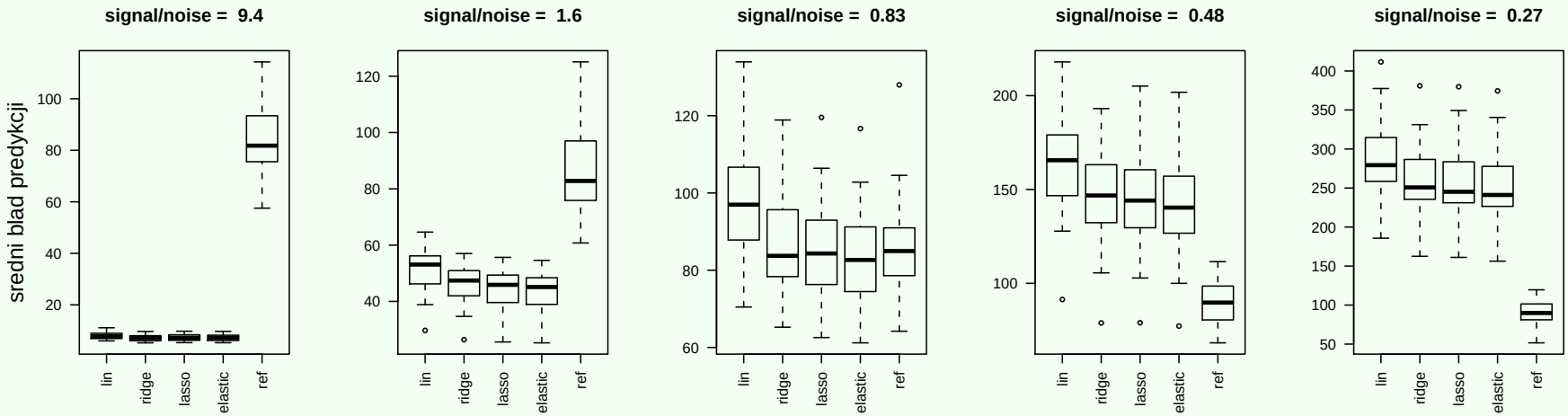
Podejście III - regularyzujemy model

Kolejna próba zbudowania dobrego predyktora oparta będzie o model regularyzowany.
W sytuacji gdy $p > n$ częstym rozwiązaniem jest regularyzacja modelu prowadząca do wprowadzenia obciążenia, które jednak znacznie redukuje wariancję estymatora $\hat{\beta}$.

Najprostszą i najstarszą techniką regularyzacji modelu liniowego jest regresja grzbietowa (równoważna dodaniu wartości λ do przekątnej macierzy $X^T X$). Również dobre recenzje zbierają jej różnorodne odmiany lub uogólnienia, takie jak regresja lasso czy ogólniejsza rodzina: sieci elastyczne. Wszystkie te metody są równoważne z minimalizacją błędu *RSS* z dodatkową karą za wielkość współczynników $\hat{\beta}$, przy czym wielkość $\hat{\beta}$ może być mierzona w normie L_1, L_2 lub mieszanej.

Rozważmy więc trzy kolejne predyktory, odpowiadające estymacji współczynników modelu liniowego z użyciem metod: regresja grzbietow, lasso, sieci elastyczne ($\alpha = 0.5$).

Podejście III - regularyzujemy model, wyniki



W zaprezentowanym scenariuszu regularyzacja zmniejsza błąd predykcji, jeżeli porównywać go do zwykłej regresji liniowej w modelu pełnym.

Znacznie dłużej (nwet dla s/n=0.8) udaje się otrzymać wyniki niegorsze niż predyktor oparty na samej średniej.

Niestety dla małych wartości s/n nawet regularyzowany model liniowy nie daje zadowalających wyników. Błąd predykcji jest większy niż błąd predykcji dla samej średniej.

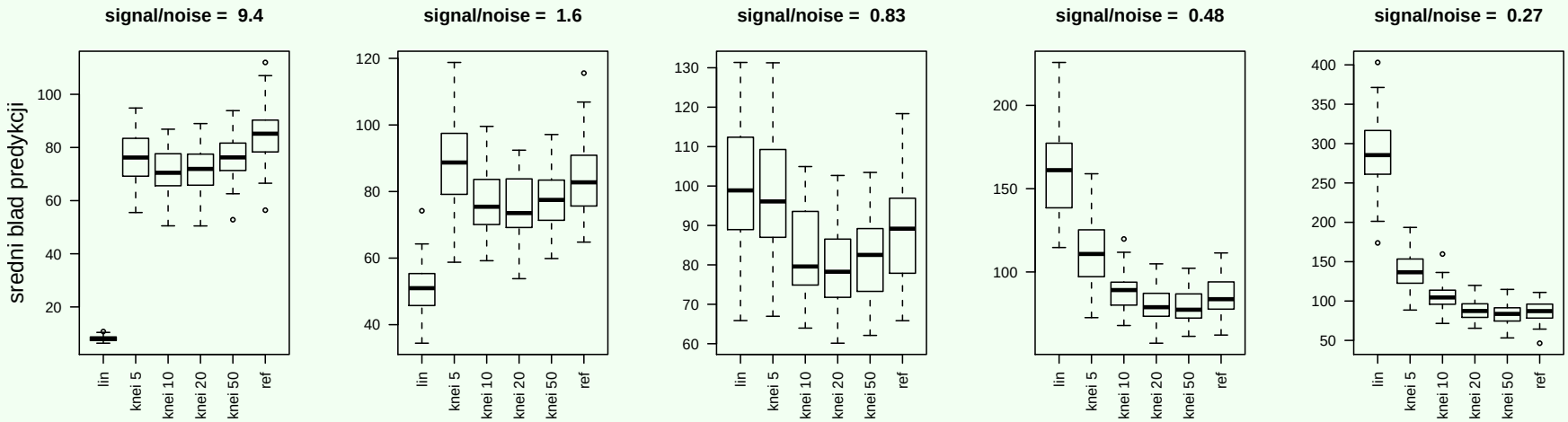
Podejśce IV - szukamy predyktorów poza modelem liniowym

Rozważmy predykor oparty o metodę k-sąsiadów. Będziemy uśredniać Y nie po wszystkich obserwacjach, ale po k najbliższych obserwacjach.

Wartość $\hat{f}(x, X, Y)$ wyznaczmy zgodnie z następujących schematem:

1. Znajdujemy k punktów ze zbioru X, których odległość $d(x, X_i)$ jest najmniejsza. Jeżeli x jest z X to usuwamy go, by „nie pomagać". Oznaczmy zbiór tych k-najbliższych sąsiadów przez X'.
2. Wyznaczmy predyktor jako średnią wartość zmiennej Y dla obserwacji ze zbioru X'
$$\hat{f}(x, X, Y)) = \frac{1}{\#X'} \sum_{i: X_i \in X'} Y_i.$$
3. Odległość $d(x, y)$ możemy wybrać dowolnie, wybierzmy ważoną odległość euklidesową. Za wagi wybierzmy wartości współczynników β^{lin}. Podsumowując
$$d(x, X_i) = \sqrt{\sum_{j=1}^{p} |\hat{\beta}_j^{lin}|(x_j - X_{ij})^2}.$$

Metoda k-sąsiadów, wyniki



Predyktor oparty o metodę k-sąsiadów pozwala na uzyskanie błędu predykcji niegorszego niż predyktor referencyjny (sama średnia) dla małych wartości signal/noise, a dla średnich wartości s/n znacznie lepszego niż pozostałe metody.

Dla dużych wartości s/n ta metoda radzi sobie jednak źle.

Błędy predykcji nawet jeżeli są mniejsze od błędu predykcji samej średniej to i tak są wysokie. Jak się jednak okaże, dla określonych β i „skorelowanych" kolumn X błąd predykcji metody k-sąsiadów jest kilkukrotnie mniejszy od błędu predykcji metody referencyjnej.

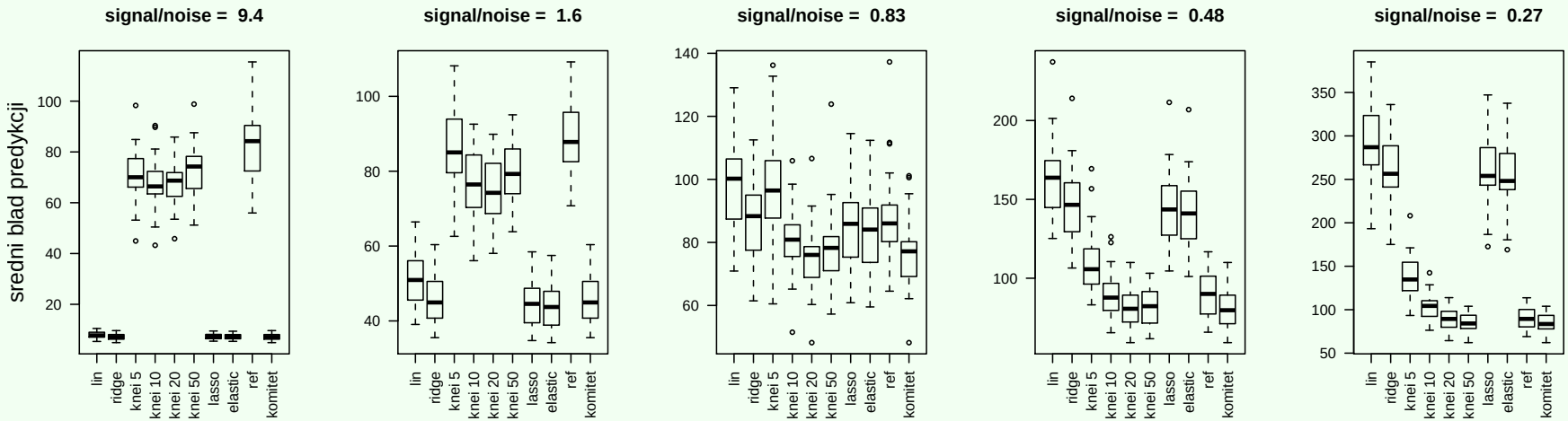
Komitet predyktorów

Zauwazyliśmy, że dla różnych wartości stosunku sygnału do szumu, oraz różnych parametrów modelu **inna** metoda konstrukcji predyktora charakteryzuje się najmniejszym będem predykcji. Naturalnym pomysłem jest próba użycia predyktora wybranego z komitetu predyktorów jako najlepszego dla zadanej macierzy X i Y. Jak wybierać?

Do wyboru kondydata wykorzystamy metodę bootstrapu parametrycznego do oceny błędu predykcji. Schemat tej metody można opisać następująco:

1. Wyznaczamy $\hat{\beta}$ korzystając z dowolnej metody estymacji (oznaczmy $\beta' := \hat{\beta}$),
2. Generujemy $B = 100$ replikacji zbioru danych, każda replikacja generowana jest z modelu $Y^{'*} = X\beta' + \varepsilon^*$, gdzie $\varepsilon^* \sim N(0, \hat{\sigma}^2)$,
3. Dla wszystkich B replikacji wyznaczamy predyktor każdą z rozważanych metod oraz liczymy błąd predkcji (dla zadanej macierzy X, ale w punkcie $\hat{\beta}$, który niestety może być daleko od β),
4. Wybieramy metodę o najmniejszym błędzie predykcji na próbach bootstrapowych i używamy jej do wykonania predykcji na oryginalnej próbie.

Komitet predyktorów, wyniki



Predyktor oparty na komitecie omawianych predyktorów i na bootstrapowym estymatorze błędu predykcji ma porównywalny błąd predykcji co lokalnie najlepszy z rozważanych predyktorów.

Więc, nawet jeżeli błąd predykcji jest badany w punkcie β' a nie β, to wciąż może być z powodzeniem wykorzystywany do wyboru predyktora.

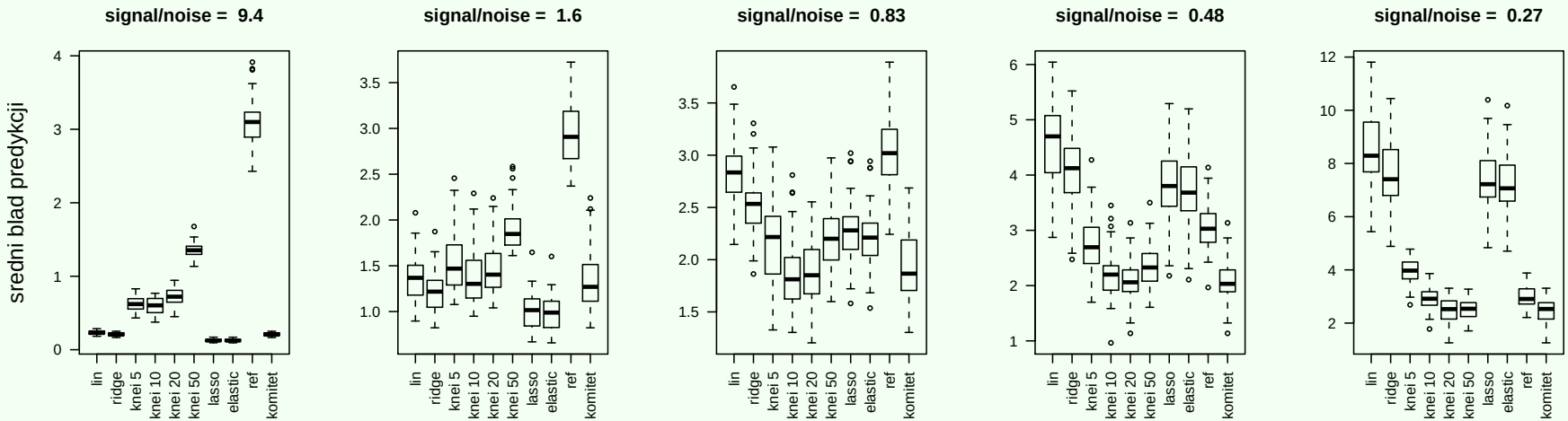
Co się dzieje jeżeli generowane są skorelowane X

W omawianym scenariuszu wartości macierzy X losowane były niezależnie, a wartości współczynników β wszystkie były równe określonej wartości różnej od zera (prawdziwym modelem był model pełny).

Okazuje się, że zarówno w sytuacji gdy w prawdziwym modelu coraz więcej współczynników β_i jest równa zero, jak i w sytuacji gdy kolumny generowane są z zadaną korelacją, w obu przypadkach otrzymujemy podobny ranking ze względu na błąd predykcji.

Jednak różnice pomiędzy najlepszym predyktorem a predyktorem liniowym stają się coraz większe, coraz częściej najlepsze wyniki uzyskuje się przy użyciu metody k-sąsiadów.

Komitet predyktorów, wyniki



W badanym scenariuszu kolumny macierzy X są skorelowane z macierzą kowariancji równą $\rho = 0.9$ poza przekątną i 1 na przekątnej a wektor β w prawdziwym modelu ma wiele współczynników równych 0.

Analizując wyniki obserwujemy znacznie lepsze zachowanie metody k-sąsiadów dla małych wartości s/n oraz bardzo dobre zachowanie metody opartej o komitet predyktorów dla rozważanych wartości s/n.

Podsumowanie

Trzy najważniejsze wnioski, które warto zabrać do domu:

1. Jeżeli stosunek sygnału do szumu jest niski (co jest bardzo częste np. dla cech o słabym podłożu genetycznym takich jak: łatwość uzależniania się, łatwość tycia, ryzyko chorób typu schizofrenia itp), to predykcja z użyciem niepustego modelu liniowego może dawać gorsze wyniki niż predykcja równa średniej z próby $\bar{Y}$,
2. Ta obserwacja zachowuje się dla różnych metod generowania macierzy X, oraz dla wielu różnych przebadanych wartości wektora β. Obserwacja ta przenosi sie również na przypadki gdy $p > n$ i $p >> n$.
3. Użycie bootstrapowego estymatora błędu predykcji (dla danego X ale w β' zamiast w nieznanym β) pozwala na skonstruowanie komitetu predyktorów o równie dobrych właściwościach predykcyjnych co najlepszy z elementów komitetu.