Wstęp do wyboru modelu liniowego i selekcji czynnikowej

Agnieszka Sołtys

Plan warsztatu

1. Definicja modelu regresji liniowej i metody najmniejszych kwadratów
2. Dlaczego model liniowy jest często używany?
3. Po co zmniejszać liczbę zmiennych w modelu?
4. Kryteria wyboru modelu (ocena błędu predykcji):
   - kroswalidacja
   - kryteria informacyjne
5. Klasyczne metody wyboru modelu liniowego:
   - metody krokowe
   - regularyzacja
   - selekcja czynnikowa

Definicja modelu regresji liniowej i metody najmniejszych kwadratów (MNK)

Model regresji liniowej

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

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix},$$

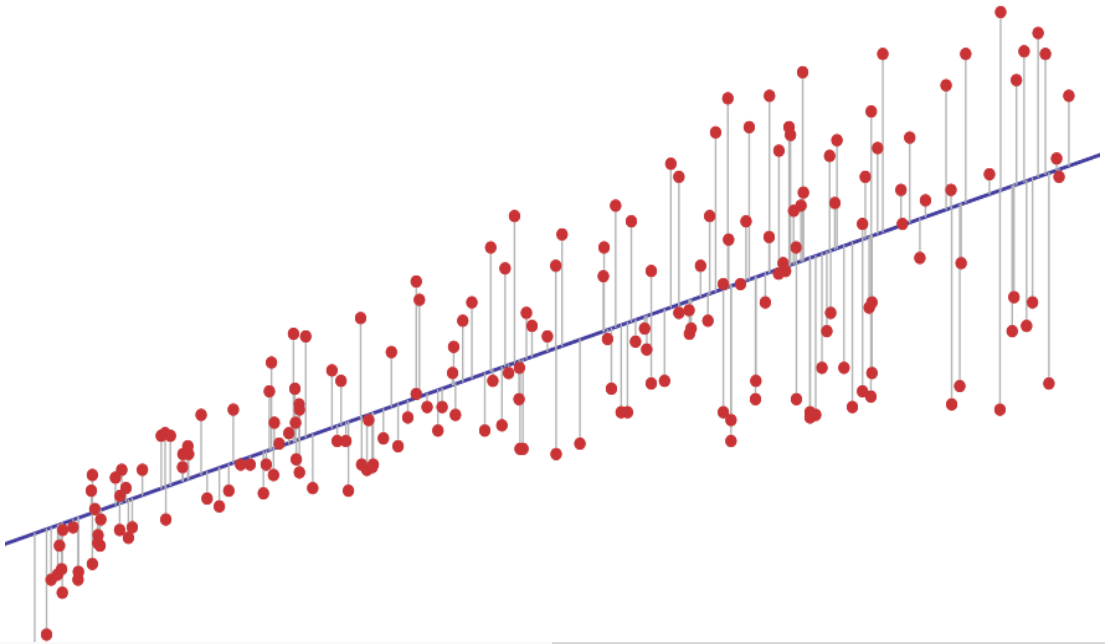
- y zmienna objaśniana,
- $\mathbb{X}$ macierz planu, $\mathbb{X} = \begin{pmatrix} X_1 \\ \dots \\ X_n \end{pmatrix}$
- β parametry, które chcemy estymować,
- ϵ wektor efektów losowych, $\mathbb{E}(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I$.

Estymator najmniejszych kwadratów

Estymator $\widehat{\beta}^{mnk}$: szukamy wartości parametru, dla której odległość (euklidesowa) danych od prostej (hiperpłaszczyzny) je przybliżającej jest najmniejsza:

$$RSS = \|y - \mathbb{X}\beta\|^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2$$

$$\widehat{\beta}^{mnk} = \arg\min_{\beta} RSS = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$$



Dlaczego model liniowy jest często używany?

Dlaczego model liniowy jest często używany?

- **Prostota** - przybliżenie danych za pomocą linii prostej (hiperpłaszczyzny).
- **Interpretowalność** - łatwo zauważyć i wyjaśnić biznesowi, jak zmienne wpływają na model.
- **Rozpoznawalność** - używany już od 1805r., dobrze znany i zbadany.
- **Szybkość** - nawet przy dużej liczbie danych wyniki dostajemy prawie od razu - wyliczamy ze wzoru.

Szerokie wykorzystanie w wielu dziedzinach wiedzy, np. w biologii, ekonomii, czy socjologii.
Często prosty model z dobrze dobranymi zmiennymi (inżynieria cech i wybór modelu) może dawać lepsze efekty niż bardziej złożone modele.

Po co zmniejszać liczbę zmiennych w modelu?

Po co zmniejszać liczbę zmiennych w modelu?

- **Interpretowalność** - dużo łatwiej zinterpretować i wyjaśnić wpływ zmiennych w modelu z 10 zmiennymi niż w tym ze 100 zmiennymi.
- **Jakość predykcji** - dekompozycja obciążenie-wariancja (bias-variance decomposition) dla estymatora w modelu z kwadratową funkcją straty:

$$PredictionError = IrreducibleError + Bias^2 + Variance$$

  - Im więcej zmiennych w modelu, tym model jest lepiej dopasowany i *Bias* jest mniejszy.
  - Im więcej zmiennych w modelu, tym model ma większą wariancję *Variance*.

Kryteria wyboru modelu (ocena błędu predykcji)

Ocena błędu predykcji

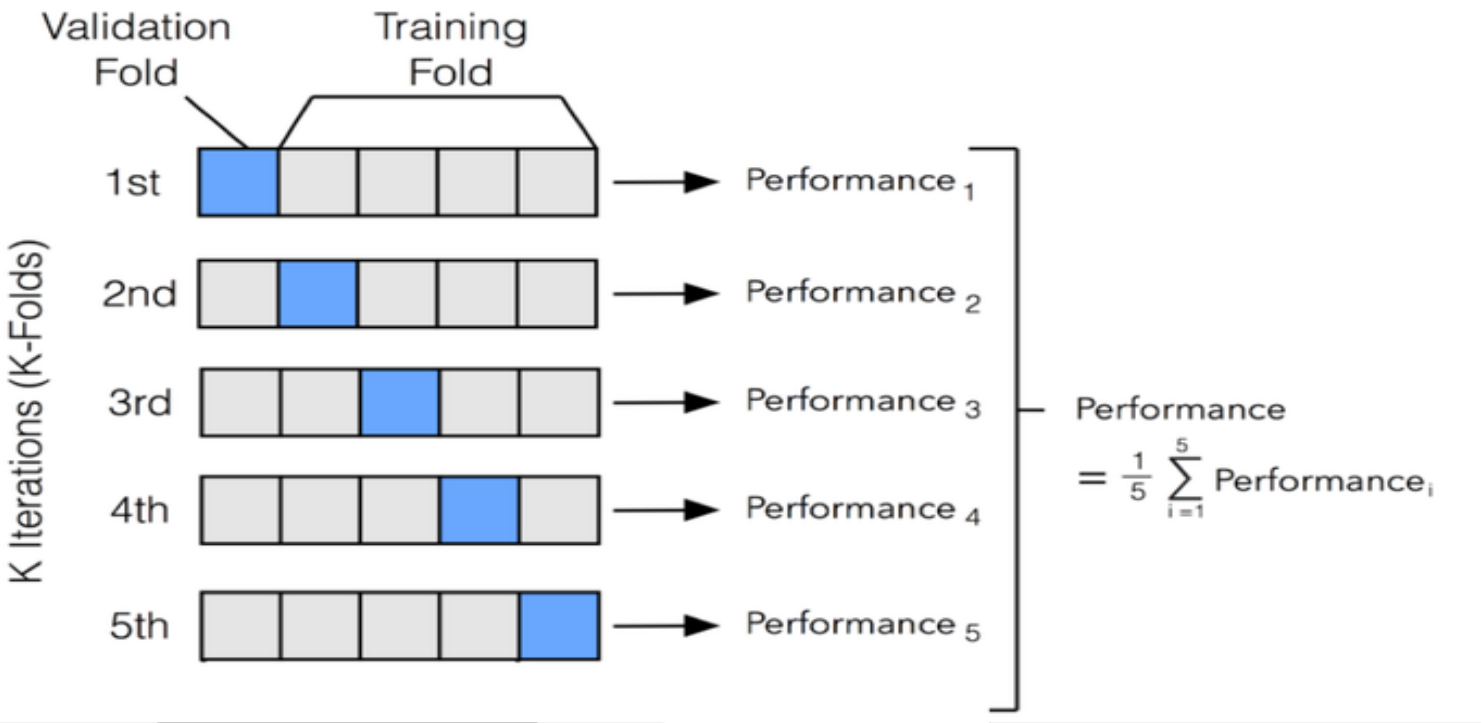
Ogólna zasada oceny błędu predykcji:
Na danych **treningowych** uczymy model (wyznaczamy $\widehat{\beta}$), a na danych **testowych** obliczamy błąd predykcji:

$$MSE = \frac{\|y^{te} - \mathbb{X}^{te}\widehat{\beta}(X^{tr}, y^{tr})\|^2}{n^{te}} = \frac{1}{n^{te}} \sum_{i=1}^{n} (y_i^{te} - X_i^{te}\widehat{\beta}(X^{tr}, y^{tr}))^2$$

$$RMSE = \sqrt{MSE}$$

Kroswalidacja

- Kroswalidacja jest bardziej stablina i dokładna niż użycie pojedynczego podziału na zestaw uczący i testowy.
- Dane dzielone są wielokrotnie (k-krotnie) i budowanych jest wiele (k) modeli.



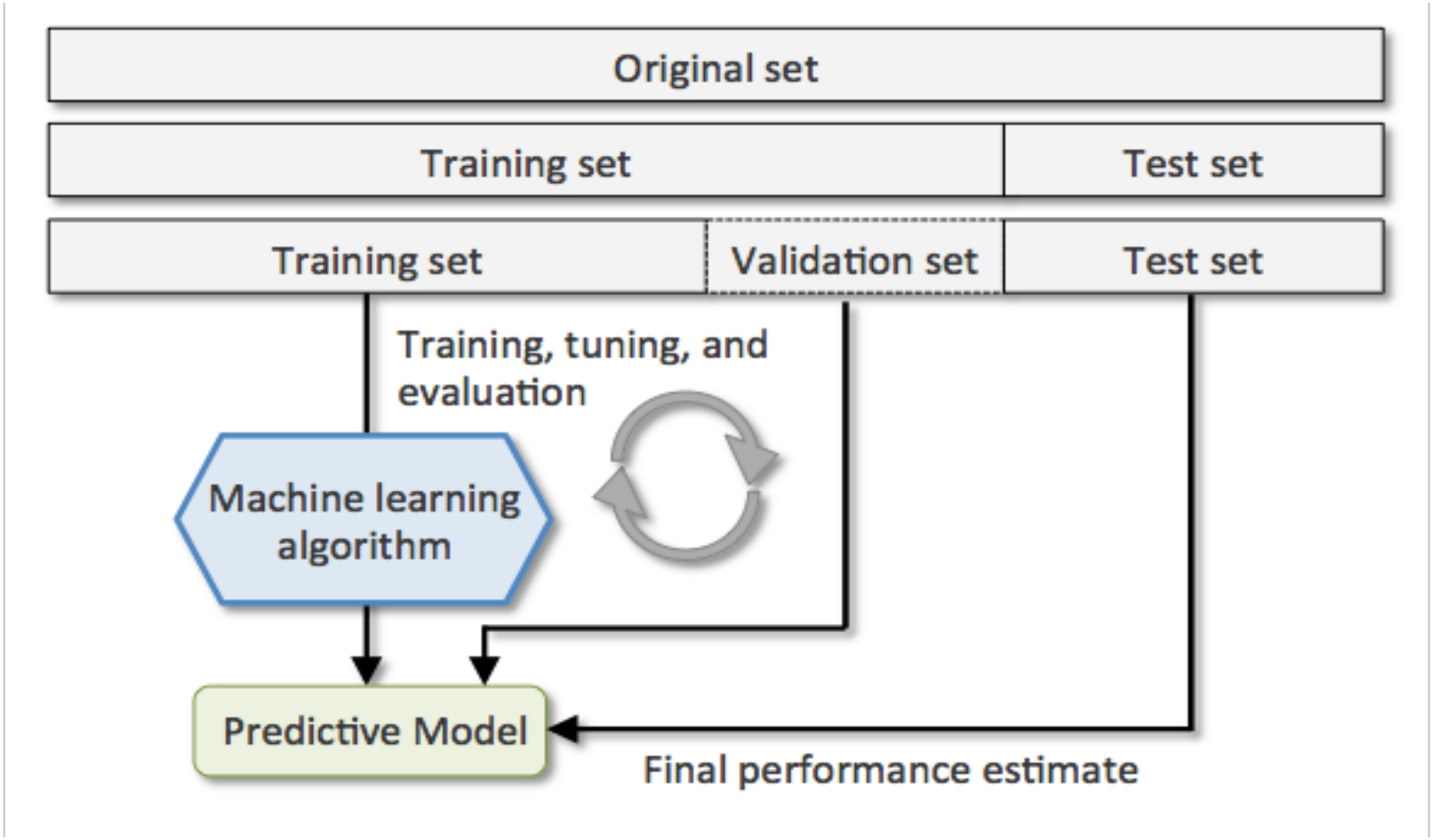
https://zitaoshen.rbind.io/project/machine_learning/machine-learning-101-cross-validation/

Kroswalidacja - uwagi

- Policzone błędy dla każdego k są na koniec uśredniane, dając ostateczne oszacownie.
- Najczęściej wybierane jest $k = 5$ lub $k = 10$.
- Każda obserwacja z danych znajdzie się w zestawie testowym dokładnie raz - uodporniamy się na „szczęśliwy" albo „pechowy" dobór próbki testowej.
- Kroswalidacja daje obraz tego, jak wrażliwy na wybór zestawu danych jest nasz model.
- Ponieważ w kroswalidacji model musi zostać nauczony k razy, dla algorytmów bardzo złożonych obliczeniowo używa się podziału na zestaw uczący i testowy.

Zestaw walidacyjny

Wybór parametrów: najlepszego zestawu cech, regularyzacji.



https://stats.stackexchange.com/questions/410118/cross-validation-vs-train-validation-test

Zestawy: treningowy, walidacyjny i testowy

- Na potrzeby warsztatu: zestawy treningowy (50%), walidacyjny (25%) i testowy (25%).
- Najczęściej: oddzielenie zestawu testowego + kroswalidacja dla zbiorów treningowego i walidacyjnego.
- Możliwa jest też zagnieżdżona kroswalidacja.

Kryteria informacyjne

- Bayesian Information Criterion - podejście Bayesowskie, maksymalizacja prawdopodobieństwa a posteriori:

$$BIC(M) = -2loglik_M + \ln(n)|M| = n \ln\left(\frac{RSS}{n}\right) + \ln(n)|M|$$

- Akaike Information Criterion - oparty na teorii informacji, wybieramy model najbliższy prawdziwemu rozkładowi danych w sensie odległości Kullbacka-Leiblera:

$$AIC(M) = -2loglik_M + 2|M| = n \ln\left(\frac{RSS}{n}\right) + 2|M|$$

Obliczamy kryterium dla wielu modeli i wybieramy ten o minimalnym kryterium.

Metody wyboru modelu liniowego, zmienne ciągłe

All subset selection

- Zakłada przejrzenie wszystkich możliwych podzbiorów zmiennych objaśniających (2^p podzbiorów), dla $p = 14$ $2^p = 16384$.
- Dla małych zbiorów danych.

Metody krokowe

Metoda step AIC (BIC)

- **Backward** : Startujemy z modelu pełnego. Dla każdej zmiennej sprawdzamy, jaki spadek/wzrost w AIC (BIC) spowoduje jej usunięcie. Usuwamy tę, która powoduje największy spadek kryterium. Powtarzamy procedurę aż AIC (BIC) przestaną się zmniejszać.
- **Forward**: Startujemy z pustego modelu. Dla każdej zmiennej sprawdzamy, jaki spadek/wzrost w AIC (BIC) spowoduje jej dodanie. Dodajemy tę, która powoduje największy spadek kryterium. Powtarzamy procedurę aż AIC (BIC) przestaną się zmniejszać.
- **Bidirectional**: w każdym kroku sprawdzamy, jaki na kryterium ma wpływ dodanie lub usunięcie jednej zmiennej

Dla dużych danych bardzo kosztowna obliczeniowo.

Metody krokowe

DMR (Delete or Merge Regressors)

- Dla pełnego modelu policz statystyki t-studenta t_i dla hipotezy $H_0 : \beta_i = 0$, i $= 1, \ldots,$ p.
- Posortuj kwadratowe statystyki t_i^2 w kolejności rosnącej.
- Akceptuj po kolei hipotezy (usuwaj zmienne z modelu) według kolejności posortowanych kwadratowych statystyk.
- Zwróć zagnieżdżoną rodzinę modeli: od modelu pełnego do modelu z jedną zmienną.

Musi dopasować tylko p modeli.

$$t_i = \frac{\widehat{\beta}_i}{sd(\widehat{\beta}_i)}$$

Regularyzacja

Dodanie ograniczeń dla parametrów β (ściąganie).

- Regresja grzbietowa: problem ze zwyczajną regresją liniową, gdy w $\mathbb{X}$ kolumny są mocno skorelowane.

$$\widehat{\beta}^{ridge} = \arg\min_{\beta} RSS, \text{ pod warunkiem } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

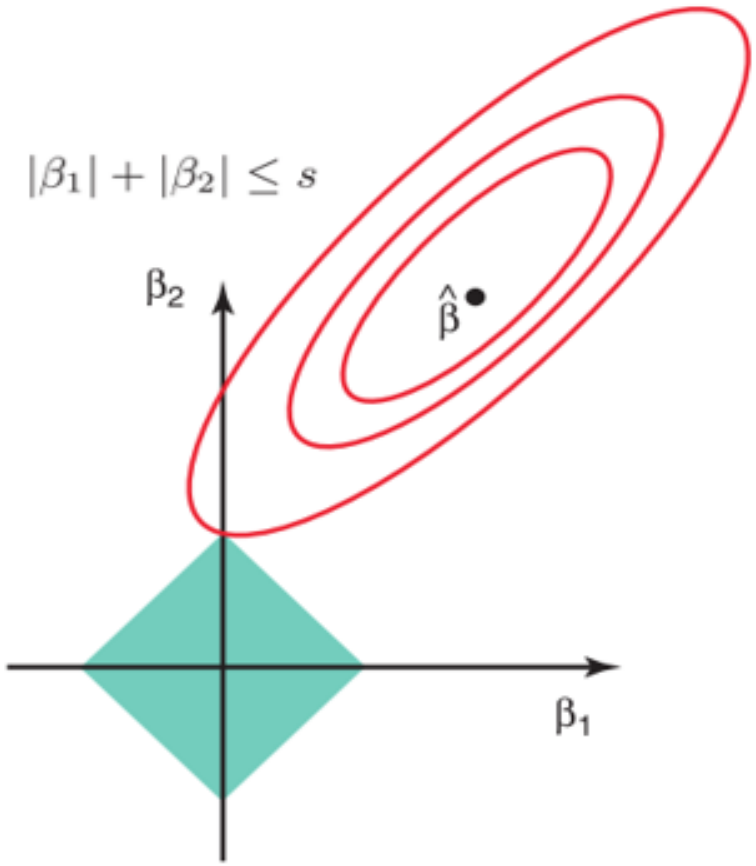
- LASSO (least absolute shrinkage and selection operator): selekcja zmiennych

$$\widehat{\beta}^{LASSO} = \arg\min_{\beta} RSS, \text{ pod warunkiem } \sum_{j=1}^{p} |\beta_j| \leq t.$$

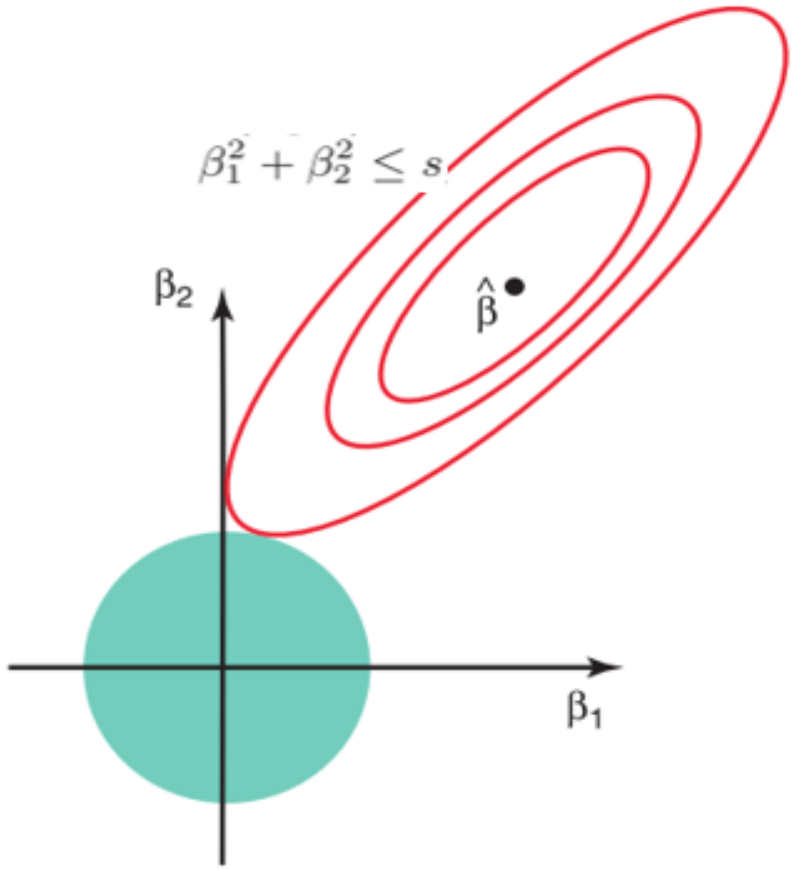
- Elastic Net: połączenie regresji grzbietowej i LASSO

Szukanie estymatorów w pakiecie glmnet: metoda spadku gradientu.

Regularyzacja



Lasso Regression



Ridge Regression

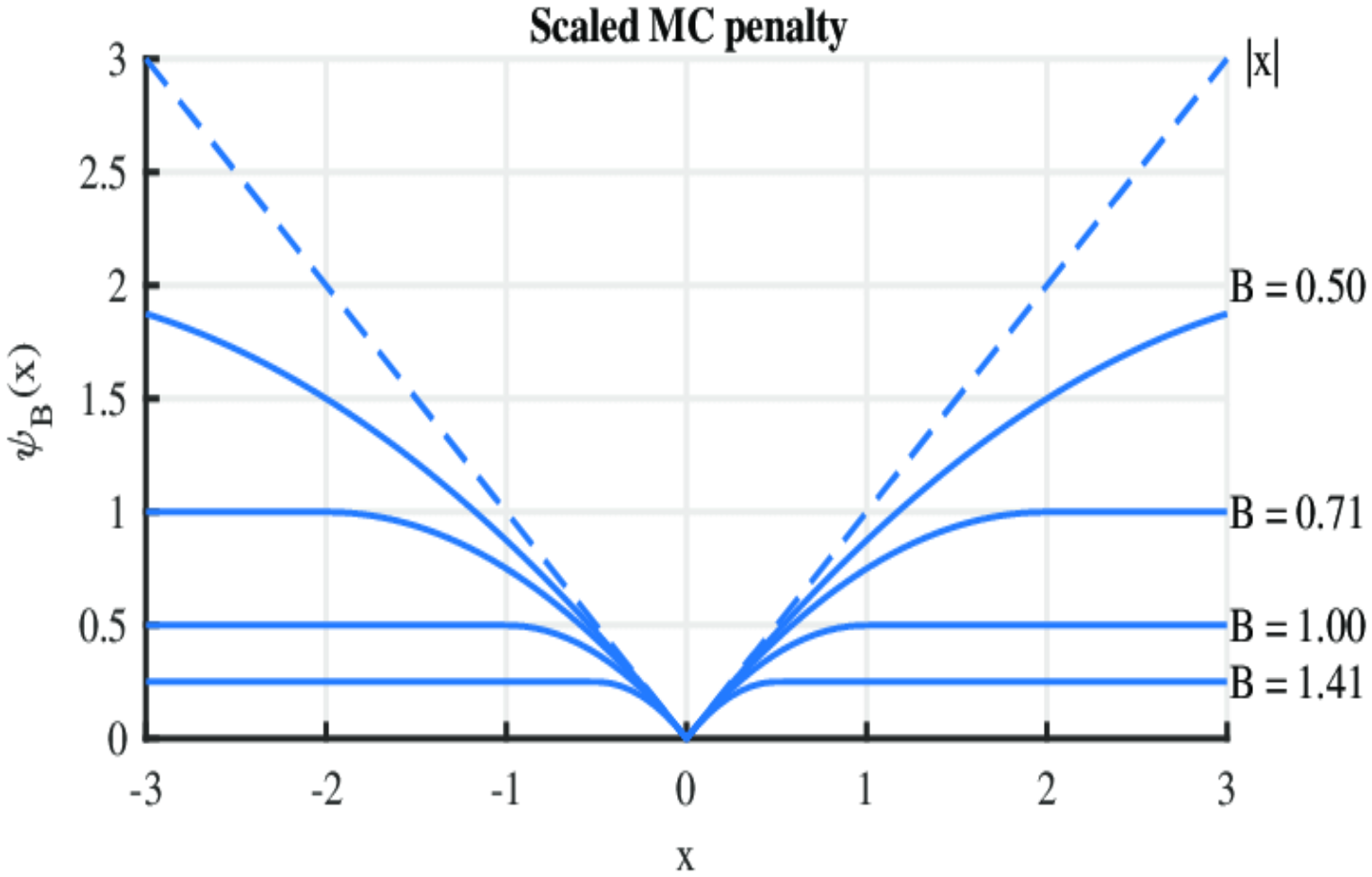
https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression

Regularyzacja

Sparsenet:

- używa regularyzacji MCP (Minimax Concave Penalty) z niewypukłą funkcją kary za parametry,
- kompromis pomiędzy LASSO i best subset selection,
- trudna optymalizacja - metoda spadku gradientu.

Regularyzacja



$$\widehat{\beta}^{LASSO} = \arg\min_{\beta} \left(RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Przypadek dużego zbioru danych, $p > n$

- DMRnet
- LASSO
- sparsenet

Metody wyboru modelu liniowego, zmienne ciągłe i jakościowe

Modele dla zmiennych jakościowych

- Wybór całych zmiennych jakościowych
- Łączenie poziomów w zmiennych jakościowych

Wybór zmiennych ciągłych i całych zmiennych jakościowych

- stepAIC, stepBIC
- group LASSO, group MCP:

$$\widehat{\beta}^{groupLASSO} = \arg\min_{\beta} \left(RSS + \lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j} \right), \|z\|_{K_j} = (z^T K_j z)^{\frac{1}{2}}$$

Np. dla jednej zmiennej ciągłej i jednej zmiennej jakościowej:

$$\widehat{\beta}^{groupLASSO} = \arg\min_{\beta} \left(RSS + \lambda \left(|\beta_1| + c \cdot \sqrt{\beta_2^2 + \beta_3^2 + \beta_4^2} \right) \right)$$

Selekcja czynnikowa

- LASSO (pakiet smurf):
  - Dla zmiennych jakościowych nakłada kary LASSO na różnice pomiędzy parametrami, wymuszając $\beta_i = \beta_j$.
  - Dla zmiennych ciągłych nakłada kary LASSO na wielkości parametrów, wymuszając $\beta_i = 0$.
- DMR (Delete or Merge Regressors) :
  - Dla zmiennych jakościowych używa kwadratowych t-statystyk dla hipotez $\beta_i = \beta_j$ i klasteryzacji hierarchicznej do posortowania kolejnych łączeń poziomów w zmiennych jakościowych.
  - Dla zmiennych ciągłych używa kwadratowych t-statystyk dla hipotez $\beta_i = 0$.
  - Sortuje hipotezy dla zmiennych ciągłych i jakościowych według kwadratowych t-statystyk i wysokości cięć w dendrogramach i po kolei akceptuje hipotezy.
  - Zwraca zagnieżdżoną rodzinę modeli: od pełnego modelu do modelu z jedną zmienną.

Źródła:

- Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Tutz, G. (2011). Regression for categorical data (Vol. 34). Cambridge University Press.
- `https://glmnet.stanford.edu/articles/glmnet.html`
- `https://statisticaloddsandends.wordpress.com/2019/12/09/the-minimax-concave-penalty-mcp/`
- `https://cran.r-project.org/web/packages/smurf/vignettes/smurf.html`
- `https://cran.r-project.org/web/packages/DMRnet/vignettes/getting-started.html`