INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

AUTOREFERAT PRZYGOTOWYWANEJ ROZPRAWY DOKTORSKIEJ
pt.

# Dobór zmiennych w modelach liniowych z wykorzystaniem indeksów wielowymiarowych

## Variable selection algorithms for linear models based on multidimensional indices

*Autor:*
mgr Barbara Żogała-Siudem

*Opiekun naukowy:*
dr hab. inż. Szymon Jaroszewicz

Warszawa, listopad 2018

# Spis treści

**1 Wprowadzenie** **2**
1.1 Cel rozprawy i opis głównych wyników . . . 2
1.2 Podstawowa hipoteza badawcza i metody badawcze . . . 2
1.3 Oznaczenia . . . 3

**2 Indeksy wielowymiarowe** **3**
2.1 Cel stosowania indeksów wielowymiarowych . . . 3
2.2 Zastosowanie indeksów do przeszukiwania w oparciu o korelację . . . 4
2.3 Procent zmiennych znajdujących się w zadanym promieniu . . . 4

**3 Model regresji liniowej** **5**
3.1 Metoda najmniejszych kwadratów . . . 5
3.2 Metody selekcji zmiennych i ich ograniczenia . . . 6
3.2.1 Regresja krokowa . . . 6
3.2.2 Lasso . . . 7

**4 Regresja krokowa z wykorzystaniem indeksów wielowymiarowych** **8**
4.1 Podstawy teoretyczne modyfikacji algorytmu . . . 8
4.2 Modyfikacja algorytmu regresji krokowej . . . 9

**5 Algorytm Lasso z wykorzystaniem indeksów wielowymiarowych** **10**
5.1 Podstawy teoretyczne do modyfikacji algorytmu . . . 10
5.2 Modyfikacja algorytmu Lasso . . . 12

**6 Eksperymenty obliczeniowe** **12**
6.1 Regresja krokowa z indeksem wielowymiarowym . . . 12
6.1.1 Opis danych rzeczywistych – baza danych Eurostat . . . 12
6.1.2 Wyniki . . . 15
6.2 Ścieżka rozwiązań dla problemu lasso z indeksem wielowymiarowym . . . 16
6.2.1 Opis danych symulacyjnych . . . 16
6.2.2 Wyniki . . . 17

**7 Podsumowanie, wnioski i dalsza planowana praca** **17**

# 1 Wprowadzenie

## 1.1 Cel rozprawy i opis głównych wyników

Celem rozprawy będzie adaptacja typowych metod doboru zmiennych w modelach liniowych na potrzeby ich zastosowań do analizy danych o **dużej liczbie zmiennych**. Przykładem takich danych może być *Linked Open Data* [1, 9, 5] – projekt mający na celu połączenie obecnie niepowiązanych kolekcji Otwartych Danych. Zawiera miliony zmiennych i choć ich liczba stale się zwiększa, to możemy przyjąć, że jest to stały zbiór, na którym, po uprzednim wstępnym ich przetworzeniu, możemy budować wiele modeli. Do dotychczasowej analizy wykorzystana została jedna z baz danych wchodząca w skład LOD – Eurostat [8]. Dostępne klasyczne metody analizy (jak np. regresja krokowa lub Lasso) mogą być w takim przypadku niewystarczające. Wynika to głównie z faktu, że przy ogromnej liczbie atrybutów, czas obliczeń szybko wzrasta. Wyklucza to typowe metody z praktycznych zastosowań, zwłaszcza ze względu na fakt, że większość zmiennych jest nieistotna i to na ich analizie algorytm traci większość czasu. Fakt ten był motywacją zastosowania **indeksu wielowymiarowego**, który pozwoli na szybką **wstępną selekcję atrybutów**. Odbędzie się to nieznacznym, jak zobaczymy, kosztem dokładności, przy dużym zysku na czasie wykonywania.

Do analizowanych w rozprawie metod należą regresja krokowa w przód (ang. *forward stepwise regresion*) i Lasso. Każdy z analizowanych algorytmów działa w sposób iteracyjny – włączając kolejno zmienne do ostatecznego modelu i w każdym z tych kroków musi przeszukać wszystkie dostępne zmienne, aby znaleźć kolejną, obecnie najlepszą. Widać zatem, że są to algorytmy o ogromnej złożoności ze względu na liczbę atrybutów. Nie stanowi to problemu w sformułowaniu klasycznym gdy ta liczba jest mała, ale dla rozważanych dużych danych $m > 10^6$ i więcej, a więc w szczególności podczas przetwarzania Otwartych Danych standardowe metody zawiodą. W sytuacji, gdy mamy do dyspozycji indeks wielowymiarowy udostępniający wyszukiwanie zmiennych w zadanym promieniu, możemy w każdym kolejnym kroku szybko filtrować zmienne i ostatecznie szukać rozwiązania jedynie pośród małego podzbioru oryginalnych danych. Odbywa się to kosztem dokładności (przybliżony indeks mógłby przeoczyć bardzo istotną zmienną), jednak jak pokażę, w praktyce **nie odbije się to niekorzystanie na ostatecznym rozwiązaniu**.

Poza opisanymi modyfikacjami wymienionych algorytmów, umożliwiającymi ich zastosowanie w opisie dużych danych, rozprawa obejmie również porównanie własności predykcyjnych tych metod oraz możliwości uzyskiwania interpretowalnych rezultatów.

Wyniki opisane w niniejszym referacie zostały opublikowane w [18] oraz [19] lub są przygotowywane do publikacji [17] lub opisane w raporcie badawczym [20].

## 1.2 Podstawowa hipoteza badawcza i metody badawcze

Podstawową hipotezą będzie założenie, że wykorzystanie przybliżonych indeksów do analizowanych metod doboru zmiennych do modelu, choć mogą zmniejszyć ich precyzję, będzie opłacalne ze względu na przyspieszenie wykonywania algorytmów oraz na zmniejszenie użycia potrzebnej pamięci. Dzięki temu umożliwi to zastosowanie ich do danych o dużej liczbie zmiennych, w tym do przetwarzania coraz większej różnorodności dostępnych otwartych danych.

Metody badawcze obejmą analizę teoretyczną z wykorzystaniem technik matematycznych dotyczących modeli liniowych. W rozprawie ściśle dowiodę, że zaproponowane modyfikacje algorytmów regresji krokowej w przód i Lasso, zwracają te same wyniki co ich oryginalne odpowiedniki, pod warunkiem, że podczas wyszukiwania zmiennych skorzystalibyśmy z dokładnego indeksu. Zastosowanie w nich przybliżonych indeksów wielowymiarowych sprawia, że otrzymujemy rozwiązanie zbliżone do dokładnego, zyskując tym jednak na wydajności.

Metody te są zatem próbą zastosowania klasycznych statystycznych narzędzi do rozwiązania problemów jakie stwarzają duże zbiory danych. Podstawową techniką umożliwiającą przyspie-

szenie analizowanych algorytmów będzie użycie wielowymiarowych indeksów (m.in. [10]), które są w stanie wyszukiwać wektory z przestrzeni $\mathbb{R}^n$ w zadanym promieniu.

W ramach analizy eksperymentalnej, wprowadzone algorytmy zostaną zaimplementowane w języku Python, a ich działanie zostanie przetestowane zarówno na danych symulowanych jak i na rzeczywistych pochodzących m. in. z bazy danych Eurostatu.

### 1.3 Oznaczenia

Przedstawię teraz podstawowe oznaczenia, wykorzystywane w tym referacie. Wektory będą oznaczone małymi literami $x$, $y$, $r$, macierze dużymi $X$, a ich transpozycje poprzez indeks górny: $x^T$, $X^T$. Macierz jednostkowa będzie oznaczona jako $I$. Normy $l_1$ i $l_2$ wektorów będą zapisywane odpowiednio jako $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ oraz $\|x\|_1 = \sum_{i=1}^n |x_i|$. Średnia wektora będzie oznaczona jako $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Będziemy zakładać, że wszystkie rozważane wektory są odpowiednio unormowane, czyli

$$\bar{x} = 0, \quad \|x\|_2 = 1. \tag{1.1}$$

Zauważmy, że współczynnik korelacji liniowej między dwoma tak unormowanymi wektorami jest równa ich iloczynowi skalarnemu i będziemy ją oznaczać

$$\operatorname{cor}(x, y) = \sum_{i=1}^{n} x_i y_i.$$

## 2 Indeksy wielowymiarowe

### 2.1 Cel stosowania indeksów wielowymiarowych

Celem stosowania indeksów wielowymiarowych jest szybkie wyszukiwanie wektorów z przestrzeni $\mathbb{R}^n$ znajdujących się blisko (względem pewnej metryki, tutaj rozważamy $l_2$) pewnego wektora $q \in \mathbb{R}^n$. Możemy wyobrazić sobie dwa zagadnienia:

- Wyszukiwanie $k$-najbliższych sąsiadów – dla danego wektora $q$ zwróć $k$ wektorów znajdujących się najbliżej.
- Wyszukiwanie w zadanym promieniu – dla danego wektora $q$ i odległości $d \geqslant 0$ zwróć wszystkie wektory leżące nie dalej niż $d$, czyli $\|x - q\|_2 \leqslant d$.

Najprostszym rozwiązaniem byłoby przeszukiwanie wszystkich wektorów w celu zwrócenia tych spełniających kryteria (*brute force search*). Ma on złożoność $\mathcal{O}(n)$. Celem indeksowania wielowymiarowego jest poprawa tego rozwiązania, kosztem poświęcenia czasu na budowę indeksu.

Rozróżniamy dwa typy indeksów: dokładne i przybliżone. Dokładne są zwykle oparte na strukturach drzewiastych i przeszukiwaniu binarnym, a ich najpopularniejszymi przykładami są np. kd-drzewa (ang. *kd-trees*) [4], drzewa kulowe (ang. *ball trees*) [12] oraz vp-drzewa (ang. *vp-trees*) [16]. Niestety ich wydajność wzrasta drastycznie wraz z długością wektorów $n$ i już dla $n = 20$ nie różni się znacznie od przeszukiwania liniowego, co oznacza, że w rozważanym przypadku nie będą przydatne.

Istnieje również wiele przybliżonych indeksów, jednak wadą większości z nich, z naszego punktu widzenia, jest to, że po pierwsze wszystkie indeksowane wektory muszą być przechowywane w pamięci RAM, a ponadto rzadko dostępna jest dobra implementacja przeszukiwania w zadanym promieniu, co jest kluczowe w naszych modyfikacjach. W eksperymentach korzystamy z biblioteki Faiss [10], która pokonuje wspomniane problemy. Ideą stojącą za tym indeksem jest podzielenie danych na dużą liczbę skupień, a następnie wyszukiwanie ich środków i szukanie rozwiązania wewnątrz znalezionych klastrów.

## 2.2 Zastosowanie indeksów do przeszukiwania w oparciu o korelację

Jeśli uwzględnimy fakt, że wszystkie zmienne są odpowiednio unormowane (zob. 1.1), to możemy zauważyć, że wyszukiwanie zmiennych o odpowiednio dużej korelacji z wektorem $q$ odpowiada wyszukiwaniu wektorów dostatecznie bliskich względem metryki euklidesowej. Zauważmy bowiem, że dla dowolnych wektorów unormowanych $x$ i $y$ zachodzi

$$\|x-y\|_2 = \sqrt{\|x\|_2^2 - 2xy + \|y\|_2^2} = \sqrt{2 - 2\mathrm{cor}(x,y)}$$
$$\mathrm{cor}(x,y) = 1 - \frac{\|x-y\|_2^2}{2}$$

Oznacza to, że dowolny indeks obsługujący odległość euklidesową, może zostać zastosowany do naszego celu.

## 2.3 Procent zmiennych znajdujących się w zadanym promieniu

W tym podrozdziale pokażę, jak duży procent wszystkich zmiennych zostanie zwrócony, jeśli ograniczymy się do wektorów skorelowanych nie mniej niż $\eta$ z pewnym wektorem $q \in \mathbb{R}^n$ dla przestrzeni o różnych wymiarach $n$.

Zauważmy, że wszystkie rozważane wektory możemy traktować jako punkty na $(n-1)$-wymiarowej sferze jednostkowej. Jeśli dodatkowo założymy, że są one równomiernie na niej rozłożone, to procent zmiennych znalezionych w zadanym promieniu możemy oszacować, obliczając pole powierzchni sfery ograniczonej przez odpowiedni kąt. Założenie o jednostajnym rozkładzie wektorów na sferze będzie nieprawdziwe w przypadku danych rzeczywistych, co można zobaczyć na przykładzie danych z Eurostatu [8] (zob. Rys. 1), jednak w praktyce ograniczenie liczby wektorów będzie i tak znaczące.

Dla danego wektora $q$ i ograniczenia na korelację $\eta$, wycinek sfery będzie postaci

$$S(\alpha) = \{x \in \mathbb{R}^n : \|x\| = 1, \sphericalangle(q,x) \leqslant \alpha\}, \tag{2.1}$$

gdzie $\alpha = \arccos(\eta)$.

Dla punktów równomiernie rozmieszczonych na sferze, proporcja punktów $P_n(\alpha)$ znajdujących się na wycinku $S(\alpha)$ będzie równa

$$P_n(\alpha) = \frac{|S(\alpha)|}{S(\pi/2)}.$$

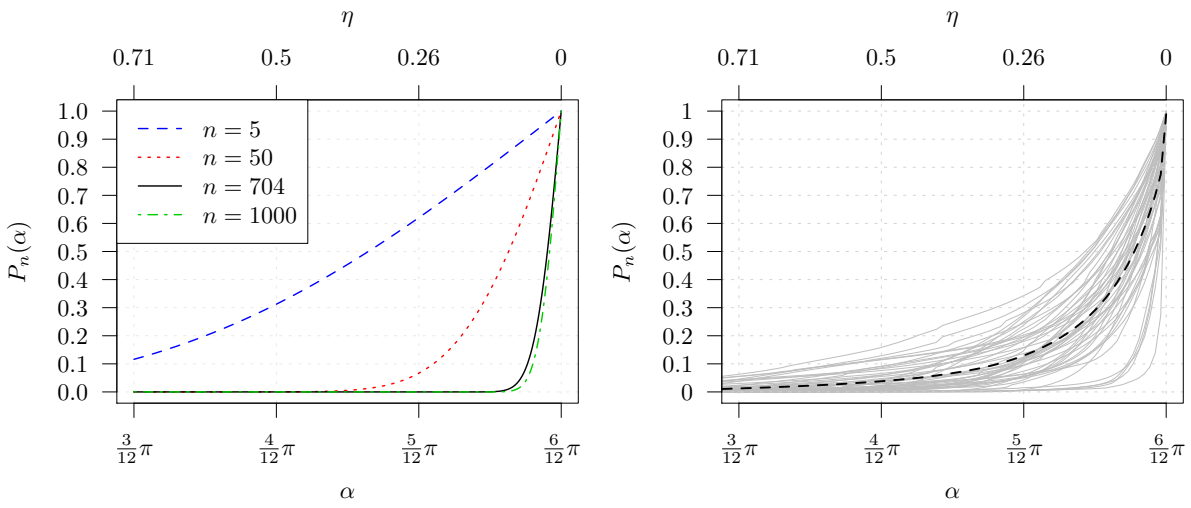
Pole powierzchni wycinka sferycznego możemy obliczyć przy użyciu poniższego wzoru (zob. [11])

$$M_n(\alpha) = \frac{2\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} \int_0^\alpha \sin^{n-2}\theta d\theta,$$

a pole powierzchni całej jednostkowej $(n-1)$-sfery wynosi $2\pi^{n/2}/\Gamma(n/2)$, zatem

$$P_n(\alpha) = \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \int_0^\alpha \sin^{n-2}\theta d\theta. \tag{2.2}$$

Rysunek 1 (lewa strona) pokazuje, jak wartość $P_n(\alpha)$ zależy od $n$ i $\alpha$. Z prawej strony znajduje się analogiczny wykres przedstawiający prawdziwe dane pochodzące z bazy danych Eurostat. Widać, że choć wymiar wektorów z Eurostatu wynosi 704, to ich zachowanie jest bardziej zbliżone do danych losowych dla $n = 50$. Jest to częściowo spowodowane faktem, że w danych rzeczywistych zawsze występować będą pewne korelacje, a spotęgowany jest przez fakt, że analizowane dane opisują pewne statystyki dla wybranych państw na przestrzeni czasu, zatem występują dodatkowe korelacje między szeregami czasowymi.




Rysunek 1: Lewa strona przedstawia procent zmiennych jakie zostaną znalezione w zadanym promieniu wyjściowego wektora $q \in \mathbb{R}^n$ w zależności od $n$ dla przypadku równomiernego rozmieszczenia punktów na sferze. Pozioma dolna oś przedstawia kąt $\alpha$ (zob. wzór 2.1), a górna korelację, czyli $\eta = \cos(\alpha)$. Prawy rysunek przedstawia 50 losowo wybranych wektorów z Eurostatu oraz procent zmiennych znajdujących się w ich sąsiedztwie, również w zależności od kąta $\alpha$ lub korelacji $\eta$. Przerywana krzywa jest średnią z tych 50 wektorów.

Możne zauważyć, że jeśli liczba obserwacji jest dostatecznie duża, to wartość $P_n(\alpha)$ staje się bardzo mała nawet dla relatywnie dużych wartości kąta $\alpha$. Oznacza to, że w algorytmach 3 i 4 przedstawionych w rozdziałach 4.2 i 5.2 będziemy istotnie wybierać mały procent wyjściowych zmiennych.

# 3 Model regresji liniowej

Przedstawię teraz krótkie wprowadzenie do budowy modeli liniowych oraz selekcji zmiennych do modelu. Ze względu na fakt, że w rozprawie uogólnię te metody, ich krótki opis jest niezbędny na potrzeby tego referatu. Więcej szczegółów można znaleźć m. in. w [6].

## 3.1 Metoda najmniejszych kwadratów

Załóżmy, że $x_1, \ldots, x_p \in \mathbb{R}^n$ są zmiennymi na podstawie których możemy budować model, a macierz $X \in \mathbb{R}^{n \times p}$ jest macierzą, której kolumny zawierają zmienne $x_i$. Niech $y \in \mathbb{R}^n$ będzie zmienną odpowiedzi, to znaczy wartością, którą chcemy modelować. Naszym zadaniem jest znalezienie liniowego związku między kolumnami $X$, a zmienną $y$

$$y = X\beta + \varepsilon, \tag{3.1}$$

gdzie $\beta$ jest wektorem współczynników, a $\varepsilon$ jest pewnym losowym błędem. Wektor współczynników możemy wyestymować przy użyciu tak zwanej metody najmniejszych kwadratów (zob. m.in.[6]) minimalizującej sumę kwadratów błędów

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|X\beta - y\|_2, \tag{3.2}$$

której rozwiązaniem jest

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{3.3}$$

Dodatkowo, niech $\hat{y} = X\hat{\beta}$ będzie predykcją zmiennej $y$, a $r = y - \hat{y}$ wektorem residuów modelu. Wtedy możemy zapisać

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = H_X y,$$
$$r = y - \hat{y} = y - X\hat{\beta} = y - X(X^T X)^{-1} X^T y = (I - H_X)y = P_X y,$$

gdzie $H_X$ jest tak zwaną *macierzą daszkową* (ang. *hat-matrix*), a $P_X = I - H_X$. Można pokazać, że obie macierze są symetryczne i idempotentne, czyli $H_X^2 = H_X$ i $P_X^2 = P_X$ [6].

Ponadto oznaczmy sumę kwadratów błędów modelu zbudowanego za pomocą zmiennych z macierzy $X$ jako

$$\mathrm{RSS}(X) = r^T r = y^T P_X^T P_X y = y^T P_X y, \tag{3.4}$$

Niech $S = \{i_1, \ldots, i_k\}$ będzie zbiorem indeksów zmiennych, a macierz $X_S$ macierzą z wybranymi kolumnami o numerach zawierających się w $S$, zatem jeśli $X = [x_1|\ldots|x_p]$, to $X_S = [x_{i_1}|\ldots|x_{i_k}]$. Spadek sumy kwadratów błędów pomiędzy modelem opartym na zmiennych $X_S$ i modelem z dodatkowo dodaną zmienną $x$ oznaczymy jako

$$\Delta\,\mathrm{RSS}(X_S, x) = \mathrm{RSS}(X_S) - \mathrm{RSS}([X_S|x]). \tag{3.5}$$

Analogicznie zdefiniujemy również *względny* spadek RSS, zwany *współczynnikiem częściowej determinacji* (zob. [13, Chapter 4g])

$$\delta\,\mathrm{RSS}(X_S, x) = \frac{\Delta\,\mathrm{RSS}(X_S, x)}{\mathrm{RSS}(X_S)}. \tag{3.6}$$

## 3.2 Metody selekcji zmiennych i ich ograniczenia

Nie zawsze budowa modelu na pełnych danych jest wskazana lub wręcz możliwa, jak na przykład, gdy $p > n$. W tej sytuacji konieczna stanie się selekcja zmiennych, na podstawie których zbudujemy model. W planowanej rozprawie doktorskiej opiszę dwa znane algorytmy służące do wyboru atrybutów – regresję krokową [7] oraz LASSO [15].

### 3.2.1 Regresja krokowa

Możemy rozróżnić trzy warianty regresji krokowej (ang. *stepwise regression*, [7]): regresję w przód, w tył oraz kombinację powyższych. Każda z nich polega na tym, żeby w kolejnych krokach minimalizować odpowiednie kryterium

$$n \log\left(\frac{\mathrm{RSS}(X_S)}{n}\right) + C(|S|), \tag{3.7}$$

gdzie $C(|S|)$ jest funkcją kary za złożoność modelu, zależną od liczby zmiennych znajdujących się w modelu.

Regresja krokowa w przód (ang. *forward stepwise regression*) polega na budowaniu modelu w sposób zachłanny, włączając do niego zmienne w kolejnych krokach, a startując z pustego modelu. W każdym kolejnym kroku dodajemy zmienną, która przy ustalonych dotychczas już dodanych daje największy spadek rozważanego kryterium (zob. punkt 3.2.1). Procedurę kontynuujemy dopóki dodanie żadnej zmiennej nie powoduje już zmniejszenia kryterium lub do momentu aż nie zostanie dodana z góry zadana liczba zmiennych $k_{max}$.

Podstawowy algorytm regresji krokowej w przód jest przedstawiony w Alg. 1. W planowanej rozprawie opiszę modyfikację tego algorytmu, aby w każdym jego kroku możliwe było filtrowanie zmiennych za pomocą indeksu wielowymiarowego na podstawie ich korelacji z już włączonymi zmiennymi oraz obecnym residuum.

**Algorithm 1** Podstawowy algorytm regresji krokowej w przód

**Require:** $y \in \mathbb{R}^{nx1}$, $X \in \mathbb{R}^{nxp}$, $k_{max} \in \mathbb{N}$

1: **procedure** FORWARD-STEPWISE($y, X, k_{max}$)
2: $S \leftarrow \emptyset$ ▷ zaczynamy od pustego modelu
3: **for** $k = 1, \ldots, k_{max}$ **do**
4: **for** $i \in \{1, \ldots, p\} \setminus S$ **do**
5: Znajdź $\Delta \mathrm{RSS}(X_S, x_i)$
6: $l_k \leftarrow \arg\max_{i \in \{1,\ldots,p\} \setminus S} \Delta \mathrm{RSS}(X_S, x_i)$ ▷ znajdź najlepszą zmienną
7: **if** model na podstawie $[X_S | x_{l_k}]$ jest *lepszy* niż ten na podstawie $X_S$ **then**
8: $S \leftarrow S \cup \{l_k\}$ ▷ dodaj $l_k$ do $S$
9: **else**
10: **return** $S$
11: **return** $S$

Regresja krokowa w tył działa podobnie, jednak tym razem zamiast startować z pustego modelu i włączać do niego kolejne zmienne, zaczynamy z pełnego modelu i w kolejnych krokach usuwamy zmienne. Z uwagi na charakter rozważanych danych nie będziemy rozważali tej wersji. Możemy również w każdym kroku rozważać zarówno dodawanie jak i usuwanie zmiennych.

Poniżej przedstawiam krótki opis wykorzystywanych kryteriów, na podstawie których możemy zadecydować, czy kolejna zmienna powinna zostać dodana do modelu, a zatem czy większy model jest *lepszy* (zob. l.7 w Alg. 1).

**Ustalone** $k_{max}$ Ustalona z góry liczba zmiennych, które włączymy do modelu. Jest to najprostsze z rozważanych kryteriów. W każdym kroku wybieramy zmienną, która daje największy spadek RSS i dodajemy w sumie $k_{max}$ zmiennych.

**AIC** Kryterium informacyjne Akaikego (AIC) [2], gdzie karą za złożoność modelu jest $2(k+1)$, zatem minimalizujemy

$$n \log\left(\frac{\mathrm{RSS}(X_S)}{n}\right) + 2(k+1). \tag{3.8}$$

**BIC** Kryterium informacyjne Schwarza (BIC) [14], gdzie funkcja kary jest postaci $(k+1)\log(n)$, zatem minimalizujemy

$$n \log\left(\frac{RSS(X_S)}{n}\right) + (k+1)\log(n). \tag{3.9}$$

**Test** $F$ $F$-test, gdzie o włączeniu kolejnej zmiennej do modelu decyduje test statystyczny $F$ [6], służący do porównania dwóch zagnieżdżonych modeli.

W każdym kolejnym kroku procedury należy zatem obliczyć wartość wybranego kryterium i następnie wybrać model minimalizujący go. W przypadku rozważanej regresji krokowej w przód będziemy zawsze porównywać model mniejszy z większym i decydować czy kolejną zmienną włączyć (i kontynuować) procedurę, czy nie istnieje już zmienna, którą możemy dodać, a tym samym kończymy działanie na dotychczasowym modelu.

### 3.2.2 Lasso

Innym sposobem doboru zmiennych do modelu liniowego może być rozwiązanie zagadnienia Lasso [15], które różni się od klasycznego problemu najmniejszych kwadratów dodaniem dodatkowej kary za złożoność modelu w postaci $\lambda\|\beta\|_1$ i ma postać problemu optymalizacyjnego

$$\beta^\star(\lambda) = \min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{3.10}$$

Nałożenie kary w postaci normy $l_1$ na współczynniki sprawia, że dla ustalonego $\lambda$ część z nich będzie miała zerowe wartości w ostatecznym rozwiązaniu. Poniżej, w Alg. 2 przedstawiam algorytm opisany w rozdziale 6.2 w [3] pozwalający na znalezienie pełnej ścieżki regularyzacyjnej dla problemu Lasso.

**Algorithm 2** Podstawowy algorytm szukający ścieżki regularyzacyjnej w problemie Lasso

**Require:** $y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$, $k_{max} \in \mathbb{N}$

1: **procedure** LASSO($y, X, k_{max}$)
2: $\lambda_1 \leftarrow \max_{j=1,\ldots,p} |\mathrm{cor}(x_j, y)|$ ▷ znajdujemy $x_j$ najbardziej skorelowany z $y$
3: $S \leftarrow \{j : |\mathrm{cor}(x_j, y)| = \lambda_1\}$
4: **for** $k = 2, \ldots, k_{max}$ **do**
5: Znajdź największą wartość $\lambda_a < \lambda_{k-1}$ i $\lambda_b < \lambda_{k-1}$ taką, że
6: (a) $\exists j_a \in S^C : |X_{j_a}^T(y - X\beta^\star(\lambda_a))| = \lambda_a$,
7: (b) $\exists j_b \in S : \beta^\star_{j_b}(\lambda_b) = 0$.
8: **if** $\lambda_a > \lambda_b$ **then**
9: $S \leftarrow S \cup \{j_a\}$ ▷ dodaj $j_a$ do zbioru aktywnych zmiennych
10: $\lambda_k \leftarrow \lambda_a$
11: **else**
12: $S \leftarrow S \setminus \{j_b\}$ ▷ usuń $j_b$ ze zbioru aktywnych zmiennych
13: $\lambda_k = \lambda_b$

W planowanej rozprawie opiszę modyfikację tego algorytmu, aby w każdym jego kroku możliwe było filtrowanie zmiennych przy użyciu indeksu wielowymiarowego na podstawie korelacji z pewnymi wektorami związanymi ze zmiennymi włączonymi do obecnego modelu oraz z residuum.

# 4 Regresja krokowa z wykorzystaniem indeksów wielowymiarowych

## 4.1 Podstawy teoretyczne modyfikacji algorytmu

Jednym z celów stawianych w planowanej rozprawie jest zmodyfikowanie Alg. 1 w taki sposób, aby w wierszu 6 nie przeszukiwać wszystkich zmiennych, ale jedynie ich podzbiór. Podstawę teoretyczną modyfikacji stanowi tw. 4.1, które mówi, że w każdym kolejnym kroku działania algorytmu, mając dany pewien proponowany wektor $x$, jesteśmy w stanie odrzucić część zmiennych, które na pewno będą gorsze od $x$. Za pomocą indeksu wielowymiarowego będziemy szukać wszystkich zmiennych, których korelacje z obecnym residuum $r$ lub z którąś zmienną już dodaną do modelu są nie mniejsze niż pewien próg. Dowód tw. 4.1 można znaleźć w artykule [18] lub raporcie badawczym [17] będącym wstępną wersją artykułu rozszerzającego wyniki przedstawione w [18].

**TWIERDZENIE 4.1.** *Załóżmy, że $S = \{i_1, \ldots, i_k\}$, zmienne $X_S = [x_{i_1}|\ldots|x_{i_k}]$ obecnie włączone do modelu są ortogonalne i znormalizowane, a $\mathrm{RSS}(X_S) > 0$. Niech $x, x'$ będą innymi zmiennymi, niewłączonymi do modelu. Wówczas*

***1.*** *spełniona jest poniższa nierówność*

$$\max\{|cor(x_{i_1}, x)|, \ldots, |cor(x_{i_k}, x)|, |cor(r, x)|\} \geqslant \frac{1}{\sqrt{\frac{1}{\delta\,\mathrm{RSS}(X_S,x)} + k}}, \tag{4.1}$$

*gdzie $r = P_{X_S}y$ jest wektorem residuów z modelu opartego na $X_S$. Dla $\delta\,\mathrm{RSS}(X_S, x) = 0$ zakładamy, że prawa strona równania jest równa 0 (zgodnie z zachowaniem granicznym).*

***2.*** *Z* $\Delta\,\mathrm{RSS}(X_S, x') \geqslant \Delta\,\mathrm{RSS}(X_S, x)$ *wynika, że*

$$\max\{|cor(x_{i_1}, x')|, \ldots, |cor(x_{i_k}, x')|, |cor(r, x')|\} \geqslant \frac{1}{\sqrt{\frac{1}{\delta RSS(X_S,x)} + k}}. \tag{4.2}$$

Twierdzenie 4.1 opisuje przypadek, gdy bierzemy pod uwagę najprostsze kryterium (zob. p. 3.2.1), t.j. włączamy z góry zadaną liczbę zmiennych. Możemy je rozszerzyć na przypadek, gdy rozważamy kryterium postaci (3.7) lub to związane z testem $F$, w tym celu spójrzmy na lemat 4.1, którego dowód można znaleźć w raporcie badawczym [17].

**LEMAT 4.1.** *Niech* $X_S = [x_{i_1}|\ldots|x_{i_k}]$ *będzie macierzą zawierającą zmienne obecnie włączone do modelu. Załóżmy, że* $x$ *jest nową zmienną, którą chcemy dodać, a* $C \in \{AIC, BIC, F\}$ *jest wykorzystywanym kryterium. Wtedy dodanie* $x$ *do modelu jest korzystne ze względu na rozważane kryterium* $C$ *wtedy i tylko wtedy gdy*

$$\delta\,\mathrm{RSS}(X_S, x) > \delta_C, \tag{4.3}$$

*gdzie*

$$\begin{aligned}
\delta_{AIC} &= 1 - \exp\left(-\frac{2}{n}\right), \\
\delta_{BIC} &= 1 - \exp\left(-\frac{\log(n)}{n}\right), \\
\delta_F &= 1 - \frac{1}{1 + \frac{1}{n-k-2}\mathbb{F}_{1,n-k-2}(1-\alpha)}.
\end{aligned}$$

*W związku z tym, przy założeniu, że* $X_S$ *jest ortogonalna, dodanie zmiennej* $x$ *może prowadzić do poprawy względem kryterium* $C$ *jeśli*

$$\max\{|cor(x_{i_1}, x)|, \ldots, |cor(x_{i_k}, x)|, |cor(r, x)|\} \geqslant \frac{1}{\sqrt{\frac{1}{\delta_C} + k}}. \tag{4.4}$$

Gdy połączmy wartości progów opisane w powyższym twierdzeniu i lemacie otrzymamy ograniczenie

$$\eta = \frac{1}{\sqrt{\frac{1}{\max(\delta_C, \delta\,\mathrm{RSS}(X_S,x))} + k}}, \tag{4.5}$$

które możemy zastosować w modyfikacji algorytmu 1.

## 4.2 Modyfikacja algorytmu regresji krokowej

Ostateczny algorytm pozwalający na znajdowanie modelu liniowego za pomocą regresji krokowej w przód przy użyciu indeksu wielowymiarowego jest przedstawiony w Alg. 3.

**Algorithm 3** Algorytm regresji krokowej z selekcją zmiennych na podstawie indeksu wielowymiarowego

**Require:** $y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$, $k_{max} \in \mathbb{N}$

1: **procedure** MI-FORWARD-STEPWISE($y, X, k_{max}$)
2: $x_{i_1} \leftarrow \text{argmax}_{i\in\{1,\ldots,p\}}|\text{cor}(y, x_i)|$ ▷ dwa zapytania o 1 sąsiada
3: $S \leftarrow \{x_{i_1}\}$ ▷ pierwsza zmienna dodana do modelu
4: **for** $k \leftarrow 1, \ldots, k_{max} - 1$ **do**
5: $r \leftarrow \text{normalize}(P_{X_S} y)$ ▷ normalizujemy wektor residuów
6: $c \leftarrow \text{argmax}_{i\in\{1,\ldots,p\}\setminus S}|\text{cor}(x_i, r)|$ ▷ dwa zapytania o 1 sąsiada, znajdujemy zmienną $x_c$ najbardziej skorelowaną z $r$
7: $\eta \leftarrow (\frac{1}{\delta RSS(X_S, x_c)} + k)^{-\frac{1}{2}}$ ▷ ograniczenie na korelacje
8: ▷ lub zmodyfikowane ograniczenie korzystające z kryteriów, zob. lemat 4.1
9: $X_S \leftarrow \text{zortogonalizuj}(X_S)$
10: $C \leftarrow \{i \in \{1, \ldots, p\} \setminus S : |\text{cor}(r, x_i)| \geqslant \eta\}$ ▷ 2 zapytania (w promieniu)
11: **for** $l \leftarrow 1, \ldots, k$ **do**
12: $C \leftarrow C \cup \{i \in \{1, \ldots, p\} \setminus S : |\text{cor}(x_{i_l}, x_i)| \geqslant \eta\}$ ▷ 2 zapytania (w promieniu)
13: $l_k \leftarrow \text{argmax}_{i\in C} \Delta\,\text{RSS}(X_S, x_i)$ ▷ znajdź najlepszą zmienną
14: **if** model oparty na $[X_S|x_{m_k}]$ jest *lepszy* niż ten oparty na $X_S$ **then**
15: $S \leftarrow S \cup \{l_k\}$
16: **else**
17: **return** $S$
18: **return** $S$

# 5 Algorytm Lasso z wykorzystaniem indeksów wielowymiarowych

## 5.1 Podstawy teoretyczne do modyfikacji algorytmu

Kolejnym celem planowanej rozprawy jest modyfikacja Alg. 2, w sposób umożliwiający filtrowanie zmiennych za pomocą indeksu wielowymiarowego, aby w wierszu 6 nie przeszukiwać wszystkich zmiennych, ale jedynie ich podzbiór. Podstawę teoretyczna modyfikacji stanowi Tw. 5.1, które mówi, że w każdym kolejnym kroku działania algorytmu, mając dany pewien proponowany wektor $x_{c_1}$, jesteśmy w stanie odrzucić część zmiennych, które na pewno nie zostaną dodane do zbioru aktywnych indeksów wcześniej niż $x_{c_1}$. Będziemy zatem szukać za pomocą indeksu wielowymiarowego wszystkich zmiennych, których korelacje z wektorem $u$ lub $s$ są nie mniejsze niż pewien próg. Dowód tw. 5.1 można znaleźć w raporcie badawczym [20] przedstawiającym częściowe wyniki na temat przedstawianej tu modyfikacji.

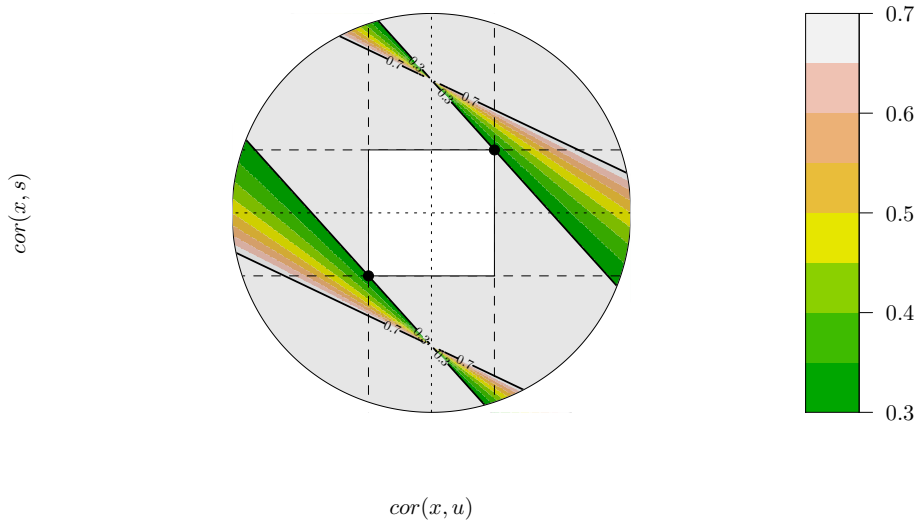
**TWIERDZENIE 5.1.** *Niech $S = \{i_1, \ldots, i_k\}$ będzie aktualnym zbiorem aktywnych indeksów, $x_{c_1}$ będzie zmienną taką, że $c_1 \notin J$, a $\lambda_{c_1}$ wartością parametru regularyzacyjnego dla którego $c_1$ zostałoby dodane do $S$. Wtedy dla dowolnej innej zmiennej $x_{c_2}$ i $\lambda_{c_2}$ zachodzi*
***jeśli***

$$|cor(x_{c_2}, s)| < \frac{\lambda_{c_1}}{\lambda_{c_1} d_s + d_u} \quad lub \quad |cor(x_{c_2}, u)| < \frac{\lambda_{c_1}}{\lambda_{c_1} d_s + d_u} \tag{5.1}$$

***to***

$$\lambda_{c_2} < \lambda_{c_1}, \tag{5.2}$$

*gdzie* $u = \frac{P_{X_S} y}{d_u}$, $d_u = \|P_{X_S} y\|_2$, $s = \frac{X_S (X_S^T X_S)^{-1} t_S}{d_s}$, $d_s = \|X_S (X_S^T X_S)^{-1} t_S\|_2$, $t_S = sign(X_S^T (y - X_S \beta_S^\star))$




Rysunek 2: Wartości parametru regularyzacyjnego $\lambda$, dla jakich $x$ zostałby włączony do zbioru aktywnych indeksów. Są one przedstawione w funkcji $\mathrm{cor}(x, s)$ i $\mathrm{cor}(x, u)$, a kolor odpowiada wartości $\lambda$ jaką parametr przyjmuje w zależności od tych korelacji. Wykres został ograniczony poprzez $\lambda = 0.7$, będącą przykładową wartością jaką przyjmuje parametr na ścieżce w momencie włączenia poprzedniej zmiennej (zatem szukane $\lambda$ musi być od niego mniejsze) oraz poprzez $\lambda_{c_1} = 0.3$ (od którego szukane $\lambda$ musi być większe). Białe obszary oznaczają miejsca gdzie rozwiązanie nie istnieje lub nie zawiera się między 0.3 i 0.7. Czarne punkty są ograniczeniami zdefiniowanymi w tw. 5.1, a szary obszar odpowiada tym ograniczeniom.

Rys. 2 stanowi ilustrację do powyższego twierdzenia. Możemy również zauważyć, że w pewnych przypadkach możliwa jest rotacja wektora $u$, tak, aby filtrowanie korelacji odbywało się jedynie względem korelacji z $u_{rot}$, co jest przedstawione na rysunku 3 i opisane dokładniej w raporcie badawczym [20].

### 5.2 Modyfikacja algorytmu Lasso

Ostateczny algorytm, uwzględniający przypadek z możliwą rotacją i bez, pozwalający na znajdowanie ścieżki regularyzacyjnej w problemie lasso z zastosowaniem indeksów wielowymiarowych jest przedstawiony w Alg. 4.

**Algorithm 4** Algorytm Lasso z wykorzystaniem indeksu wielowymiarowego

**Require:** $y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$, $k_{max} \in \mathbb{N}$

**procedure** MI-Lasso$(y, X, k_{max})$
  $\lambda_1 \leftarrow \max_{j=1,\dots,p} |\mathrm{cor}(x_j, y)|$ ▷ zaczynamy od $x_j$ najbardziej skorelowanego z $y$
  $S \leftarrow \{j : |\mathrm{cor}(x_j, y)| = \lambda_1\}$
  **for** $k = 2, \dots, k_{max}$ **do**
    Znajdź $C = (c_1, c_2)$ ▷ zob. rys. 3
    **if** $c_1 > 0$ i $c_2 > 0$ **then**
      Znajdź $u_{rot}$ i $v = |OC|$ ▷ zob. rys. 3
      Znajdź $S_{search} = \{j \in S^C : |\mathrm{cor}(x_j, u_{rot})| > v\}$
    **else**
      Znajdź $v$ ▷ zob. wzór 5.1 w tw. 5.1
      Znajdź $S_{search} = \{j \in S^C : |\mathrm{cor}(x_j, u)| > v\} \cup \{j \in S^C : |\mathrm{cor}(x_j, s)| > v\}$
    Znajdź największą wartość $\lambda_a < \lambda_{k-1}$ i $\lambda_b < \lambda_{k-1}$ taką, że
    (a) $\lambda_a = \max_{j\in S_{search}} \left( \frac{d_u \mathrm{cor}(x_j,u)}{1-d_s \mathrm{cor}(x_j,s)}, \frac{-d_u \mathrm{cor}(x_j,u)}{1+d_s \mathrm{cor}(x_j,s)} \right)$
    (b) $\lambda_b = \max_{i\in S} \frac{((X_S^T X_S)^{-1})_i X_S^T y}{((X_S^T X_S)^{-1})_i t_S}$
    **if** $\lambda_a > \lambda_b$ **then**
      $S \leftarrow S \cup \{j_a\}$ ▷ $j_a$ jest indeksem zmiennej odpowiadającej $\lambda_a$
      $\lambda_k \leftarrow \lambda_a$
    **else**
      $S \leftarrow S \setminus \{j_b\}$ ▷ $j_b$ jest indeksem zmiennej odpowiadającej $\lambda_b$
      $\lambda_k = \lambda_b$

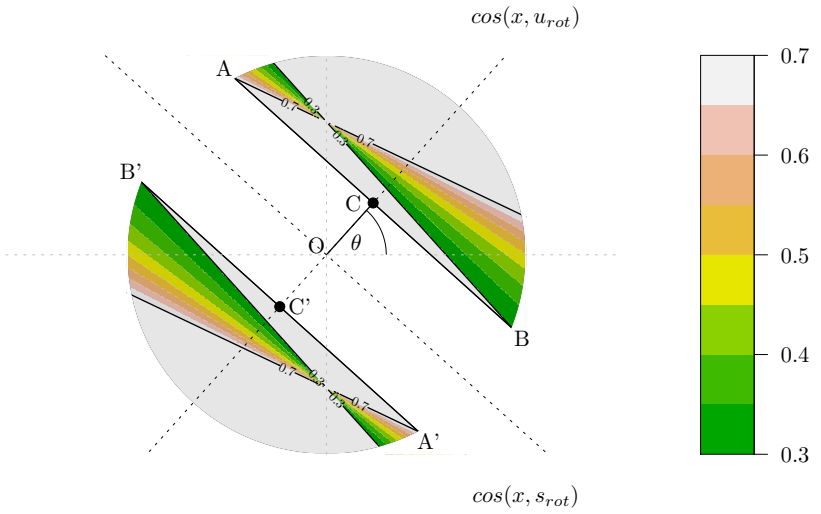
## 6 Eksperymenty obliczeniowe

### 6.1 Regresja krokowa z indeksem wielowymiarowym

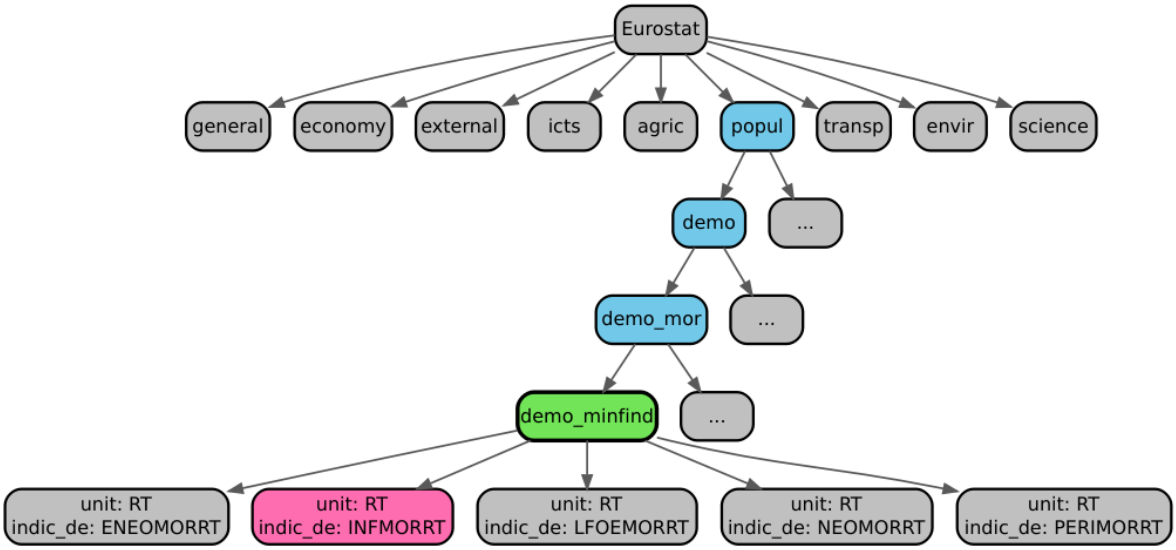
#### 6.1.1 Opis danych rzeczywistych – baza danych Eurostat

Poniżej znajduje się krótki opis danych, na jakich przeprowadzone były dotychczasowe eksperymenty – bazę danych Eurostat [8]. Zawiera ona tysiące zbiorów podzielonych na kategorie. Każda z kategorii składa się ostatecznie z kilku milionów zmiennych opisujących pewne charakterystyki europejskich państw na przestrzeni lat. Rysunek 4 przedstawia część schematu bazy z zaznaczoną ścieżką do przykładowej zmiennej dotyczącej śmiertelności noworodków.

Jedną z podkategorii kategorii `popul`, dotyczącej ludności i warunków społecznych jest `demo` dotyczące demografii i migracji, następnie `demo_mor` opisująca śmiertelności. Do tej podkategorii należy zbiór `demo_minfind` o śmiertelności noworodków (*Infant mortality rates*). Na każdy ze zbiorów składają się dane opisujące pewne zjawisko w krajach europejskich na przestrzeni lat, z podziałem na pewne wybrane cechy. Przykładowa zmienna (zaznaczona na rys. 4 na różowo)




Rysunek 3: Wartości parametru regularyzacyjnego $\lambda$ dla jakich $x$ zostałby włączony do zbioru aktywnych indeksów. Zostały on przedstawione w funkcji $\text{cor}(x, s_{rot})$ i $\text{cor}(x, u_{rot})$, a kolor odpowiada wartości $\lambda$ jaką przyjmuje w zależności od tych korelacji. Wykres został ograniczony poprzez $\lambda = 0.7$, będącą przykładową wartością jaką przyjmuje parametr na ścieżce w momencie włączenia poprzedniej zmiennej (zatem szukane $\lambda$ musi być od niego mniejsze) oraz poprzez $\lambda_{c_1} = 0.3$ (od którego szukane $\lambda$ musi być większe). Białe obszary oznaczają miejsca gdzie rozwiązanie nie istnieje lub nie zawiera się między 0.3 i 0.7. Czarne punkty są ograniczeniami jakie można nałożyć na $\text{cor}(x, u_{rot})$, zamiast na $\text{cor}(x, u)$ i $\text{cor}(x, s)$ i wtedy szary obszar odpowiada tym ograniczeniom.



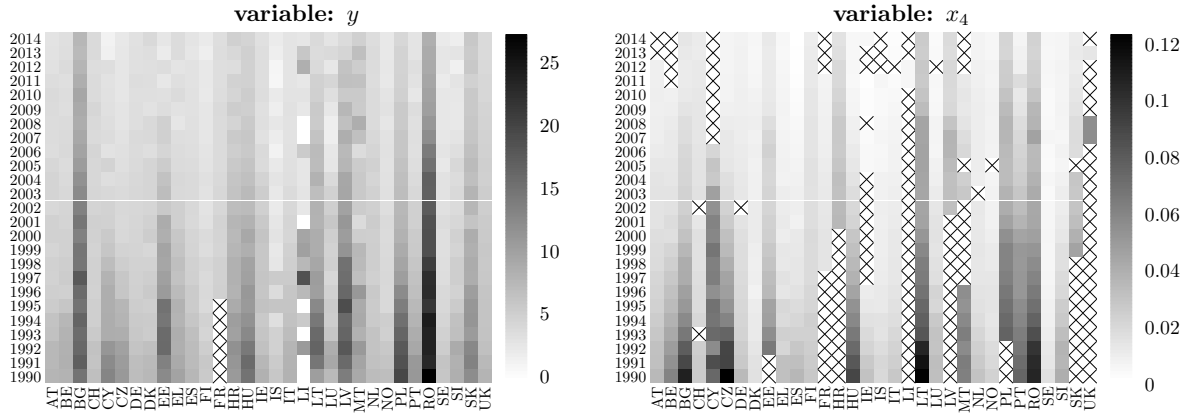

Rysunek 4: Część bazy danych Eurostat (szczegółowy opis można znaleźć w [8]). Graf pokazuje ścieżkę wiodącą do przykładowej zmiennej opisującej śmiertelność noworodków (zaznaczonej na różowo). Niebieskie wierzchołki opisują kategorie, a zielony to konkretny zbiór zawierający rozważaną zmienną.

jest przedstawiona za pomocą wymiarów `unit`: `RT` (oznaczającą stosunek na 1000 urodzeń) oraz `indic_de`: `INFMORRT` (opisywana wielkość: śmiertelność noworodków).

Każda ze zmiennych jest opisywana przez `geo` (kraj lub region) i `time` (okres czasu, zwykle są to lata, ale niektóre zmienne są opisane za pomocą danych miesięcznych lub kwartalnych). Do dotychczasowej analizy zostały wykorzystane dane, które można było opisać za pomocą par `(kraj, rok)` i w całej bazie jest ich około 8 milionów. Wybranymi krajami były wszystkie państwa Unii Europejskiej oraz dodatkowo 4 państwa należące do Europejskiego Stowarzyszenia Wolnego Handlu. Ograniczyliśmy się dodatkowo do 25 lat ($1990-2014$), zatem ostateczna liczba obserwacji dla każdej zmiennej wyniosła 800. Dodatkowo podczas testów dzielimy zmienne na część uczącą (lata $1990-2010$) i testową (lata $2011-2014$).

Jak w przypadku każdych danych rzeczywistych trzeba odpowiednio uwzględnić brakujące dane, które zostały uzupełnione w taki sposób, że osobno dla każdego kraju środkowe wartości w szeregach czasowych zostały uzupełnione za pomocą liniowej interpolacji a skrajne najbliższą niebrakującą wartością.

Rysunek 5 przedstawia strukturę braków danych dla dwóch przykładowych zmiennych – opisanej wcześniej śmiertelności noworodków i jednej ze zmiennych znalezionych przez model opisany w podrozdziale 6.1.2.




Rysunek 5: Struktura braków danych na dwóch przykładowych zmiennych. × oznacza brak danych, a natężenie koloru mówi o wartości danej obserwacji.

Baza danych Eurostat jest regularnie uaktualniana, a dane wykorzystane do dotychczasowej analizy zostały pobrane 28.02.2018.

### 6.1.2 Wyniki

Ten punkt rozpocznę od krótkiego przykładu zastosowania algorytmu 3 na danych pochodzących z Eurostatu. Jako zmienną odpowiedzi przyjmiemy śmiertelność noworodków `popul: demo_minfind, indec_de: INFMORRT, unit: RT`. Do modelu zostanie dobranych w sposób automatyczny 5 zmiennych (wersja nieinteraktywna) i porównane zostaną pierwiastki błędów średniokwadratowych między kolejnymi modelami na zbiorze uczącym i testowym. Szczegółową analizę przykładów zarówno w wersji interaktywnej (z częściowym ręcznym wyborem zmiennych w kolejnych krokach), jak i w pełni automatycznej można znaleźć w pracy [17].

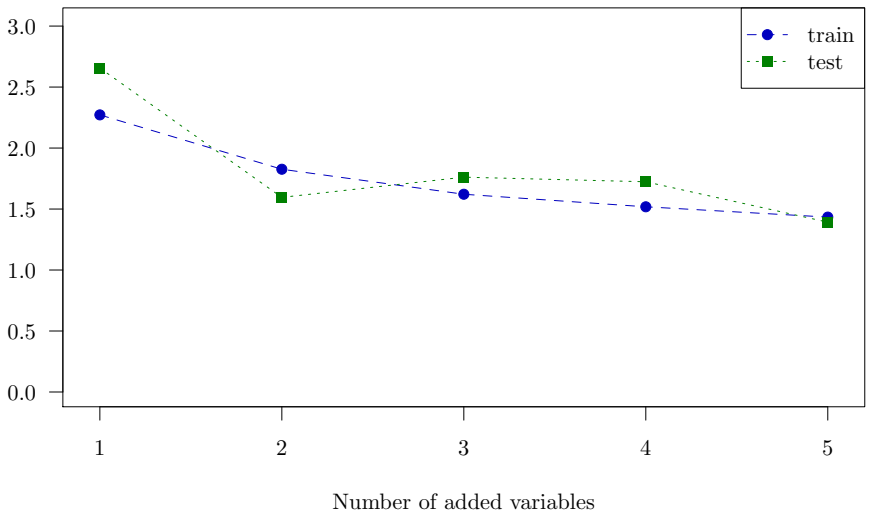
Rysunek 6 przedstawia, jak zmienia się pierwiastek z sumy kwadratów błędów podczas dodawania kolejnych zmiennych. Okazuje się, że w pełni automatyczny dobór zmiennych z zastosowaniem przybliżonego indeksu do filtrowania danych w kolejnych krokach daje dosyć dobre wyniki.

Następnie przeprowadzona została symulacja, w której w każdym kroku tworzymy wektor odpowiedzi $y$ jako liniową kombinację losowo wybranych zmiennych z bazy Eurostat oraz pewnego normalnego błędu. Rozważamy dwa warianty – wyjściowe modele z 2 oraz z 3 zmiennymi w następującej postaci:

$$\begin{aligned} &(a) \quad y = 5x_1 + 2x_2 + \varepsilon, \\ &(b) \quad y = 10x_1 + 3x_2 + x_3 + \varepsilon, \end{aligned} \tag{6.1}$$

gdzie $\varepsilon$ jest gaussowskim szumem (z odchyleniem standardowym równym 0.02) Symulacje powtarzamy 100 razy.

Za każdym razem budujemy model z maksymalnie 5 zmiennymi, z zastosowaniem kryterium (test $F$ z $\alpha = \frac{0.01}{k(k+1)}$, uwzględniający poprawkę na porównania wielokrotne). W idealnym przypadku kryterium powinno zakończyć budowę modelu po dodaniu 2 lub 3 zmiennych, jednak z powodu dodanego szumu nie musi się tak stać.




Rysunek 6: Zmiana pierwiastka z sumy kwadratów błędów pomiędzy modelami tworzonymi w kolejnych krokach.

Tabela 1: Jak często kryterium oparte na teście $F$ kończyło budowę modelu dla danych opisanych wzorami (6.1).

| wielkość modelu | $y \sim x_1 + x_2$ | $y \sim x_1 + x_2 + x_3$ |
|---|---|---|
| 1 zmienna | 1% | 0% |
| 2 zmienne | 29% | 0% |
| 3 zmienne | 30% | 20% |
| 4 zmienne | 6% | 18% |
| 5 lub więcej | 34% | 62% |

W tabeli 1 możemy porównać jak często kryterium oparte na teście $F$ zakończyło budowę modelu w przypadku, gdy $y$ był utworzony jako kombinacja 2, a kiedy jako kombinacja 3 zmiennych.

Rysunek 7 przedstawia jak zmieniała się wartość minimalnej korelacji ($\eta$) w zależności od numeru szukanej zmiennej i rodzaju danych.

## 6.2 Ścieżka rozwiązań dla problemu lasso z indeksem wielowymiarowym

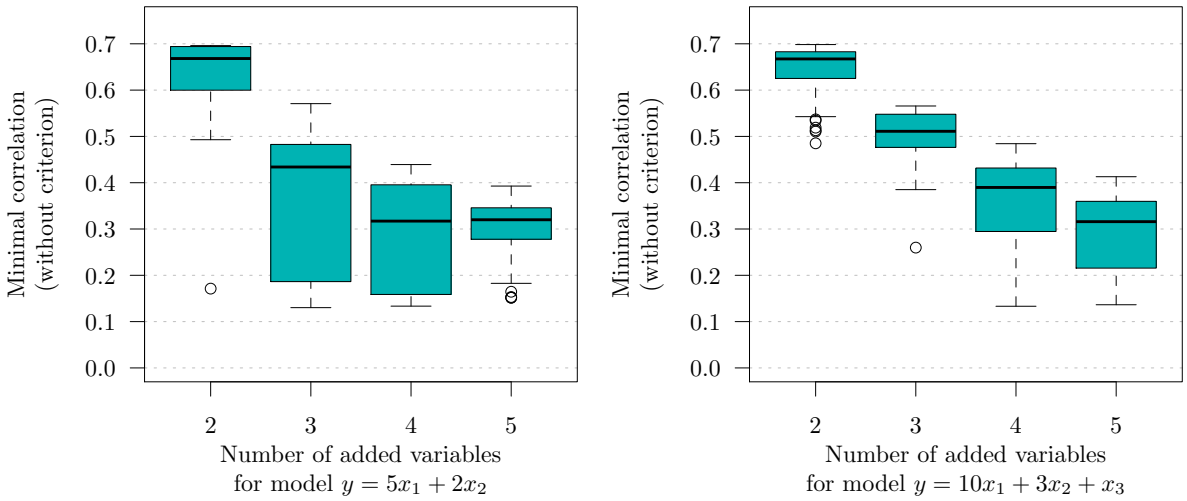
### 6.2.1 Opis danych symulacyjnych

Algorytm 4 został przetestowany na danych symulacyjnych. Rozważone zostały zmienne $x_1, \ldots, x_m$ wygenerowane z rozkładu jednostajnego takiego, że $x_i \sim \mathcal{U}_n(0, 1)$, liczba zmiennych wyniosła $m = 100000$, a zmienna odpowiedzi $y$ została wygenerowana jako

$$y = 10x_{i_1} + 5x_{i_2} + 3x_{i_3} + 2x_{i_4} + x_{i_5} + \varepsilon$$

gdzie $\varepsilon \sim \mathcal{N}_n(0, 0.02)$.

Wszystkie wektory $x_i$ oraz $y$ zostały odpowiednio unormowane (zob. równanie (1.1)), a symulacje zostały wykonane dla różnych wartości $n \in (50, 100, 300, 500, 1000)$ i powtórzone dla każdej z nich 50-krotnie.




Rysunek 7: Minimalna korelacja dla modeli z 2 i 3 zmiennymi.

### 6.2.2 Wyniki

W tej sekcji przeanalizujemy jak duży w kolejnych krokach będzie zbiór $S_{search}$ (zob. wiersz 8 i 11 w alg. 4).

Rysunek 8 przedstawia wykresy skrzynkowe dla liczby zmiennych znalezionych podczas filtrowania ich w każdym kroku algorytmu (lewa strona) oraz minimalną korelację, czyli ograniczenie opisane w tw. 5.1 (prawa strona).

Wartości te zostały przedstawione w zależności od liczby obserwacji $n$ oraz obecnej liczby zmiennych w modelu $k$. Możemy zauważyć, że im większa jest liczba obserwacji $n$ tym mniej zmiennych zostaje znalezionych. W szczególności dla $n \geqslant 300$ i liczby zmiennych w modelu mniejszej niż 5 (należy przypomnieć, że $y$ został utworzony właśnie na podstawie 5 zmiennych) $S_{search}$ było zwykle równe 1 lub 2, zatem znaleziono poniżej 0.01% wszystkich zmiennych.

## 7 Podsumowanie, wnioski i dalsza planowana praca

Wprowadzone modyfikacje i zastosowanie przybliżonych indeksów wielowymiarowych do szybkiego filtrowania zmiennych pozwalają na zastosowanie klasycznych metod do selekcji zmiennych w modelach liniowych takich jak regresja krokowa i Lasso na danych o dużej liczbie predyktorów.

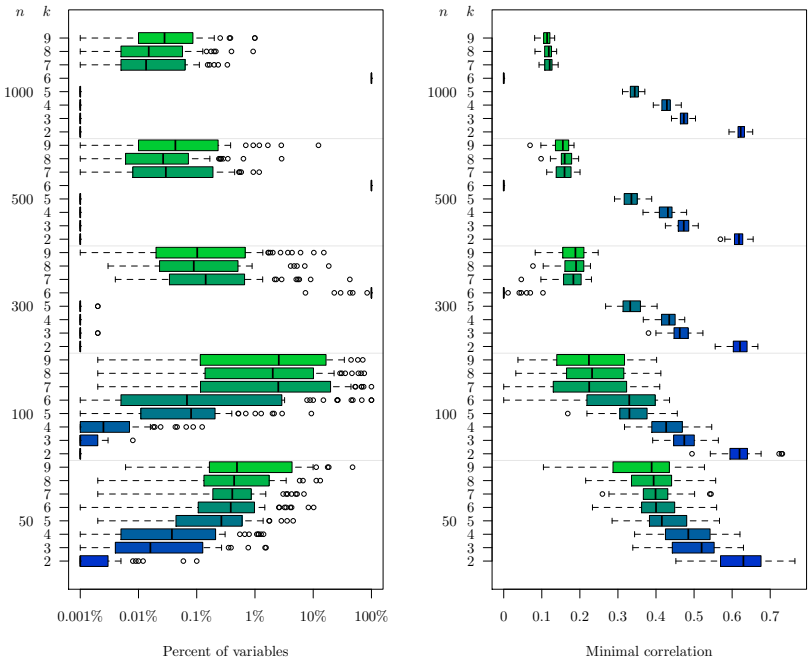
W planowanej rozprawie ściśle dowiodę, że zaproponowane modyfikacje algorytmów stepwise i Lasso, zwracają te same wyniki co ich oryginalne odpowiedniki, pod warunkiem, że podczas wyszukiwania zmiennych skorzystalibyśmy z dokładnego indeksu. Zastosowanie w nich przybliżonych indeksów wielowymiarowych sprawia, że otrzymujemy rozwiązanie zbliżone do oryginalnego, zyskując tym jednak na wydajności. Wyniki te opisane są w artykułach [18] i [19] oraz raporcie badawczym [20] i artykule przygotowanym do publikacji [17].

Dotychczasowe wyniki działania algorytmów są bardzo obiecujące i pokazują, że metody te można z powodzeniem zastosować na danych o dużej liczbie zmiennych.

## Literatura

[1] Linked Data.

[2] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.




Rysunek 8: Wykresy skrzynkowe dla liczby zmiennych znalezionych podczas filtrowania ich w każdym kroku algorytmu 4 (lewa strona) oraz minimalną korelację, czyli ograniczenie opisane w tw. 5.1 (prawa strona). $n$ jest liczbą obserwacji dla danego modelu, a $k$ liczbą zmiennych należących do aktualnego zbiorów aktywnych zmiennych, czyli liczbą zmiennych w obecnym modelu.

[3] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski. Optimization with Sparsity-Inducing Penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, Stycze/n 2012.

[4] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[5] C. Bizer, T. Heath, T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[6] Heumann C., Nittner T., Rao C.R., Scheid S., Toutenburg H. *Linear Models: Least Squares and Alternatives*. Springer New York, 2013.

[7] M. A. Efroymson. *Multiple Regression Analysis*. Wiley, 1960.

[8] Eurostat database. http://ec.europa.eu/eurostat.

[9] T. Heath, Bizer C. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, wydanie 1, 2011.

[10] Jeff Johnson, Matthijs Douze, Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[11] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4:66–70, 2011.

[12] Stephen M Omohundro. Five balltree construction algorithms. Raport instytutowy, International Computer Science Institute Berkeley, 1989.

[13] C.R. Rao. *Linear Statistical Inference and its Applications*. Wiley Series in Probability and Statistics. Wiley, 2009.

[14] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, strony 267–288, 1996.

[16] Peter N Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. *SODA*, wolumen 93, strony 311–321, 1993.

[17] Barbara Żogała-Siudem, Szymon Jaroszewicz. Stepwise regression on large collections of open data. a practical approach. artykuł przygotowany do publikacji.

[18] Barbara Żogała-Siudem, Szymon Jaroszewicz. Fast stepwise regression on Linked Data. *1st Workshop on Linked Data for Knowledge Discovery (LD4KD)*, 2014.

[19] Barbara Żogała-Siudem, Szymon Jaroszewicz. Geometric approach to stepwise regression. *Computational Methods in Data Analysis. Information Technologies: Research and their Interdisciplinary Applications ITRIA 2015*, strony 213–223, 2015.

[20] Barbara Żogała-Siudem, Szymon Jaroszewicz. Lasso regularization path on large collections of data using multidimensional indexing. *Raport Badawczy IBS PAN*, (9), 2018.