

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregów czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Mariusz Kubus

Politechnika Opolska

PROPOZYCJA MODYFIKACJI METODY ZŁAGODZONEGO LASSO

Streszczenie: Regularyzowana regresja liniowa (np. LASSO [Tibshirani 1996]) zyskała duże zainteresowanie jako narzędzie selekcji zmiennych. Meinshausen [2007] zaproponował modyfikację metody LASSO, wprowadzając parametr łagodzący, który kontroluje variancję parametrów strukturalnych niezależnie od etapu eliminacji zmiennych. Metoda ta jest rekomendowana dla dużych wymiarów i dla dużego stosunku variancji zmiennej objaśnianej do variancji składnika losowego. W artykule zaproponowano modyfikację metody złagodzonego LASSO. Przeprowadzone symulacje pokazały, że nowe podejście daje bardziej stabilne wyniki i skuteczniej eliminuje zmienne nieistotne (tj. takie, które nie mają wpływu na zmienną objaśnianą).

Słowa kluczowe: regularyzowana regresja liniowa, złagodzone LASSO, selekcja zmiennych.

1. Wstęp

Problematyka selekcji zmiennych w metodach statystycznego uczenia z nauczycielem cieszy się obecnie dużym zainteresowaniem. Uzyskana w ten sposób redukcja wymiaru przestrzeni cech nie tylko ma walory interpretacyjne, ale często pozwala uzyskać model, który cechuje się większą dokładnością przewidywania nieznanymi wartościami (lub kategoriami) zmiennej objaśnianej dla nowych obiektów. Zmienne objaśniające, które nie mają wpływu na zmienną objaśnianą – tzw. zmienne nieistotne (*irrelevant variables*) – wywołują efekt nadmiernego dopasowania do danych (*overfitting*), przez co model traci na zdolności generalizacji. Wśród obecnie wymienianych trzech podejść do selekcji zmiennych [zob. np. Guyon i in. 2006] dużą popularnością cieszą się metody, w których estymacja modelu i dobór zmiennych odbywa się jednocześnie (*embedded methods*). Inaczej mówiąc selekcja zmiennych jest integralną częścią algorytmu uczącego. Przykładami takich metod są drzewa klasyfikacyjne (lub regresyjne) oraz modele liniowe z regularyzacją. Tym drugim poświęcony jest ten artykuł.

Główną ideą regularyzacji jest możliwość uzyskania modeli o różnym stopniu złożoności. W praktyce wiąże się to z lepszym lub gorszym dopasowaniem do

danych ze zbioru uczącego. W przypadku modeli liniowych złożoność rozumiana jest jako liczba parametrów lub norma wektora parametrów. Pierwszą propozycją regularyzacji w liniowym modelu regresji, która wywoływała efekt selekcji zmiennych, było LASSO [Tibshirani 1996]. Pomysł ten doczekał się wielu rozszerzeń i modyfikacji (np. elastyczna sieć [Zou, Hastie 2005], zgrupowane LASSO [Yuan, Lin 2007], złagodzone LASSO [Meinshausen 2007]). W porównaniu z większością metod doboru zmiennych przed etapem uczenia (*filters*) w regularyzowanej regresji liniowej uwzględniany jest kontekst oddziaływania wielu zmiennych objaśniających na zmienną objaśnianą. Z kolei w porównaniu z regresją krokową uważana jest za metodę mniej skłoną do nadmiernego dopasowania do danych. Ponadto metody estymacji w regularyzowanej regresji liniowej cechują się relatywnie małą złożonością obliczeniową. Główną trudnością w praktycznym stosowaniu LASSO (jak i innych metod regularyzacji) jest ustalenie wartości parametru kary. Meinshausen [2007] w metodzie złagodzonego LASSO (*relaxed LASSO*) zaproponował modyfikację polegającą na wprowadzeniu dodatkowego parametru regularyzacji. Metoda ta jest zalecana szczególnie w przypadku wysokich wymiarów przestrzeni cech oraz dużej wariancji zmiennej objaśnianej w stosunku do wariancji składnika losowego (*signal-to-noise ratio*). W przeprowadzonych symulacjach, mających na celu porównanie różnych metod selekcji zmiennych w regresji dla dużych wymiarów przestrzeni cech [zob. Kubus 2013a], okazało się, że metoda Meinshausena [2007] jest konkurencyjna, lecz cechuje się małą stabilnością, tzn. czasem włącza do modelu dużą liczbę zmiennych nieistotnych, sztucznie wprowadzonych do zbioru danych. W artykule zaproponowana będzie modyfikacja metody złagodzonego LASSO, polegająca na konstrukcji ciągu modeli zagnieżdżonych i wyborze optymalnego. Skuteczność nowego podejścia zostanie potwierdzona empirycznie za pomocą badań symulacyjnych.

2. Regularyzacja w liniowym modelu regresji

Przedmiotem rozważań będzie zbiór wielowymiarowych obserwacji (tzw. zbiór uczący):

$$\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N) : \mathbf{x}_i \in \mathbf{X} = (X_1, \dots, X_p), y_i \in Y, i \in \{1, \dots, N\}\}, \quad (1)$$

gdzie Y (ilościowa zmienna objaśniana) reprezentuje zjawisko, które chcemy wyjaśnić na podstawie obserwowanych cech X_1, \dots, X_p , o których zakładać będziemy, że są ilościowe lub binarne. Na podstawie informacji, jaką niesie zbiór uczący, estymowane będą parametry liniowego modelu regresji:

$$y = b_0 + b_1 x_1 + \dots + b_p x_p + \varepsilon. \quad (2)$$

W przypadku, gdy w zbiorze danych znajdują się zmienne nieistotne, estymatory klasycznej metody najmniejszych kwadratów (MNK), choć nieobciążone, nie

gwarantują dokładnych predykcji dla obiektów nowych (spoza zbioru uczącego). Sytuacja ta sprawia, że model odzwierciedla nie tylko proces generowania danych, ale też zawarty w nich szum, przez co jest mało stabilny i traci zdolność generalizacji. Wprowadzenie składnika kary za duże wartości bezwzględne parametrów w kryterium wykorzystywanym do estymacji daje możliwość uzyskania estymatorów o mniejszej wariancji, choć obciążonych. Tak dzieje się w metodzie LASSO [Tibshirani 1996]. Kryterium ma postać:

$$\hat{\mathbf{b}}^{LASSO} = \arg \min_b \left(\sum_{i=1}^N \left(y_i - b_0 - \sum_{j=1}^p b_j x_{ij} \right)^2 + \lambda \cdot \sum_{j=1}^p |b_j| \right), \quad (3)$$

gdzie λ jest parametrem regularyzacji. Wartości (bezwzględne) estymatorów LASSO są mniejsze od odpowiadających im estymatorów MNK. W skrajnym przypadku niektóre współczynniki się zerują, co daje efekt selekcji zmiennych. Rozmiar zmniejszania wartości (bezwzględnych) współczynników oraz liczba zerujących się współczynników zależy od wartości parametru λ . Duże wartości λ drastycznie zmniejszają wartości (bezwzględne) współczynników i więcej z nich zostaje wyzerowanych. Zmniejsza się wtedy wariancja, lecz rośnie obciążenie. Celem jest uzyskanie pewnego stanu kompromisu między obciążeniem a wariancją, a w efekcie modelu o optymalnej zdolności generalizacji, czyli dokładności przewidywania dla nowych obiektów [zob. np. Hastie i in. 2009; s. 219-224]. Ustalenie parametru λ jest zatem kluczowym zadaniem w stosowaniu regresji liniowej z regularyzacją. Zwykle w tym celu dla różnych wartości λ stosuje się ocenę błędu predykcji przez sprawdzanie krzyżowe lub kryteria informacyjne. Studium porównawcze tych kryteriów można znaleźć w pracy Kubusa [2013b].

Zaproponowana przez Meinshausena [2007] modyfikacja metody LASSO – nazwana złagodzonym LASSO (*relaxed LASSO*) – związana jest z ustaleniem rozmiaru kary za duże wartości (bezwzględne) parametrów. Punktem wyjścia jest postawienie pytania, czy kontrolowanie efektu zmniejszania wartości bezwzględnych parametrów (a co za tym idzie wariancji) oraz efektu selekcji zmiennych przez tylko jeden parametr regularyzacji jest rozwiązaniem optymalnym. Na przykład Efron i in. [2004] stosują połączenie LARS z MNK polegające na tym, że LARS dokonuje selekcji zmiennych, a pozostałe współczynniki estymowane są klasyczną MNK. Meinshausen [2007] wprowadza dodatkowy parametr regularyzacji i proponuje dwukrokową procedurę. Najpierw dokonuje się selekcji zmiennych przez klasyczne zastosowanie LASSO. Ten etap kontrolowany jest parametrem λ . Następnie dla zredukowanego zbioru predyktorów jeszcze raz stosuje się LASSO, a rozmiar zmniejszania wartości bezwzględnych współczynników (a co za tym idzie wariancji) kontrolowany jest parametrem φ , nazywanym parametrem łagodzącym (*relaxation parameter*).

Estymatory złagodzonego LASSO (metoda ta oznaczana będzie przez rLASSO) definiuje się następująco. Oznaczmy przez $A_\lambda \subseteq \{1, \dots, p\}$ niepusty podzbiór indek-

sów zmiennych objaśniających, dla których oszacowano niezerowe współczynniki w metodzie LASSO. Przez $\tilde{\mathbf{b}}$ oznaczymy nowy wektor parametrów modelu liniowego, który będzie estymowany w drugim kroku procedury. Ma on postać:

$$\tilde{b}_j = \begin{cases} b_j & \text{dla } j \in A_\lambda \\ 0 & \text{dla } j \notin A_\lambda \end{cases}. \quad (4)$$

Ponadto $\tilde{b}_0 = b_0$. Wówczas:

$$\hat{\mathbf{b}}^{rLASSO} = \arg \min_b \left(\sum_{i=1}^N \left(y_i - \tilde{b}_0 - \sum_{j=1}^p \tilde{b}_j x_{ij} \right)^2 + \varphi \cdot \lambda \cdot \sum_{j=1}^p |\tilde{b}_j| \right). \quad (5)$$

Jeżeli $\varphi = 1$, to estymatory rLASSO są identyczne z estymatorami LASSO. Z kolei jeśli $\varphi = 0$, to uzyskuje się estymatory MNK dla wyselekcjonowanych wstępnie przez LASSO zmiennych. W tym przypadku model podobny jest do wspomnianego już połączenia LARS z MNK [Efron i in. 2004]. Dla $\varphi \in (0;1)$ można uzyskać continuum modeli pośrednich między tymi skrajnymi przypadkami. W przeprowadzonym przez Meinshausena [2007] eksperymencie, przy założeniu niezależności zmiennych, metoda złagodzonego LASSO wykazała przewagę nad klasycznym LASSO oraz nad połączeniem LARS z MNK. Przy porównywalnych błędach predykcji złagodzone LASSO wprowadzało do modelu mniej zmiennych nieistotnych. Różnice były wyraźne zwłaszcza w przypadku dużej wariancji zmiennej objaśnianej w porównaniu z wariancją składnika losowego (*signal-to-noise ratio*).

3. Propozycja modyfikacji

W trakcie badań własnych nad algorytmem rLASSO okazało się, że jego ponowne zastosowanie dla zredukowanej przestrzeni cech prowadziło do dalszej redukcji wymiaru. Co więcej, empirycznie stwierdzono, że kontynuacja takiego postępowania wykazuje się zbieżnością, tzn. przestrzeń zawsze zredukowana jest dość szybko do jednej zmiennej. Wobec przytoczonych obserwacji proponujemy, by za pomocą wielokrotnego zastosowania metody rLASSO skonstruować ciąg modeli zagnieżdżonych, a następnie wybrać optymalny. Proponowany algorytm oznaczany będzie nrLASSO, a jego formalny zapis przedstawiono w tabeli 1.

Do oceny jakości modeli zagnieżdżonych w trzecim kroku algorytmu wykorzystano kryterium informacyjne EDC [Bai i in. 1986], które dało obiecujące rezultaty w symulacjach przeprowadzonych w pracy [Kubus 2013b]. Ma ono postać:

$$Q(k) = N \cdot \ln \left(\frac{RSS}{N} \right) + (k+1) \cdot P(N), \quad (6)$$

gdzie: k jest liczbą zmiennych w modelu, RSS sumą kwadratów reszt, a $P(N)$ funkcją spełniającą warunki:

$$\lim_{N \rightarrow \infty} \frac{P(N)}{N} = 0 \quad \text{oraz} \quad \lim_{N \rightarrow \infty} \frac{P(N)}{\ln \ln N} = \infty, \quad (7)$$

która decyduje o rozmiarze kary za złożoność modelu. W proponowanym algorytmie przyjęto $P(N) = \sqrt{N}$.

Tabela 1. Algorytm nrLASSO

Niech $A_i \subseteq \{1, \dots, p\}$ będzie podzbiorem indeksów zmiennych objaśniających X_1, \dots, X_p w i -tym kroku algorytmu.

1. Ustal wartości początkowe: $i = 1$ oraz $A_1 = \{1, \dots, p\}$.
2. Dopóki A_i zawiera więcej niż jeden indeks zmiennej, wykonuj:
 - 2a. Zbuduj model rLASSO (M_i) dla predyktorów z indeksami ze zbioru A_i .
 - 2b. Oblicz wartość kryterium oceny Q dla modelu M_i .
 - 2c. Zwiększ numer iteracji o jeden: $i \leftarrow i + 1$.
 - 2d. Zmodyfikuj zbiór indeksów A_i (usuń indeksy zmiennych, dla których współczynniki w modelu M_{i-1} były równe zero).
3. Z ciągu modeli zagnieżdżonych $M_1 < \dots < M_{i-1}$ wybierz model optymalny na podstawie minimalnej wartości kryterium oceny Q .

Źródło: opracowanie własne.

4. Badania symulacyjne

W celu zweryfikowania zaproponowanej modyfikacji metody złagodzonego LASSO przeprowadzono badania symulacyjne. W każdym z generowanych zbiorów danych dla liniowych modeli regresji (2) dokonano podziału na próbę uczącą (100 obserwacji) oraz testową (także 100 obserwacji). Szum gaussowski ε na poziomie 0,4 odchylenia standardowego zmiennej objaśnianej dodawano tylko w próbach uczących. W ten sposób uzyskano stosunek wariancji y do wariancji składnika losowego równy 6,25. Zbiory testowe, niewykorzystane w etapie uczenia, służyły do oszacowania błędu predykcji estymowanych modeli.

W każdym eksperymencie dołączano też 20 zmiennych nieistotnych, niemających wpływu na zmienną objaśnianą. Pierwsze 10 było parami skorelowane według formuły:

$$X_{2k} = k \cdot X_{2k-1} + e, \quad (8)$$

gdzie $k \in \{1, \dots, 5\}$. Realizacje zmiennych o numerach nieparzystych generowano z rozkładu $N(0; 1)$, a szum e z rozkładu $N(0; 0,1)$. Kolejne 10 zmiennych nieistot-

nych generowano z rozkładu zero-jedynkowego: pięć z frakcją jedynek 0,5 oraz pięć z frakcją jedynek 0,25. Każdy z trzech opisanych poniżej eksperymentów był przeprowadzony 100 razy. Wszystkich obliczeń dokonano za pomocą programu R, wykorzystując pakiety `lars` i `relaxo` oraz własne procedury. Po selekcji zmiennych metodami LASSO, rLASSO i nrLASSO współczynniki estymowano klasyczną MNK.

Eksperyment 1

Rozważany będzie model liniowy (2) z pięcioma zmiennymi objaśniającymi ($p = 5$). Realizacje zmiennych X_1, \dots, X_5 oraz współczynniki b_0, b_1, \dots, b_5 generowano niezależnie z rozkładu $N(0; 1)$.

Eksperyment 2

Ponownie rozważać będziemy model liniowy (2) dla $p = 5$. Tym razem wprowadzone będą zależności między zmiennymi z modelu. Zmienne X_1, \dots, X_3 oraz współczynniki b_0, b_1, \dots, b_5 generowano z rozkładu $N(0; 1)$ natomiast:

$$\begin{aligned} X_4 &= X_2 + X_3 + e \\ X_5 &= X_1 + e \end{aligned} \quad (9)$$

gdzie e oznacza szum gaussowski, który generowano z rozkładu $N(0; 0,1)$.

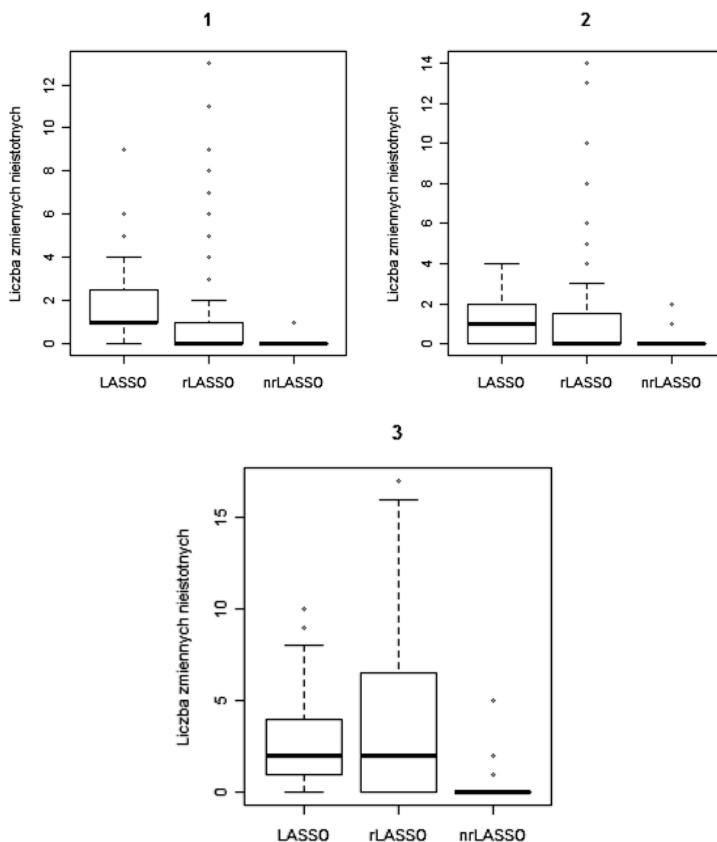
Eksperyment 3

Rozważany będzie model liniowy (2) z 10 zależnymi zmiennymi objaśniającymi ($p = 10$). Wszystkie współczynniki oraz realizacje zmiennych objaśniających (z wyjątkiem X_5 oraz X_{10}) były generowane niezależnie z rozkładu $N(0; 1)$. Zależności wprowadzono wg formuły:

$$x_{5+k*5} = \alpha_{k1}x_{1+k*5} + \alpha_{k2}x_{2+k*5} + \alpha_{k3}x_{3+k*5} + \alpha_{k4}x_{4+k*5} + e_k, \quad (10)$$

dla $k \in \{0, 1\}$. Tu również współczynniki losowane były z jednowymiarowego standaryzowanego rozkładu normalnego, a poziom szumu dobierano losowo: $e_k \sim N(0, s_k)$, gdzie $s_k = m \cdot sd(x_{5+k*5})$ i $m \in \{0,1; 0,2; 0,3; 0,4\}$.

Liczby zmiennych nieistotnych wprowadzanych do modeli w 100 symulacjach przedstawia rys. 1. Ewidentnie widać przewagę zaproponowanej modyfikacji (oznaczonej przez nrLASSO). W każdym z trzech modeli zmienne nieistotne były wprowadzane w niewielkiej liczbie, co najwyżej 3 razy na 100. Zbadano też błędy średniokwadratowe szacowane na zbiorach testowych, a ich rozkłady porównano testem Friedmana (zob. tab. 2). W przypadku dwóch pierwszych eksperymentów hipoteza zerowa o jednakowych rozkładach została odrzucona. Mediany błędów były mniejsze dla metody nrLASSO, a dalsza analiza post-hoc wskazała istotne różnice dla par nrLASSO – LASSO oraz nrLASSO – rLASSO w pierwszym eks-



Rys. 1. Liczby zmiennych nieistotnych wprowadzanych do modeli w eksperymentach 1-3. Dla każdego eksperymentu wykonano 100 symulacji

Źródło: obliczenia własne.

perymentie oraz dla pary nrLASSO – LASSO w drugim eksperymencie. Dla trzeciego modelu mediana błędów w nrLASSO jest nieco większa, ale różnica ta nie jest statystycznie istotna.

Tabela 2. Mediany błędów średniokwadratowych szacowanych na zbiorach testowych w 100 symulacjach przeprowadzonych dla każdego eksperymentu oraz wyniki testu Friedmana

Nr eksperymentu	LASSO	rLASSO	nrLASSO	Test Friedmana (wartości <i>p</i>)
1	0,0802	0,0682	0,0529	0,00000
2	0,0570	0,0519	0,0391	0,00121
3	0,5348	0,5149	0,5980	0,87370

Źródło: obliczenia własne.

5. Podsumowanie

Przeprowadzone symulacje pokazały, że z punktu widzenia zdolności eliminacji zmiennych nieistotnych (tj. takich, które nie mają wpływu na zmienną objaśnianą) zaproponowany w artykule algorytm jest znacznie stabilniejszy od oryginalnego złagodzonego LASSO. Zdecydowanie lepiej je identyfikował przy porównywalnych lub nawet niższych błędach predykcji. Rezultat ten został potwierdzony testami statystycznej istotności. Warto podkreślić, że eksperymenty przeprowadzono w sytuacji ogólniejszej niż w artykule Meishausena [2007], mianowicie dla zmiennych zależnych oraz binarnych zmiennych nieistotnych.

Literatura

- Bai Z.D., Krishnaiah P.R., Zhao L.C. (1986), *On the detection of the number of signals in the presence of white noise*, „Journal of Multivariate Analysis” 20, s.1-25.
- Efron B., Hastie T., Johnstone I., Tibshirani R. (2004), *Least Angle Regression*, „Annals of Statistics” 32 (2), s. 407-499.
- Guyon I., Gunn S., Nikravesh M., Zadeh L. (2006), *Feature Extraction: Foundations and Applications*, Springer, New York.
- Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, Springer, New York.
- Kubus M. (2013a), *Feature selection in high dimensional regression problem*, [w:] C. Domański (red.), *Methods and Applications of Multivariate Statistical Analysis*, Acta Universitatis Lodzianis, „Folia Oeconomica” 286, s. 139-146.
- Kubus M. (2013b), *On model selection in some regularized linear regression methods*, [w:] Cz. Domański, A. Kupis-Fijałkowska (red.), *Multivariate Statistical Analysis – Theory and Practice*, Acta Universitatis Lodzianis, „Folia Oeconomica” 285, s. 115-123.
- Meinshausen N. (2007), *Lasso with relaxation*, „Computational Statistics and Data Analysis” 52(1), s. 374-293.
- Tibshirani R. (1996), *Regression shrinkage and selection via the lasso*, „Journal of the Royal Statistical Society” Series B 58, s. 267-288.
- Yuan M., Lin Y. (2007), *Model selection and estimation in regression with grouped variables*, „Journal of the Royal Statistical Society” Series B. 68(1), s. 49-67.
- Zou H., Hastie T. (2005), *Regularization and variable selection via the elastic net*, „Journal of the Royal Statistical Society” Series B. 67(2), s. 301-320.

THE PROPOSITION OF MODIFICATION OF THE RELAXED LASSO METHOD

Summary: Regularized linear regression (i.e. LASSO [Tibshirani 1996]) has reached a lot of interest as a feature selection tool. Meinshausen [2007] proposed a modified version of the LASSO by introducing a relaxation parameter which controls the variances of the parameters, regardless of the feature elimination stage. This method is recommended in high dimensions, and for the high signal-to-noise ratio. The modification of the relaxed LASSO method is proposed in this paper. The simulations show that the new approach provides more stable results, and more effectively discards noisy variables.

Keywords: regularized linear regression, relaxed LASSO, feature selection.