Integracja metody LARS z pakietem ROI i analiza jej wydajności

Szymon Kowalik

Kosma Grochowski

Definicja zadania

$$\min_{\beta}\left[\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right]$$

$$\text{gdzie } X \in \mathbb{R}^{m\times n}, y \in \mathbb{R}^m, \beta \in \mathbb{R}^n, 0 < \lambda \in \mathbb{R}$$

LARS – Least-angle regression

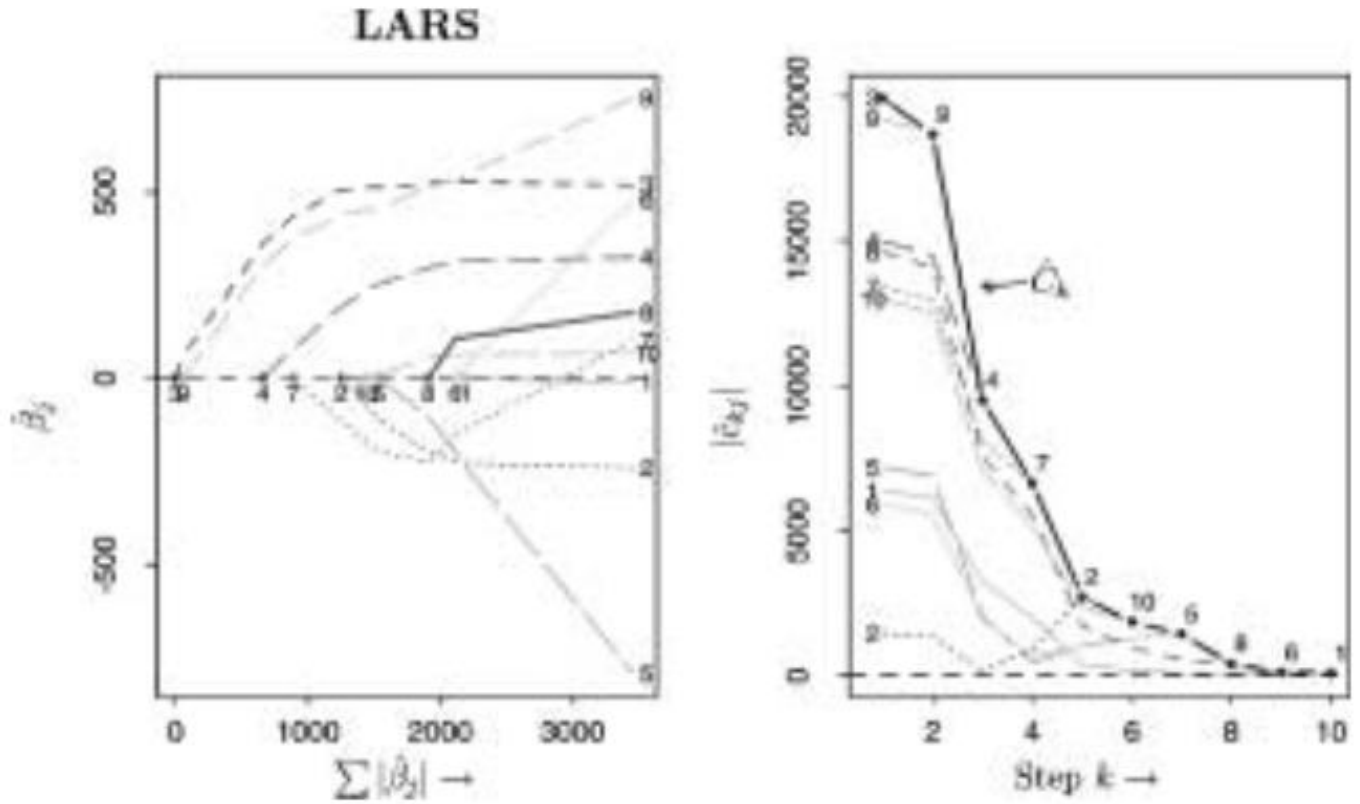


FIG. 3. *LARS analysis of the diabetes study*: (left) *estimates of regression coefficients* $\hat{\beta}_j$, $j = 1, 2, \ldots, 10$; *plotted versus* $\sum|\hat{\beta}_j|$; *plot is slightly different than either Lasso or Stagewise, Figure* 1; (right) *absolute current correlations as function of LARS step; variables enter active set* (2.9) *in order* $3, 9, 4, 7, \ldots, 1$; *heavy curve shows maximum current correlation* $\hat{C}_k$ *declining with* k.

ROI – R Optimization Infrastructure

Integracja pakietu LARS z ROI

Rozpatrywany problem w równoważnej postaci kwadratowej:

$$\min_{(\beta,\gamma,t)} \left[\frac{1}{2}\gamma^T\gamma + \lambda \mathbf{1}^T t \right]$$

$$\text{przy ograniczeniach } y - X\beta = \gamma$$

$$-t \leq \beta \leq t$$

$$\text{gdzie } \gamma \in \mathbb{R}^n, t \in \mathbb{R}^n.$$

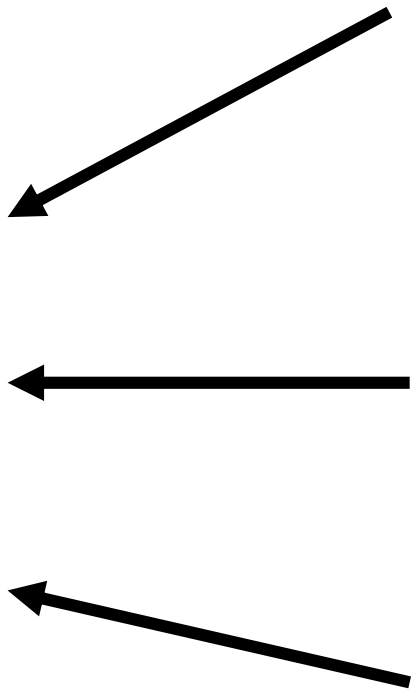
Integracja pakietu LARS z ROI

$$\min_{(\beta,\gamma,t)} \left[\frac{1}{2}\gamma^T\gamma + \lambda \mathbf{1}^T t \right]$$

przy ograniczeniach $y - X\beta = \gamma$

$$-t \leq \beta \leq t$$

gdzie $\gamma \in \mathbb{R}^n, t \in \mathbb{R}^n$.



`Q_objective`

`L_constraint`

$X\beta + \gamma = y, 0 \leq \beta + t$ oraz $\beta - t \leq 0$

`V_bound`

$(-\infty, \infty)$

TABLE 1

Diabetes study: 442 diabetes patients were measured on 10 baseline variables; a prediction model was desired for the response variable, a measure of disease progression one year after baseline

	AGE	SEX	BMI	BP	Serum measurements						Response
Patient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Solver	qpOASES	LARS
Długość działania [s]	10.268	0.046
Uzyskane minimum	5 821 850	5 821 850

Tabela 1: Zestawienie wyników działania metod qpOASES i LARS dla zbioru diabetyków

Wyniki - zbiór diabetyków

```r
sampleSize = 10000
parameterSize = 1000
relevantParameters = 5
# input data
set.seed(42)
x <-data.frame(row.names=1:sampleSize)
for (i in 1:parameterSize){
  x[i] = rnorm(sampleSize,sample(1:5,1),sample(1:5,1))
}
x <-  as.matrix(x)
expected_beta = c(1,2,3,4,5,6, integer(parameterSize - relevantParameters))
set.seed(1337)
y =  expected_beta[1] + rnorm(sampleSize,0,5) # random error
for(i in 1:relevantParameters)
  y = y + expected_beta[i+1]*x[1:sampleSize,i]
```

```
> lars_solved$solution
   [1]  1.995875e+00  3.008961e+00  3.975857e+00  5.009184e+00  5.993435e+00  1.667706e-02  4.261603e-02 -1.210468e-02 -6.640612e-02
  [10]  2.189152e-04  2.188099e-02  1.568798e-02  1.409422e-02 -2.246953e-02 -1.160973e-02 -1.485064e-02  1.939044e-02  9.835514e-03
  [19]  4.285601e-03  1.752309e-03  7.061594e-03 -2.191137e-03 -3.639575e-04 -9.637074e-03 -6.492898e-03 -1.687570e-02  1.696625e-02
  [28] -9.833943e-03 -2.596959e-02 -2.602807e-02  4.090208e-02 -4.462902e-03  2.902277e-03  3.747867e-03 -7.485697e-02  1.967416e-02
  [37] -1.115977e-02  1.302357e-02 -4.980054e-03 -1.072638e-02 -2.070503e-02  8.666690e-03 -1.293163e-02  1.191716e-01 -4.056124e-02
  [46] -9.295164e-02  5.399942e-02  1.614551e-03 -5.134621e-03 -4.791745e-03  2.233274e-02  2.450839e-03  2.016117e-02  1.788676e-02
  [55]  1.515971e-02 -3.866086e-02  2.968536e-02  4.128974e-02 -1.548843e-03 -2.799268e-02  2.335621e-02  2.019262e-02  1.474969e-02
  [64] -8.893443e-03 -2.992327e-03  2.033466e-02 -4.780453e-03  4.042948e-02  4.579709e-03 -3.289773e-03  2.617861e-02 -2.515173e-02
  [73] -4.213640e-05  1.074158e-02  7.471648e-04 -3.958956e-03  2.369591e-03 -7.484342e-04  2.457282e-03  1.097061e-02 -5.142730e-02
  [82]  1.011753e-02 -1.686706e-02 -4.373651e-03  1.194226e-02 -3.892990e-02 -3.928281e-02  5.250893e-02 -4.295586e-03 -8.825925e-03
  [91]  1.404986e-02  3.581753e-02  2.528565e-03 -2.855080e-02 -1.754234e-02 -7.347552e-02  1.533609e-02 -1.119319e-02  7.911222e-02
```

```
> lars_solved$solution
   [1]  1.984225e+00  3.006116e+00  3.769562e+00  4.998976e+00  5.937082e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [10]  0.000000e+00  7.123507e-03  0.000000e+00  8.739340e-03  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [19]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [28]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  8.681202e-03
  [37]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [46]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  1.763945e-03  0.000000e+00
  [55]  2.589573e-04  0.000000e+00  0.000000e+00  2.437127e-02  0.000000e+00  0.000000e+00  3.258853e-03  0.000000e+00  0.000000e+00
  [64]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [73]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
  [82]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  2.146420e-02  0.000000e+00  0.000000e+00
  [91]  3.445155e-03  0.000000e+00  0.000000e+00 -1.545400e-02 -7.555467e-03  0.000000e+00  0.000000e+00  0.000000e+00  4.029913e-02
```

Wyniki - zbiór sztucznie wygenerowany

Wyniki - Milion Song Dataset

Solver	qpOASES	LARS
Długość działania [s]	–	52.218
Uzyskane minimum	–	4 855 338 257

Tabela 2: Zestawienie wyników działania metod qpOASES i LARS dla zbioru Million Song Dataset. Wyniki metodą qpOASES nie zostały uzyskane ze względu na konieczność alokacji 1.98 TB pamięci, co nie było możliwe na dostępnym sprzęcie.

```
> lars_solved$solution
 [1] 37.3559561521 -1.6492771718 -0.0764143876  0.0260697824 -0.4862708458 -4.3935856715 -2.3670519570 -0.0641517548 -2.5538759843
[10]  2.2322219931 -7.9036106368  3.6006216701  1.1516981402  0.0121752567  0.0217104520  0.0011421394  0.0239042479  0.0761736594
[19]  0.0515965569 -0.0232086561  0.0727317738  0.0985964690  0.0008668678  0.1886522000 -0.2886676595  0.0005560200 -0.0128450373
[28]  0.0125541121  0.0126541251  0.0432787295  0.0480023564 -0.0003870457  0.0071250983  0.1168424827  0.1246476924  0.0115271259
[37] -0.0133575340 -0.0099549843  0.0110775638 -0.0036024597  0.0585208117  0.0285495880  0.0514585387  0.0016100830 -0.0444386952
[46]  0.2698362827  0.0011606104 -0.0349459764  0.0002122285  0.0499779112  0.0692937997  0.0214057775 -0.0478259350 -0.0028348575
[55]  0.0602524214 -0.0070885328 -0.0323221190 -0.0136372152  0.0301034178  0.0446487951 -0.0131394446 -0.0524214290  1.1118005321
[64]  0.0227850008 -0.0268747435 -0.0159599035 -0.0805296612 -0.0102387241 -0.0528189120 -0.0316184212 -0.0144420883 -0.0195038765
[73]  0.0159409156 -0.0602471592 -0.0063217442  0.5778799256  0.0327854095  0.0483628772  0.0389492687 -0.0055737319  0.5107689048
[82] -0.0323711377 -0.0849152856 -0.0599779835 -0.4040065907  0.1074068547  0.0091730145  0.9893902110 -0.0106387498  0.0351111141
```

Rysunek 3: Przedstawienie predyktorów dla $\lambda = 0$

```
> lars_solved$solution
 [1] 37.8422134050 -1.2455856329 -0.0963855568 -0.8958187010  0.0000000000 -0.7729077892  0.0000000000  0.0000000000  0.0000000000
[10]  0.0000000000  0.0000000000  0.0000000000  1.0075649415  0.0189157376  0.0150267264  0.0055158094  0.0322303557  0.0756651841
[19]  0.0464350044  0.0000000000  0.1159091095  0.0000000000  0.0000000000  0.2166998300 -0.3210318660  0.0000000000  0.0000000000
[28]  0.0000000000  0.0002353114  0.0138865961  0.0099284308  0.0000000000  0.0000000000  0.0000000000  0.0341434195  0.1100487605
[37] -0.0081218971 -0.0005829536  0.0122885031  0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000
[46]  0.2977076910  0.0001587481 -0.0008093558  0.0000000000  0.0474093119  0.0287685051  0.0577611153 -0.0503492363  0.0000000000
[55]  0.0000000000 -0.0255900964 -0.0202475280  0.0071626587  0.0000000000  0.0314986619 -0.0281474739 -0.0347120242  0.7506768756
[64]  0.0194019596  0.0000000000  0.0000000000 -0.0231306031 -0.0102399615 -0.0384741738 -0.0386408378 -0.0016141190  0.0023216355
[73]  0.0288944459 -0.0211423299  0.0000000000  0.0000000000  0.0431383444  0.0089855693  0.0408852188 -0.0206150917  0.1685361412
[82] -0.0589551274 -0.0193331268 -0.0459084594  0.0000000000  0.0516776937  0.0000000000  0.0000000000 -0.0092742202  0.0000000000
```

Rysunek 4: Przedstawienie predyktorów dla $\lambda = 10e8$

Wnioski

Metoda LARS daje szybkie i dokładne wyniki nawet dla dużych zbiorów

Zwiększanie współczynnika lambda powoduje redukcję wariancji modelu

Pakiet ROI zapewnia standaryzację użycia implementacji algorytmów optymalizacyjnych

Dziękujemy za uwagę :-)