Using of Ellipsoid Method and Linear Regression with L_1-regularization on Medical Data Investigation

P. Stetsyuk, **V. Stovba**, I. Senko, I. Chaikovsky

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine

4th International Workshop of IT-professionals on Artificial Intelligence
(ProfIT AI 2024)

September 25-27, 2024

Content

1. Criterion of Least Moduli powered to $1 \leq p \leq 2$
2. The emlmpr algorithm with L_1-regularization
3. Computational experiments ($\lambda = 0$)
4. Medical data investigation ($\lambda \geq 0$)

Content

1. Criterion of Least Moduli powered to $1 \leq p \leq 2$
2. The emlmpr algorithm with L_1-regularization
3. Computational experiments ($\lambda = 0$)
4. Medical data investigation ($\lambda \geq 0$)

Linear regression and LMP criterion

m observations $(\mathbf{a}_i, y_i)$, $\mathbf{a}_i = (a_{i1}, \ldots, a_{in}) \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = \overline{1, m}$

n unknown parameters $x_1, \ldots, x_n$ to be evaluated, $m > n$

$$y_i = \sum_{j=1}^{n} a_{ij} x_j + \varepsilon_i, \quad i = \overline{1, m} \tag{1}$$

Criterion of Least Moduli powered to $1 \leq p \leq 2$ (LMP) is a mathematical programming problem

$$f_p^* = f_p(x_p^*) = \min_{x \in \mathbb{R}^n} \left\{ f_p(x) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} a_{ij} x_j \right|^p \right\} \tag{2}$$

LMP with $p = 1, 2$

Least Moduli Method is the problem (2) with $p = 1$

$$f_1^* = \min_{x \in \mathbb{R}^n} \left\{ f_1(x) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} a_{ij} x_j \right| \right\} \tag{3}$$

Least Square Method is the problem (2) with $p = 2$

$$f_2^* = \min_{x \in \mathbb{R}^n} \left\{ f_2(x) = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} a_{ij} x_j \right)^2 \right\} \tag{4}$$

LMP with L_1-regularization

corresponds to the following mathematical programming problem

$$f_p^* = f_p(x_p^*) = \min_{x \in \mathbb{R}^n} \left\{ f_p(x) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} a_{ij} x_j \right|^p + \lambda \sum_{j=1}^{n} |x_j| \right\} \quad (5)$$

- Problem (5) is a problem of unconditional minimization of a convex nonsmooth function $f_p(x)$ if $p > 1$ and convex piecewise-linear function if $p = 1$
- $\lambda \geq 0$ is a regularization parameter

Content

1. Criterion of Least Moduli powered to $1 \leq p \leq 2$
2. The emlmpr algorithm with L_1-regularization
3. Computational experiments ($\lambda = 0$)
4. Medical data investigation ($\lambda \geq 0$)

The emlmpr algorithm based on ellipsoid method

Input parameters: $B_0 := I_n$, x_0, r_0, $\varepsilon_f > 0$

Step 1. Calculate $f_p(x_k)$ and $g_{f_p}(x_k)$. If $r_k \left\| B_k^T g_{f_p}(x_k) \right\| \leq \varepsilon_f$, then STOP: $k^* = k$ and $x_p^* = x_k$. Otherwise move to Step 2.

Step 2. Let $\xi_k := \frac{B_k^T g_{f_p}(x_k)}{\left\| B_k^T g_{f_p}(x_k) \right\|}$.

Step 3. Calculate the next point

$$x_{k+1} := x_k - h_k B_k \xi_k, \text{ where } h_k = \frac{1}{n+1} r_k.$$

Step 4. Calculate

$$B_{k+1} := B_k + \left(\sqrt{\frac{n-1}{n+1}} - 1 \right) (B_k \xi_k) \xi_k^T \text{ and } r_{k+1} := r_k \frac{n}{\sqrt{n^2 - 1}}.$$

Step 5. Move to iteration $k+1$ with values x_{k+1}, r_{k+1}, B_{k+1}.

Computational capabilities of emlmpr

Theorem. Sequence of points $\{x_k\}_{k=0}^{k^*}$ satisfy the following inequalities:

$$\left\| B_k^{-1}(x_k - x_p^*) \right\| \leq r_k, \quad k = 0, 1, \ldots, k^*.$$

On each iteration $k > 0$ the value of decreasing of volume of the ellipsoid $E_k = \left\{ x \in \mathbb{R}^n : \left\| B_k^{-1}(x_k - x) \right\| \leq r_k \right\}$, which localizes point x_p^*, is constant and equal to

$$q = \frac{vol(E_k)}{vol(E_{k-1})} = \sqrt{\frac{n-1}{n+1}} \left(\frac{n}{\sqrt{n^2-1}} \right)^n < exp\left\{ -\frac{1}{2(n+1)} \right\} < 1.$$

The **emlmpr** algorithm can be successfully run on modern computers if $n = 10 \div 30$ and $m = 100 \div 1000$

Content

1. Criterion of Least Moduli powered to $1 \leq p \leq 2$
2. The emlmpr algorithm with L_1-regularization
3. Computational experiments ($\lambda = 0$)
4. Medical data investigation ($\lambda \geq 0$)

Test example 1: solving time estimation

- $\mathbf{n} = 30$, $\mathbf{m} = 300$
- `A = 10 * rand(m,n),  y = A * xstar(n,1)`
- `xstar(n,1) = round(10 * rand(n,1) + 0.5)`
- `x0(n,1) = round(5 * rand(n,1))`
- `r0 = 5 * norm(x0 - xstar)`
- Accuracy $\varepsilon_f = (10^{-6})^p$

$$n = 30, \quad m = 300, \quad 1 \le p \le 2, \quad \lambda = 0$$

p	ε_f	$time\ (sec)$	itn	f_p	dx
1.0	1.0e-06	5.17	45375	1.71062e-08	2.3e-11
1.2	3.2e-08	6.99	42148	3.58266e-10	1.2e-10
1.5	1.0e-09	5.39	40061	9.87425e-12	4.1e-10
1.8	3.2e-11	5.97	38260	1.81563e-13	7.2e-10
2.0	1.0e-12	3.91	37216	7.45098e-15	1.8e-09

Test example 2: robustness of LMM

- $\mathbf{n} = 30$, $\mathbf{m} = 300$
- `A = 10 * rand(m,n),  y = A * xstar(n,1)`
- `xstar(n,1) = round(10 * rand(n,1) + 0.5)`
- `x0(n,1) = round(5 * rand(n,1))`
- `r0 = 5 * norm(x0 - xstar)`
- even components of **y** are multiplied by the value
  `q = (1.0 + 1.0 * sign(0.5 - rand))`

$$n = 30, \quad m = 300, \quad 1 \le p \le 2, \quad \lambda = 0$$

p	ε_f	*time (sec)*	*itn*	f_r	dx
1.0	1.0e-06	3.60	43337	1.34006e+05	**1.1e-10**
1.2	3.2e-08	2.80	23909	7.26497e+05	2.8e+01
1.5	1.0e-09	3.41	28017	3.85135e+06	5.4e+01
1.8	3.2e-11	3.77	32560	2.04598e+07	6.4e+01
2.0	1.0e-12	3.03	37360	1.09408e+08	6.9e+01

Content

1. Criterion of Least Moduli powered to $1 \leq p \leq 2$
2. The emlmpr algorithm with L_1-regularization
3. Computational experiments ($\lambda = 0$)
4. Medical data investigation ($\lambda \geq 0$)

Medical data description

- 90 patients, 16 features (cardiological and basic)
- Beck anxiety scale is predicted
- Sequential Feature Selector (R^2 quality metric, Scikit-Learn v1.2)
- Isolation Forest (Scikit-Learn v1.2)
- $p = 1$ та $p = 2$

Solving time ($\lambda = 0$)

Beck anxiety scale, $m = 90$, $n = 16$, $p = 1, 2$

p	ε_f	*time (sec)*	*itn*	f_p
1.0	1.0e-06	**0.56**	7640	2.61732e+02
1.0	1.0e-20	**0.69**	8383	2.61732e+02
2.0	1.0e-12	**0.95**	11700	1.38880e+03
2.0	1.0e-40	**2.15**	29719	1.38880e+03

Model parameters ($\lambda = 0$)

$p = 1$		$p = 2$	
$\varepsilon = 10^{-6}$	$\varepsilon = 10^{-20}$	$\varepsilon = 10^{-12}$	$\varepsilon = 10^{-40}$
-8.9453e-02	-8.9453e-02	-1.1440e-01	-1.1440e-01
-2.5798e-03	-2.5798e-03	-7.0401e-03	-7.0401e-03
-3.2033e-02	-3.2033e-02	-2.4929e-02	-2.4929e-02
2.0159e-02	2.0159e-02	2.7608e-02	2.7608e-02
2.4191e-01	2.4191e-01	2.6629e-01	2.6629e-01
6.8819e-03	6.8820e-03	3.5181e-02	3.5181e-02
1.4368e+08	**1.4868e+08**	**7.5756e+06**	**6.3818e+06**
-1.4482e-02	-1.4482e-02	-1.2610e-02	-1.2610e-02
3.7213e-02	3.7213e-02	3.7520e-02	3.7520e-02
-8.9244e-03	-8.9244e-03	-4.5045e-02	-4.5045e-02
1.4087e+00	1.4087e+00	2.1146e+00	2.1146e+00
-1.0175e+00	-1.0175e+00	-2.4560e+00	-2.4560e+00
-3.2982e-01	-3.2982e-01	-1.0611e-01	-1.0611e-01
-3.9263e+08	**-3.7817e+08**	**-7.8529e+06**	**-4.2703e+07**
1.4368e+08	**1.4868e+08**	**7.5756e+06**	**6.3818e+06**
5.5243e+08	**-4.9760e+08**	**-4.5914e+08**	**9.8456e+08**

Model parameters ($\lambda = 0.1$)

$p = 1$		$p = 2$	
$\varepsilon = 10^{-6}$	$\varepsilon = 10^{-20}$	$\varepsilon = 10^{-12}$	$\varepsilon = 10^{-40}$
-8.9453e-02	-8.9453e-02	-1.14**38**e-01	-1.14**38**e-01
-2.5798e-03	-2.5798e-03	-7.04**85**e-03	-7.04**85**e-03
-3.2033e-02	-3.2033e-02	-2.4929e-02	-2.4929e-02
2.0159e-02	2.0159e-02	2.76**17**e-02	2.76**17**e-02
2.4191e-01	2.4191e-01	2.662**0**e-01	2.662**0**e-01
6.8819e-03	6.8820e-03	3.51**99**e-02	3.51**99**e-02
4.0311e-02	**4.0311e-02**	**5.6342e-02**	**5.6342e-02**
-1.4482e-02	-1.4482e-02	-1.2**595**e-02	-1.2**595**e-02
3.7213e-02	3.7213e-02	3.752**7**e-02	3.752**7**e-02
-8.924**6**e-03	-8.924**6**e-03	-4.**49**4**8**e-02	-4.**49**4**8**e-02
1.4087e+00	1.4087e+00	2.1**069**e+00	2.1**069**e+00
-1.0175e+00	-1.0175e+00	-2.4**451**e+00	-2.4**451**e+00
-3.2982e-01	-3.2982e-01	-1.0**552**e-01	-1.0**552**e-01
5.7126e-08	**7.8108e-16**	**2.6027e-13**	**9.1628e-28**
1.3304e-01	**1.3304e-01**	**1.6186e-01**	**1.6186e-01**
1.4384e-08	**-2.0000e-16**	**4.6263e-14**	**-5.3458e-28**

Conclusions

1. The **emlmpr** algorithm solves linear regression problem with *dozens* of unknown parameters and *hundreds* of observations within 3–6 seconds

2. L_1-regularization is an effective tool for finding parameters of linear regression model if its features are linearly dependent

Acknowledgements

- National Research Foundation of Ukraine (grant №2021.01/0136 and grant №2023.04/0094)
- Volkswagen Foundation grant №97 775
- Project of Research Works of Young Scientists №07-02/03-2023
- NASU grant for research laboratories/groups of young scientists №02/01-2024(5)
- DTT TS KNU NASU grant №2M-2024 (0124U002162)

Questions?

Thank you for your attention

vik.stovba@gmail.com