Практическая эконометрика. Лекция 7. Большая размерность

авторы: Георгий Калашнов, Ольга Сучкова,
преподаватели 2021: Ольга Сучкова, Алексей Замниус,
Анна Ставнийчук

20 октября 2021 г.

План на сегодня

Ridge Regression, LASSO

Надо выбрать λ. Это можно делать с помощью валидационной выборки. Оптимальная $\lambda = \sigma^2/\theta^2$

$$||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2 \to \min_{\beta}$$

$$\hat{\beta} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

или

$$||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1 \to \min_{\beta}$$

- вопросы большого количества параметров (и мультиколлинеарность)
- очень много объясняющих переменных, из которых мы не можем выбрать иначе
-

Почему сжимающие оценки помогают?

Они сужают множество моделей в котором мы ищем

- Это будто мы себе добавили λ бесплатных бесшумных, но смещенных наблюдений $y = 0, X = 1$
- Байесовский взгляд: мы верим в простые модели больше, чем в сложные и придаем им больший вес (априорное распределение) $\beta \sim N(0, \frac{\sigma^2}{\lambda})$.

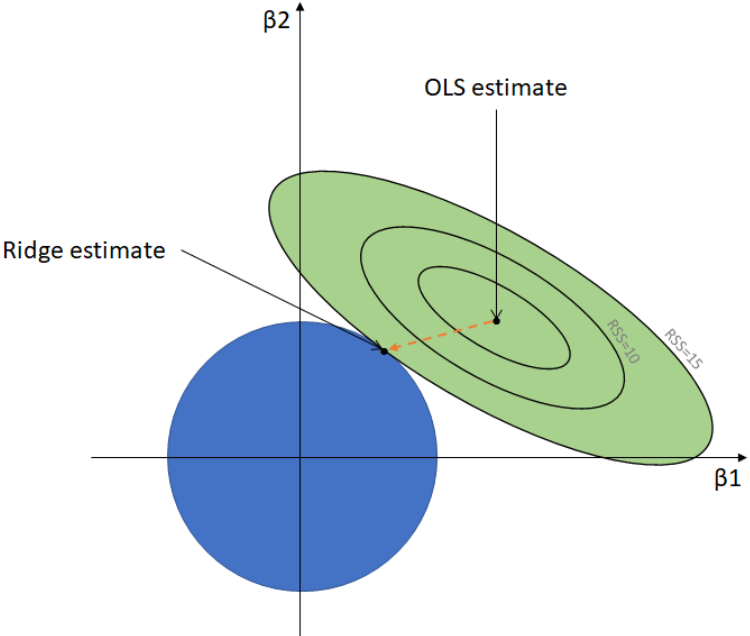
$$P(\beta|X, y) \propto P(X, y|\beta)P(\beta)$$

Или с логарифмами:

$$\text{Log-Likelihood} + \log(P(\beta))$$

- Оптимизационный взгляд: мы накладываем ограничение на β, чтобы они были не очень большими (λ – множитель лагранжа)

Графически

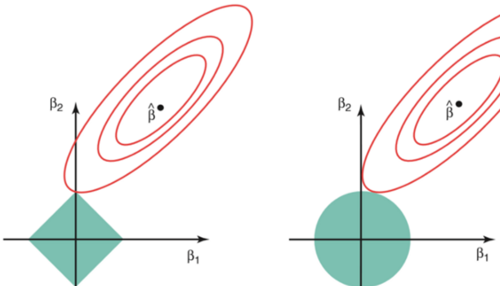


Lasso

Что если мы особенно верим в модели, в которых почти все переменные равны 0. Давайте дадим им еще больше веса

Пример: мы исследуем дискриминацию. Какая разница между фамилиями Кузнецов и Сидоров? Может, занулить все частые «обычные» фамилии?

$$||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1 \to \min_\beta$$



Регуляризация

Где ещё применяется?

Регуляризация

Как выбрать λ?

Double Lasso, вариант double selection

- Выбрать набор переменных с помощью Lasso $Y \sim X$
- Выбрать набор переменных с помощью Lasso о $T \sim X$
- Использовать в финальной регрессии только эти переменные

Double LASSO (Belloni et al., 2014)

Модель:

- $Y_1, \ Y_0$ – потенциальные исходы
- T – 1, если наблюдение в treatment и 0 иначе (**treatment variable**)
- X – Независимые переменные
- Пусть X имеет большую размерность

Мы хотим средний эффект воздействия (**average treatment effect**): ATE $= \mathbb{E}\tau$

$$\frac{1}{N_1}\sum Y_1 - \frac{1}{N_0}\sum Y_0 \xrightarrow{p} \mathbb{E}\tau$$

Double LASSO (Belloni et al., 2014)

Модель:

- Шаг 1. Лассо-регрессия $_i$, на все X.
- Шаг 2. Лассо-регрессия Y_i, на все X.
- Шас 3. Линейная регрессия Y_i, на $_i$ и все отобранные X с ненулевыми коэффициентами на шаге 1 и 2.

Если $_i$ – рандомизированный в ходе эксперимента, то ... ?

Double LASSO

Partialling-out - второй способ Double LASSO

Влияние родительства на удовлетворённость жизнью (Bhargava et al, 2014):

Table 1. Parents and Well-Being: Replication and Reanalysis of Nelson, Kushlev, English, Dunn, and Lyubomirsky (2013)

Model and predictor	Study 1 (World Values Survey)			Study 2 (Data from Carstensen et al., 2011)			
	Satisfaction	Happiness	Thoughts about meaning in life	Happiness	Positive emotion	Depression	Thoughts about meaning in life
Main effect of parenthood on well-being from Study 1 and Study 2							
Original analysis (Nelson et al., 2013, pp. 4, 7)							
Parenthood	Positive effect, $p < .001$, $N = 6{,}846$	Positive effect, $p = .004$, $N = 6{,}793$	Positive effect, $p < .001$, $N = 6{,}807$	Positive effect, $p = .008$, $N = 327$	Positive effect, $p < .001$, $N = 329$	Negative effect, $p = .003$, $N = 239$	Positive effect, $p = .01$, $N = 178$
Replication with no controls							
Parenthood	$\hat{\beta} = 0.224$, $p < .001$, $N = 6{,}846$	$\hat{\beta} = 0.050$, $p = .004$, $N = 6{,}793$	$\hat{\beta} = 0.081$, $p < .001$, $N = 6{,}807$	$\hat{\beta} = 0.271$, $p = .02$, $N = 328$	$\hat{\beta} = 0.502$, $p < .001$, $N = 329$	$\hat{\beta} = -2.135$, $p = .05$, $N = 252$	$\hat{\beta} = 0.433$, $p = .02$, $N = 178$
Replication controlling for marital status							
Parenthood	$\hat{\beta} = -0.052$, $p = .36$	$\hat{\beta} = -0.045$, $p = .02$	$\hat{\beta} = 0.100$, $p < .001$	$\hat{\beta} = 0.273$, $p = .03$	$\hat{\beta} = 0.473$, $p = .001$	$\hat{\beta} = -1.826$, $p = .14$	$\hat{\beta} = 0.420$, $p = .04$
Replication controlling for marital status and age							
Parenthood	$\hat{\beta} = -0.138$, $p = .04$	$\hat{\beta} = -0.042$, $p = .04$	$\hat{\beta} = 0.096$, $p = .03$	$\hat{\beta} = 0.061$, $p = .68$	$\hat{\beta} = 0.100$, $p = .52$	$\hat{\beta} = -0.888$, $p = .57$	$\hat{\beta} = 0.186$, $p = .43$
Replication controlling for marital status, age, and gender							
Parenthood	$\hat{\beta} = -.144$, $p = .03$	$\hat{\beta} = -.044$, $p = .04$	$\hat{\beta} = .073$, $p < .01$	$\hat{\beta} = 0.038$, $p = .80$	$\hat{\beta} = 0.134$, $p = .40$	$\hat{\beta} = -1.142$, $p = .47$	$\hat{\beta} = 0.135$, $p = .54$
Replication controlling for marital status, age, gender, and income							
Parenthood	$\hat{\beta} = -0.065$, $p = .34$	$\hat{\beta} = -0.019$, $p = .38$	$\hat{\beta} = 0.067$, $p = .02$	—	—	—	—
Reanalysis of moderation from Study 1 and Study 2							
Moderation							
Parenthood × Male	$\hat{\beta} = 0.051$, $p = .62$	$\hat{\beta} = 0.053$, $p = .13$	$\hat{\beta} = 0.073$, $p = .11$	$\hat{\beta} = 0.375$, $p = .15$	$\hat{\beta} = 0.117$, $p = .65$	$\hat{\beta} = -1.851$, $p = .48$	$\hat{\beta} = 0.394$, $p = .28$
Parenthood × Marital Status	$\hat{\beta} = 0.193$, $p = .11$	$\hat{\beta} = 0.110$, $p = .006$	$\hat{\beta} = 0.033$, $p = .53$	$\hat{\beta} = -0.351$, $p = .50$	$\hat{\beta} = 0.220$, $p = .69$	$\hat{\beta} = 5.449$, $p = .03$	$\hat{\beta} = -0.345$, $p = .68$
Parenthood × Age	$\hat{\beta} = 0.062$, $p < .01$	$\hat{\beta} = 0.002$, $p = .72$	$\hat{\beta} = 0.002$, $p = .84$	$\hat{\beta} = 0.017$, $p = .65$	$\hat{\beta} = 0.059$, $p = .10$	$\hat{\beta} = -0.511$, $p = .14$	$\hat{\beta} = -0.013$, $p = .79$
Parenthood × Age^2	$\hat{\beta} = -0.0005$,	$\hat{\beta} = 0.000$,	$\hat{\beta} = 0.000$,	$\hat{\beta} = 0.000$,	$\hat{\beta} = 0.000$,	$\hat{\beta} = 0.004$,	$\hat{\beta} = 0.000$,

Влияние родительства на удовлетворённость жизнью (Urminksy, Hansen, Chernozhukov)

- У Bhargava (2014) Без контрольных переменных статус родителя положительно влияет на удовлетворённость жизнью β=-0.224, p<0.001
- С контролем на семейное положение, пол и возраст связь становится отрицательной β=-0.144, p=0.04, с добавлением доход - незначимой β=-0.144, p=0.04
- Контрольные переменные должны быть обоснованы по теории и из логики, но что делать технически? Разница в результатах Bhargava et al, 2014 может быть ложной из-за multiple testing и большого количества возможных ковариатов.

Отбор из >400 параметров по double-lasso (Urminksy, Hansen, Chernozhukov)

Regression of parenthood on life satisfaction, with double-lasso selected covariates.

Variable	β	SE	t	p	Low CI	High CI
Primary variables:						
Constant	6.750	0.128	52.57	.000	6.498	7.001
Parent	-0.196	0.071	-2.75	.006	-0.336	-0.056
Main effect covariates:						
Married (including living together as married)	0.513	0.157	3.27	.001	0.206	0.821
Income (3 point scale)	0.582	0.119	4.90	.000	0.349	0.815
Age	0.912	0.235	3.88	.000	0.451	1.373
Age=18	0.300	0.213	1.41	.159	-0.117	0.717
Age=19	0.129	0.213	0.61	.544	-0.288	0.547
Age=20	0.521	0.191	2.73	.006	0.147	0.896
Age=21	0.175	0.175	1.00	.319	-0.169	0.518
Age=22	0.545	0.169	3.23	.001	0.214	0.876
Age=23	0.187	0.171	1.09	.274	-0.148	0.523
Gender (Male)	-0.142	0.063	-2.24	.025	-0.267	-0.018
Employment: Housewife	0.066	0.123	0.54	.590	-0.175	0.308
Chief wage earner	0.144	0.070	2.07	.038	0.008	0.281
Interaction covariates:						
Married x Age	0.143	0.352	0.41	.685	-0.547	0.833
Married x Age to fourth power	0.491	0.608	0.81	.419	-0.701	1.684
Married x Income rating (3 point)	-0.106	0.181	-0.59	.557	-0.461	0.248
Married x Income rating (11 point)	0.303	0.221	1.37	.170	-0.130	0.737
Employment: Student x Male	0.344	0.269	1.28	.202	-0.184	0.871

Про вопрос Лизы с прошлой лекции

Что делать, если X влияет только на T ? (см. 4 схемы на доске)

Ориентироваться можно по Backdoor-критерию (J. Pearl)

Для упорядоченной пары переменных (T, Y) в ориентированном ациклическом графе G, набор переменных X удовлетворяет критерию backdoor относительно пары (T, Y), если ни один узел из X не является «потомком» от T, и X блокирует все цепочки между T и Y, которые содержат стрелку, входящую в узел T.

Backdoor-критерий (J. Pearl)

- Post-treatment не удовлетворяет этому критерию. Это «плохие контрольные переменные»
- Pre-treatment - под вопросом: зависит от структуры графа G
- Вывод: перед оценкой регрессий нарисуйте схему взаимодействия между показателями

Double Robustness

А еще можно сделать и то и другое (выкладки на доске)

Double Robustness

- $e(X) = P(T = 1|X)$ – по определению настоящая (postulated) мера склонности к попаданию в тритмент-группу.
- $m_1(X) = E[Y|T = 1, X]$, $m_0(X) = E[Y|T = 0, X]$ – истинная зависимость Y от X в двух группах.
- $ATE = E[Y_1 - Y_0]$ – по определению истинный эффект воздействия.
- $\widehat{ATE_{DR}} = \frac{1}{n}[\sum_{i=1}^{n}(\frac{T_i * Y_i}{e(X_i,\hat{\alpha})} - \frac{(T_i - e(X_i,\hat{\alpha})) * m_1(X_i,\hat{\beta}_1)}{e(X_i,\hat{\alpha})}] - \frac{1}{n}[\sum_{i=1}^{n}(\frac{(1-T_i) * Y_i}{(1-e(X_i,\hat{\alpha}))} + \frac{(T_i - e(X_i,\hat{\alpha})) * m_0(X_i,\hat{\beta}_0)}{(1-e(X_i,\hat{\alpha}))}]$
- $e(X, \hat{\alpha})$ – оценённая на данных вероятность попадания в тритмент-группу в зависимости от характеристик.
- $m_1(X, \hat{\beta}_1)$ и $m_0(X, \hat{\beta}_0)$– оценённые на данных зависимости Y в тритмент- и контрольной группе от характеристик.

Double Robustness

- $\widehat{ATE_{DR}} = \frac{1}{n}[\sum_{i=1}^{n}(\frac{T_i * Y_i}{e(X_i,\hat{\alpha})} - \frac{(T_i - e(X_i,\hat{\alpha})) * m_1(X_i,\hat{\beta}_1)}{e(X_i,\hat{\alpha})}] - \frac{1}{n}[\sum_{i=1}^{n}(\frac{(1-T_i) * Y_i}{(1-e(X_i,\hat{\alpha}))} + \frac{(T_i - e(X_i,\hat{\alpha})) * m_0(X_i,\hat{\beta}_0)}{(1-e(X_i,\hat{\alpha}))}]$
- Достаточно оценить правильно либо $e(X)$, либо $m_1(X) = E[Y|T = 1, X]$, $m_0(X) = E[Y|T = 0, X]$.

Литература: книжки и образовательные материалы I

Литература: статьи I