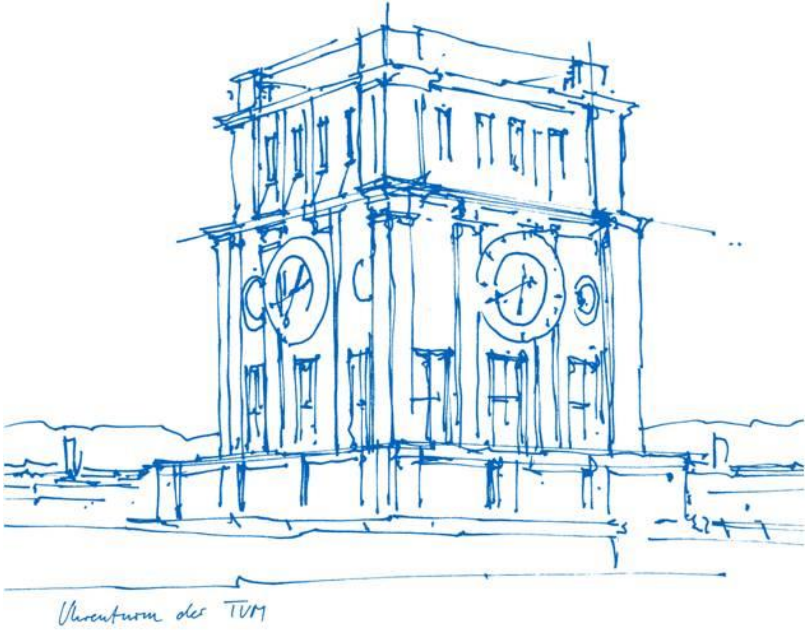
Regression Shrinkage

Jan-Niklas Dörr

Technische Universität München



Why to use Shrinkage Methods?

- Dealing with **“fat“ data**, i.e., large number of regressors p compared to the number of observations n
- Two cases:
- $\boldsymbol{p < n}$:
  - OSL/2SLS is prone to overfitting based on the comparable large number of regressors
  - Lasso and Ridge reduce the overfitting tendency (by Regularization)
  - Lasso additionally improves interpretation and helps with the construction of instruments

- $\boldsymbol{p \geq n}$:
  - OLS/2SLS not feasible (underdetermined system)
  - Ridge and Lasso do still work

Main Concept of Shrinkage Methods

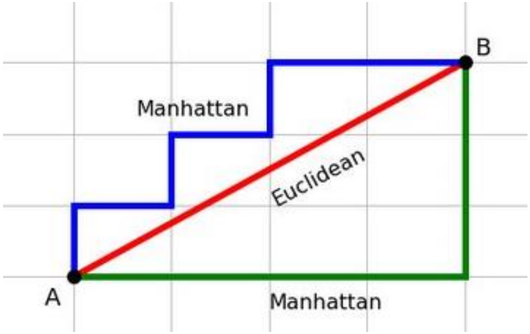
- Constrain or shrink coefficients by introducing penalties in the objective
- Normalize variables. (standardize by $z = \frac{x - \mu}{\sigma}$)
  - Centered means, i.e., $\mu = 0$
  - Unit variance, i.e., one unit is σ
  - Required for comparability of penalties
- Partial out the constant (we don't want to shrink it)

- Results of Shrinkage methods are different to OLS due to error terms (unless $\lambda = 0$)
- To compare, use Post-Lasso, i.e.:
  1. Lasso to select regressors
  2. Use preselected regressors in OLS

Excursion: Norms

- Norms are used to measure the magnitude of vectors
- Ridge and Lasso are utilizing a norm for their penalty

- Assume we have a vector $\vec{x} = [x_1, x_2, \ldots, x_n]'$, then the following norms are defined as:
  - L^0 is the zero norm $||x||_0 \coloneqq \sum_{i=1}^{n} |x_i|^0$ (not a real norm)
  - L^1 is the absolute norm $||x||_1 \coloneqq |x| = \sum_{i=1}^{n} |x_i|$, (Manhattan distance)
  - L^2 is the Euclidean norm $||x||_2 \coloneqq \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$
  - L^{infty} is the infinity norm $||x||_\infty \coloneqq \max_n |x_n|$



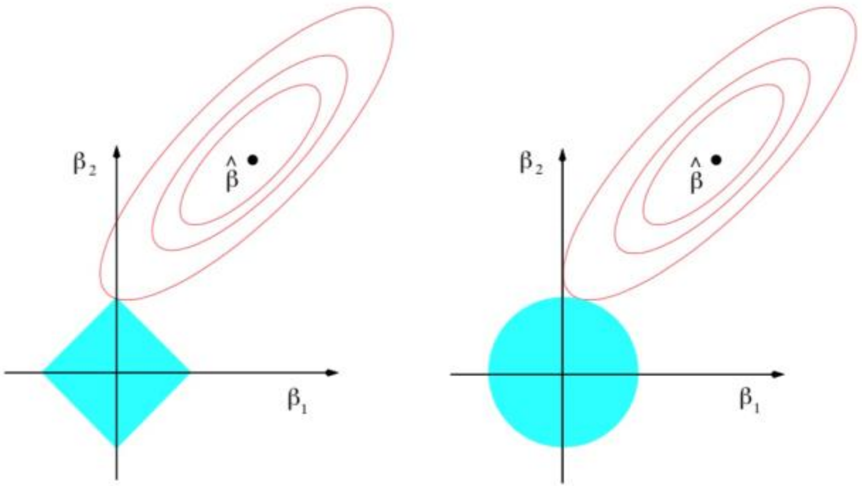
To generalize $\boldsymbol{L^p, p \geq 1}$ $\boldsymbol{||x||_p \coloneqq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}}$

Available Shrinkage Methods

- Best Subset $:= \frac{RSS}{n} + \lambda \cdot ||\beta||_0$ (not really a shrinkage method, computational expensive)
- Lasso $:= \frac{RSS}{n} + \lambda \cdot ||\beta||_1$
- Ridge $:= \frac{RSS}{n} + \lambda \cdot ||\beta||_2$

,with:

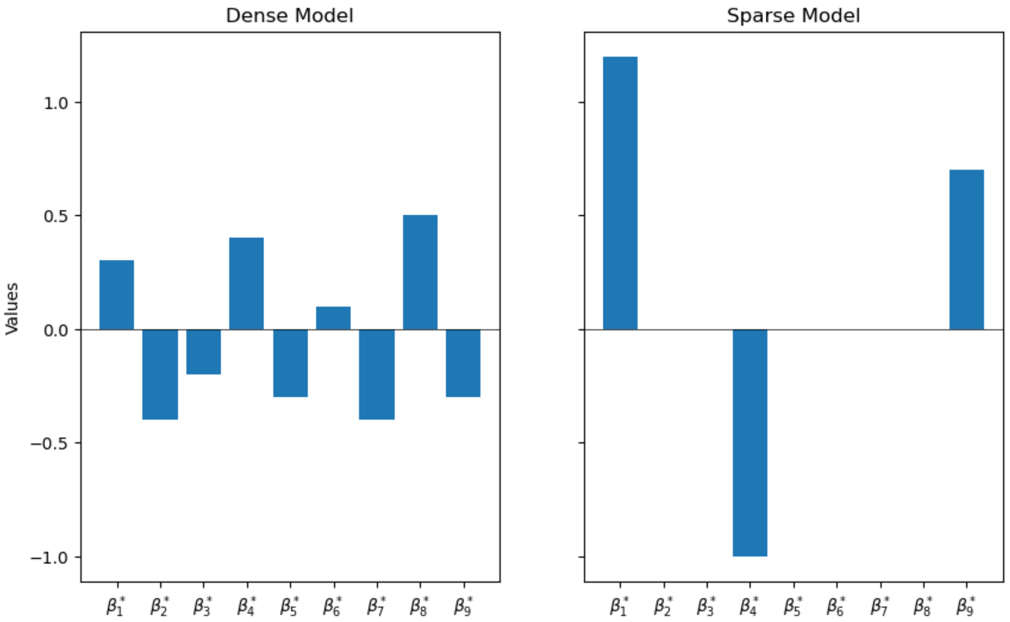
- $RSS := \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$
- and λ being a tuning (or hyper) parameter
- Find best λ by:
  - K-fold cross validation
  - Rigorous lasso



Source: Hastie et al. (2009) - The elements of statistical learning, p. 71

When to use which?

- Ridge works well in a **dense** setting, i.e., many true values $\beta_j^* \neq 0, \frac{||\beta^*||_0}{n} \gg 0$
- Lasso works well in a **sparse** setting, i.e., many true values $\beta_j^* = 0, \frac{||\beta^*||_0}{n} \to 0$ or $||\beta^*||_0 \ll \mathrm{n}$
- However, if a model is sparse or dense is not known at the beginning. Has to be estimated (by experience?)



Exercises

Exercise 1:

Assuming our whole dataset looks like the following. What is the issue with OLS? What are suitable alternatives?

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
0	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
1	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
2	104.593	7075	514	4	71	11	Male	No	No	Asian	580
3	148.924	9504	681	3	36	11	Female	No	No	Asian	964
4	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331
5	80.180	8047	569	4	77	10	Male	No	No	Caucasian	1151
6	20.996	3388	259	2	37	12	Female	No	No	African American	203
7	71.408	7114	512	2	87	9	Male	No	No	Asian	872
8	15.125	3300	266	5	66	13	Female	No	No	Caucasian	279

Exercise 2:

Luckily, the dataset contains more observations. (see separate file)

a) Perform Lasso to determine the impact of the variables on the variable ”balance”. What are the regressors with the biggest impact?
b) Tune λ by using 10-fold cross-validation. What are the implications of the tuning process?
c) Conduct Post-Lasso with $\lambda = 10$ as a penalty term for Lasso.

Additional Material

- You will find the code for the exercises here:
  https://github.com/Jan-Niklas-Doerr/Regression-Shrinkage

- If you are interested in more material, you may want to check out this GitHub repository with sessions to multiple topics:
  https://matteocourthoud.github.io/course/ml-econ/

Regression Shrinkage

Jan-Niklas Dörr

München, 25. April 2024