Regularization methods

Ajda Marjanovič

November 22, 2023

Introduction to Machine Learning

Machine learning is neither an empirical panacea nor a substitute for economic theory and the structure it lends to empirical work. In other words, finance domain knowledge remains an indispensable component of statistical learning problems in asset markets.

- Giglio et al. (2022)

Introduction to Machine Learning

Machine learning is neither an empirical panacea nor a substitute for economic theory and the structure it lends to empirical work. In other words, finance domain knowledge remains an indispensable component of statistical learning problems in asset markets.

- Giglio et al. (2022)

- book: Introduction to Statistical Learning with R / Python (James et al., 2023) - link
- a nice introduction to many ML topics, offers a good intuition without complex math, has coding labs
- some of these slides are based on the material from this book

Linear models

- recall the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- usually estimated using ordinary least squares fitting
- why consider alternatives to OLS?
  - **Prediction Accuracy**: when there are a lot of predictors $p > n$ (especially when predictors are highly correlated)
  - **Model Interpretability**: By removing irrelevant features — by setting the corresponding coefficient estimates to 0 — we can obtain a model that is more easily interpreted

Example: Asset Pricing

Feng et al. (2020). Taming the Factor Zoo: A Test of New Factors. *Journal of Finance*.

> THE SEARCH FOR FACTORS THAT explain the cross section of expected stock returns has produced hundreds of potential candidates, as noted by Cochrane (2011) and more recently by Harvey, Liu, and Zhu (2015), McLean and Pontiff (2016), and Hou, Xue, and Zhang (2017). A fundamental task facing the asset pricing field today is to bring more discipline to the proliferation of factors. In particular, a question that remains open is: how to judge whether a new factor adds explanatory power for asset pricing, relative to the hundreds of factors the literature has so far produced?

Shrinkage Methods

- ridge regression and lasso (elastic net as a combination of these two)
- we fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero

Ridge regression

- LS fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using values that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- fit a model containing all predictors with a technique that constrains (regularizes) the coefficient estimates or shrinks them towards zero
- ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a *tuning* parameter, to be determined separately

Ridge regression

- as with least squares, ridge regression seeks coefficient estimates that fit the data well, by minimizing the RSS
- the second term $\lambda \sum_j \beta_j^2$ called a shrinkage penalty, is small when $\beta_1, \ldots, \beta_p$ are close to zero, so it has the effect of shrinking the estimates of β_j towards zero
- the tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates
- selecting a good value for λ is critical (usually cross-validation is used)

Cross validation

Figure 1: k-fold CV

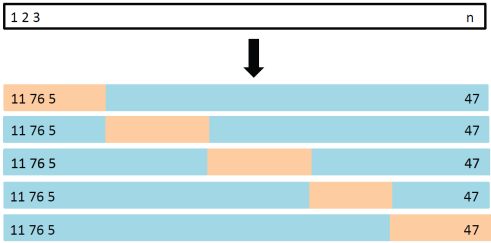
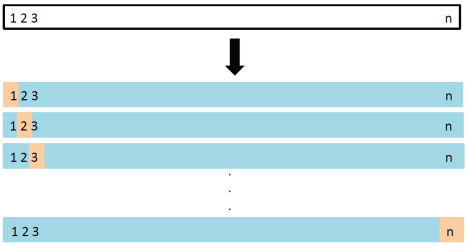


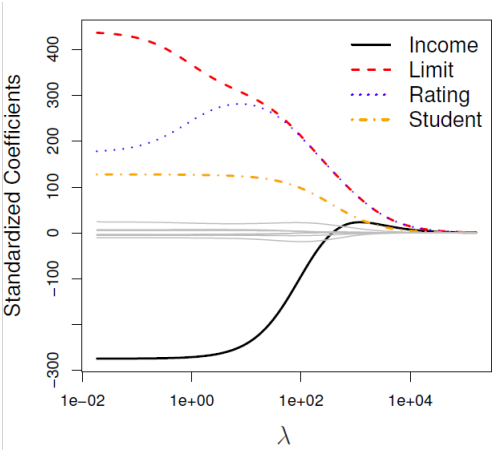
Figure 2: leave-one-out CV



Split the sample in the training (blue) and validation (orange) set on which MSE is computed. Repeat k-times so each group acts as a validation set once. Compute the average MSE over the k iterations.

Ridge regression

Example on a credit dataset:



- each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ
- the larger the λ (the more we are strengthening the regularization effect), the more coefficients are shrinked towards 0

LASSO

Ridge regression

- ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function
- solution: apply ridge regression after normalizing the predictors: $\tilde{x} = \frac{x - \mu_x}{\sigma_x^2}$

Lasso regression

- ridge regression does have one obvious disadvantage: it will include all p predictors in the final model
- the Lasso is an alternative to ridge regression that overcomes this disadvantage. The lasso coefficients $\hat{\beta}_\lambda^L$ minimize the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

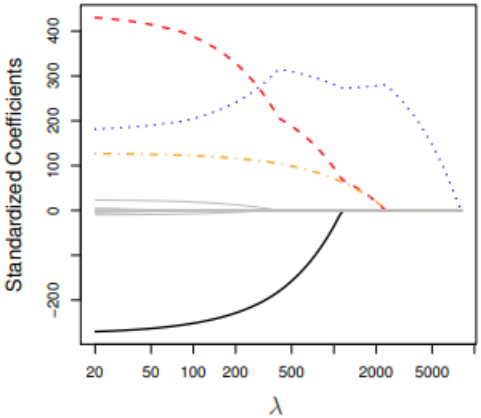
- lasso uses an ℓ_1 penalty instead of an ℓ_2 penalty. The ℓ_1 norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$. Instead the ℓ_2 norm of a coefficient vector β is given by $\|\beta\|_2 = \sqrt{\sum\beta_j^2}$ (sometimes simplified to $\sum\beta_j^2$)

Lasso regression

- as with ridge regression, the lasso shrinks the coefficient estimates towards zero
- in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large $\rightarrow$ variable selection
- we say that the lasso yields sparse models — that is, models that involve only a subset of the variables
- as in ridge regression, selecting a good value of λ for the lasso is critical (cross-validation)

Lasso regression

Example on a credit dataset:



- each curve corresponds to the lasso regression coefficient estimate for one of the ten variables, plotted as a function of λ
- the larger the λ (the more we are strengthening the regularization effect), the more coefficients are set to 0

RIDGE

Ridge vs Lasso

- in general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors
- however, the number of predictors that is related to the response is never known a priori for real data sets

Selecting the tuning parameter λ

- we require a method for selecting a value for the tuning parameter λ
- cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ
- we then select the tuning parameter value for which the cross-validation error is smallest
- finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter

Application example
Data

- we will be using a (subset) of the data from the following paper:

  Chen et al. (2024). Deep Learning in Asset Pricing. *Management Science*.
- sample period: 1/1/2010 - 1/12/2016, monthly frequency
- data on returns and 46 firm characteristics for many firms
- which characteristics (factors or features) are better at explaining returns?

Features

	Past returns			Value	
(1)	r2_1	Short-term momentum	(26)	A2ME	Assets to market cap
(2)	r12_2	Momentum	(27)	BEME	Book to market ratio
(3)	r12_7	Intermediate momentum	(28)	C	Ratio of cash and short-term investments to total assets
(4)	r36_13	Long-term momentum	(29)	CF	Free cash flow to book value
(5)	ST_Rev	Short-term reversal	(30)	CF2P	Cashflow to price
(6)	LT_Rev	Long-term reversal	(31)	D2P	Dividend yield
			(32)	E2P	Earnings to price
	Investment		(33)	Q	Tobin's Q
(7)	Investment	Investment	(34)	S2P	Sales to price
(8)	NOA	Net operating assets	(35)	Lev	Leverage
(9)	DPI2A	Change in property, plants, and equipment			
(10)	NI	Net share issues		Trading frictions	
			(36)	AT	Total assets
	Profitability		(37)	Beta	CAPM beta
(11)	PROF	Profitability	(38)	IdioVol	Idiosyncratic volatility
(12)	ATO	Net sales over lagged net operating assets	(39)	LME	Size
(13)	CTO	Capital turnover	(40)	LTurnover	Turnover
(14)	FC2Y	Fixed costs to sales	(41)	MktBeta	Market Beta
(15)	OP	Operating profitability	(42)	Rel2High	Closeness to past year high
(16)	PM	Profit margin	(43)	Resid_Var	Residual variance
(17)	RNA	Return on net operating assets	(44)	Spread	Bid-ask spread
(18)	ROA	Return on assets	(45)	SUV	Standard unexplained volume
(19)	ROE	Return on equity	(46)	Variance	Variance
(20)	SGA2S	Selling, general and administrative expenses to sales			
(21)	D2A	Capital intensity			
	Intangibles				
(22)	AC	Accrual			
(23)	OA	Operating accruals			
(24)	OL	Operating leverage			
(25)	PCM	Price to cost margin			

References

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14(1):337–368.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.