Prediction, Model Selection and Causal Inference with Regularized Regression

Methods and Applications of Machine Learning

Elisa Tosetti

2023-2024

Outline of the lesson

- Model selection in high dimensional models $\longrightarrow$ regularization
- Regularization and causal inference
- Some examples

Resources:

- Hastie et al, Section 3.4
- Belloni A., Chernozhukov V., and Hansen C. (2014), *High-Dimensional Methods and Inference on Structural and Treatment Effects*, Journal of Economic Perspectives, 28, 29-50

Model selection in high dimensional models

Consider the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

where y_i are observations of the response variable, $x_{i1}, ..., x_{ik}$ are observations on k regressors, and u_i are zero-mean disturbances

It often happens that there are a large number of regressors k, possibly much larger than the sample size n

If k is close to or even larger than n, we say that data are **high-dimensional**:

- If $k > n$, the model is not identified
- If $p = n$, perfect fit
- If $p < n$ but large, overfitting is likely: some of the predictors are only significant by chance (false positives), but perform poorly on new (unseen) data

Model selection in high dimensional models

Two very common problems in applied work in economics:

- Selecting controls to address omitted variable bias when many potential controls are available
- Selecting instruments when many potential instruments are available

Model selection in high dimensional models

The standard approach for model selection in economics is hypothesis testing:

- Researchers try many combinations of regressors, looking for statistical significance
- "it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance' and to then report only what 'worked' " (Simmons et al., 2011)
- This often leads to p-hacking...

Regularization for linear regression under sparsity

Estimation of

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

with very large k becomes manageable if we assume **sparsity**

Sparsity: only a relatively small number (s) of these regressors are important for capturing accurately the main features of the regression function:

$$s : \sum_{j=1}^{k} 1\{\beta_j \neq 0\} << n$$

- In other words: most of the true coefficients β_j are actually zero. But we don't know which ones are zeros and which ones aren't
- We can also use the weaker assumption of **approximate sparsity**: some of the β_j coefficients are well-approximated by zero, and the approximation error is sufficiently 'small'

Regularization for linear regression

Rather than OLS

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$$

Fit the constrained problem:

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 + \lambda||\beta||$$

where where $\lambda \geq 0$ is a tuning parameter, to be determined separately and $||.||$ is usually assumed to be:

- 2-norm (L2): $||\beta||_2^2 = \sum_{j=1}^{k} \beta_j^2 \longrightarrow$ **Ridge regression**
- 1-norm (L1): $||\beta||_1 = \sum_{j=1}^{k} |\beta_j| \longrightarrow$ **Lasso regression**

Important message of regularized regression: *There's a cost to including lots of regressors, and we can reduce the objective function by throwing out the ones that contribute little to the fit*

Regularization for linear regression

- The second term, $\lambda||\beta||$, called a **shrinkage penalty**, is small when $\beta_1, ..., \beta_k$ are close to zero, $\rightarrow$ it has the effect of shrinking the estimates of β_j towards zero
- The tuning parameter λ controls the relative impact of the term $||\beta||$ on the regression coefficient estimates
- When $\lambda = 0$ the penalty term has no effect, and the regularized regression will produce the least squares estimates
- As $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the regularized regression coefficient estimates will approach zero

Ridge regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_j)^2 + \lambda \sum_{j=1}^{k}\beta_j^2$$

Note that the shrinkage penalty is applied to $\beta_1, \beta_2, ..., \beta_k$, but not to the intercept β_0

As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias

Ridge regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij})^2 \ s.t. \ \sum_{j=1}^{k}\beta_j^2 < \tau$$

- Blue circle is the constraint region $\beta_1^2 + \beta_2^2 < \tau$ while red lines are RSS contour lines
- β_0 is the OLS estimate, while β_R is the Ridge estimate
- $\beta_{1,R} \neq 0$ and $\beta_{2,R} \neq 0$ implying that both regressors are included

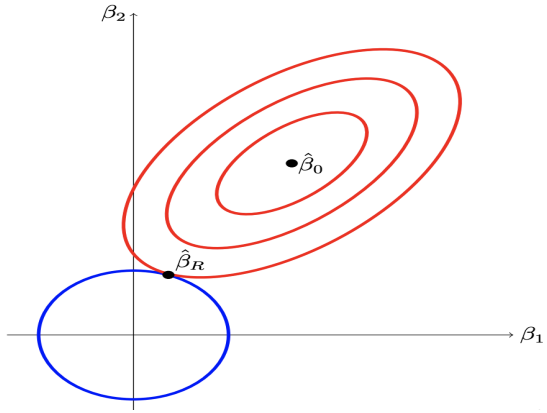


Figure 1: Example: the case k=2

Ridge regression

Disadvantages of the ridge regression:

- The penalty $\lambda \sum_{j=1}^{k} \beta_j^2$ will shrink all coefficients towards zero, but it will not set any of them exactly to zero
- This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables k is quite large

Lasso regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_j)^2 + \lambda \sum_{j=1}^{k}|\beta_j|$$

- The L1 penalty here has the effect of forcing some coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large
- Hence the Lasso performs variable selection
- We say that the Lasso yields **sparse models** — that is, **models that involve only a subset of the variables**

The Lasso regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij})^2 \ s.t. \ \sum_{j=1}^{k}|\beta_j| < \tau$$

- Blue diamond is the constraint region $\beta_1^2 + \beta_2^2 < \tau$, while red lines are RSS contour lines
- β_0 is the OLS estimate, while β_L is the LASSO estimate
- $\beta_{1,L} = 0$ implying that the LASSO omits regressor 1 from the model

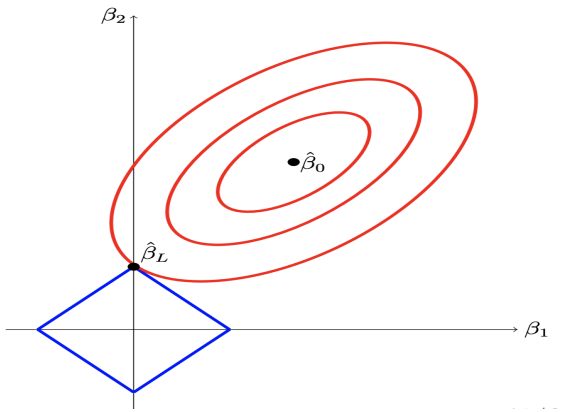


Figure 2: Example: the case k=2

An application to the Credit data set

Consider the following data set on 400 clients from a bank, taken from the *ISLR* package in *R*

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.593	7075	514	4	71	11	Male	No	No	Asian	580
148.924	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

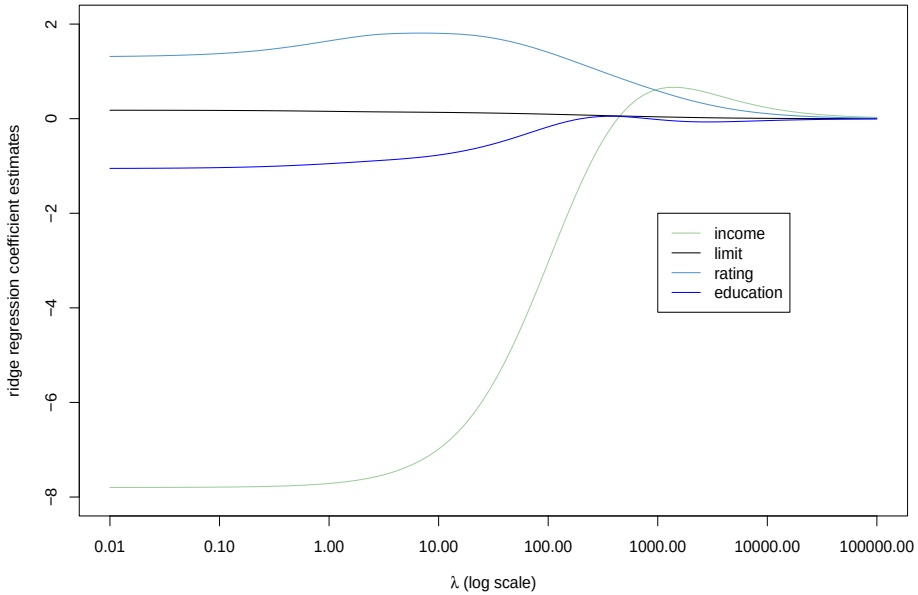
An application to the Credit data set

Simple OLS regression of Balance on all other variables

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
## ------------------------------------------------
## Income                     -7.803*** (0.234)
## Limit                      0.191*** (0.033)
## Rating                      1.137** (0.491)
## Cards                      17.724*** (4.341)
## Age                        -0.614** (0.294)
## Education                   -1.099 (1.598)
## GenderMale                  10.653 (9.914)
## StudentYes                 425.747*** (16.723)
## MarriedYes                  -8.534 (10.363)
## EthnicityAsian              16.804 (14.119)
## EthnicityCaucasian          10.107 (12.210)
## Constant                  -489.861*** (35.801)
## ------------------------------------------------
## Observations                    400
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

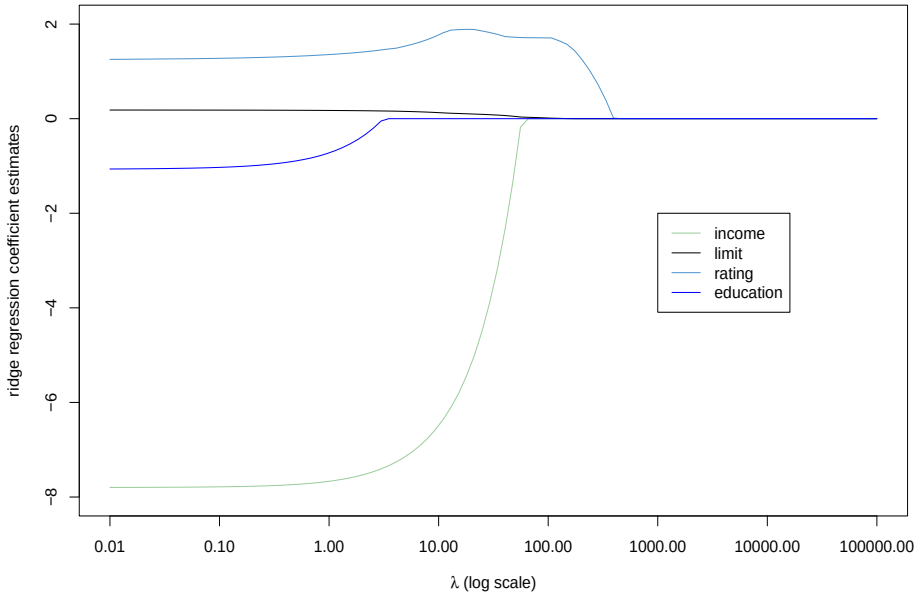
An application to the Credit data set

Plot of $\hat{\beta}_{income}$, $\hat{\beta}_{limit}$, $\hat{\beta}_{rating}$ and $\hat{\beta}_{education}$ for λ varying between 0.01 and 10,0000 in *ridge regression*:



An application to the Credit data set

Plot of $\hat{\beta}_{income}$, $\hat{\beta}_{limit}$, $\hat{\beta}_{rating}$ and $\hat{\beta}_{education}$ for λ varying between 0.01 and 10,0000 in *lasso regression*:



Regularization for causal inference: the Post Lasso estimator

The main strength of the Lasso is prediction (rather than model selection). But the Lasso's strength as a prediction technique can also be used to aid causal inference:

- Regularization by the L1-norm naturally helps the Lasso estimator to avoid overfitting, but it also shrinks the fitted coefficients towards zero, causing a potentially significant bias
- In order to remove some of this bias, consider the **Post-Lasso estimator** that applies ordinary least squares to the model selected by Lasso

The Post-Lasso estimator

β_{Post} is the solution to the following minimization problem:

$$\min_{\beta_0,..,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_j)^2 \text{ s.t. } \beta_j = 0 \text{ if } \tilde{\beta}_j = 0$$

where $\tilde{\beta}_j$ is a sparse first-step estimator such as the Lasso

Thus, post-estimation OLS treats the first-step estimator as a genuine model selection technique

See *Belloni and Chernozhukov (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli 19(2): 521–547*

Choosing the tuning parameter

Implementing the Ridge and Lasso regressions requires a method for selecting a value for the tuning parameter λ

Three approaches are available:

1. Cross validation (CV)
2. AIC/BIC information criteria
3. Theory grounded choice of λ

Choosing the tuning parameter by CV

In 'K-fold' CV, the data set is split into K portions or 'folds'; each fold is used once as the validation sample and the remainder are used to fit the model for some value of λ:

1. Define a grid of λ values
2. For each λ calculate the validation MSE within each fold and the overall cross-validation error
3. Select the value λ for which the overall cross-validation error is smallest

Finally, the model is re-fit using all the available observations and the selected value for λ

This is a data driven approach that maximizes out-of-sample prediction

An application to the Credit data set

Optimal λ using CV with 10-fold in Lasso estimation is 0.9376

....and Lasso estimation using the optimal λ is:

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## Income               -7.6804578
## Limit                 0.2027930
## Rating                0.9290433
## Cards                18.1835706
## Age                  -0.5933681
## Education            -0.8531460
## GenderMale            8.2851084
## StudentYes          423.2997215
## MarriedYes           -4.9660415
## EthnicityCaucasian    .
```

An application to the Credit data set

Post-Lasso results are:

```
##
## ===========================================
##                   Dependent variable:
##               ----------------------------
## -------------------------------------------
## Income               -7.802*** (0.234)
## Limit                0.193*** (0.033)
## Rating                1.102** (0.489)
## Cards                17.923*** (4.332)
## Age                  -0.635** (0.293)
## Education             -1.115 (1.596)
## GenderMale            10.407 (9.904)
## StudentYes          426.469*** (16.678)
## MarriedYes            -7.019 (10.278)
## Constant           -478.810*** (34.343)
## -------------------------------------------
## Observations                400
```

The Rigorous Lasso

- Cross-validation can help selecting λ for prediction but it is not theory grounded and leads to overfitting bias $\longrightarrow$ CV does not guarantee any Post-Lasso properties
- When the final goal is **inference** we also care about the efficiency of the estimator in the Post-Lasso regression $\longrightarrow$ what is the right choice of λ in this case?
- Belloni et al (2012) propose the **Rigorous Lasso** which uses feasible estimation of theoretically-grounded optimal λ under heteroskedastic and non-Gaussian errors

The Rigorous Lasso

The theory of the 'rigorous' LASSO has two main ingredients:

1. Restricted eigenvalue condition (REC): OLS requires full rank condition, which is too strong in the high-dimensional context. REC is much weaker
2. Penalization level: We need λ to be large enough to 'control' the noise in the data. At the same time, we want the penalty to be as small as possible (due to shrinkage bias)

$$\min_{\beta_0,..,\beta_k} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_j)^2 + \frac{\lambda}{n}\sum_{j=1}^{k}\psi_j|\beta_j|$$

The penalty loadings are chosen to insure basic equivariance of coefficient estimates to rescaling of x_j and can also be chosen to address heteroskedasticity in model errors

The Rigorous Lasso

$$\text{under homoskedasticity: } \lambda = 2c\sqrt{n}\hat{\sigma}\Phi^{-1}\left(1-\frac{\gamma}{2k}\right), \psi_j = \sqrt{\frac{1}{n}\sum_i x_{ij}^2}$$

$$\text{under heteroskedasticity: } \lambda = 2c\sqrt{n}\Phi^{-1}\left(1-\frac{\gamma}{2k}\right), \psi_j = \sqrt{\frac{1}{n}\sum_i x_{ij}^2 \hat{u}_i^2}$$

where

- c is a slack parameter that in most applications is set to $c = 1.1$
- $\hat{\sigma}$ is an estimate of the standard deviation of u
- Φ denotes the cumulative standard normal distribution
- γ is the probability level, which is set to $\gamma = 0.1$ by default
- n is the number of observations and k is the number of predictors

The Rigorous Lasso

The search for optimal λ starts with some:

- First initial guess of $\hat{\sigma}$
- Which provides new estimates
- Which provides new residuals
- Which gives new updated guess for $\hat{\sigma}$
- We continue iterating until convergence, i.e., until the new guess for $\hat{\sigma}$ doesn't change from one iteration to another

The iterative approach is automatically performed by the *rlasso()* function in R

The Rigorous Lasso

Under the above choice of λ, Belloni et al (2012) are able to provide an asymptotic bound for the prediction error and for the bias in estimating the target parameters β

The Rigorous approach places a **high priority on controlling overfitting**, thus often producing parsimonious models

This strong focus on containing overfitting is of theoretical benefit for selecting control variables or instruments, but also implies that the approach may be outperformed by cross-validation techniques for pure prediction tasks

Which approach is most appropriate depends on the type of data at hand and the purpose of the analysis

An application to the Credit data set: Post-lasso estimation, CV vs Rigorous Lasso

```
##
##                     (1)                    (2)
## Income       -7.802*** (0.234)     -7.795*** (0.233)
## Limit         0.193*** (0.033)      0.194*** (0.032)
## Rating         1.102** (0.489)       1.091** (0.485)
## Cards        17.923*** (4.332)     18.212*** (4.319)
## Age           -0.635** (0.293)      -0.624** (0.292)
## Education      -1.115 (1.596)
## GenderMale     10.407 (9.904)
## StudentYes  426.469*** (16.678)  425.610*** (16.510)
## MarriedYes    -7.019 (10.278)
## Constant   -478.810*** (34.343) -493.734*** (24.825)
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Regularization for causal inference

Basic setup: we know which is the causal variable of interest. No variable selection needed for this. But the Lasso can be used to select other variables or instruments included in the estimation

Consider the following **approximate sparse model**

A partial linear control function model specifies

$$y_i = \beta_0 + \tau D_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

where k can be very large

D_i is our policy or treatment variable of interest

Interest lies in estimating τ (our ATE)

$x_{i1}, x_{i2}, ..., x_{ik}$ are control variables **not directly of interest**

Choosing controls: Post-Double-Selection Lasso

We want to obtain an estimate of the parameter β_D

The problem are the controls: we want to include controls because we are worried about omitted variable bias, but which ones do we use?

- If we use too many, we run into a version of the overfitting problem. We could even have $k > n$, so using them all is just impossible
- If we use too few, or use the wrong ones, then OLS gives us a biased estimate of β_D because of omitted variable bias

Researchers may consciously or unconsciously choose controls to generate the results they want

Which controls do we use?

Naive approach: estimate the model using the Lasso (imposing that D_i is not subject to selection), and use the controls selected by the LASSO

- Badly biased. Reason: we might miss controls that have a strong predictive power for D_i, but only small effect on y_i
- Similarly, if we only consider the regression of D_i against the controls, we might miss controls that have a strong predictive power for y_i, but only a moderately sized effect on D_i

Post-Double-Selection (PDS) LASSO

- **Step 1:** Use the LASSO to estimate

$$y_i = \beta_0\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

i.e., without D_i as a regressor. Denote the set of LASSO-selected controls by A

- **Step 2:** Use the LASSO to estimate

$$D_i = \gamma_1 x_{i1} + \gamma_2 x_{i2} + ... + \gamma_k x_{ik} + \epsilon_i$$

i.e., where the causal variable of interest is the dependent variable. Denote the set of LASSO-selected controls by B

- **Step 3:** Estimate using OLS

$$y_i = \beta_D D_i + \beta' \mathbf{w}_i + v_i$$

where $\mathbf{w}_i = A \cup B$, i.e., the union of the selected controls from Steps 1 and 2

See *Alexandre Belloni, Victor Chernozhukov, and Christian Hansen 2014, Inference on treatment effects after selection among high-dimensional controls. Review of Economic Studies, 81: 608–650*

An application to the Credit data set: using the Double Lasso

Suppose we take $D = Student$. Our double selection approach leads to the following results for β_D:

```
## [1] "Estimates and significance testing of the effect of target v
##             Estimate. Std. Error t value Pr(>|t|)
## StudentYes    422.67      15.72   26.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```