Linear Regression IV: Penalized Regression Models

Paul Goldsmith-Pinkham

February 25, 2021

Today's topic - penalized regression, e.g. Lasso

- Today: Machine learning methods of a particular kind
- Specifically, we will be focus on linear models that use penalization to select relevant right hand side variables
- Will mainly focus on Lasso (Least Absolute Shrinkage and Selection Operator), coined by Tibshiriani in 1996
- Key concept underlying these methods – *model selection*
  - This is typically not a topic great for causal inference
- Ends up being potentially very valuable!

What's the big idea?

- There are many circumstances when we have a problem like the following:
  1. Many variables (too many) that we would like to use as regressors
  2. A unknown and potentially complicated function of many variables

- Two simple versions of this, in a setting where we have data (Y_i, X_i), and the dimension of $X_i = p$
  1. $Y_i = X_{i,0}\beta_0 + \epsilon_i$, where $X_{i,0}$ is a vector of $p_0 \leq p$ covariates, but unfortunately you don't necessarily know which variables in your data X_i (which have dimension p) are $X_{i,0}$. Alternatively, can write as $Y_i = X_i\beta + \epsilon_i$, where $\beta_k = 0$ for a subset of variables
  2. $Y_i = f(X_i) + \epsilon_i$,, and we want to approximate f as best we can – we can do complicated functions of X_i to approximate it (a la semiparametrics) but this gets hard when p grows.

- A key idea which will come up in our later results is *sparsity* – e.g. p_0 is small, or for f that is can be approximated by a small number of variables (combinations of X

What is Lasso? Tibshiriani (1996)

- What is Lasso? Let's stay in our simple linear model , and ignore issues of endogeneity
- Recall that the "true" model (which we prespecified) had only a subset of non-zero entries
  - E.g., there are irrelevant right hand side variables
  - We would like to know which are the true right ones for purposes of interpretation
- As our dataset grows, if *p* stays fixed, OLS will eventually figure out which β are zeros
  - But it's not immediate – only in the limit do the ("wrong") estimates converge to zero!
  - Worse yet, if the variables are correlated or noisy, it's hard to get good estimates that don't make the model poorly fit!
- In finite samples, we would like to have an approach that selects the "right" variables to focus on , and fits the outcome well
  - This is a model selection problem
  - It's *also* a regularization problem

What is Lasso? Tibshiriani (1996)

- Tibshiriani (1996) proposed Lasso to do both things – identify the non-zero covariates and shrink estimates accordingly

- Let's compare OLS and Lasso's objective functions:

$$\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \qquad \text{(OLS)}$$

$$\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \qquad \text{(Lasso)}$$

- In essence, Lasso added one "thresholding" penalty, where λ is a tuning parameter chosen by the researcher.

- Since we don't want things to get big, Lasso will choose to push coefficient values down
  - Most importantly, due to the L_1 norm, this will tend to push coefficients to *zero*
  - Why? Intuitively, if it was worth decreasing β_k slightly, it will continue to be worthwhile until it hits zero

What is Lasso? Graphically

- Note that the first term in the minimization

$$n^{-1} \sum_{i=1}^{n} (Y_i - X_i\beta)^2$$

is equivalent to

$$(\beta - \beta_{ols})' X' X (\beta - \beta_{ols})$$

- Hence given a λ constraint we're finding the isoquant closest to β_{ols}

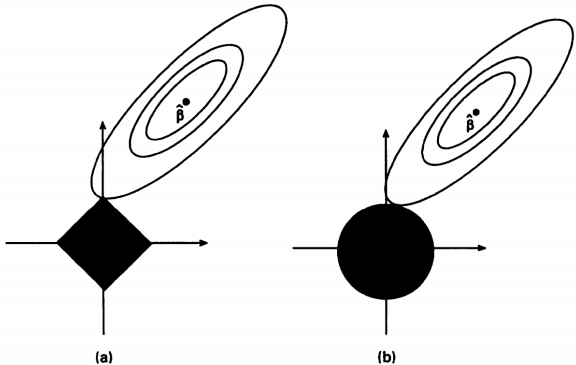


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

What's so great about Lasso? A quick aside on MSE

- Well, it (and modified versions of it) have two very nice properties:
  - very efficient estimators – they predict Y well
  - pick a subet of covariates, making model interpretaiton easier!

- A quick aside on regularized estimators. Recall that for a given estimator $\hat{\theta}$ of θ, we care a lot about the mean squared error, $MSE(\hat{\theta})$, especially for predictors
  - Recall that $MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$
  - In most estimation, we've cared a *lot* about Bias being zero (or being small!)
  - Regularized estimators give up a little bit of bias in order to reduce overall MSE

- So a nice feature of Lasso is that is has lower MSE than OLS in most cases, but the terms can be biased
  - This is true of all ML approaches generally

The magical Oracle property of Lasso

- One particularly interesting property of Lasso (and derivative approaches) is that it has what is called the “oracle property” under certain conditions

- In essence, we get the “right” model and asymptotic normality!

- Your reaction may be “this seems too good to be true”
  - Don’t worry, it is
  - But it still does well!

Zou (2006):

of variable selection.

Let us consider model estimation and variable selection in linear regression models. Suppose that $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, are the linearly independent predictors. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be the predictor matrix. We assume that $E[y|\mathbf{x}] = \beta_1^* x_1 + \cdots + \beta_p^* x_p$. Without loss of generality, we assume that the data are centered, so the intercept is not included in the regression function. Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and further assume that $|\mathcal{A}| = p_0 < p$. Thus the true model depends only on a subset of the predictors. Denote by $\hat{\boldsymbol{\beta}}(\delta)$ the coefficient estimator produced by a fitting procedure δ. Using the language of Fan and Li (2001), we call δ an *oracle* procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate, $\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta)_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}}) \rightarrow_d \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is the covariance matrix knowing the true subset model.

An aside on Pointwise vs. Uniform convergence

- Recall that all our asymptotic results are about approximations to finite sample distributions
- In fact, Penalized and ML methods improve finite sample performance!
  - OLS does very well with infeasibly **infinite** data
- Recall from your econometrics courses the *pointwise* convergence of an estimator
  - Given a true estimand θ_0, we can consider the convergence of $\hat{\theta}$ to that θ_0
- But, this holds fixed our vlaue of θ_0 – we typically want *uniform* convergence
  - E.g. the convergence can be done across all values of θ_0 simultaneously
- Why does this matter? Uniform convergence matters for our asymptotic approximations to do a good job in approximating finite samples

Leeb and Potscher (2008)

We have shown that sparsity of an estimator leads to undesirable risk properties of that estimator. The result is set in a linear model framework, but easily extends to much more general parametric and semiparametric models, including time series models. Sparsity is often connected to a so-called "oracle property". We point out that this latter property is highly misleading and should not be relied on when judging performance of an estimator. Both observations are not really new, but worth recalling: Hodges' construction of an estimator exhibiting a deceiving pointwise asymptotic behavior (i.e., the oracle property in today's parlance) has led mathematical statisticians to realize the importance uniformity has to play in asymptotic statistical results. It is thus remarkable that today—more than 50 years later—we observe a return of Hodges' estimator in the guise of newly proposed estimators (i.e., sparse estimators). What is even more surprising is that the deceiving pointwise asymptotic properties of these estimators (i.e., the oracle property) are now advertised as virtues of these methods. It is therefore perhaps fitting to repeat Hajek's (1971, p. 153) warning:

> Especially misinformative can be those limit results that are not uniform. Then the limit may exhibit some features that are not even approximately true for any finite n.

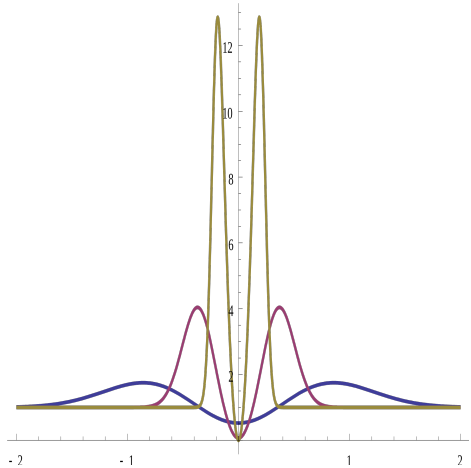
The discussion in the present paper as well as in Leeb and Pötscher (2005) shows in particular that distributional or risk behavior of consistent post-model-selection estimators is not as sometimes believed, but is much worse.

Back to lasso's oracle property

- Leeb and Potscher (2005, 2008) say: "Wait, hold on." (They are much punchier than that)
- The implication is that we should not rely heavily on the oracle property of Lasso (and other penalized methods).
- This ability to select elements can be misleading
- This is not a new fact, and one that econometricians/statisticians should have been aware of
  - E.g. Hodges' Estimator

Example with Hodges' Estimator

- Consider an estimator $\hat{\theta}_n$ for θ. Now we construct our new estimator,
  - $\hat{theta}_{n,hodge} = \hat{\theta}_n$ if $|\hat{theta}_n| \geq n^{-1/4}$
  - $\hat{theta}_{n,hodge} = 0$ if $|\hat{theta}_n| < n^{-1/4}$
- This is a quasi-shrunk estimator, with superefficient convergence of the estimator when $\theta = 0$ and normal asymptotic convergence everywhere else
- But, the convergence is not uniform, and creates very weird properties near zero
  - Blue = $n = 5$, purple = $n = 50$, olive = $n = 500$



Irrepresentability

- Important note – the convergence results for Lasso hold under an important condition known as the irrepresentability condition.

- Many regressions have collinear regressors, as this is a natural feature of lots of statistical problems

- One very awkward property of Lasso is that having right-hand side variables that are highly correlated can create very weird problems

- If the covariates that should be excluded are correlated with the relevant covariates in a meaningful way, then it's possible that lasso will pick the irrelevant covariate, even for large sample

- This problem is very solveable – simply orthogonalize the covariates manually!
  - But that kind of defeats the "interpretability" point...
  - But this is fixable!

Puffer Transformation (Jia and Rohe (2015))

- Key insight in linear model is that for a given $n \times n$ matrix F, we can premultiply and estimate

$$FY = FX\beta + F\epsilon \qquad \text{vs.} \qquad Y = X\beta + \epsilon$$

- Notatbly this will give us the β. However, if we use the appropriate tranformation (preconditioning) matrix, we can ensure that the consistency of the estimates hold

- Let $F = UD^{-1}U'$ be the Puffer transformation, where U and D come from the Singular Value Decomposition of $X = UDV'$.
  - Under this transformation, we can ensure consistent estimates of β, and most important, it's the *same* β
  - If we had orthogonalized our X, we woudl have a different linear combination of the underlying β

- The tradeoff with this method is it can increase variance of the estimators. See the paper for details on implementation

The Geometry of Puffer

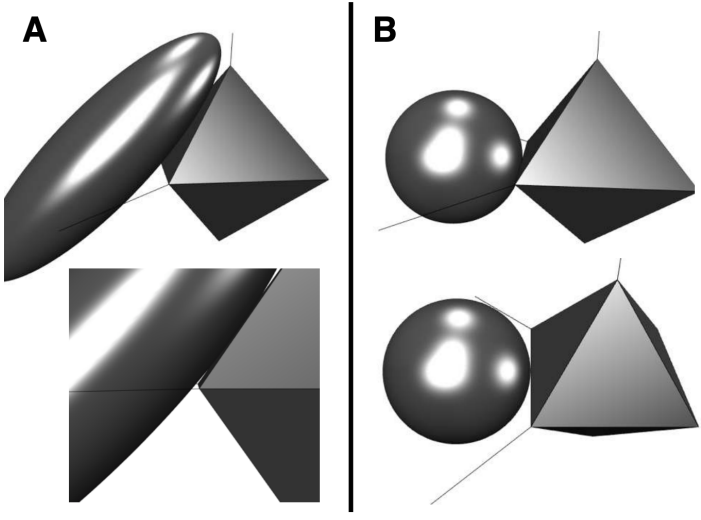


FIG 2. *Panel A illustrates the Lasso without preconditioning. In panel B, preconditioning turns the ellipse from the ℓ_2 loss into a sphere. Here, the Lasso correctly selects the true model. These figures were drawn with the R library RGL (Adler et al., 2003).*

Punchline for Lasso

- Remember, asymptotics are approximations. Uniformity matters a lot! The thresholding criterion for Lasso creates weird behavior that can be unsmooth

- E.g. there's no free lunch!

- Or *is* there?? Key point that we will revisit shortly – Lasso wanted to find all effects, no matter how small. What if we relax this?
  - E.g., what if our goal is not the parameters themselves, but to approximate something in a way that does not create issues?

Generalizations + Sparsity

- There are a number of other types of linear regularization methods
  - We'll cover more non-linear methods later in the course
- These include Ridge regression, Group Lasso, Elastic Net, Others
- These all revolve around methods to shrink with either L1 or L2 norms
  - Many are dealing with highly correlated regressors
- Today, will continue to focus on Lasso, which has been a major focus on econometrics research
  - Why? Model selection aspect of Lasso + Sparsity assumption is very powerful

Other Linear Methods:

- Ridge Regression
- Elastic Net
- Group Lasso
- Fused Lasso
- Adaptive Lasso
- Bridge regression
- Bayesian Lasso
- Prior Lasso

So what? How does an applied economist use this?

- With that under our belt, let's discuss applications.
- Most direct historical purposes have been for prediction
  - Prediction is extremely useful!
- Mullanaithan and Spiess (2017) discuss various uses of general ML
  - Prediction in decision problems, e.g. bail decisions (or lending, Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2020))
  - Prediction in forecasting – e.g. asset pricing. (For example, see Feng, Giglio and Xiu)
  - Testing a model or predictor – e.g. creating an ML benchmark
- What we'll discuss rest of today: how Lasso methods can be used in causal inference

Lasso and Nuisance Parameters

- Concise way to remember the relative merits of LASSO/ML vs. standard model: $\hat{y}$ vs $\hat{\beta}$ (Mullanaithan and Spiess (2017))
  - Lasso is best for constructing a lower MSE estimate of something.
  - You shouldn't always trust it for partiicular estimates of all the underlying parameters in the model (and inference is challenging for all of them, e.g. Leeb and Potscher's critique)

- Remember the problem of semiparametric models and nuisance parameters? Turns out this is a great problem for us to solve!
  - We have a function we need to estimate
  - We don't care about the parameters of the function per se

Lasso vs. OLS

$\hat{y}$ vs. $\hat{\beta}$

Partial linear model

- Consider a partially additive model:

$$Y_i = D_i\tau + g_0(X_i) + U_i$$

where D_i is randomly assigned, and X are pre-treatment covariates , and g_0 is some unknown function.

- A simpler version of this could be

$$Y_i = D_i\tau + X_{i,0}\beta + U_i$$

where we don't know which X_i are in the model

- Note that the estimation of g_0 (or the β) are *nuisance* parameters – we'd like them to get good / better estimates of τ, but we don't care about them *per se*

Partial linear model and causal inference

- There are a lot of results in this space, heavily influenced by Victor Chernozhukov.
- Going to touch on two points:
  1. How did they address the Leeb and Potscher issue
  2. How to address the bias variance problem if interested in causal parameters?
- Many of the insights here carry over into the linear IV case
  - Today, we'll focus on exogeneous regressors (e.g. random experiment)
- This discussion riffs heavily on Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2017,2018)

Why would we need ML if RCT?

Regarding ML and RCTs:

> At first blush, those two sets of methods may seem to have very different applications: in the most basic randomized controlled experiment, there is a sample with a single treatment and a single outcome. Covariates are not necessary and even linear regression is not the best way to analyze the data (Imbens and Rubin, 2015). In practice however, applied researchers are often confronted with more complex experiments. For example, there might be accidental imbalances in the sample, which require select control variables in a principled way. ML tools, such as the lasso method proposed in Belloni et al. (2014, 2017) or the double machine learning method proposed in Chernozhukov et al. (2017), have proven useful for this purpose. Moreover, some complex RCT

- Subsequent discussion regards *subgroup analysis* – a topic for our ML discussion at the end of the course!
  - In essence, what if you have lots of treatment combinations / groups?

The DML (Double/De-biased) Machine Learning Cookbook

$$Y_i = D_i\tau + g_0(X_i) + U_i, \qquad D_i, X_i \text{ are demeaned}$$

- Will first outline the approach, then we can discuss details

- Key ingredients / assumptions:
  1. *Sparsity*; estimation error in g_0 is orthogonal to the moments that help estimate D_i
  2. Sample splitting; account for the overfitting bias from high-dimensional approaches

- Start with the "naive" approach
  1. Split the sample in half
  2. Estimate $\hat{g}_0$ using one half of the sample using the regression
  3. Use this estimate $\hat{g}_0$ in the second half of the sample to construct:

$$\hat{\tau} = \left(n^{-1} \sum_i D_i^2 \right)^{-1} \left(n^{-1} \sum_i D_i (Y_i - \hat{g}(X_i) \right)$$

The DML (Double/De-biased) Machine Learning Cookbook

$$\hat{\tau} = \left(n^{-1} \sum_i D_i^2 \right)^{-1} \left(n^{-1} \sum_i D_i (Y_i - \hat{g}(X_i)) \right)$$

- What happens here? If $D_i = m(X_i) + V_i$, and m is a nontrivial function, then $|\sqrt{n}(\hat{\tau} - \tau)| \to \infty$

- Why?

$$\sqrt{n}(\hat{\tau} - \tau) = \left(n^{-1} \sum_i D_i^2 \right)^{-1} \left(n^{-1/2} \sum_i D_i U_i \right) + \underbrace{\left(n^{-1} \sum_i D_i^2 \right)^{-1} \left(n^{-1/2} \sum_i D_i (g_0 - \hat{g}_0) \right)}_{\text{This term can blow up}}$$

The DML (Double/De-biased) Machine Learning Cookbook

- In a correctly specified RCT, this term shouldn't matter. However, in finite samples (or with issues with controls), this could create poor performance (e.g. unbalance in treatment)
  - Belloni, Chernozhukov, and Hansen (2014) discuss results where you can use lasso to directly choose the relevant controls
  - I don't encourage this approach – a data-driven approach to choosing your controls is challenging for causal inference

- What's the solution? Double lasso! In this setting, we need to do Frisch-Waugh style orthogonalization. E.g., also estimate $\hat{m}(X_i)$
  - E.g. $\hat{V}_i = D_i - \hat{m}(X_i)$ and then

$$\hat{\tau} = \left(n^{-1} \sum_i \hat{V}_i D_i \right)^{-1} \left(n^{-1} \sum_i \hat{V}_i (Y_i - \hat{g}(X_i) \right)$$

What are the pieces of the DML estimator?

- Three pieces to this estimator in the limiting distribution
  1. The standard distribution
  2. Regularization bias – assumed to small
  3. Remainder term – sample splitting helps here

- How does this estimator get around the Leeb and Potscher critique?
  - Estimation is not about nailing every piece of $g(X_i)$ or $m(X_i)$
  - Instead, uniformity in $\hat{\tau}$ is achieved by having the estimation error in g and m be *orthogonal* to $\hat{\theta}$'s estimation
  - The key crucial (untestable) assumption: sparsity

- Chernozhukov and co-authors also emphasize that the sample splitting is very effective at ensuring that many types of data processes can be incorporated
  - Can simply sample split many times and then average over the estimates

The DML (Double/De-biased) Machine Learning Cookbook

- In steps, the algorithm is as follows (see their AER P&P for greater detail0
  1. Split the sample up into K pieces
  2. In each split sample, use the complement of the data to estimate the nuisance parameters using ML
  3. Estimate the parameter of interest using the split data
  4. Average the K estimators