Part 6: Model Selection and Intro to ML

Chris Conlon

March 6, 2021

Applied Econometrics II

Penalized Regression

Penalized Regression

Suppose we fit a regression model and penalized extra variables all in one go, what would that look like?

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]$$

- We can consider the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|^q$ as penalizing models where β gets further away from zero.
- Similar to placing a prior distribution on β_j centered at 0.
- We definitely want to standardize our inputs before using penalized regression methods.
- Usually you fix q and then look at how estimates respond to γ.
- There are two famous cases $q = 1$ (Lasso) and $q = 2$ (Ridge) though in practice there are many possibilities.

LASSO Regression

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \left[\frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{K} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{K} |\beta_j| \right]$$

- Penalty is L_1 norm on β.
- Can re-write as a constriant $\sum_{j=1}^{K} |\beta_j| \leq s$
- If X is orthonormal then $\hat{\beta}_j^{LASSO} = sign(\hat{\beta}_j) \cdot (|\hat{\beta}_j| - \lambda)_+$
- In words: we get coefficients that are closer to zero by λ, but coefficients within λ of zero are shrunk to zero.
- This leads people to describe LASSO as a shrinkage estimator. It produces models that are sparse.
- Instead of a discrete parameter such as the number of lags p we can continuously penalize additional complexity with λ.

LASSO Regression

But... is choosing λ any easier than choosing p?

- We call λ the regularization parameter.
- We can choose λ in a way that minimizes expected prediction error (EPE).
- Recall $EPE(\lambda) = E_x E_{y|x}([Y - g(X, \lambda)]^2 | X)$.
- In practice most people look at out of sample prediction error rate on a cross validated sample.

LASSO Path

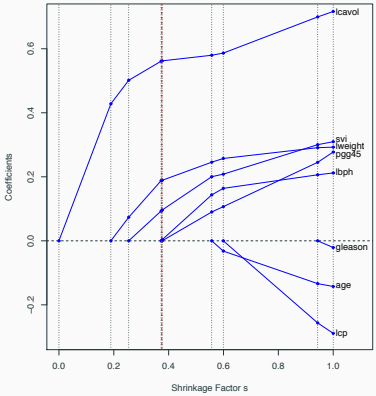


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Ridge Regression

Another popular alternative is the $q = 2$ case

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left[\frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{K} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{K} |\beta_j|^2 \right]$$

- Penalty is L_2 norm on β.
- Can re-write as a constriant $\sum_{j=1}^{K} |\beta_j|^2 \leq s$
- $\hat{\beta}^{Ridge} = (X'X + \lambda I)^{-1} X'Y$.
- If X is orthonormal then $\hat{\beta}_j^{Ridge} = \hat{\beta}_j/(1 + \lambda)$
- In words: everything gets dampened by a constant factor λ (we don't get zeros).
- Adding a constant to the diagonal of $(X'X)$ ensures that the matrix will be invertible even when we have multicollinearity.

Ridge Path

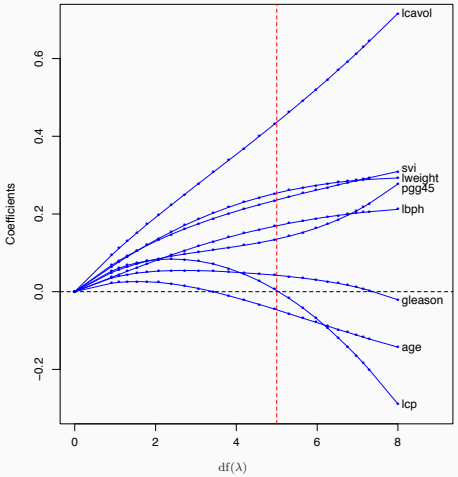


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

LASSO vs Ridge

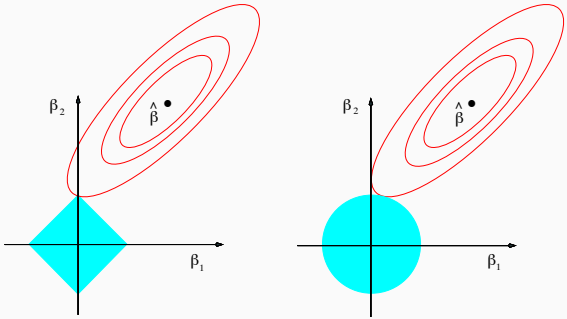


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

What is the point?

Ridge:

- Ridge doesn't provide sparsity which can be a good thing.
- It is most helpful (relative to OLS) when X's are highly correlated with one another.
- OLS can set large but imprecise coefficients when it cannot disentangle effects.

LASSO:

- LASSO is useful for variable/feature selection.
- LASSO does not generally posses the oracle property though variants such as adaptive LASSO may.
- LASSO sometimes has the oracle property for $p \gg N$ and cases where the true β's are not too large.
- People sometimes use LASSO to choose components and then OLS for unbiased coefficient estimates

Elastic Net

We can actually combine them using elastic net regression:

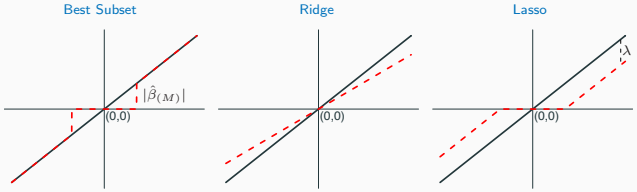
$$P(\lambda_1, \lambda_2, \beta) = \lambda_1 \sum_{j=1}^{K} |\beta_j| + \lambda_2 \sum_{j=1}^{K} |\beta_j|^2$$

This includes both the L_1 and L_2 penalties.

LASSO vs. Ridge

TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of* $\mathbf{X}$. *M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The 45° line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



LAR: Least Angle Regression

Remember Forward Stagewise Regression, consider this alternative:

1. Start with $r = y - \overline{y}$ and $(\beta_1, \ldots, \beta_p) = 0$. (Standardize first!)
2. Find the predictor x_j most correlated with r.
3. Move β_j from 0 to its least-squares estimate $\langle < x_j, r \rangle$ slowly
4. Update $r \leftarrow r - \delta_j \cdot x_j$.
5. Keep moving x_j in same direction until x_k has as much correlation with updated r,
6. Continue updating (β_j, β_k) in direction of joint least-squares coefficients until some other competitor x_l has as much correlation with r.
7. Continue until all p predictors have entered. After $\min[N-1, p]$ steps we arrive at full OLS solution.

- Optional: If a current least-squares estimate hits zero drop it from the active set and re-estimate the joint least squares direction without it.

LAR: Least Angle Regression

Why do we need LAR?

- It turns out that with the optional step from the previous slide: LAR gives us an easy algorithm to compute the LASSO estimate.
- Actually it does even better – it gives us the full path of LASSO estimates for all values of λ!
- This is actually a relatively new result.

LASSO vs LAR

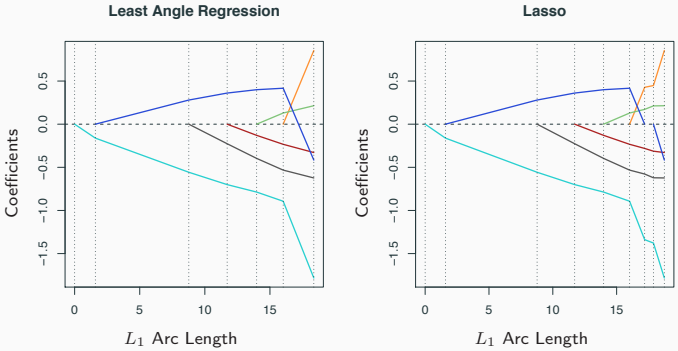
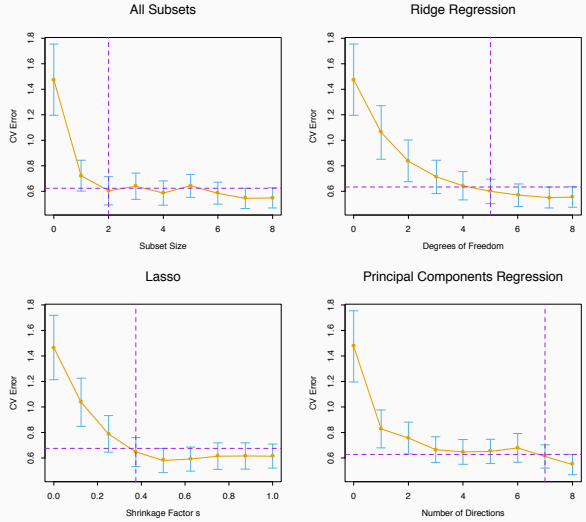


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about* 18.

Overall Comparison



Overall Comparison

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Overall Comparison

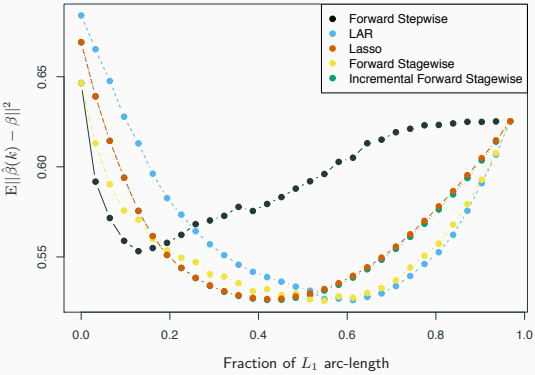


FIGURE 3.16. *Comparison of LAR and lasso with forward stepwise, forward stagewise (FS) and incremental forward stagewise (FS_0) regression. The setup is the same as in Figure 3.6, except $N = 100$ here rather than 300. Here the slower FS regression ultimately outperforms forward stepwise. LAR and lasso show similar behavior to FS and FS_0. Since the procedures take different numbers of steps (across simulation replicates and methods), we plot the MSE as a function of the fraction of total L_1 arc-length toward the least-squares fit.*

Oracle Property

- An important question with LASSO is whether or not it produces consistent parameter estimates (Generally no).
- We think of asymptotics as taking both $N, p \to \infty$.
- Donohu (2006) shows that for $p > N$ case, when the true model is sparse, LASSO identified correct predictors with high probability (with certain assumptions on $\mathbf{X}$) as we slowly relax the penalty.
- Condition looks like ("good" variables are not too correlated with "bad" variables).

$$\max_{j \in S^c} ||x_j' X_S (X_S' X_S)^{-1}||_1 \leq (1 - \epsilon)$$

- Where X_S are columns corresponding to nonzero coefficients, and S^c are set of columns with zero coefficients (at true value).

Other Extensions

- *Grouped LASSO* for penalzing groups of coefficients at once (like fixed effects)
- *Relaxed LASSO* run LASSO to select coefficients and then run a non-penalized subset regression or LASSO with a less stringent penalty on the subset. (Here CV tends to pick a less strong penalty term λ leading to less shrinkage and bias).
- *SCAD: Smoothly Clipped Absolute Deviation*: do less shrinkage on big coefficients but preserve sparsity

$$\frac{dJ_a(|beta, \lambda)}{d\beta} = \lambda \cdot sgn(\beta) \left[I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right]$$

- *Adaptive LASSO* uses a weighted penalty of the form $\sum_{j=1}^{p} w_j |\beta_j|$ where $W_j = 1/|\hat{\beta}_j|^\nu$ using the OLS estimates as weights. This yields consistent estimates of parameters while retaining the convexity property.

Penalty Comparisons

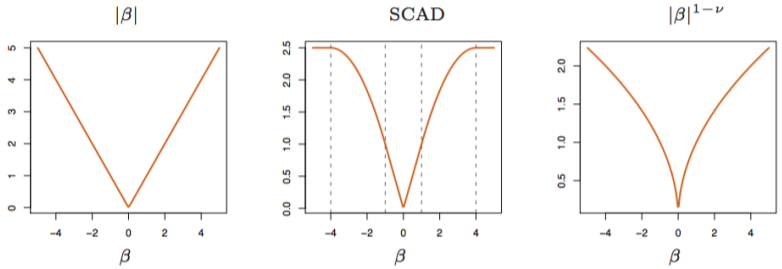


FIGURE 3.20. *The lasso and two alternative non-convex penalties designed to penalize large coefficients less. For SCAD we use* $\lambda = 1$ *and* $a = 4$*, and* $\nu = \frac{1}{2}$ *in the last panel.*

Implementation

- Routines are highly specialized: there are lots of tricks
- No hope of coding this up on your own!
- In R you can use `glmnet` or `lars`.
- In Python you can use `scikit-learn`
- In most recent Matlab in Stats toolbox you have `lasso` and `ridge`.
- In STATA you can download `.ado` files from Christian Hansen's website (Chicago).