

# Part 6: Model Selection and Intro to ML

---

Chris Conlon

April 7, 2020

Applied Econometrics II

- Chapters 2-7 of *Elements of Statistical Learning*

# Intro

---

How many components should we include in our model?

- Too few: under-fitting and large residuals.
- Too many: over-fitting and poor out of sample prediction.

How do we choose?

- $X$  variables.
- Instrumental Variables.

# When do we have too much data?

- On the internet!
- Hedonics: What really determines the price of your house?
- Prediction: What really determines loan defaults?
- Consideration Sets: How many products do consumers really choose among on the shelf?
- Which elements of financial filings really matter?

## What do people mostly do in practice?

- Regress  $Y$  on  $X$  with all variables included.
- Drop some variables if they aren't significant?
- Re-run with some things dropped
- Add in some other things that may or may not be significant.

# Nested and Non-nested Models

What makes a model **nested** or **non-nested**?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$$

A nested model can be written as a restricted version of the larger model

- ie: all of the following are nested within the model above

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_3 z_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_2 w_i + \beta_3 z_i + \varepsilon_i$$

- ie: this model is non-nested (because of  $s_i$ )

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \gamma s_i + \varepsilon_i$$

# What we teach undergrads

Two traditional ways to select the number of components in a model:

$$\overline{R}^2 = 1 - SSR(p)/TSS - SSR(p)/TSS \cdot \frac{p}{N - p - 1}$$

$$AIC(p) = \ln \left( \frac{SSR(p)}{N} \right) + (p + 1) \frac{2}{N}$$

$$BIC(p) = \ln \left( \frac{SSR(p)}{N} \right) + (p + 1) \frac{\ln N}{N}$$

- Commonly employed in macroeconometric or time-series context for things like selecting lags of an autoregression

$$y_t = \alpha_0 + \sum_{k=1}^p \alpha_k y_{t-k} + \varepsilon_t$$



- AIC tends to select larger models than BIC since it penalizes the number of parameters less heavily.
- These usually depend on ordering potential models by  $p$  the number of components and then sequentially fitting them.
- AIC is not consistent: as  $N \rightarrow \infty$  it may still select too many parameters.
- BIC is consistent: as  $N \rightarrow \infty$  it will select the correct number of parameters.
- Of course for finite-sample  $N < \infty$  anything can happen.

# Where does it come from?

How do we come up with these penalized regressions?

- AIC/BIC arise from considering the likelihood ratio test (LRT) of a maximum likelihood estimator and making a lot of assumptions.
- AIC arises from minimizing the Expected KLIC.

$$\begin{aligned} KLIC(f, g) &= \int f(y) \log(f(y)) \partial y - \int f(y) \log(g(y)) \partial y \\ &= C_f - E \log(g(y)) \end{aligned}$$

- Low values of KLIC mean the models are similar.

## Where does it come from?

How do we come up with these penalized regressions?

- Recall that OLS is a ML estimator in the case where  $\varepsilon$  is normally distributed.

$$D = -2 \ln \left( \frac{\text{Likelihood } H_0}{\text{Likelihood } H_a} \right) = -2 \ln \underbrace{\left( \frac{(\sup L(\theta|x) : \theta \in \Theta_0)}{(\sup L(\theta|x) : \theta \in \Theta)} \right)}_{\Lambda(x)}$$

- If the models are **nested** then  $\Theta_0 \subset \Theta$  and  $\dim(\Theta) - \dim(\Theta_0) = q$  then as  $N \rightarrow \infty$  we have that  $D \rightarrow^d \chi^2(q)$ .

Many cases we are interested in are **not strictly nested**

- Should I include  $x_2$  OR  $x_3$  in my regression? (partially overlapping)
- Is the correct distribution  $f(y|x, \theta)$  normal or log-normal? (non-overlapping)

## Non-nested cases

- Cox (1961) suggested the following (often infeasible solution) by assuming that  $F_\theta$  is the true model.

$$LR(\hat{\theta}, \hat{\gamma}) = L_f(\hat{\theta}) - L_g(\hat{\gamma}) = \sum_{i=1}^N \ln \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})}$$

- Depending on which the true model is you could reject  $F_\theta$  for  $G_\gamma$  and vice versa!
- Deriving the test statistic is hard (and specific to  $F_\theta$ ) because we must obtain  $E_f[\ln \frac{f(y_i|x_i, \hat{\theta})}{g(y_i|x_i, \hat{\gamma})}]$ .
- Similar to AIC in that we are minimizing KLIC over  $F_\theta$ .

$$\begin{aligned} H_0 : E_{h(y|x)} \left[ \frac{f(y|x, \theta)}{g(y|x, \gamma)} \right] &= 0 \\ \rightarrow E_h[\ln(h/g)] - E_h[\ln(h/f)] &= 0 \end{aligned}$$

- Instead of taking expectation with respect to one of two distributions, we take it with respect to  $h(y|x)$  the unknown but **true distribution**.
- Same as testing whether two densities  $(f, g)$  have same KLIC.
- The main result is that (details in 8.5 of CT):

$$\frac{1}{\sqrt{N}} LR(\hat{\theta}, \hat{\gamma}) \rightarrow^d N[0, \omega_*^2], \quad \omega_*^2 = V_0 \left[ \ln \frac{f(y|x, \hat{\theta})}{g(y|x, \hat{\gamma})} \right]$$

## Back to the real world...

- We have some theoretical benchmark which lets us discern which of two model we prefer (under certain assumptions).
- In practice we often start with a functional form like:
$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{i,k} + \varepsilon_i$$
- Which  $x$ 's do we include?
- Which  $x$ 's do we leave out?
- It is not clear that BIC/AIC or Vuong test tells us what we should do in practice.
- If you have  $K$  potential regressors you could consider all  $2^K$  possible regressions.
- Or you could consider all  $\binom{K}{p}$  possible combinations with  $p$  parameters.
- This sounds very time consuming

# Things to keep in mind

Two major (related) problems:

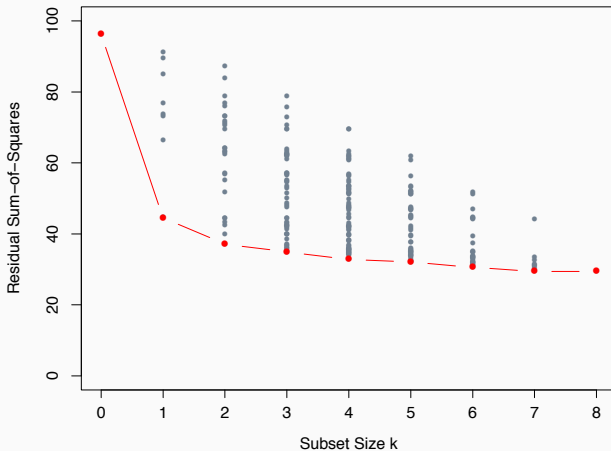
- Regressors are correlated with one another:
  - small changes in the sample:  $\beta_1$  goes up,  $\beta_2$  goes down.
  - large coefficients can lead to wild predictions.
  - If relationship between  $y_i$  and  $x_i$  nonlinear, and  $x_i, z_i$  are highly correlated then we may attribute some of this nonlinearity to  $z_i$ , even when it has no effect.
- Lots of imprecisely estimated parameters can make prediction tricky
  - Small changes in the sample can lead to large changes in  $\hat{y}_i|x_i$ .

The big idea: maybe we tolerate some **bias** to greatly reduce **variance**.

- This is where ML and Econometrics diverge!
- Everything you learned so far has been focused on **unbiasedness**



## Minimizing SSR/AIC: all possible regressions



**FIGURE 3.5.** All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.

# What is orthogonality?

- We can think about a world where  $\langle x_j, x_k \rangle = 0$  for  $j \neq k$ .
- In this world I can get  $\beta_j$  by regressing  $y$  on  $x_j$  by simple linear regression.
- I could do this for each  $j$  and the resulting vector  $\beta$  would be the same as running multiple regression.
- We could try and transform  $X$  so that it forms an **orthogonal basis**.
- Unless we are running regressions by hand this doesn't seem tremendously helpful.
- However, in practice this is often what your software does!

# Gram-Schmidt/QR Decomposition

1. Let  $x_0 = z_0 = 1$
2. For  $j = 1, 2, \dots, p$ : Regress  $x_j$  on  $z_0, z_1, \dots, z_{j-1}$  to give you  $\hat{\gamma}_{jl} = \langle z_l, x_j \rangle / \langle z_l, x_l \rangle$  and residual  $z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k$ .
3. With your transformed orthogonal basis  $\mathbf{z}$  you can now regress  $y$  on  $z_p$  one by one to obtain  $\hat{\beta}_p$ .

What does this do?

- The resulting vector  $\hat{\beta}$  has been adjusted to deliver the marginal contribution of  $x_j$  on  $y$  after adjusting for all  $x_{-j}$ .
- If  $x_j$  is highly correlated with other  $x_k$ 's then the residual  $z_j$  will be close to zero and the coefficient will be unstable.
- This will be true for any variables  $x_l$  within a set of correlated variables.
- We can delete any one of them to resolve this issue.

# QR Decomposition (Technical Details)

QR Decomposition has a matrix form which regression software uses:

$$\begin{aligned}\mathbf{X} &= \mathbf{Z}\mathbf{\Gamma} \\ &= \underbrace{\mathbf{Z}\mathbf{D}^{-1}}_{\mathbf{Q}} \underbrace{\mathbf{D}\mathbf{\Gamma}}_{\mathbf{R}} \\ \hat{\beta} &= \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{Q}\mathbf{Q}'\mathbf{y}\end{aligned}$$

- $\mathbf{Z}$  is the matrix of the orthogonalized residuals  $z_j$ 's.
- $\mathbf{\Gamma}$  is upper triangular matrix with entries  $\hat{\gamma}_{kj}$
- $\mathbf{D}$  is diagonal matrix with entries  $\|z_j\|$ .
- $\mathbf{Q}$  is  $N \times ((p+1))$  orthogonal matrix  $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$
- $\mathbf{R}$  is  $(p+1) \times (p+1)$  upper triangular matrix.

# What happens in practice?

What are people likely doing in practice:

- Start with a single  $x$  variable and then slowly add more until additional  $x$ 's were insignificant
- Start with all possible  $x$  variables and drop those where  $t$ -statistics were insignificant.
- These procedures actually make some sense if the columns of  $X$  are linearly independent or orthogonal.
- In practice our regressors are often correlated (sometimes highly so).

# Forward Stepwise Regression

Consider the following **greedy algorithm**

1. Start with an empty model and add a constant  $\bar{y}$ .
2. Then run  $K$  single-variable regressions, choose the  $x_k$  with the highest  $t$ -statistic call this  $x^{(1)}$ .
3. Now run  $K - 1$  two variable regressions where the constant and  $x^{(1)}$  and choose  $x^{(2)}$  as regression where  $x_k$  has the highest  $t$ -statistic.
4. Now run  $K - 2$  three variable regressions where the constant and  $x^{(1)}, x^{(2)}$
5. You get the idea!

We stop when the  $x_k$  with the highest  $t$ -statistic is below some threshold (often 20% significance).

# Backwards Stepwise Regression

1. Start with an full model.
2. Remove the  $x$  variable with the lowest  $t$ -statistic. Call this  $x^{(k)}$ .
3. Re-run the regression without  $x^{(k)}$ .
4. Repeat until the smallest  $t$ -statistic exceeds some threshold.

## Comparison

- Backwards and forwards stepwise regression tend to give similar choices (but not always).
- Everything is trivial if  $X$ 's columns are orthogonal (computer has some tricks otherwise-  $QR$ ).
- Forward stepwise works when we have more regressors than observations  $K > N$ .
- I proposed the  $t$ -stat here but some packages use AIC/BIC as the criteria.
- We should also be careful to **group dummy variables together** as a single regressor.
- These are implemented in `step` in R and `stepwise` in Stata.
- We probably want to adjust our standard errors for the fact that we have run many regressions in sequence before arriving at our model. **In practice not enough people do this!**



# Multiple Testing Problem

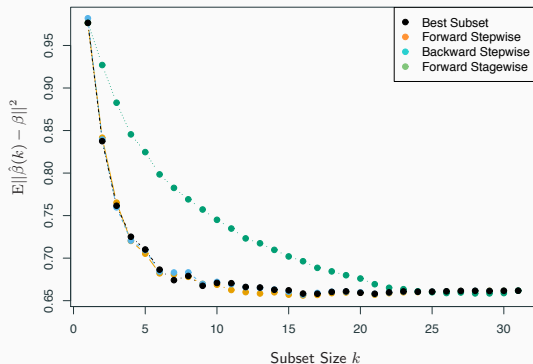
- A big deal in Econometrics frequently ignored in applied work is the **Multiple Testing Problem**
- You didn't just pick the regression in your table and run that without considering any others.
- This means that your  $t$  and  $F$  stats are going to be too large!! (Standard errors too small!)
- How much bigger should they be?
  - Analytic size corrections can be tricky and data dependent
  - Bootstrap/Monte-Carlo studies should give you a better idea.

## (Incremental) Forward Stagewise Regression

As an alternative consider:

1. Start with  $r = y$  and  $(\beta_1, \dots, \beta_p) = 0$ .
  2. Find the predictor  $x_j$  most correlated with  $r$ .
  3. Update  $\beta_j \leftarrow \beta_j + \delta_j$  where  $\delta_j = \epsilon \cdot \text{sgn}\langle r, x_j \rangle$ .
  4. Update  $r \leftarrow r - \delta_j \cdot x_j$  and repeat for  $S$  steps.
- Alternative  $\delta_j = \langle r, x_j \rangle$
  - We can continue until no regressors have correlation with residuals
  - This is very slow (it takes many many  $S$ ).
  - Sometimes slowness can be good – in high dimensions to avoid overfitting.

# Stepwise selection procedures



**FIGURE 3.6.** Comparison of four subset-selection techniques on a simulated linear regression problem  $Y = X^T \beta + \varepsilon$ . There are  $N = 300$  observations on  $p = 31$  standard Gaussian variables, with pairwise correlations all equal to 0.85. For 10 of the variables, the coefficients are drawn at random from a  $N(0, 0.4)$  distribution; the rest are zero. The noise  $\varepsilon \sim N(0, 6.25)$ , resulting in a signal-to-noise ratio of 0.64. Results are averaged over 50 simulations. Shown is the mean-squared error of the estimated coefficient  $\hat{\beta}(k)$  at each step from the true  $\beta$ .

# 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  $\beta$  gets further away from zero.
- Similar to placing a **prior distribution** on  $\beta_j$  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  $\gamma$ .
- There are two famous cases  $q = 1$  (Lasso) and  $q = 2$  (Ridge) though in practice

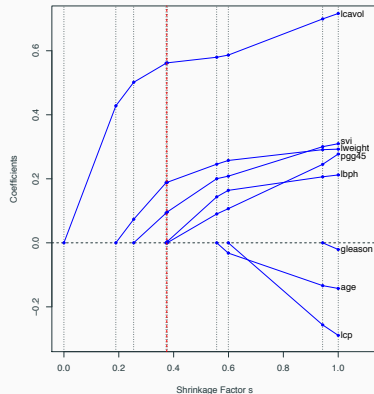
# LASSO Regression

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

- Penalty is  $L_1$  norm on  $\beta$ .
- Can re-write as a constraint  $\sum_{j=1}^K |\beta_j| \leq s$
- If  $X$  is orthonormal then  $\hat{\beta}_j^{LASSO} = \text{sign}(\hat{\beta}_j) \cdot (|\hat{\beta}_j| - \lambda)_+$
- In words: we get coefficients that are closer to zero by  $\lambda$ , but coefficients within  $\lambda$  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  $\lambda$ .

But... is choosing  $\lambda$  any easier than choosing  $p$ ?

- We call  $\lambda$  the **regularization** parameter.
- We can choose  $\lambda$  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**.



**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

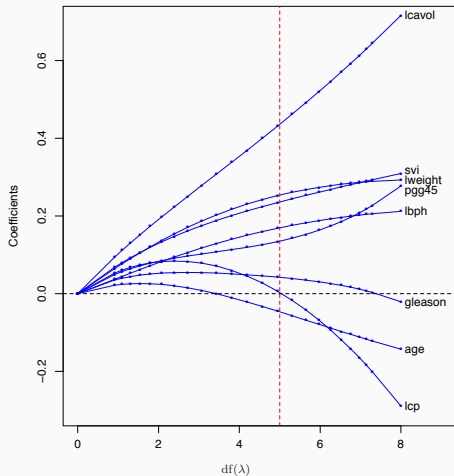
Another popular alternative is the  $q = 2$  case

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

- Penalty is  $L_2$  norm on  $\beta$ .
- Can re-write as a constraint  $\sum_{j=1}^K |\beta_j|^2 \leq s$
- $\hat{\beta}^{Ridge} = (X'X + \lambda I)^{-1} X'Y$ .
- If  $X$  is orthonormal then  $\hat{\beta}_j^{Ridge} = \hat{\beta}_j / (1 + \lambda)$
- In words: everything gets dampened by a constant factor  $\lambda$  (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.

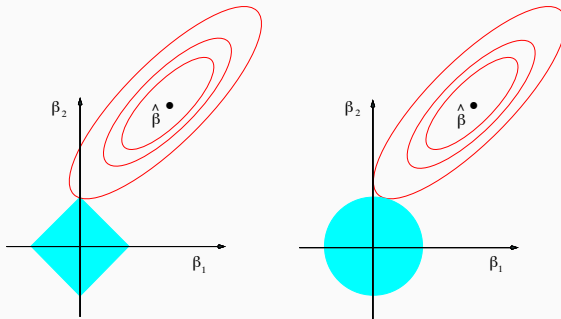


# Ridge Path



**FIGURE 3.8.** Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter  $\lambda$  is varied. Coefficients are plotted versus  $df(\lambda)$ , the effective degrees of freedom. A vertical line is drawn at  $df = 5.0$ , 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 possess 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  $\beta$ 's are not too large.
- People sometimes use LASSO to choose components and then OLS for unbiased coefficient estimates

We can actually combine them using **elastic net regression**:

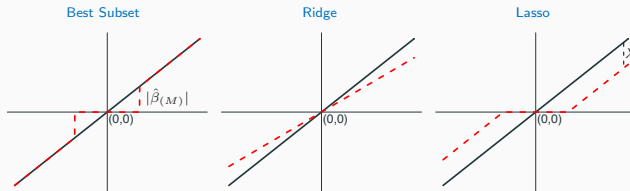
$$P(\lambda_1, \lambda_2, \beta) = \lambda_1 \sum_{j=1}^K |\beta_j| + \lambda_2 \sum_{j=1}^K |\beta_j|^2$$

This includes both the  $L_1$  and  $L_2$  penalties.

# LASSO vs. Ridge

**TABLE 3.4.** Estimators of  $\beta_j$  in the case of orthonormal columns of  $\mathbf{X}$ .  $M$  and  $\lambda$  are constants chosen by the corresponding techniques; sign denotes the sign of its argument ( $\pm 1$ ), 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 $M$ ) | $\hat{\beta}_j \cdot I( \hat{\beta}_j  \geq  \hat{\beta}_{(M)} )$ |
| Ridge                   | $\hat{\beta}_j / (1 + \lambda)$                                   |
| Lasso                   | $\text{sign}(\hat{\beta}_j)( \hat{\beta}_j  - \lambda)_+$         |



# LAR: Least Angle Regression

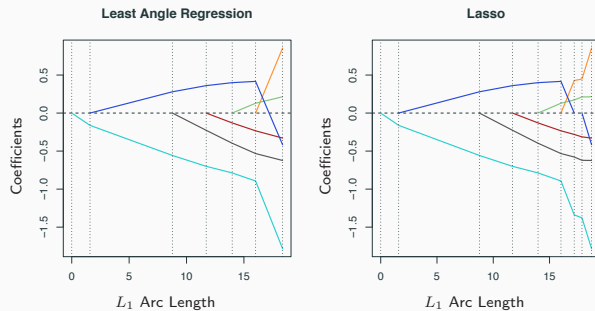
Remember Forward Stagewise Regression, consider this alternative:

1. Start with  $r = y - \bar{y}$  and  $(\beta_1, \dots, \beta_p) = 0$ . (Standardize first!)
  2. Find the predictor  $x_j$  most correlated with  $r$ .
  3. Move  $\beta_j$  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  $(\beta_j, \beta_k)$  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.

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  $\lambda$ !
- This is actually a relatively new result.

# LASSO vs LAR

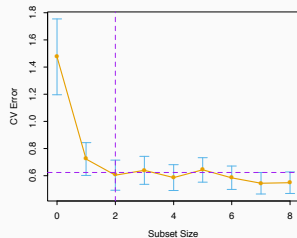


**FIGURE 3.15.** Left panel shows the LAR coefficient profiles on the simulated data, as a function of the  $L_1$  arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.

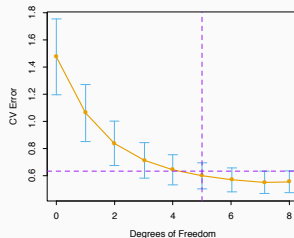


# Overall Comparison

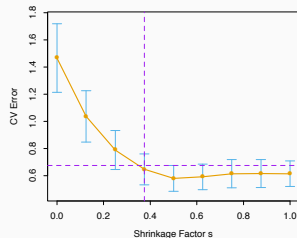
All Subsets



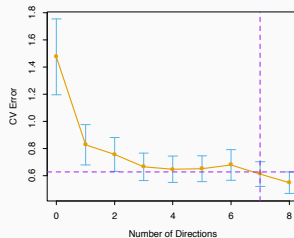
Ridge Regression



Lasso



Principal Components Regression

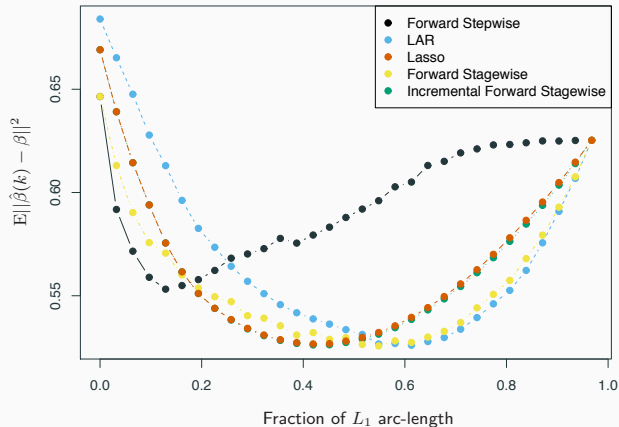


# Overall Comparison

**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<sub>0</sub>) 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<sub>0</sub>. Since the procedures take different numbers of steps (across simulation replications and methods), we plot the MSE as a function

# 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).

We have other data reduction techniques with a long history in Econometrics

- Principal Components
- Factor Analysis
- Partial Least Squares

# Principal Components

Suppose we have a very high dimensional  $X$  where we have a high degree of correlation among the components  $x_j$ .

- We can start by computing the appropriate correlation matrix  $C = E[\tilde{X}'\tilde{X}]$  where  $\tilde{X}$  denotes we have standardized each column  $x_j$  to have mean zero and variance 1.
- Diagonalize  $C$  via the eigen-decomposition  $V^{-1}CV = D$  where  $D$  is the diagonal matrix of eigenvalues.
- Sort  $D$  and the corresponding columns of  $V$  in decreasing order of the eigenvalues  $d_j$ .
- Choose a subset of  $m < K$  eigenvalues and eigenvectors and call that  $V_m$  and  $\lambda_m$
- Compute transformed data:  $Z_m = V_m\tilde{X}$  which is of dimension  $N \times m$  instead of  $N \times K$ .

# Principal Components

- If  $m \ll K$  then we can substantially reduce the dimension of the data.
- The idea is to choose  $m$  so that  $(Z'Z)$  spans approximately the same space that  $(X'X)$  does.
- This works because we use the **principal eigenvectors** (those with the largest eigenvalues).
- The first eigenvector explains most of the variation in the data, the second the most of the remaining variation, and so on.
- You may also recall that eigenvectors form an **orthonormal basis**, so each dimension is linearly independent of the others.
- As eigenvalues decline, it means they explain less of the variance.

# Principal Components

Output from software will include

- Coefficients: these transform from  $X \rightarrow Z$
- Score: these are the transformed  $Z$ 's
- Latent/Eigenvalue: the corresponding Eigenvalue
- Explained/Cumulative: cumulative explained variance  $\sum_{j=1}^m (\lambda_j / \sum_k \lambda_k)$
- Stata: `pca`, Matlab: `pca`, R/stats: `princomp`.

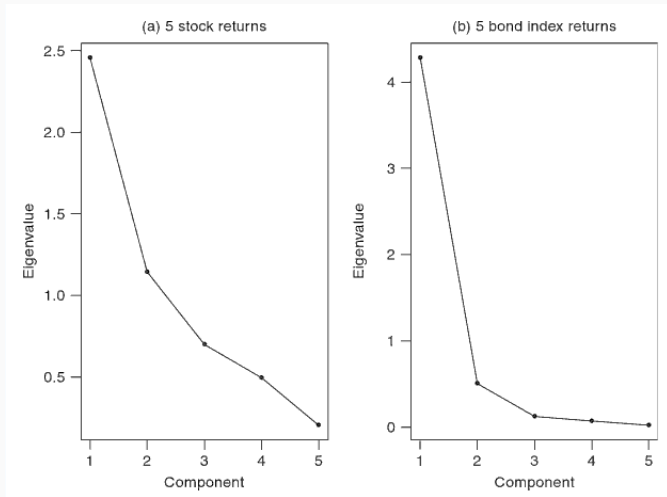


# Principal Components

How many components?

- Choose # of components by the eigenvalues or % of variance explained
- Common cutoffs are 90-95% of variance.
- Eigenvalue based cutoff rules (only take eigenvalues  $> 1$ ).
- Most common method is to eyeball the scree plot.

# Principal Components



# Principal Components

Table 1 Principal Components

| Principal Component | Eigenvalue | Proportion of Variance | Cumulative Variance |
|---------------------|------------|------------------------|---------------------|
| 1                   | 75         | 26.95%                 | 26.95%              |
| 2                   | 43         | 15.45%                 | 42.40%              |
| 3                   | 30         | 10.78%                 | 53.18%              |
| 4                   | 21         | 7.55%                  | 60.73%              |
| 5                   | 19         | 6.83%                  | 67.55%              |
| 6                   | 18         | 6.47%                  | 74.02%              |
| 7                   | 17         | 6.11%                  | 80.13%              |
| 8                   | 11         | 3.95%                  | 84.08%              |
| 9                   | 10         | 3.59%                  | 87.67%              |
| 10                  | 10         | 3.41%                  | 91.09%              |
| 11                  | 9          | 3.16%                  | 94.25%              |
| 12                  | 5          | 1.80%                  | 96.05%              |
| 13                  | 4          | 1.58%                  | 97.63%              |

# Principal Components

- We can run a regression on principal components  $Z$ 's and then recover the betas of the  $X$ 's

$$\hat{y}_{(M)}^{pcr} = \bar{y}\mathbf{1} + \sum_{m=1}^M \hat{\theta}_m \mathbf{z}_m$$

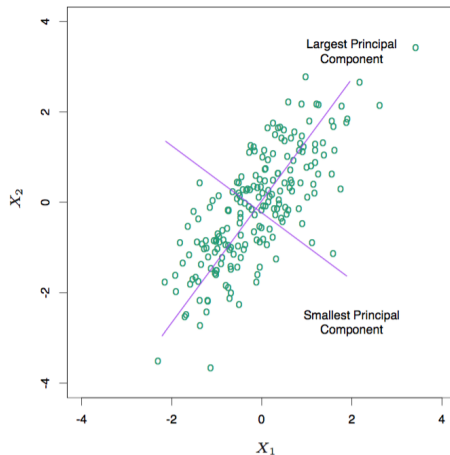
- Because principal components are orthogonal we can find coefficients using univariate regression  $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$ .
- We can recover the  $x$  coefficients because the PCA is a linear transformation:

$$\hat{\beta}^{PCR} = \sum_{m=1}^M \hat{\theta}_m v_m$$

## Principal Components (and Ridge)

- If  $M = P$  (we use all components) then PCR = OLS.
- If  $M < P$  then we discard the  $p - M$  smallest eigenvalue components
- This is similar to ridge which shrinks  $\beta$ 's for components with small eigenvalues.
- Think about the variance matrix  $X'X/n$  or  $X'X = VD^2V'$ .
- First component (largest eigenvalue) is  $\mathbf{z}_1 = \mathbf{X}v_1 = \mathbf{u}_1d_1$ .
- Variance is  $Var(\mathbf{z}_1) = Var(\mathbf{X}v_1) = \frac{d_1^2}{N}$  ( $\mathbf{z}_1$  is first principal component of  $\mathbf{X}$ ).

# Principal Components



**FIGURE 3.9.** Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression

# Principal Components And Ridge

Consider the objective function that Ridge minimizes:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$$

And the solution (which addresses multicollinearity!)

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

# Principal Components And Ridge

Just like we can diagonalize (some) square matrices, we can take the **singular value decomposition** (SVD) of any matrix  $\mathbf{X}$  that is  $N \times p$

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

- $\mathbf{U}, \mathbf{V}$  are  $N \times p$  and  $p \times p$  orthonormal matrices ( $\mathbf{U}$  spans the column space, and  $\mathbf{V}$  spans the row space of  $\mathbf{X}$ .)
- $\mathbf{D}$  is a diagonal matrix  $p \times p$  with elements corresponding to the singular values of  $\mathbf{X}$ .
- If  $\mathbf{X}$  is a square, diagonalizable matrix the singular values are equal to the eigenvalues.



# Principal Components And Ridge

Now the least squares solution is simple

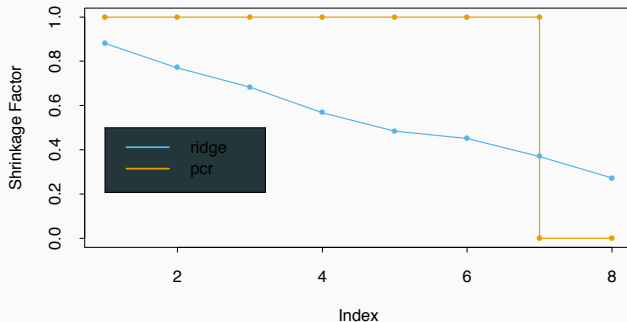
$$\mathbf{X}\hat{\beta}^{ols} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{U}\mathbf{U}'\mathbf{y}.$$

- $\mathbf{U}'\mathbf{y}$  are the values of  $\mathbf{y}$  mapped into the orthonormal basis  $\mathbf{U}$ .
- This looks a lot like  $\mathbf{QR}$  except that we have chosen a different basis.

Ridge is simple too:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{ridge} &= \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}'\mathbf{y} \\ &= \sum_{j=1}^p \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j' \mathbf{y}\end{aligned}$$

# Principal Components



**FIGURE 3.17.** Ridge regression shrinks the regression coefficients of the principal components, using shrinkage factors  $d_j^2/(d_j^2 + \lambda)$  as in (3.47). Principal component regression truncates them. Shown are the shrinkage and truncation patterns corresponding to Figure 3.7, as a function of the principal component index.

- This is the original GMM example, though it comes from macro-finance not microeconometrics

$$\begin{aligned} \max_{c_{t+i}, A_{t+i}} \quad & E_t \sum_{i=0}^{\infty} \frac{U(c_{t+i})}{(1+\delta)^i} \quad \text{subject to} \\ A_{t+i} \quad &= (1+r)A_{t+i-1} + y_{t+i} - c_{t+1} \\ 0 \quad &= \lim_{i \rightarrow \infty} E_t A_{t+i} (1+r)^{-i} \end{aligned}$$

- $A_t$  are your investment assets with return  $r$  and discount factor  $\delta$
- $y_t$  is your income in period  $t$ ,  $c_t$  is your consumption

## Hansen Singleton (1982)

- Assume CRRA utility with risk aversion  $\gamma$

$$U(c_{t+i}) = \frac{c_{t+i}^{1-\gamma}}{1-\gamma}$$

- We can take the first-order/Euler conditions and get:

$$E_t \left( \underbrace{\frac{1+r}{1+\delta} c_{t+1}^{-\gamma} - c_t^{-\gamma}}_{g(x_{t+i}, \theta)} \right) = 0$$

- We want to estimate the “deep parameters”  $\theta \equiv (\gamma, \delta)$ .

- We can solve for  $\theta$  without actually solving the dynamic programming problem!
- We just need some instruments  $z_t$  that are conditionally independent/orthogonal to  $g(x_{t+i}, \theta)$  so that

$$E_t[g(x_{t+i}, \theta)|z_t] = 0 \Rightarrow E_t[g(x_{t+i}, \theta)z_t] = 0$$

- This is **nonlinear GMM**. We need a matrix of instruments  $z_t$  with dimension  $N \times Q$  where  $Q \geq \dim(\theta)$ .
- Where do we get  $z_t$ ?  $\rightarrow$  Economic Theory!

## Hansen Singleton (1982)

Consider  $E_t[g(x_{t+i}, \theta)|z_t] = 0$ .

- The error arises from the error in the Euler equation: deviations between observed behavior and expected behavior.
- If the model is true this optimization error should be independent of anything known to the agent at the time the decision was made.
- We often write:  $E_t[g(x_{t+i}, \theta)|z_t, \Omega_t] = 0$  where  $\Omega_t$  is everything known by the agent up until time  $t$  (including the full history).
- If we have some potential instrument  $z_t$  and use the full history then  $z_{t-1}, z_{t-2}, \dots$  are all valid instruments
- If we use the conditional moment restriction  $E_t[g(x_{t+i}, \theta)|z_t, \Omega_t] = 0$  then any nonlinear function of  $z_t$  is also an instrument

$$E_t[g(x_{t+i}, \theta)f(z_{t,t-1}, \dots, 0)] = 0$$

## Hansen Singleton (1982)

- We have literally infinite possibilities to construct instruments  $z_t, z_t^2, z_t^3, z_t \cdot z_{t-1}, \dots$
- But our instruments could be **weak** even though we have many of them.
- And they might be highly correlated with each other.
- Carrasco (2012) suggests **regularization** on the instruments first.
- One possibility for  $f(z_t, z_{t-1}, z_{t-2}, \dots)$  is to take several higher order interactions and take the first  $Q$  principal components.
- Even though we might have 100 instruments, after running PCA we might find that 99% of our variation is only in 6 components. In that case we should not try and identify more than 6 parameters!
- Conlon (2014) suggests this as a test of non-identification in nonlinear BLP-type GMM problems.

# Factor Models

- Related to PCA is the **factor model**
- These are frequently used in finance for asset pricing.

$$r_i = b_0 + b_1 f_1 + b_2 f_2 + \dots b_p f_p + e_i$$

- Typically we choose factors so that  $E[f_i] = 0$  and  $E[f_i e_i] = 0$  and that  $Cov(f_i, f_j) = 0$  for  $i \neq j$ .
- That is we choose scaled factors to form an orthogonal basis (which makes pricing assets easier).
- Instead of choosing  $f$  to best explain  $X'X$  we choose it to best explain  $r$  by taking linear combinations of our  $X$ 's.
- CAPM is a single factor model (where the factor is the **market return**).
- Fama-French have expanded to a 5 factor model (book-to-market, market-cap, profitability, momentum)



## Factor Models: Other Examples

- *Eigenfaces* reduces your face to the first few eigenvalues – this is how face detection works!
- In psychometrics they use data from multiple tests to measure different forms of intelligence (mathematical reasoning, verbal, logical, spatial, etc.)
  - An old literature searched for general intelligence factor  $g$
  - Nobody can tell what the GMAT measured!
- In marketing PCA/factor analyses are used in the construction of *perceptual maps*
- Marketing practitioners use FA/PCA more than academics these days (guess: maybe?)

# 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  $N, p \rightarrow \infty$ .
- 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).

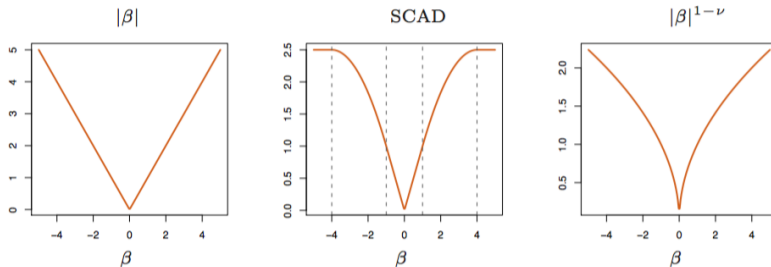
## Other Extensions

- *Grouped LASSO* for penalizing 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  $\lambda$  leading to less shrinkage and bias).
- *SCAD: Smoothly Clipped Absolute Deviation*: do less shrinkage on big coefficients but preserve sparsity

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

- *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.*