Chapter 15

Selection, Shrinkage, and Distillation

We start with more on selection ("hard threshold" – variables are either kept or discarded), and then we introduce shrinkage ("soft threshold" – all variables are kept, but parameter estimates are coaxed in a certain direction), and then lasso, which blends selection and shrinkage.

15.1 All-Subsets Model Selection I: Information Criteria

All-subsets model selection means that we examine every possible combination of K regressors and select the best. Examples include SIC and AIC.

Let us now discuss SIC and AIC in greater depth, as they are tremendously important tools for building forecasting models. We often could fit a wide variety of forecasting models, but how do we select among them? What are the consequences, for example, of fitting a number of models and selecting the model with highest R^2? Is there a better way? This issue of **model selection** is of tremendous importance in all of forecasting.

It turns out that model-selection strategies such as selecting the model with highest R^2 do *not* produce good out-of-sample forecasting models. Fortunately, however, a number of powerful modern tools exist to assist with model selection. Most model selection criteria attempt to find the model

with the smallest out-of-sample 1-step-ahead mean squared prediction error. The criteria we examine fit this general approach; the differences among criteria amount to different penalties for the number of degrees of freedom used in estimating the model (that is, the number of parameters estimated). Because all of the criteria are effectively estimates of out-of-sample mean square prediction error, they have a negative orientation – the smaller the better.

First consider the **mean squared error**,

$$MSE = \frac{\sum_{t=1}^{T} e_t^2}{T},$$

where T is the sample size and $e_t = y_t - \hat{y}_t$. MSE is intimately related to two other diagnostic statistics routinely computed by regression software, the **sum of squared residuals** and R^2. Looking at the MSE formula reveals that the model with the smallest MSE is also the model with smallest sum of squared residuals, because scaling the sum of squared residuals by $1/T$ doesn't change the ranking. So selecting the model with the smallest MSE is equivalent to selecting the model with the smallest sum of squared residuals. Similarly, recall the formula for R^2,

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = 1 - \frac{MSE}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

The denominator of the ratio that appears in the formula is just the sum of squared deviations of y from its sample mean (the so-called "total sum of squares"), which depends only on the data, not on the particular model fit. Thus, selecting the model that minimizes the sum of squared residuals – which as we saw is equivalent to selecting the model that minimizes MSE – is also equivalent to selecting the model that maximizes R^2.

Selecting forecasting models on the basis of MSE or any of the equivalent forms discussed above – that is, using in-sample MSE to estimate the out-of-sample 1-step-ahead MSE – turns out to be a bad idea. In-sample

MSE *can't* rise when more variables are added to a model, and typically it will fall continuously as more variables are added, because the estimated parameters are explicitly chosen to *minimize* the sum of squared residuals. Newly-included variables could get estimated coefficients of zero, but that's a probability-zero event, and to the extent that the estimate is anything else, the sum of squared residuals must fall. Thus, the more variables we include in a forecasting model, the lower the sum of squared residuals will be, and therefore the lower MSE will be, and the higher R^2 will be. Again, the sum of squared residuals can't rise, and due to sampling error it's very unlikely that we'd get a coefficient of exactly zero on a newly-included variable even if the coefficient is zero in population.

The effects described above go under various names, including **in-sample overfitting**, reflecting the idea that including more variables in a forecasting model won't necessarily improve its out-of-sample forecasting performance, although it will improve the model's "fit" on historical data. The upshot is that in-sample MSE is a downward biased estimator of out-of-sample MSE, and the size of the bias increases with the number of variables included in the model. In-sample MSE provides an overly-optimistic (that is, too small) assessment of out-of-sample MSE.

To reduce the bias associated with MSE and its relatives, we need to penalize for degrees of freedom used. Thus let's consider the mean squared error corrected for degrees of freedom,

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K},$$

where K is the number of degrees of freedom used in model fitting.[1] s^2 is just the usual unbiased estimate of the regression disturbance variance. That is, it is the square of the usual standard error of the regression. So selecting the model that minimizes s^2 is equivalent to selecting the model that minimizes

[1] The degrees of freedom used in model fitting is simply the number of parameters estimated.

the standard error of the regression. s^2 is also intimately connected to the R^2 adjusted for degrees of freedom (the "**adjusted** R^2," or $\bar{R}^2$). Recall that

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T-K)}{\sum_{t=1}^{T}(y_t - \bar{y})^2/(T-1)} = 1 - \frac{s^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2/(T-1)}.$$

The denominator of the $\bar{R}^2$ expression depends only on the data, not the particular model fit, so the model that minimizes s^2 is also the model that maximizes $\bar{R}^2$. In short, the strategies of selecting the model that minimizes s^2, or the model that minimizes the standard error of the regression, or the model that maximizes $\bar{R}^2$, are equivalent, and they do penalize for degrees of freedom used.

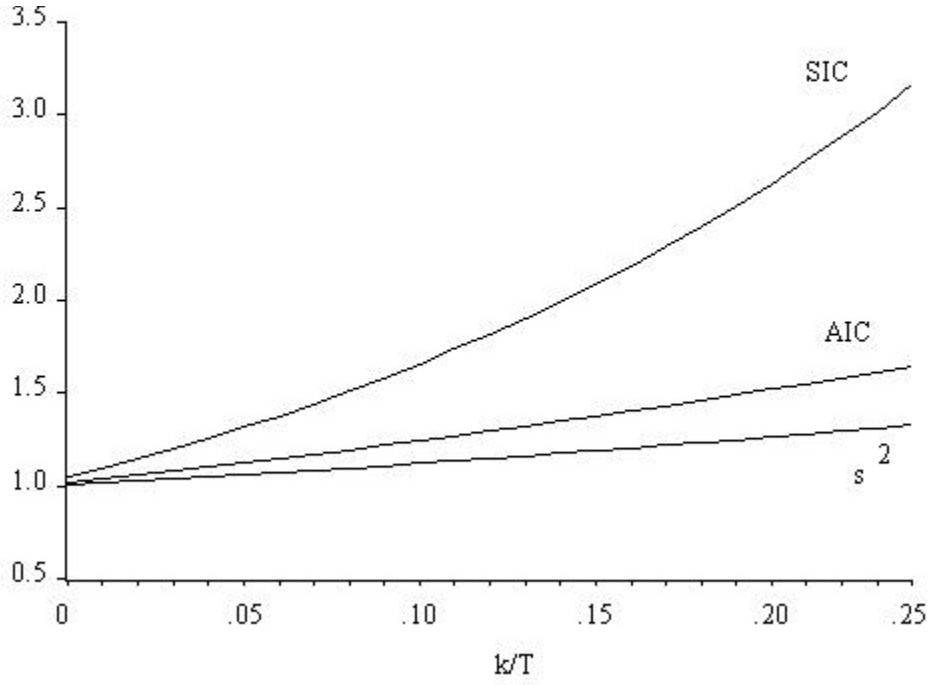
To highlight the degree-of-freedom penalty, let's rewrite s^2 as a penalty factor times the MSE,

$$s^2 = \left(\frac{T}{T-K}\right)\frac{\sum_{t=1}^{T} e_t^2}{T}.$$

Note in particular that including more variables in a regression will not necessarily lower s^2 or raise $\bar{R}^2$ – the MSE will fall, but the degrees-of-freedom penalty will rise, so the product could go either way.

As with s^2, many of the most important forecast model selection criteria are of the form "penalty factor times MSE." The idea is simply that if we want to get an accurate estimate of the 1-step-ahead out-of-sample forecast MSE, we need to penalize the in-sample residual MSE to reflect the degrees of freedom used. Two very important such criteria are the **Akaike Information Criterion** (AIC) and the **Schwarz Information Criterion** (SIC). Their formulas are:

$$AIC = e^{\left(\frac{2K}{T}\right)}\frac{\sum_{t=1}^{T} e_t^2}{T}$$



and

$$SIC = T^{\left(\frac{K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}.$$

How do the penalty factors associated with MSE, s^2, AIC and SIC compare in terms of severity? All of the penalty factors are functions of K/T, the number of parameters estimated per sample observation, and we can compare the penalty factors graphically as K/T varies. In Figure *** we show the penalties as K/T moves from 0 to .25, for a sample size of $T = 100$. The s^2 penalty is small and rises slowly with K/T; the AIC penalty is a bit larger and still rises only slowly with K/T. The SIC penalty, on the other hand, is substantially larger and rises much more quickly with K/T.

It's clear that the different criteria penalize degrees of freedom differently. In addition, we could propose many other criteria by altering the penalty. How, then, do we select among the criteria? More generally, what properties might we expect a "good" model selection criterion to have? Are s^2, AIC and SIC "good" model selection criteria?

We evaluate model selection criteria in terms of a key property called **consistency**, also known as the **oracle property**. A model selection criterion is consistent if:

a. when the true model (that is, the **data-generating process, or DGP**) is among a fixed set models considered, the probability of selecting the true DGP approaches one as the sample size gets large, and

b. when the true model is *not* among a fixed set of models considered, so that it's impossible to select the true DGP, the probability of selecting the best *approximation* to the true DGP approaches one as the sample size gets large.

We must of course define what we mean by "best approximation" above. Most model selection criteria – including all of those discussed here – assess goodness of approximation in terms of out-of-sample mean squared forecast error.

Consistency is of course desirable. If the DGP is among those considered, then we'd hope that as the sample size gets large we'd eventually select it. Of course, all of our models are false – they're intentional simplifications of a much more complex reality. Thus the second notion of consistency is the more compelling.

MSE is inconsistent, because it doesn't penalize for degrees of freedom; that's why it's unattractive. s^2 does penalize for degrees of freedom, but as it turns out, not enough to render it a consistent model selection procedure. The AIC penalizes degrees of freedom more heavily than s^2, but it too remains inconsistent; even as the sample size gets large, the AIC selects models that are too large ("overparameterized"). The SIC, which penalizes degrees of freedom most heavily, *is* consistent.

The discussion thus far conveys the impression that SIC is unambiguously superior to AIC for selecting forecasting models, but such is not the

case. Until now, we've implicitly assumed a fixed set of models. In that case, SIC *is* a superior model selection criterion. However, a potentially more compelling thought experiment for forecasting may be that we may want to expand the set of models we entertain as the sample size grows, to get progressively better approximations to the elusive DGP. We're then led to a different optimality property, called **asymptotic efficiency**. An asymptotically efficient model selection criterion chooses a sequence of models, as the sample size get large, whose out-of-sample forecast MSE approaches the one that would be obtained using the DGP at a rate at least as fast as that of any other model selection criterion. The AIC, although inconsistent, *is* asymptotically efficient, whereas the SIC is not.

In practical forecasting we usually report and examine both AIC and SIC. Most often they select the same model. When they don't, and despite the theoretical asymptotic efficiency property of AIC, this author recommends use of the more parsimonious model selected by the SIC, other things equal. This accords with the parsimony principle of Chapter 2 and with the results of studies comparing out-of-sample forecasting performance of models selected by various criteria.

The AIC and SIC have enjoyed widespread popularity, but they are not universally applicable, and we're still learning about their performance in specific situations. However, the general principle that we need somehow to inflate in-sample loss estimates to get good out-of-sample loss estimates *is* universally applicable.

The versions of AIC and SIC introduced above – and the claimed optimality properties in terms of out-of-sample forecast MSE – are actually specialized to the Gaussian case, which is why they are written in terms of minimized SSR's rather than maximized lnL's.[2] More generally, AIC and SIC are written not in terms of minimized SSR's, but rather in terms of

[2]Recall that in the Gaussian case SSR minimization and lnL maximization are equivalent.

maximized lnL's. We have:

$$AIC = -2lnL + 2K$$

and

$$SIC = -2lnL + KlnT.$$

These are useful for any model estimated by maximum likelihood, Gaussian or non-Gaussian.

15.2 All-Subsets Model Selection II: Cross Validation

Cross validation (CV) proceeds as follows. Consider selecting among J models. Start with model 1, estimate it using all data observations except the first, use it to predict the first observation, and compute the associated squared prediction error. Then estimate it using all observations except the second, use it to predict the second observation, and compute the associated squared error. Keep doing this – estimating the model with one observation deleted and then using the estimated model to predict the deleted observation – until each observation has been sequentially deleted, and average the squared errors in predicting each of the T sequentially deleted observations. Repeat the procedure for the other models, $j = 2, ..., J$, and select the model with the smallest average squared prediction error.

Actually this is "$T-fold$" CV, because we split the data into T parts (the T individual observations) and predict each of them. More generally we can split the data into M parts ($M < T$) and cross validate on them ("$M-fold$" CV). As M falls, M-fold CV eventually becomes consistent. $M = 10$ often works well in practice.

It is instructive to compare SIC and CV, both of which have the oracle property. SIC achieves it by penalizing in-sample residual MSE to obtain an approximately-unbiased estimate of out-of-sample MSE. CV, in contrast,

achieves it by directly obtaining an unbiased estimated out-of-sample MSE.

CV is more general than information criteria insofar as it can be used even when the model degrees of freedom is unclear. In addition, non-quadratic loss can be introduced easily. Generalizations to time-series contexts are available.

15.3 Stepwise Selection

All-subsets selection, whether by AIC, SIC or CV, quickly gets hard as there are 2^K subsets of K regressors. Other procedures, like the stepwise selection procedures that we now introduce, don't explore every possible subset. They are more ad hoc but very useful.

15.3.1 Forward

Algorithm:

– Begin regressing only on an intercept

– Move to a one-regressor model by including that variable with the smallest t-stat p-value

– Move to a two-regressor model by including that variable with the smallest p-value

– Move to a three-regressor model by including that variable with the smallest p-value

Often people use information criteria or CV to select from the stepwise sequence of models. This is a "greedy algorithm," producing an increasing sequence of candidate models. Often people use information criteria or CV to select from the stepwise sequence of models. No guaranteed optimality properties of the selected model.

"forward stepwise regression"

– Often people use information criteria or cross validation to select from the stepwise sequence of models.

15.3.2 Backward

Algorithm:

– Start with a regression that includes all K variables

– Move to a $K-1$ variable model by dropping the variable with the largest t-stat p-value

– Move to a $K-2$ variable model by dropping the variable with the largest p-value

Often people use information criteria or CV to select from the stepwise sequence of models. This is a "greedy algorithm," producing a decreasing sequence of candidate models. Often people use information criteria or CV to select from the stepwise sequence of models. No guaranteed optimality properties of the selected model.

15.4 One-Shot Estimation: Bayesian Shrinkage

Shrinkage is a generic feature of Bayesian estimation. The Bayes rule under quadratic loss is the posterior mean, which is a weighted average of the MLE and the prior mean,

$$\hat{\beta}_{bayes} = \omega_1 \hat{\beta}_{MLE} + \omega_2 \beta_0,$$

where the weights depend on prior precision. Hence the the Bayes rule pulls, or "shrinks," the MLE toward the prior mean.

A classic shrinkage estimator is **ridge regression**,,[3]

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

$\lambda \to 0$ produces OLS, whereas $\lambda \to \infty$ shrinks completely to 0. λ can be chosen by CV. (Notice that λ can *not* be chosen by information criteria, as K regressors are included regardless of λ. Hence CV is a more general

[3]The ridge regression estimator can be shown to be the posterior mean for a certain prior and likelihood.

selection procedure, useful for selecting various "tuning parameters" (like λ) as opposed to just numbers of variables in hard-threshold procedures.

15.5 One-Shot Estimation: Selection *and* Shrinkage

15.5.1 Penalized Estimation

Consider the penalized estimator,

$$\hat{\beta}_{PEN} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i|^q \right),$$

or equivalently

$$\hat{\beta}_{PEN} = argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2$$

s.t.

$$\sum_{i=1}^{K} |\beta_i|^q \leq c.$$

Concave penalty functions non-differentiable at the origin produce selection. Smooth convex penalties produce shrinkage. Indeed one can show that taking $q \to 0$ produces subset selection, and taking $q = 2$ produces ridge regression. Hence penalized estimation nests those situations and includes an intermediate case ($q = 1$) that produces the lasso, to which we now turn.

15.5.2 The Lasso

The lasso solves the L1-penalized regression problem of finding

$$\hat{\beta}_{LASSO} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right)$$

or equivalently

$$\hat{\beta}_{LASSO} = argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2$$

s.t.

$$\sum_{i=1}^{K} |\beta_i| \leq c.$$

Ridge shrinks, but the lasso shrinks *and* selects. Figure ?? says it all. Notice that, like ridge and other Bayesian procedures, lasso requires only *one* estimation. And moreover, the lasso uses minimization problem is convex (lasso uses the smallest q for which it is convex), which renders the single estimation highly tractable computationally.

Lasso also has a very convenient d.f. result. The effective number of parameters is precisely the number of variables selected (number of non-zero β's). This means that we can use info criteria to select among "lasso models" for various λ. That is, the lasso is another device for producing an "increasing" sequence of candidate models (as λ increases). The "best" λ can then be chosen by information criteria (or cross-validation, of course).

Elastic Net

$$\hat{\beta}_{EN} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left(\alpha |\beta_i| + (1-\alpha)\beta_i^2 \right) \right)$$

- A mixture of Lasso and Ridge regression; that is, it combines L1 and L2 penalties.
- Unlike Lasso, it moves strongly correlated predictors in or out of the model together, hopefully producing improving prediction accuracy relative to Lasso.
- Unlike Lasso, there are two tuning parameters in the elastic net λ and α.

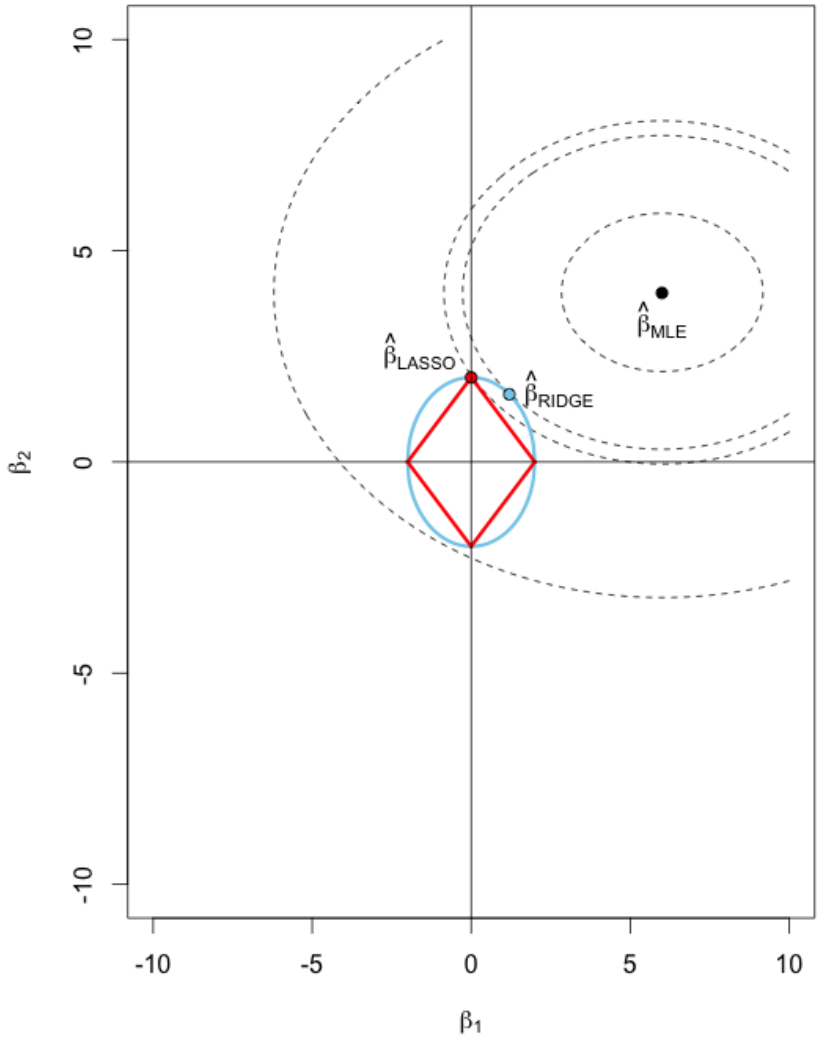


Figure 15.1: Lasso and Ridge Comparison

For $\alpha = 1$ elastic net turns into a Lasso model, For $\alpha = 0$ it is equivalent to ridge regression.

Adaptive Lasso

$$\hat{\beta}_{ALASSO} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i |\beta_i| \right),$$

where $w_i = 1/\hat{\beta}_i^{\nu}$, $\hat{\beta}_i$ is the OLS estimate, and $\nu > 0$.

- Every parameter in the penalty function is weighted differently, in contrast to the "regular" Lasso.
- The weights are calculated by OLS.
- Oracle property.

Adaptive Elastic Net

$$\hat{\beta}_{AEN} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_i \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i \left(\alpha |\beta_i| + (1-\alpha)\beta_i^2 \right) \right),$$

where $w_i = 1/\hat{\beta}_i^{\nu}$, $\hat{\beta}_i$ is the OLS estimate, and $\nu > 0$.

- A combination of elastic net and adaptive Lasso.
- Oracle property.

15.6 Distillation: Principal Components

15.6.1 Distilling "X Variables" into Principal Components

Data Summarization. Think of a giant (wide) X matrix and how to "distill" it.

$X'X$ eigen-decomposition:

$$X'X = VD^2V'$$

The j^{th} column of V, v_j, is the j^{th} eigenvector of $X'X$
Diagonal matrix D^2 contains the descending eigenvalues of $X'X$

First principal component (PC):
$z_1 = Xv_1$
$var(z_1) = d_1^2/T$
(maximal sample variance among all possible l.c.'s of columns of X)

In general:
$z_j = Xv_j \ \perp z_{j'}, j' \neq j$
$var(z_j) \leq d_j^2/T$

15.6.2 Principal Components Regression

The idea is to enforce parsimony with little information loss by regressing not on the full X, but rather on the first few PC's of X. We speak of "Principal components regression" (PCR), or "Factor-Augmented Regression".

Ridge regression and PCR are both shrinkage procedures involving PC's. Ridge effectively includes all PC's and shrinks according to sizes of eigenvalues associated with the PC's. PCR effectively shrinks some PCs completely to zero (those not included) and doesn't shrink others at all (those included).

15.7 Exercises, Problems and Complements

1. Information criteria in time-series environments.

   This chapter, and hence it's discussion of information criteria, emphasizes cross-section environments. We motivated SIC and AIC in terms of out-of-sample forecast MSE. Everything goes through in time-series

environments, but in time series there is also a horizon issue. SIC and AIC are then linked to *1-step-ahead* out-of-sample forecast MSE. Modifications for multi-step time-series forecasting are also available.

15.8 Notes