Statistics 8330: Data Analysis III
Linear Regression: Model Selection and Regularization (Shrinkage)

Suggested Reading: James, Witten, Hastie, and Tibshirani (JWHT), 2013, Chapter 6

Supplemental Reading: Hastie, Tibshirani, and Friedman (HTF), 2009, Chapter 3

Christopher K. Wikle

University of Missouri
Department of Statistics

Linear Regression

Recall, the point of linear regression is that we assume $E(Y|X)$ is linear in the inputs $X_1, \ldots, X_p$ (and linear in the parameters):

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

and we assume additive independent errors,

$$Y = f(X) + \epsilon.$$

In DA I we studied this model in quite some detail – particularly in the context of least squares (LS) solutions. In cases where the true relationship between inputs and outputs is approximately linear, this model has low bias and, if $n \gg p$, it also has low predictive variance and performs well with test data.

In situations where $n \approx p$, there is substantial variability and the model can over fit and perform poorly with test data. When $n < p$ there is no unique solution and the variance is infinite.

Improving Prediction Accuracy and/or Interpretability

In cases where the number of predictors is large, highly dependent, or we are interested in improving interpretability, we can apply modifications to the linear regression modeling procedure. Here, we consider the following three approaches, which are (perhaps) surprisingly quite related.

1. **subset selection:** choose a subset of the p predictors
2. **shrinkage:** fit all p predictors but force ("shrink") their parameters towards zero; this shrinkage (also known as "regularization") reduces variance and can also be used for variable selection in some cases
3. **dimension reduction:** project the p predictors into an M-dimensional subspace, $M < p$; i.e., compute M different linear combinations or "projections" of the p predictors and use these new variables as predictors

Subset Selection

CKW

We covered subset selection approaches in DA I. We briefly discuss some of these ideas again here to better establish their connection to the shrinkage and dimension reduction approaches.

Best Subset Selection: in this case we fit all p possible one predictor models, all $p(p-1)/2$ possible two predictor models, etc. Ultimately, there are 2^p possible models and it can be challenging to find the "best" one (using C_p, AIC, BIC, Adjusted R^2, or some cross-validation error measure.) An efficient algorithm (e.g., the *leaps and bounds algorithm* of Furnival and Wilson, 1974) allows one to consider best subsets for p as large as 30 or 40.

In general, it is typically computationally prohibitive to use cross-validation methods directly in the leaps-and-bounds approach, but one selects models based on training sample bias adjustments (e.g., C_p, AIC, etc.) and/or can apply cross-validation to the "best" k-variable model (as chosen from the training data) for $k = 0, \ldots, p$, AFTER running the best-subsets procedure.

Subset Selection (cont.)

Stepwise Selection: In cases with very large numbers of predictor variables, *stepwise* procedures may be a good alternative to best subset selection.

The **forward stepwise selection** procedure is one of the most common stepwise methods. It considers a much smaller number of models than the best subsets approach. This procedure starts with a model containing no predictors, then adds predictors to the model one-at-a-time, until all of the predictors are in the model. Critically, the variable added at each step gives the largest *additional* improvement to model fit. A total of $1 + p(p+1)/2$ models is fit – orders of magnitude smaller than for best subsets.

Again, this procedure is typically not implemented in the context of cross-validiation directly, but cross-validation is sometimes used on each of the best models (for $k = 0, \ldots, p$) that were selected from the training sample based on C_p, AIC, R^2_{adj}, etc.

This procedure is not guaranteed to give the "best" k-variable model for each k (nor, obviously, the best overall model).

Subset Selection (cont.)

CKW

Note that the forward stepwise approach can be used when $p > n$, but only models up to order $n - 1$ can be considered.

One can also consider *backward stepwise* procedures, that start with the LS model with all p predictors and then iteratively removes them one-at-a-time. (This procedure can't be applied when $p > n$).

Also, as we discussed in DA I, one can apply hybrid stepwise algorithms that consider removing variables that are no longer helpful after having added a new variable.

When applying cross-validation in the context of such model selection, one can account for the variability in the procedure by applying the *one-standard-error rule*. That is, one evaluates their cross-validation criteria (e.g., MSE) for each model size, and then selects the smallest model for which the estimated test error is within one standard error of the lowest point on the curve (i.e., select the more parsimonious model).

Cross-Validation Model Selection (HTF, 2009, Ch. 7.10)

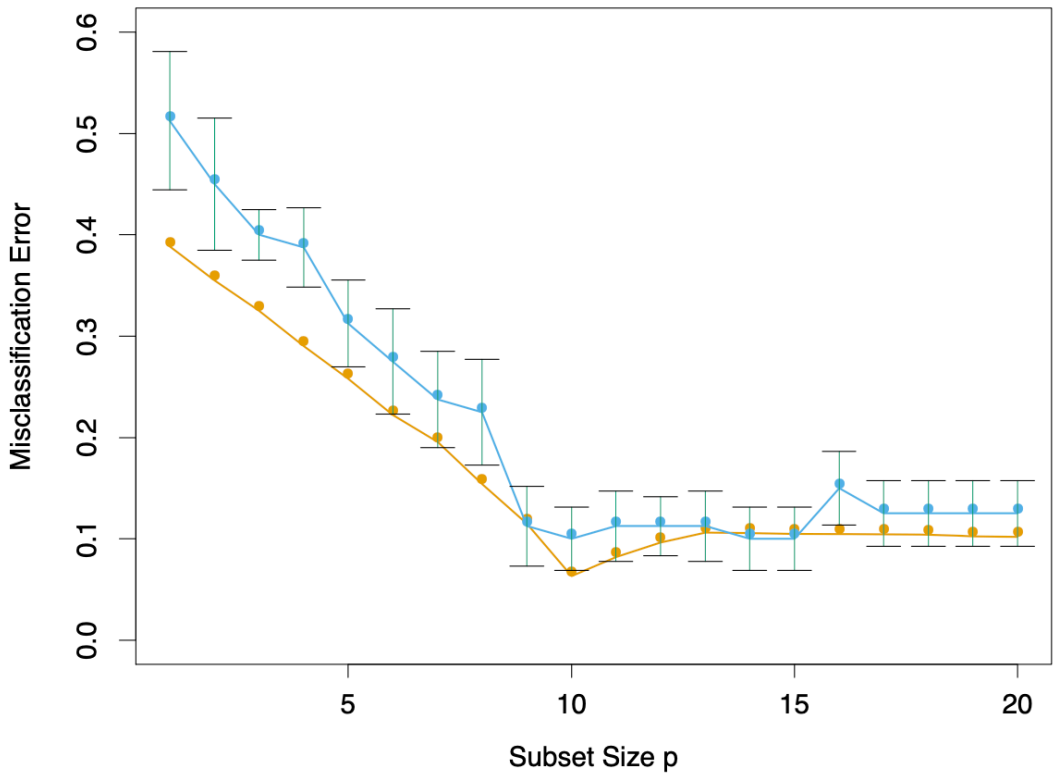

FIGURE 7.9. *Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3.*

Cross-Validation Model Selection (JWHT, 2013, Ch. 6.1)

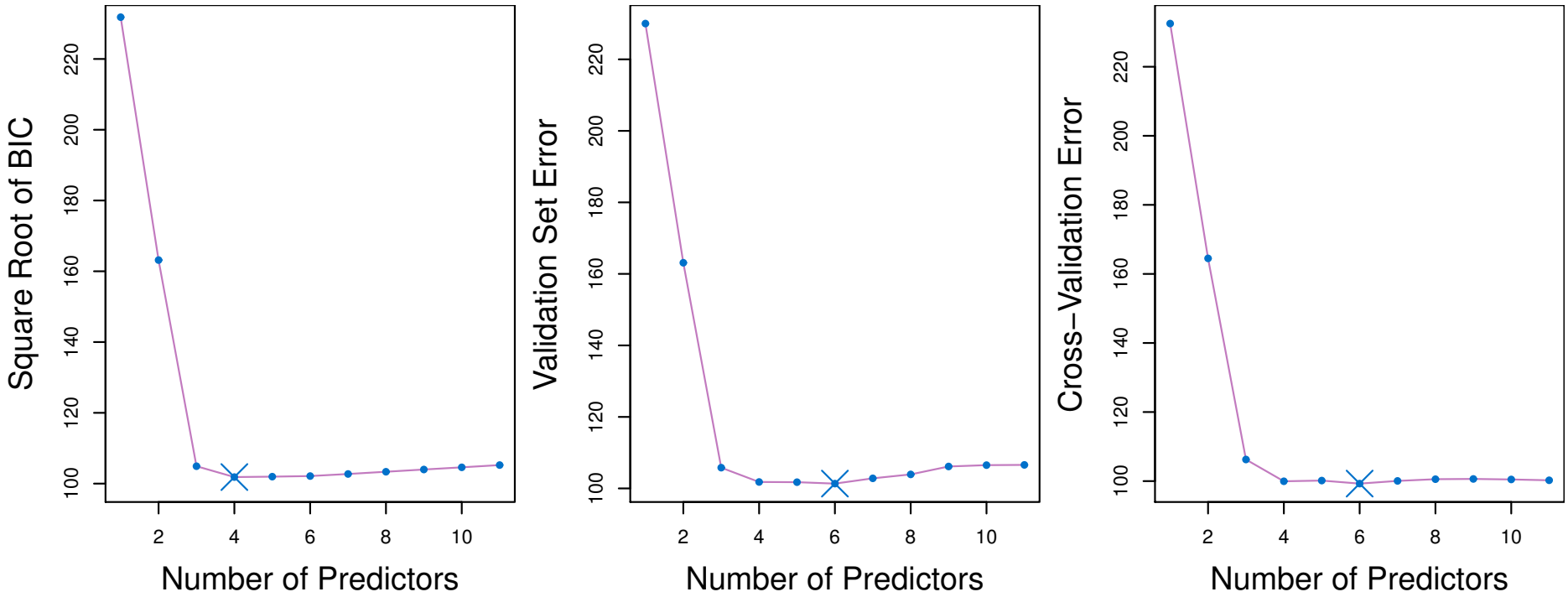

FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Shrinkage Methods

CKW

The subset selection methods attempt to trade some bias for variance reduction by removing variables. Another way to do this is to constrain the LS solution in a way that regularizes (shrinks) the coefficients towards zero. This leads to biased estimates, but tends to reduce variance. The two most famous such methods are *ridge regression* and the *lasso*.

Ridge Regression: You may recall from DA I that we can use ridge regression as a way to deal with multicollinearity in multiple regression.

Specifically, adding a constant to the diagonal of the $\mathbf{X}^T\mathbf{X}$ matrix allows us to get well-behaved (low variance) regression estimates even when $\mathbf{X}^T\mathbf{X}$ is nearly singular. Here, we will see that its use as a shrinkage estimator is more general than just addressing issues of multicollinearity.

Shrinkage Methods: Ridge Regression (cont.)

Recall that LS regression finds the estimates $\hat{\beta}$ that minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

In the case of ridge regression, we instead seek the estimates $\hat{\beta}^R$ that minimize:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where λ is a *tuning parameter*.

The second term in the objective function ($\lambda \sum_{j=1}^{p} \beta_j^2$) is an ℓ_2-norm *shrinkage penalty* that is small if the β are near zero. So, minimizing this objective function subject to this constraint forces $\hat{\beta}^R$ to be closer to zero than the LS estimates. Clearly, when $\lambda = 0$ we get the LS estimates and when $\lambda \to \infty$, the coefficients would go to zero.

Shrinkage Methods: Ridge Regression (cont.)

Note that we *don't shrink the intercept* and so we typically center the columns of the data matrix $\mathbf{X}$ by removing the mean of each column before we do the regression. In this case, note that $\hat{\beta}_0 = \bar{y}$. Recall, from DA I, we get

$$\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

and so the ridge solution is also a linear function of $\mathbf{y}$. In practice, one can use cross-validation to estimate λ as we will discuss below.

Note that the ridge estimator is NOT scale-invariant and so coefficients can change quite substantially depending on the scale of the covariates. For this reason, ridge regression is performed with scaled (and centered) covariates. That is, after centering we also divide each column of $\mathbf{X}$ by the standard deviation of that column.

Shrinkage Methods: Ridge Regression (cont.)

The key to ridge regression is that λ allows for a continuous range of possible bias-variance trade-offs. In general, as λ increases, the variance decreases but the bias increases. Thus, ridge regression typically works best when LS estimates have high variance. Consider the example from JWHT (2013, Chp. 6.2):

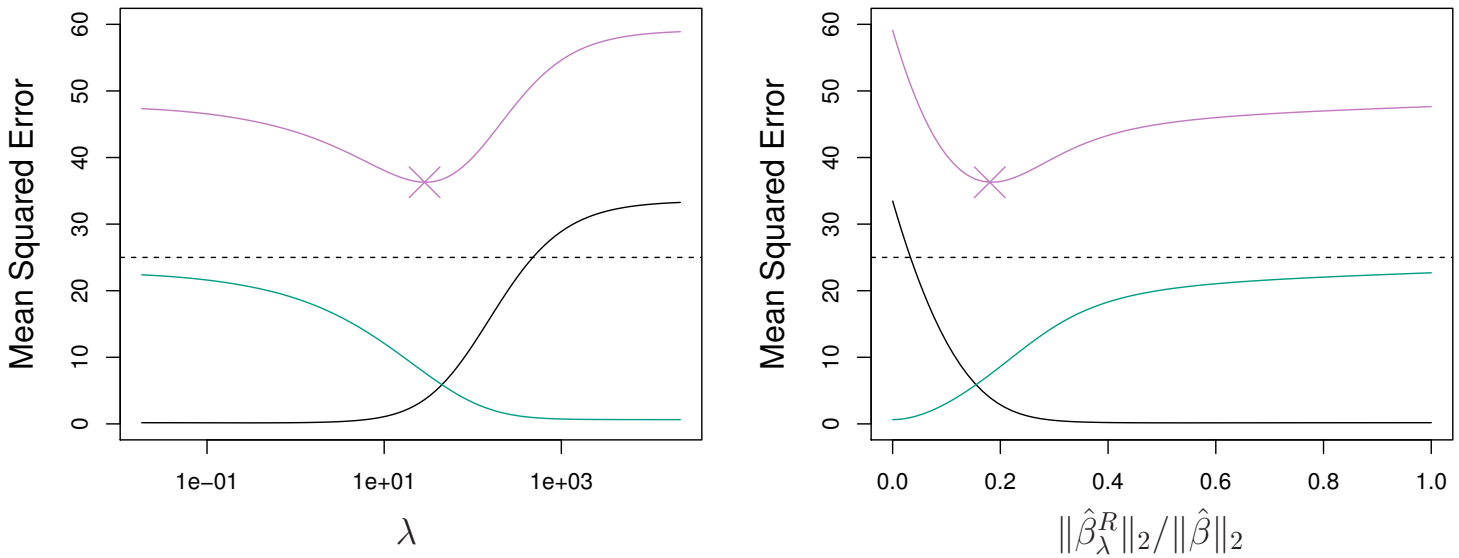

FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$*. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Shrinkage Methods: Ridge Regression (cont.)

A few final points about ridge regression:

- Like forward stepwise methods, ridge regression can be used in situations when $p > n$
- Ridge regression can be implemented very efficiently on a computer
- One can show that ridge regression corresponds to the posterior mean/mode of a normal error Bayesian regression model. That is, if $y_i \sim N(\beta_0 + x_i^T \beta, \sigma^2)$ is the "likelihood" in a Bayes setting, and the parameters have the independent prior distributions $\beta_j \sim N(0, \tau^2)$ (with τ^2, σ^2 assumed known), then one can show that $\lambda = \sigma^2/\tau^2$ (the ratio of regression to prior variances) and the ridge estimate is the posterior mode (and mean, since the posterior is Gaussian).
- The ridge estimates are also connected to principal component regression (as we will see). In particular, when one considers the principal components associated with the predictors, ridge regression shrinks the low variance principal components the most; see HTF (2009, 3.4.1) for details.

Shrinkage Methods: The Lasso

CKW

Although ridge regression shrinks the parameters to 0, unless $\lambda = \infty$ (unrealistic) it cannot set any of them *exactly* equal to 0. As in the subset selection case, we may actually want to facilitate model parsimony and interpretation by removing some variables (i.e., those that have a $\beta = 0$). The **lasso** (Tibshirani, 1996) is a fairly recent shrinkage-based alternative to ridge regression that allows for this.

The lasso coefficients, $\hat{\beta}^L$, are those that minimize the objective function:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Critically, the difference between the lasso and ridge regression is that the penalty term now includes $|\beta_j|$ as opposed to β_j^2. That is, lasso considers an ℓ_1 norm penalty instead of an ℓ_2 penalty. (Note: the ℓ_1 norm is given by: $||\beta||_1 = \sum_j |\beta_j|$; the ℓ_2 norm is given by: $||\beta||_2 = \sqrt{\sum_j \beta_j^2}$.)

Shrinkage Methods: The Lasso (cont.)

As does ridge regression, the lasso shrinks the regression coefficients towards 0. But, it also has the effect of forcing some coefficients to be exactly zero when the tuning parameter is large enough. Thus, it performs *variable selection* and produces models that are more *sparse* (or, parsimonious) than the LS fit.

Unfortunately, unlike ridge regression, there is no closed form for the lasso estimates $\hat{\beta}^L$ and these must be found through numerical optimization. As with ridge regression, the choice of λ is important and can be facilitated by cross-validation (see below).

Shrinkage Methods: The Lasso (cont.)

It is instructive to consider how the lasso, ridge regression and best subset regression are related. Note that we can write all three as the following constrained minimization problems, respectively:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s,$$

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s,$$

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s,$$

Thus, these all seek to minimize the RSS subject to constraints on the parameters. (Note: the last constraint is that no more than s coefficients can be nonzero and is not computationally feasible for large p).

Shrinkage Methods: The Lasso (cont.)

Consider the illustration from JWHT (2013, 6.2) for why the lasso can give 0 estimates for the case of $p = 2$:

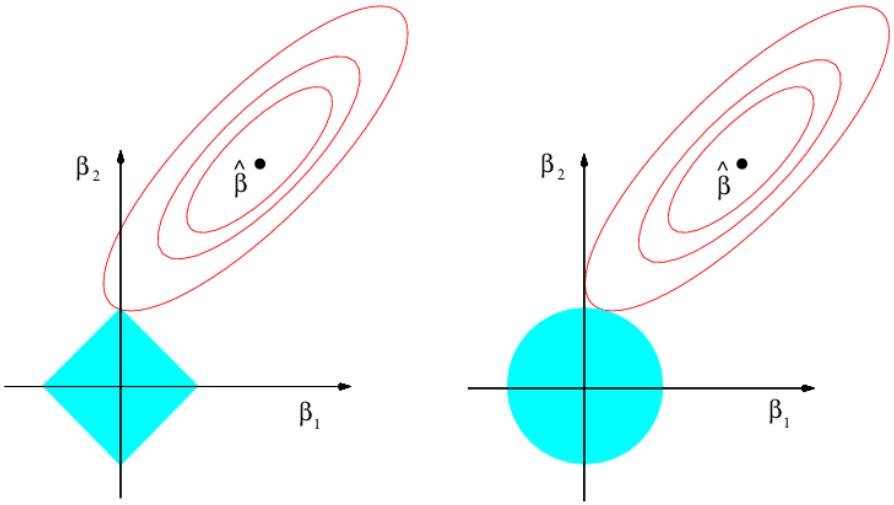

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

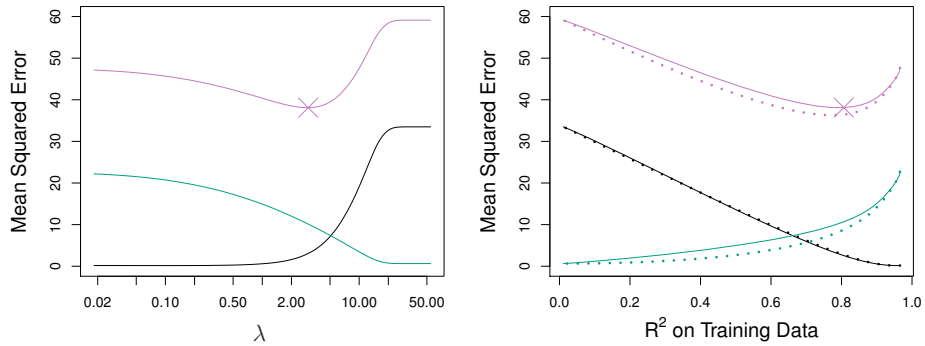

FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Shrinkage Methods: The Lasso (cont.)

CKW

Some additional comments about the lasso.

- In cases where some predictors are not needed, the lasso typically outperforms ridge regression; in situations where all predictors should be in the model, ridge regression typically outperforms the lasso
- The lasso tends to shrink each parameter by approximately the same amount, so that small coefficients get zero'd out, whereas ridge regression tends to shrink proportionally
- Like ridge regression, the lasso algorithm can be made computationally quite efficient
- Like ridge regression, the lasso also has an equivalent Bayesian interpretation; if the prior distribution on the parameters β_j are independent double exponential (Laplace) distributions with mean zero and scale parameter λ, then the posterior mode for β is the lasso solution (but not the posterior mean!); the Laplace prior puts more "mass" at zero, giving more *a priori* belief that the coefficients are zero

Shrinkage Methods: Cross-Validation Fitting $\mathcal{CKW}$

Cross-validation is a useful way to choose the shrinkage parameter (λ) in ridge regression and the lasso. The procedure is simple and given below.

1. Choose a "grid" of λ values
2. Compute the cross-validation error for each value of λ
3. Select the λ for which the cross-validation error is smallest (recall the one-standard error rule)
4. Refit the model using all available observations for the selected tuning parameter

Shrinkage Methods: Generalization

We considered an ℓ_2 and ℓ_1 penalty for ridge regression and the lasso, respectively. Not surprisingly, there is no reason why we couldn't generalize this to minimize the following objective function:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q,$$

for $q \geq 0$, where $q = 0$, $q = 1$, $q = 2$ correspond to subset selection, the lasso, and ridge regression, respectively.

In principle, one could estimate q from the data. However, HTF (2009, 3.4.3) suggest that it isn't really worth it. Rather, they suggest alternative hybrid penalties (e.g., the *elastic net penalty*, which we will see later) that are compromises between ridge regression and the lasso.

Shrinkage Methods: Generalization (cont.)

As we mentioned, the penalty component in the objective function is analogous to putting a prior distribution on the β parameters in a Bayesian regression (e.g., a normal prior is related to ridge regression and a Laplace prior is related to lasso).

There is a separate literature that has evolved that considers model selection in the Bayesian context. One such approach is **stochastic search variable selection (SSVS)** (e.g., George and McCulloch, 1993,1997).

In its classic form, SSVS corresponds to a hierarchical mixture of normal distributions:

$$\beta_j|\gamma_j \sim \gamma_j N(0, c_j\tau_j^2) + (1-\gamma_j)N(0,\tau_j^2),$$

$$\gamma_j \overset{iid}{\sim} \text{Bern}(\pi_j),$$

where $\gamma_j = 1$ indicates that the jth variable is in the model (with probability π_j). Typically, one wants τ_j to be small so that when $\gamma_j = 0$ it is reasonable to estimate β_j close to zero. Similarly, we want c_j to be large so that we are more likely to get a non-zero β_j when $\gamma_j = 1$. This is analogous to ridge regression.

An alternative SSVS procedure replaces the second mixture distribution with a delta function at 0 (more analogous to lasso).