Model selection, shrinkage, and regularization

December 29, 2024

Table of contents

What is model selection?

- So far, we have fit *linear models* using *least squares*.
- Extensions are possible in the direction of
  - using *nonlinear models* or
  - using fitting procedures *other than least squares*
  - or both.
- This lecture is about improving *linear models* with *alternative fitting procedures* that are jointly referred to as *linear model selection*.

We focus on two classes of model selection methods

Let p be the number of *predictors/independent variables/features/inputs*[1] in a linear model of the type $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$.

1. *Subset selection* methods

Select a subset of the p predictors. Then estimate the reduced model using *least squares*.

2. *Shrinkage/regularization* methods

Estimate the model on all p predictors using a fitting procedure that shrinks coefficients towards zero. Some such methods set very small coefficient estimates to zero exactly, effectively removing variables from the model (*variable selection/feature selection*).

[1] Recall that these are all broadly the same—and that so are *response/dependent variable/labels/output*—but have different nuances.

We focus on two classes of model selection methods

Let p be the number of *predictors/independent variables/features/inputs*[1] in a linear model of the type $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$.

1. *Subset selection* methods

Select a subset of the p predictors. Then estimate the reduced model using *least squares*.

2. *Shrinkage/regularization* methods

Estimate the model on all p predictors using a fitting procedure that shrinks coefficients towards zero. Some such methods set very small coefficient estimates to zero exactly, effectively removing variables from the model (*variable selection/feature selection*).

[1] Recall that these are all broadly the same—and that so are *response/dependent variable/labels/output*—but have different nuances.

Subset selection methods

Best subset and *stepwise* model selection

Within the *subset selection* class of methods, we cover the *best subset* and the *stepwise* model selection methods.

Subset selection methods

Best subset selection

The *best subset selection* algorithm

Algorithm

1. Start with a model[2] containing no predictors and predicting simply the sample mean for each observation, so that $Y = \bar{Y} + \epsilon$. Call this the *null model* and represent it by $\mathcal{M}_0$.
2. For $k = 1, 2, \ldots p$:
   - 2.1 Fit all $\binom{p}{k}$ models that contain exactly k predictors.[3]
   - 2.2 Pick from among the $\binom{p}{k}$ models the one with the largest *coefficient of determination* R^2 and call it $\mathcal{M}_k$.
3. From among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, select a single best model using model selection criteria.[4]

[2] We will only apply best subset selection for least squares estimation of linear regression models. The method can be extended to other types of models.

[3] Recall that the binomial coefficient $\binom{p}{k} = \frac{p!}{k!(p-k)!}$, which should help you recall that this number can get large.

[4] C_p, *AIC*, *BIC*, and *adjusted* R^2 (see Slide 18). For now, it is enough that these are goodness-of-fit measures.

Example: *Best subset selection* applied to `Credit` data set

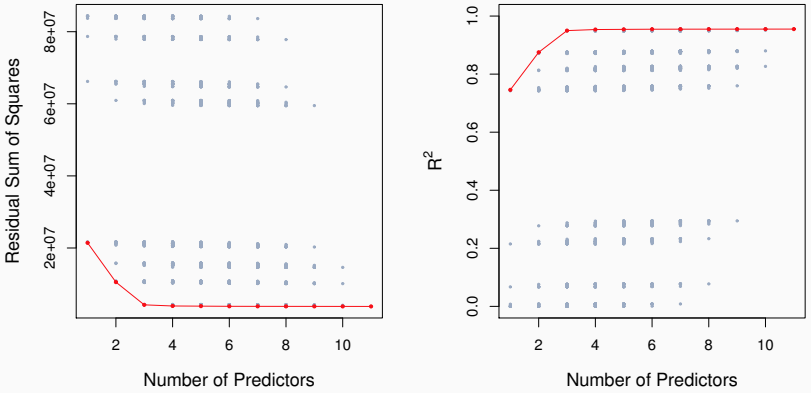

Figure 1: For each possible model containing a subset of the predictors in the `Credit` data set, the *RSS* and R^2 are displayed.

Fill in the blanks

1. The number of total predictors in the original dataset is denoted by _.
2. The model containing none of the predictors is called the _____.
3. The higher the residual sum of squares (RSS), the _____ the model's fit.

True or false?

1. __ Subset selection methods do not use least squares.
2. __ The null model ($\mathcal{M}_0$) has zero predictors and uses only the mean of the response variable to predict the response variable.

Fill in the blanks

1. The number of total predictors in the original dataset is denoted by p.
2. The model containing none of the predictors is called the null model.
3. The higher the residual sum of squares (RSS), the worse the model's fit.

True or false?

1. F Subset selection methods do not use least squares. *Subset selection methods do not use least squares on the full set of predictors. They do use least squares on a subset of predictors.*
2. T The null model ($\mathcal{M}_0$) has zero predictors and uses only the mean of the response variable to predict the response variable.

Subset selection methods

Stepwise selection

Problems with *best subset selection* (1/2)

Combinatorial explosion

If *best subset selection* estimates all possible models of a given size, why can't we always use it?

- For each value of $k = 1, 2, \ldots, p$, the best subset selection algorithm estimates $\binom{p}{k}$ models, for a sum total of $\sum_{k=1}^{p} \binom{p}{k} = 2^p - 1$ models.
- This is only computationally feasible for a small or moderate number of total predictors p because of *combinatorial explosion*.

p	0	1	2	3	5	10	25	100	300	...
$2^p - 1$	0	1	3	7	31	1,023	33,554,431	1.27×10^{30}	2×10^{90}	...

Table 1: Combinatorial explosion in the *best subset selection* algorithm. There are about 6×10^{79} atoms in the universe.

Problems with *best subset selection* (2/2)

Overfitting

- When p is large, best subset selection is more likely to lead to overfitting.
- This leads to a high variance of the coefficient estimates.
- For this reason, and because of the *combinatorial explosion* discussed on the previous slide, best subset selection is not always suitable.

Alternatives

Stepwise selection methods provide algorithms that explore a much smaller set of models, which helps alleviate these problems.

The intuition of *forward stepwise selection*

- *Forward stepwise selection* starts with the null model $\mathcal{M}_0$ and progressively adds predictors to it until all available predictors have been included.
- At each step, the variable that contributes the most to model fit is added.

The *forward stepwise selection* algorithm

Algorithm

1. Start with the *null model* $\mathcal{M}_0$ containing no predictors and predicting simply the sample mean for each observation, so that $Y = \bar{Y} + \epsilon$.
2. For $k = 0, \ldots, p-1$:
   - 2.1 Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ by one additional predictor.
   - 2.2 Select from among the $p - k$ candidate models the one with the largest *coefficient of determination* R^2 and call it $\mathcal{M}_{k+1}$.
3. From among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, select a single best model.

Advantages and disadvantages

Stepwise selection is very efficient, but not guaranteed to find the best possible from among the 2^p possible model specifications.

Example: *Best subset* and *forward stepwise* selection

p	Best subset selection	Forward stepwise selection
1	`rating`	`rating`
2	`rating`, `income`	`rating`, `income`
3	`rating`, `income`, `student`	`rating`, `income`, `student`
4	`cards`, `income`, `student`, `limit`	`rating`, `income`, `student`, `limit`

Table 2: Variable sets selected using *best subset* and *forward stepwise* selection for the `Credit` data set for model sizes of $p \in \{1, 2, 3, 4\}$ predictors. The models selected by the two methods are identical for $p \in \{1, 2, 3\}$ and start to differ for $p = 4$.

The intuition of *backward stepwise selection*

- *Backward stepwise selection* solves the problems associated with *best subset selection* in much the same way that *forward stepwise selection* does; that is, by operating on a reduced set of candidate models.
- Instead of proceeding from the null model and adding predictors, *backward stepwise selection* begins with the full least-squares model with all p predictors and then progressively removes predictors.
- At each step, the predictor that contributes the least to model fit is removed.

The *backward stepwise selection* algorithm

Algorithm

1. Start with the *full model* $\mathcal{M}_p$ containing all p predictors, so that $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_P + \epsilon$.
2. For $k = p, p-1, \ldots, 1$:
   - 2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - 2.2 Select from among the k models the one with the largest *coefficient of determination* R^2 and call it $\mathcal{M}_{k-1}$.
3. From among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, select a single best model.

More on *backward stepwise selection*

- Like *Forward stepwise selection*, *backward stepwise selection* estimates only $1 + p(p+1)/2$ models instead of the 2^p models estimated using *best subset selection*.
- It, too, is not guaranteed to result in the *best* model containing subsets of the p predictors.
- Because *backward stepwise selection* proceeds from the full model where $k = p$, it requires that the sample size n be larger than the number of variables p so that the full model can be estimated. By contrast, *forward stepwise selection*, which proceeds from $k = 1$ can even be used when $n < p$, in which case $k = 1, \ldots, p$ becomes $k = 1, \ldots, n$.

True or false?

1. __ Stepwise selection and best subset selection will *always* yield different models for a given model size p.
2. __ Stepwise selection and best subset selection will *often* yield different models for a given model size p.
3. __ When best subset selection is computationally feasible, it should be preferred to stepwise selection.

True or false?

1. F Stepwise selection and best subset selection will *always* yield different models for a given model size p.
2. T Stepwise selection and best subset selection will *often* yield different models for a given model size p.
3. T When best subset selection is computationally feasible, it should be preferred to stepwise selection.

Subset selection methods

Choosing the optimal model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$

Choosing the optimal model

- The full model $\mathcal{M}_p : Y = \beta_0 + \sum_{k=1}^{p} \beta_k X_k + \epsilon$ always has the largest R^2 because it includes all information contained in the predictor set.
- The R^2 is not a suitable criterion for selecting from a set of models with different numbers of predictors.
- The inclusion of additional variables always decreases the *training error* but leads to overfitting. The *training error* is not a good estimate of *test error*, which is what we really care about.
- *Test error* can be estimated either *indirectly* by *adjusting* the *training error* to account for the additional variables or *directly* using a validation set or cross-validation approach.

Estimating test error *indirectly*

Model selection criteria: C_p, *AIC*, *BIC*, and *adjusted* R^2

- Model selection criteria can be used in Step 3 of the *best subset selection*, *forward stepwise selection* and *backward stepwise selection* algorithms introduced above.[5]
- They adjust the *training error* upward to account for model size, for an estimate of the *test error*.

[5] See Slides 5, 11, and 14.

Example: Model selection criteria the `Credit` data set.

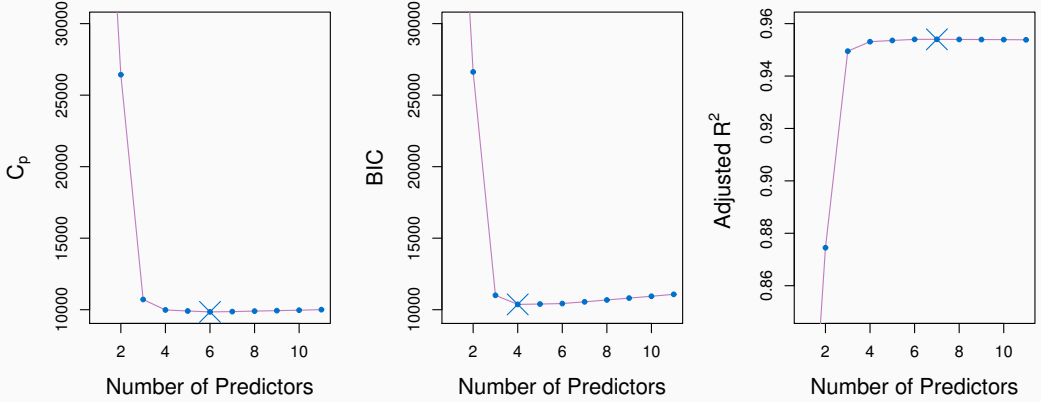

Figure 2: Model selection criteria for models produced by *best subset selection* on the `Credit` data set. The cross indicates the number of predictors in the final model selected under each criterion.

Definitions of model selection criteria: C_p

Mallow's C_p

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}\right)$$

where d is the number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.

Note that $\uparrow \text{RSS} \implies \uparrow C_p$, so that a high C_p implies a high RSS and thus a poor model fit.

We select the model with the lowest C_p.

Definitions of model selection criteria: C_p

Mallow's C_p

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}\right)$$

where d is the number of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.

Note that $\uparrow \text{RSS} \implies \uparrow C_p$, so that a high C_p implies a high RSS and thus a poor model fit.

We select the model with the lowest C_p.

Definitions of model selection criteria: AIC

Akaike Information Criterion (AIC)

$$\text{AIC} = -2 \log L + 2d$$

where L is the maximized value of the likelihood function for a model estimated using maximum likelihood.

Note that $\uparrow L \implies \downarrow \text{AIC}$, so that a lower AIC implies a higher likelihood L. Because the likelihood relates to the probability of drawing the sample obtained for a given set of parameters, a higher likelihood implies a better fit.

We select the model with the lowest AIC.

In the case of a linear model with Gaussian errors, *maximum likelihood* and *least squares* are equivalent and C_p and AIC lead to identical results.

Definitions of model selection criteria: AIC

Akaike Information Criterion (AIC)

$$\text{AIC} = -2 \log L + 2d$$

where L is the maximized value of the likelihood function for a model estimated using maximum likelihood.

Note that $\uparrow L \implies \downarrow \text{AIC}$, so that a lower AIC implies a higher likelihood L. Because the likelihood relates to the probability of drawing the sample obtained for a given set of parameters, a higher likelihood implies a better fit.

We select the model with the lowest AIC.

In the case of a linear model with Gaussian errors, *maximum likelihood* and *least squares* are equivalent and C_p and AIC lead to identical results.

Definitions of model selection criteria: AIC

Akaike Information Criterion (AIC)

$$\text{AIC} = -2 \log L + 2d$$

where L is the maximized value of the likelihood function for a model estimated using maximum likelihood.

Note that $\uparrow L \implies \downarrow \text{AIC}$, so that a lower AIC implies a higher likelihood L. Because the likelihood relates to the probability of drawing the sample obtained for a given set of parameters, a higher likelihood implies a better fit.

We select the model with the lowest AIC.

In the case of a linear model with Gaussian errors, *maximum likelihood* and *least squares* are equivalent and C_p and AIC lead to identical results.

Definitions of model selection criteria: BIC

Bayesian Information Criterion (BIC)

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n) d\hat{\sigma}^2\right)$$

Note that $\uparrow$ RSS $\implies \uparrow$ BIC, so that a high BIC implies a high RSS and thus a poor model fit. We select the model with the lowest BIC.

BIC replaces the $2d\hat{\sigma}^2$ used by C_p with $\log(n)d\hat{\sigma}^2$, where n is the number of observations. Because $\log(n) > 2$ for $n > 7$, the BIC generally places a heavier penalty on larger models, resulting in smaller models than C_p, as can be seen in Figure 2.

Definitions of model selection criteria: BIC

Bayesian Information Criterion (BIC)

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n) d\hat{\sigma}^2\right)$$

Note that $\uparrow$ RSS $\implies \uparrow$ BIC, so that a high BIC implies a high RSS and thus a poor model fit. We select the model with the lowest BIC.

BIC replaces the $2d\hat{\sigma}^2$ used by C_p with $\log(n)d\hat{\sigma}^2$, where n is the number of observations. Because $\log(n) > 2$ for $n > 7$, the BIC generally places a heavier penalty on larger models, resulting in smaller models than C_p, as can be seen in Figure 2.

Definitions of model selection criteria: Adjusted R^2

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

where TSS is the *total sum of squares* and the model has been estimated using *least squares*.

Note that $\uparrow$ RSS $\implies \downarrow$ Adjusted R^2, so that a high Adjusted R^2 implies a low RSS and thus a good model fit. We select the model with the highest Adjusted R^2.

Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease due to the presence of d in the denominator. The adjusted R^2 *pays a price* for the inclusion of unnecessary variables in the model.

Definitions of model selection criteria: Adjusted R^2

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

where TSS is the *total sum of squares* and the model has been estimated using *least squares*.

Note that $\uparrow$ RSS $\implies \downarrow$ Adjusted R^2, so that a high Adjusted R^2 implies a low RSS and thus a good model fit. We select the model with the highest Adjusted R^2.

Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease due to the presence of d in the denominator. The adjusted R^2 *pays a price* for the inclusion of unnecessary variables in the model.

Definitions of model selection criteria: Adjusted R^2

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

where TSS is the *total sum of squares* and the model has been estimated using *least squares*.

Note that $\uparrow$ RSS $\implies \downarrow$ Adjusted R^2, so that a high Adjusted R^2 implies a low RSS and thus a good model fit. We select the model with the highest Adjusted R^2.

Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease due to the presence of d in the denominator. The adjusted R^2 *pays a price* for the inclusion of unnecessary variables in the model.

Direct test error estimates

Validation, cross-validation

- Each of the procedures returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots$. Our job is to select the optimal number of predictors $\hat{k}$ and return $\mathcal{M}_{\hat{k}}$.
- We compute the *validation set error* or the *cross-validation error* for each candidate model $\mathcal{M}_k$ and then select the k for which the estimated test error is smallest.
- This procedure has an advantage relative to C_p, AIC, BIC, and adjusted R^2 in that it provides a direct estimate of the test error, and *doesn't require an estimate of the error variance* σ^2.
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g., the number of predictors in the model) or hard to estimate the error variance σ^2.

Example: Validation and cross validation

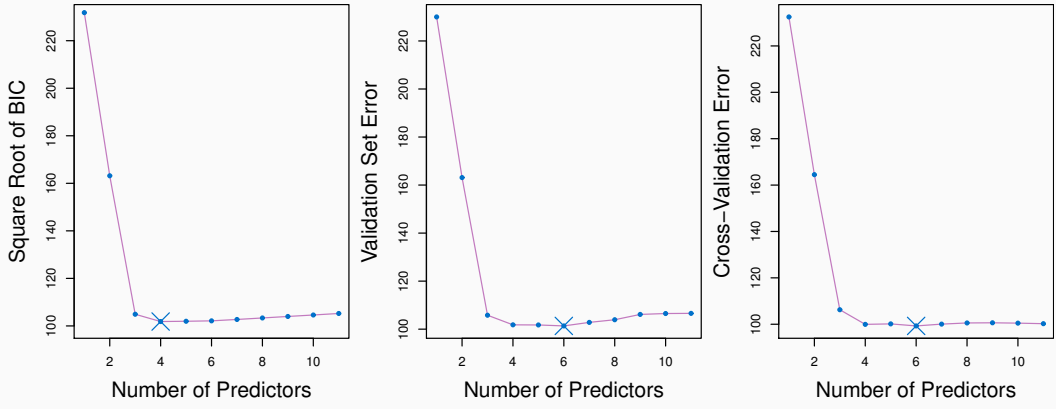

Figure 3: $\sqrt{\text{BIC}}$, validation set error, and cross-validation error for models resulting from *best subset selection* for `Credit` data.

Example: Validation and cross validation (explanations)

- The *validation set errors* were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set.
- The *cross-validation errors* were computed using $k = 10$ folds. In this case the validation and cross-validation methods both result in a six-variable model, but this need not always be the case.
- All three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Fill in the blanks

1. The ___ is the proportion of total variation in the dependent variable explained by the independent variables through the model.
2. The methods previously discussed produce one model per model size p. Using the R^2 to select from among them will always yield the _____ model.

True or false?

1. ___ Model selection criteria are an indirect way to calculate test error from training error.
2. ___ When using the AIC, the model with the highest AIC is the best.
3. ___ When using the Adjusted R^2, the model with the highest Adjusted R^2 is the best.

Fill in the blanks

1. The R^2 is the proportion of total variation in the dependent variable explained by the independent variables through the model.
2. The methods previously discussed produce one model per model size p. Using the R^2 to select from among them will always yield the largest model.

True or false?

1. T Model selection criteria are an indirect way to calculate test error from training error.
2. F When using the AIC, the model with the highest AIC is the best.
3. T When using the Adjusted R^2, the model with the highest Adjusted R^2 is the best.

Shrinkage methods

Shrinkage methods

- The subset selection methods covered above use *least squares* at each step of the algorithm to fit a model for a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that *constrains* or *regularizes* the coefficient estimates by *shrinking* them towards zero.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Shrinkage methods

Ridge regression

Ridge regression: Least squares revisited

Recap: *Least squares* estimation

$$\beta_0, \beta_1, \ldots, \beta_p = \arg\min_{\beta} RSS, \text{where}$$

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

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

That is, the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_p$ are those values that minimize the sum of the squared errors, or equivalently, the sum of the squared deviations of observed from predicted values of Y.

Ridge regression: Augmenting *least squares* by a penalty

Ridge regression coefficient estimates

$$\hat{\beta}_\lambda^R = \arg\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$= \arg\min_\beta \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter that can be chosen freely and controls the amount of shrinkage.

Ridge regression: Augmenting *least squares* by a penalty

Ridge regression coefficient estimates

$$\hat{\beta}_\lambda^R = \arg\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$= \arg\min_\beta \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is a tuning parameter that can be chosen freely and controls the amount of shrinkage.

Ridge regression: Interpretation of coefficients

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by minimizing the RSS.
- However the term $\lambda \sum_j \beta_j^2$, called a *shrinkage penalty*, is small when $\beta_1, \ldots, \beta_p$ are close to zero, which has the effect of *shrinking* the estimates of β_j towards zero.
- The *tuning parameter* λ controls the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value of λ is critical and is done using cross-validation.

Example: *Ridge regression* on the `Credit` data set

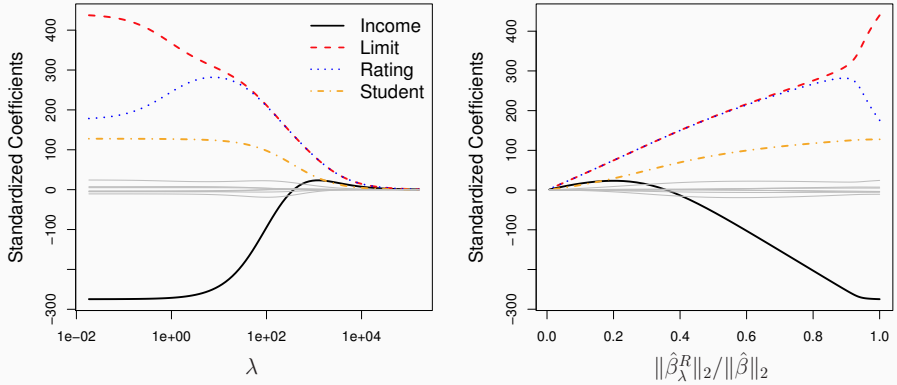

Figure 4: Standardized estimates of *ridge* regression coefficients for different values of λ for a model fit on all p variables of the `Credit` data set.

Example: *Ridge regression* on the `Credit` data set

Explanations

- In the left-hand panel, each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ.
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we now display $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$, where $\hat{\beta}$ denotes the vector of least squares coefficient estimates.
- The notation $||\beta||_2$ denotes the ℓ_2 norm (pronounced "ell two") of a vector, and is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

Ridge regression: scaling of predictors

- The least squares coefficient estimates are *scale equivariant*: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant, due to the sum of squares coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula

$$\tilde{x} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}} \tag{1}$$

Why does ridge regression yield improved estimates?

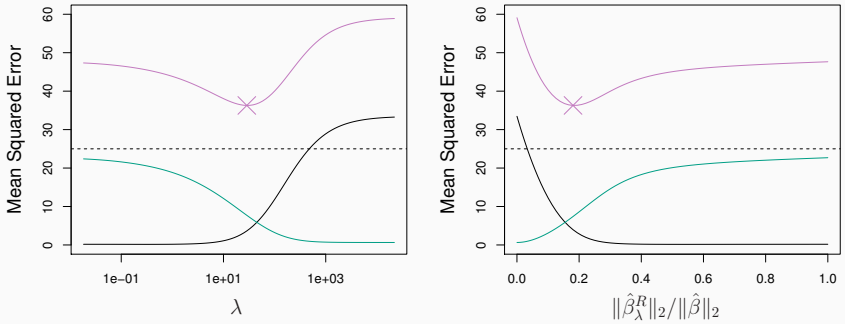

Figure 5: Simulated data with $n = 50$ observations, $p = 45$ predictors, all having nonzero coefficients. Squared bias (black), variance (green), and test mean squared error (purple) for ridge regression predictions as a function of λ and $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.

Shrinkage methods

The lasso

The *lasso*:[6] A sparse alternative to *ridge* regression

Lasso regression coefficient estimates

Unlike *subset selection*, *ridge* regression results in a model with all p predictors. The lasso overcomes this weakness by using an ℓ_1 penalty instead of an ℓ_2 penalty.

$$\hat{\beta}_\lambda^L = \arg\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= \arg\min_\beta RSS + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where the ℓ_1 ("ell one") norm of β is given by $||\beta||_1 = \sum |\beta_j|$.

[6] **L**east **a**bsolute **s**hrinkage and **s**election **o**perator

The *lasso*:[6] A sparse alternative to *ridge* regression

Lasso regression coefficient estimates

Unlike *subset selection*, *ridge* regression results in a model with all p predictors. The lasso overcomes this weakness by using an ℓ_1 penalty instead of an ℓ_2 penalty.

$$\hat{\beta}_\lambda^L = \arg\min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= \arg\min_\beta RSS + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where the ℓ_1 ("ell one") norm of β is given by $||\beta||_1 = \sum |\beta_j|$.

[6] **L**east **a**bsolute **s**hrinkage and **s**election **o**perator

The *lasso*: Interpretation of coefficients

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs *variable selection*.
- We say that the lasso yields *sparse* models—that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

Example: *Lasso regression* on the `Credit` data set

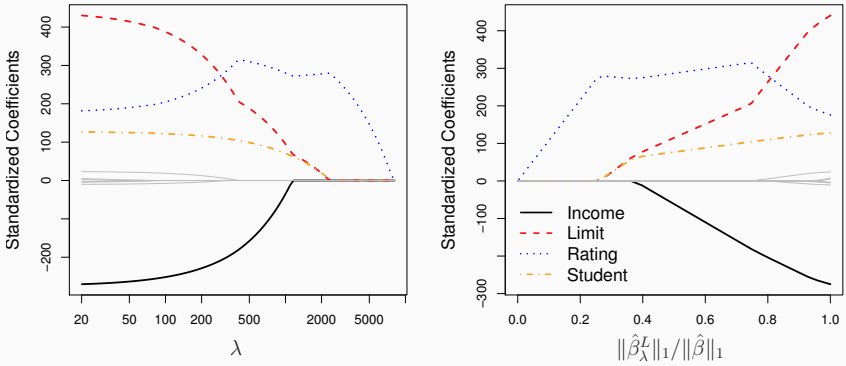

Figure 6: Standardized estimates of *lasso* coefficients for different values of λ for a model fit on all p variables of the `Credit` data set. Note how for large enough values of λ the coefficient estimates are set to zero exactly; a feature that estimates in Figure 4 lack.

Why does the lasso perform variable selection?

Why is is that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero? One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{s.t.} \quad \begin{cases} \sum_{j=1}^{p} |\beta_j| \leq s & \text{for } \textit{lasso} \\ \sum_{j=1}^{p} \beta_j^2 \leq s & \text{for } \textit{ridge} \end{cases}$$

The lasso picture

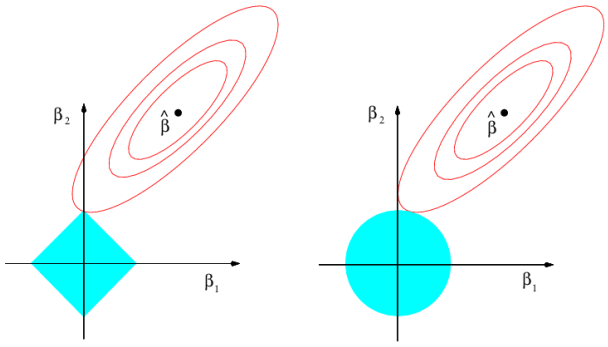

Figure 7: For $p = 2$ predictors, the constraint function implied by the penalty term and the RSS can be shown visually. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The blue areas are the constraint regions defined by $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$. The red ellipses are the contours of the RSS.

Comparing lasso and ridge regression

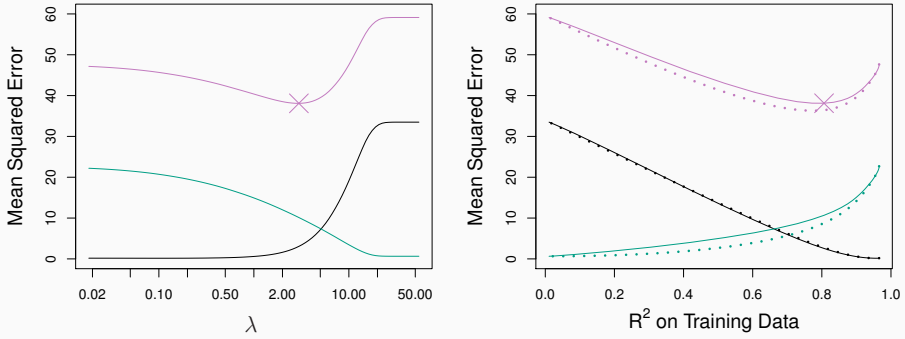

Figure 8: Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on simulated data set of Slide 35. Right: Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.

Comparing lasso and ridge regression (continued)

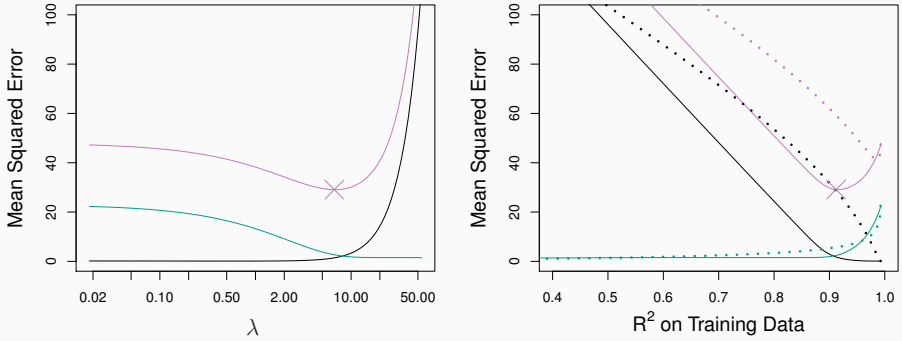

Figure 9: The simulated data is similar to that in Figure 8, except that now only two predictors are related to the response.

Conclusions

- These two examples illustrate that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known *a priori* for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

Fill in the blanks

1. Shrinkage methods are also called _____ methods.
2. The fact of limited model size is called _____.

True or false?

1. __ Instead of selecting a subset of the available variables, shrinkage uses a different estimator on the full set of variables.
2. __ Ridge regression provides shrinkage and model selection in one step by setting very small coefficients to zero.
3. __ A tuning parameter λ allows us to decide how much shrinkage we want to apply to the model. A lower λ will yield a model closer to a non-regularized model. A larger λ will yield a sparser model.

Fill in the blanks

1. Shrinkage methods are also called regularization methods.
2. The fact of limited model size is called sparsity.

True or false?

1. T Instead of selecting a subset of the available variables, shrinkage uses a different estimator on the full set of variables.
2. F Ridge regression provides shrinkage and model selection in one step by setting very small coefficients to zero.
3. T A tuning parameter λ allows us to decide how much shrinkage we want to apply to the model. A lower λ will yield a model closer to a non-regularized model. A larger λ will yield a sparser model.

Shrinkage methods

Selecting the tuning parameter

Selecting λ for ridge regression and lasso

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best.
- That is, we require a method for selecting a value for the tuning parameter λ or equivalently, the value of the constraint s.
- *Cross-validation* provides a simple way to tackle this problem. We choose a grid of λ values and compute the cross-validation error rate for each value of λ.
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Example: Selecting λ for *ridge* on the `Credit` data set

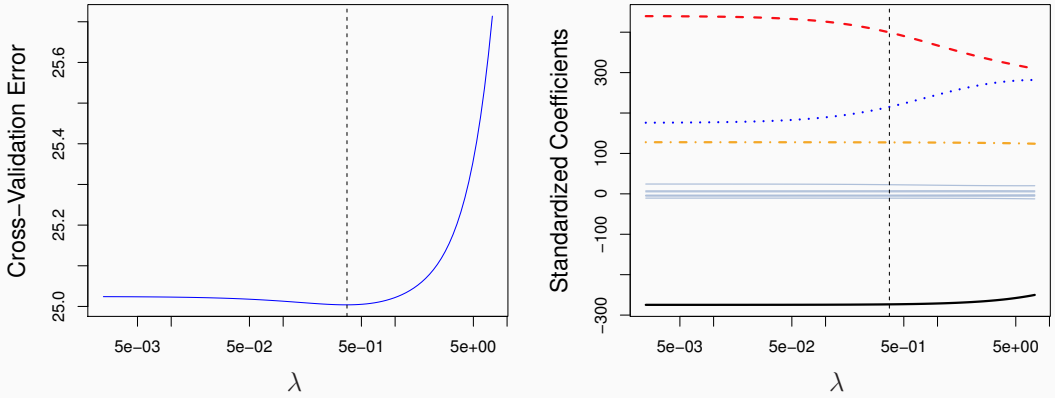

Figure 10: Left: Cross-validation errors that result from applying ridge regression to the `Credit` data set with various values of λ. Right: The coefficient estimates as a function of λ. The vertical dashes lines indicate the value of λ selected by cross-validation.

Example: Selecting λ for *lasso* on simulated data

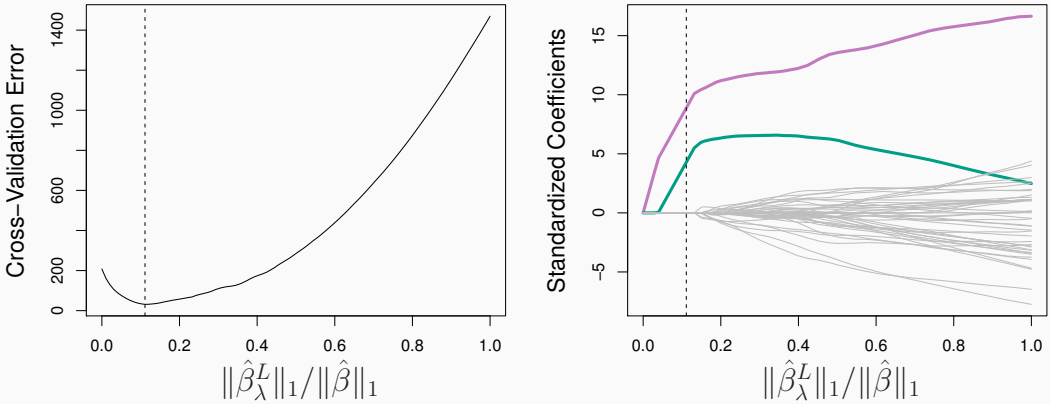

Figure 11: Left: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 9. Right: The corresponding lasso coefficient estimates are displayed. The vertical dashes lines indicate the lasso fit for which the cross-validation error is smallest.

Summary

Summary (1/2)

Model selection

- *Best Subset Selection*: This is the most exhaustive approach. It considers all possible combinations of predictors and selects the model with the best fit (usually evaluated using criteria like Akaike Information Criterion or Bayesian Information Criterion). While it's comprehensive, it's computationally expensive and practically infeasible with a large number of predictors due to combinatorial explosion.
- *Forward Stepwise Selection*: This is a more computationally manageable method. It starts with no predictors and adds them one at a time, each time choosing the predictor that provides the best fit to the model. It's less computationally intensive than best subset selection but can miss interactions between variables since it never evaluates all possible models.
- *Backward Stepwise Selection*: The opposite of forward selection. It starts with all predictors and systematically removes the least significant one at each step. This approach is good when you have a large number of predictors to start with, but like forward selection, it can miss important interactions.

Summary (2/2)

Shrinkage and Regularization (Lasso and Ridge Regression)

- *Ridge Regression*: It adds a penalty equal to the square of the magnitude of coefficients to the loss function. This method shrinks coefficients towards zero but doesn't set any to zero, which means it does not do feature selection but reduces overfitting.
- *Lasso Regression*: It uses an absolute value penalty which can shrink coefficients all the way to zero, thus performing feature selection. Lasso is useful when we believe many features are irrelevant or when we want a sparse model.

Each of these methods has its trade-offs. Best subset is comprehensive but impractical for many variables. Forward and backward selections are more computationally feasible but might miss important predictors. Ridge and Lasso introduce bias to reduce variance and overfitting, with Lasso providing the added advantage of feature selection.

The choice of method often depends on the specific context, like the number of predictors, computational resources, and the need for interpretability (sparse models like those from Lasso are easier to interpret). In practice, cross-validation is crucial to assess the performance of these models and avoid overfitting.

This material draws extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the lecture slides available from these authors.