QBUS3820: Statistical Learning and Data Mining

Lecture 4: Variable Selection and Regularization in Linear Regression

Semester 1, 2019

Discipline of Business Analytics, The University of Sydney Business School

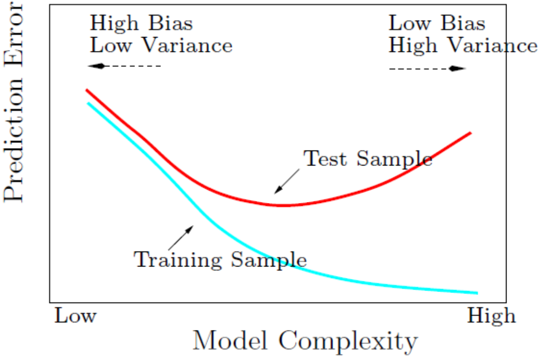
Lecture 4: Variable Selection and Regularisation

1. Review of the bias-variance trade-off
2. Model Selection vs Variable Selection
3. Subset selection methods
4. Regularisation methods
5. Comparisons and extensions

Review of the bias-variance trade-off

Bias-variance trade-off

In Lecture 1 we discussed the fundamental concept of the bias-variance trade-off for estimation.



Bias-variance trade-off

- Increasing model complexity brings greater flexibility and, therefore, lower bias. However, this comes at a cost of higher variance. Overfitting can be a problem.
- Decreasing model complexity leads to lower variance. However, simpler models may not be sufficiently flexible to capture the underlying patterns in the data, leading to higher bias.

Examples

Linear regression. Adding predictors increases model complexity. Least squares estimates have greater variability when the number of predictors is large. On the other hand, excluding relevant predictors leads to bias.

KNN regression. Reducing the number of neighbours increases model complexity. Closer neighbours means lower bias. However, the fact that we average fewer observations leads to increased variance.

Model selection methods

In the last lecture, we considered model selection methods: K-fold cross-validation, LOOCV and analytical criteria (C_p, AIC, BIC).

Recall:

- LOOCV and AIC generally pick similar models, especially when the sample size is large.
- The advantage of AIC over LOOCV is mainly computational.
- Cross-validation is universally applicable, while this is not the case for AIC.
- Cross-validation should be preferred to AIC when the assumptions of the model (e.g. constant error variance) are likely to be wrong.

Limitations of model selection

- Standard statistical inference (e.g. confidence intervals, p-values) is no longer valid after model selection.
- This is because standard inference assumes a fixed model, whereas model selection will by definition pick a specific model that fits the data well.
- In our context the way around this difficulty would be data splitting: using one part of the sample for model selection, and another for inference.

Model Selection vs Variable Selection

Equity premium prediction data example

Quarterly data from Goyal and Welch (2008), updated to 2015.

```
Response: quarterly S&P 500 returns minus treasury bill rate

Predictors (lagged by one quarter):
1. dp        Dividend to price ratio
2. dy        Dividend yield
3. ep        Earnings per share
4. bm        Book-to-market ratio
5. ntis      Net equity expansion
6. tbl       Treasury bill rate
7. ltr       Long term rate of return on US bods
8. tms       Term spread
9. dfy       Default yield spread
10.dfr       Default return spread
11.infl      Inflation
12.ik        Investment to capital ratio

Number of observations: 275 (1947-2015)
```

Complete subset regressions

- Suppose that we want to use linear regression to predict the equity premium. One option is to include all the $p = 12$ available predictors and estimate the model by OLS. However, the data are very noisy and this will lead to overfitting.

- The **complete subset regressions** (CSR) method is a simple and easy-to-understand algorithm that we can use to reduce overfitting.

- The CSR method fixes the model size k, then predicts the response by taking a simple average of the predictions produced by all possible linear regression models containing exactly k predictors. The number of such models is $\binom{p}{k}$, which is the number of all possible combinations of k out of p predictors.

Complete subset regressions

Algorithm Complete subset regressions

Set k.

Generate all the $S = \binom{p}{k}$ possible predictor subsets of size k.

for *subset* in all subsets **do**

  Estimate the model by OLS based on the predictors included in *subset*. Denote the estimated regression function by $\widehat{f}_{\text{subset}}$ (note: the function only uses the values of the predictors in *subset*).

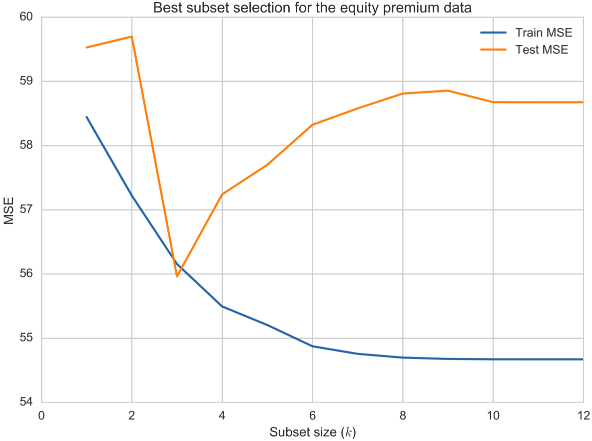
end for

The prediction for a new input vector $\boldsymbol{x}_0$ is

$$\widehat{y}_0 = (1/S) \sum_{\text{all subsets}} \widehat{f}_{\text{subset}}(\boldsymbol{x}_0)$$

Equity premium prediction

We use leave-one-out cross validation to select the optimal model complexity (k) for the CSR method.



Equity premium prediction

- This data is characterised by a low signal-to-noise ratio. The evidence in the literature shows that the predictability of the equity premium is low.
- The optimal subset size according to leave-one-out cross validation is $k = 4$. This value of k also gives the lowest test MSE.
- Using all the inputs leads to poor predictions: the test R^2 is only 0.014 (compared to 0.04 for the CSR method).
- CSR is a method of **model selection**, not a method of **variable selection**, as we use all possible ways of picking k out of all p variables and average over all those ways for each k.

Subset selection methods

OLS (review)

In the OLS method, we select the coefficient values that minimise the residual sum of squares

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ols}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Why might we not be satisfied with OLS?

Prediction accuracy. When the number of predictors, p, is large (e.g. comparable to n) OLS has high variance and tends to overfit. We can improve performance by setting some coefficients to zero or shrinking them. In other words, we will accept some bias in order to reduce variance.

Interpretability. A regression model with too many variables (including those not actually related to the response) is hard or impossible to interpret. By removing some of the variables, we can get a model that is more interpretable.

Possible alternatives to OLS

Subset selection. Identify a subset of $k < p$ predictors to use. Estimate the model by using OLS on the reduced set of variables.

Regularisation (shrinkage). Fit a model involving all p predictors, but shrink the coefficients towards zero relative to OLS. Depending on the type of shrinkage, some estimated coefficients may be zero, in which case the method also performs variable selection.

Best subset selection

The **best subset selection** method considers all possible subsets of predictors, estimates all of the corresponding models, and selects the best one according to a model selection criterion (cross-validation, AIC or BIC).

Given p predictors, there are 2^p possible models to choose from.

Best subset selection

For example, if $p = 3$ we would estimate $2^3 = 8$ models:

$$\begin{aligned}
k = 0: \quad Y &= \beta_0 + \varepsilon \\
k = 1: \quad Y &= \beta_0 + \beta_1 X_1 + \varepsilon \\
Y &= \beta_0 + \beta_2 X_2 + \varepsilon \\
Y &= \beta_0 + \beta_3 X_3 + \varepsilon \\
k = 2: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\
Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \\
Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\
k = 3: \quad Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon
\end{aligned}$$

Best subset selection

Algorithm Best subset selection

```
Estimate the null model $\mathcal{M}_0$, which contains only the intercept.
for $k = 1, 2, \ldots, p$ do
    Fit all $\binom{p}{k}$ possible models with exactly $k$ predictors.
    Pick the model with the lowest RSS and call it $\mathcal{M}_k$.
end for
Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross-validation, AIC, or BIC.
```

Computational considerations

The best subset method suffers from a problem of combinatorial explosion, because it requires estimation of 2^p different models. The computational requirement is therefore extremely high, except in low dimensions.

For example, for $p = 30$ we would need to fit a over 1 billion models!

Stepwise selection

Stepwise selection methods form a family of search algorithms that find promising subsets by sequentially adding predictors (**Forward** selection) or removing predictors (**Backward** selection).

These sequential approaches dramatically reduce the computational cost when compared to estimating all possible models.

Conceptually, they are approximations to best subset selection.

Forward selection

Algorithm Forward selection

Estimate the null model $\mathcal{M}_0$, which contains only the intercept.
for $k = 0, 1, \ldots, p - 1$ **do**
    Fit all the $p - k$ models that add **one** predictor to $\mathcal{M}_k$.
    Choose the best of these $p - k$ models in terms of RSS and call it $\mathcal{M}_{k+1}$.
end for
Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross-validation, AIC, or BIC.

Backward selection

Algorithm Backward selection

1: Estimate the full model $\mathcal{M}_p$ by OLS.

2: **for** $k = p, \ldots, 2, 1$ **do**

3: Fit all the k models that remove **one** predictor from $\mathcal{M}_k$.

4: Choose the best of these k models in terms of RSS and call it $\mathcal{M}_{k-1}$.

5: **end for**

6: Select the best model among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ according to cross-validation, AIC, or BIC.

Stepwise selection

- Compared to best subset selection, the forward and backward stepwise algorithms reduce the number of estimated models from 2^p to $1 + p(p+1)/2$. For example, for $p = 30$ the number of the models is 466.

- The disadvantage is that the final model selected by stepwise selection is not guaranteed to optimise any selection criterion among the 2^p possible models.

Subset selection

Advantages

- Accuracy relative to OLS. When the number of predictors is large, subset selection generally leads to better predictions compared to estimating a model with all the predictors.
- Interpretability. The final model is a linear regression model based on a reduced set of predictors.

Disadvantages

- Computational cost (in the case of best subset selection).
- By making binary decisions on whether to include or exclude particular variables, subset selection may exhibit higher variance than regularisation approaches.

Data: Equity Premium Prediction

Quarterly data from Goyal and Welch (2008), updated to 2015.

```
Response: quarterly S&P 500 returns minus treasury bill rate

Predictors (lagged by one quarter):
1. dp        Dividend to price ratio
2. dy        Dividend yield
3. ep        Earnings per share
4. bm        Book-to-market ratio
5. ntis      Net equity expansion
6. tbl       Treasury bill rate
7. ltr       Long term rate of return on US bods
8. tms       Term spread
9. dfy       Default yield spread
10.dfr       Default return spread
11.infl      Inflation
12.ik        Investment to capital ratio

Number of observations: 275 (1947-2015)
```

Illustration: Equity Premium Prediction

We select the following models in the equity premium dataset based on the AIC:

Best subset selection: (dy, bm, tms, dfr)

Forward selection: (ik, tms, dfr)

Backward selection: (dy, tms, dfr)

Illustration: Equity Premium Prediction

Table 1: Equity Premium Prediction Results

	Train R^2	Test R^2
OLS	0.108	0.014
Best Subset	0.095	0.038
Forward	0.083	0.042
Backward	0.084	0.060
CSRs	0.078	0.039

A potentially problematic way of doing variable selection

Removing statistically insignificant predictors (based on the p-values) is not a well-justified variable selection approach.

- A statistically significant coefficient means we can reliably say that it is not *exactly* zero. This does not directly translate into a statement about the predictive performance of a particular model.

- Furthermore, there are multiple-testing issues when predictors are removed sequentially (similar to the problem of doing inference after model selection, discussed earlier).

Illustration: Equity Premium Prediction

OLS Regression Results

Dep. Variable:	ret	R-squared:	0.108
Model:	OLS	Adj. R-squared:	0.051
Method:	Least Squares	F-statistic:	1.901
Date:		Prob (F-statistic):	0.0421
Time:		Log-Likelihood:	-629.21
No. Observations:	184	AIC:	1282.
Df Residuals:	172	BIC:	1321.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	26.1369	14.287	1.829	0.069	-2.064	54.337
dp	0.3280	8.247	0.040	0.968	-15.951	16.607
dy	3.3442	7.941	0.421	0.674	-12.330	19.019
ep	0.3133	2.345	0.134	0.894	-4.315	4.942
bm	-3.2443	6.719	-0.483	0.630	-16.507	10.018
ntis	-46.9566	38.911	-1.207	0.229	-123.762	29.848
tbl	-2.8651	20.922	-0.137	0.891	-44.162	38.432
ltr	10.2432	14.468	0.708	0.480	-18.314	38.800
tms	13.1083	11.129	1.178	0.240	-8.859	35.076
dfy	-156.8202	213.943	-0.733	0.465	-579.111	265.471
dfr	71.0710	29.099	2.442	0.016	13.634	128.508
infl	-36.9489	82.870	-0.446	0.656	-200.521	126.623
ik	-208.4868	242.844	-0.859	0.392	-687.824	270.851

Regularisation methods

Regularisation methods

Regularisation, or **shrinkage**, methods for linear regression follow the general framework of **regularised empirical risk minimisation**:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left[\underbrace{\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))}_{\text{empirical risk}} + \lambda \, C(\boldsymbol{\theta}) \right]$$

We will use the squared error loss, and the complexity function will involve a norm of the regression coefficient vector, $\boldsymbol{\beta}$.

Ridge regression

The **ridge regression** method solves the following problem:

$$\widehat{\beta}_{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where λ is a tuning parameter.

$\lambda = 0$ gives the OLS solution

The penalty has the effect of shrinking the OLS coefficients towards zero.

The solution goes to zero as $\lambda \to \infty$.

Ridge regression

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

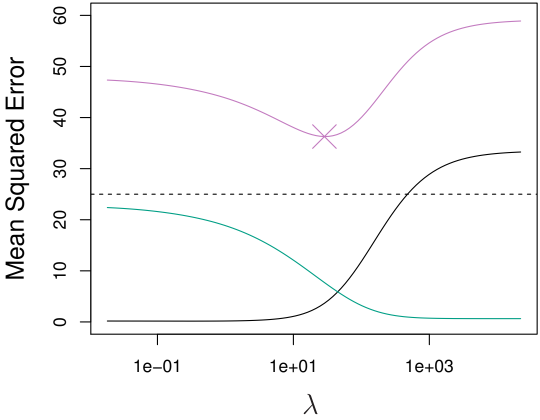
The criterion penalizes $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2$, which is the squared Euclidean norm (a.k.a. the ℓ_2 norm) of $\boldsymbol{\beta}$.

This type of approach is referred to as $\boldsymbol{\ell_2}$ **regularisation**.

Bias-Variance trade-off for Ridge regression

Simulated data with $n = 50$ and $p = 45$

Bias-squared **Variance** **Test MSE**



Ridge regression: Bayesian interpretation

It turns out that if we define $\lambda = \sigma^2/\tau^2$, then the ridge regression estimator is the posterior mode for a Bayesian Gaussian linear regression model with prior on the regression coefficients $\beta_1, ..., \beta_p$, under which β_j are independent $N(0, \tau^2)$.

Note that decreasing the τ (i.e. making the prior more concentrated around zero) increases the λ, which has the effect of increasing the penalty on the coefficients, resulting in greater shrinkage towards zero.

Ridge regression

The ridge estimator has an equivalent formulation as a constrained minimisation problem:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

for some $t > 0$.

Tuning parameters λ and t control the amount of shrinkage. There is a one-to-one correspondence between them.

Practical details

1. To make the procedure invariant to the scale of the inputs, we *standardise the predictors* before implementing the Ridge method (i.e. for each predictor we subtract its mean and divide by its standard deviation).

2. The intercept coefficient is not penalised. Once the predictors have been standardised, the Ridge estimate of the intercept has a very simple formula: $\widehat{\beta}_0 = \overline{y} = \sum_{i=1}^{n} y_i/n$.

   In fact, for this formula to hold, it is sufficient for the predictors to be *centered* (i.e. for each predictor we subtract its mean), and not necessarily standardized.

3. The above point about the estimate of the intercept is also valid for OLS and Lasso (this method is discussed later in the lecture).

Ridge regression

To simplify the presentation on the next two slides, we will assume that the predictors, as well as the response, have been centered.

By the discussion on the previous slide, this means that $\widehat{\beta}_0$ has been subtracted from each y_i, and thus the intercept has been removed from the estimation procedure:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Ridge regression

In matrix form, the optimization problem is:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \, \|\boldsymbol{\beta}\|^2.$$

As we've removed the intercept from the estimation procedure, matrix $\boldsymbol{X}$ no longer contains the column of ones and is simply the matrix of predictor values.

The formula for the ridge solution (which follows via a derivation analogous to the one we did for OLS) is:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\,\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Orthonormal predictors

We say that two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthonormal when

$$||\boldsymbol{u}|| = \sqrt{\boldsymbol{u}^T\boldsymbol{u}} = 1, \quad ||\boldsymbol{v}|| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}} = 1, \quad \text{and} \quad \boldsymbol{u}^T\boldsymbol{v} = 0.$$

For illustration, consider the special case of orthonormal predictors (i.e. the columns of $\boldsymbol{X}$ are orthonormal). In this case, $\boldsymbol{X}^T\boldsymbol{X} = I$, where I is a p by p identity matrix (with ones on the diagonal and zeroes everywhere else).

Thus, the ridge estimate is just a scaled version of the OLS estimate:

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{I} + \lambda\,\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}_{\text{ols}}$$

Ridge regression

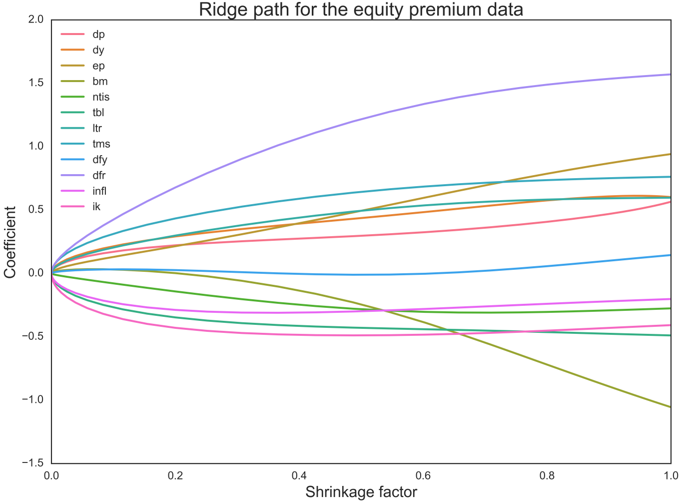
For each value of the tuning parameter λ, and the corresponding ridge estimate $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$, we define the ridge shrinkage factor as

$$\frac{||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||}{||\widehat{\boldsymbol{\beta}}_{\text{ols}}||}$$

As λ decreases from an infinitely large value down to zero, the ridge shrinkage factor increases from zero to one.

The next slide illustrates the effect of varying the shrinkage factor on the estimated coefficients.

Ridge coefficients as functions of the shrinkage factor

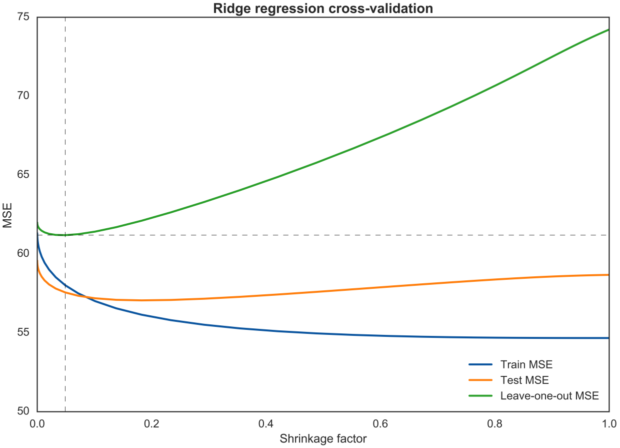


Selecting λ

The ridge regression method leads to a range of models corresponding to different values of λ. We select λ by cross validation.

Like in the case of OLS, a shortcut is available for computing the LOOCV MSE.

Selecting λ (equity premium data)



The Lasso

The **Lasso** (least absolute shrinkage and selection operator) method solves the following problem

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where λ is a tuning parameter.

$\sum_{j=1}^{p} |\beta_j|$ is the ℓ_1 norm of $\boldsymbol{\beta}$, denoted by $\|\boldsymbol{\beta}\|_1$.

Thus, Lasso performs $\boldsymbol{\ell_1}$ **regularisation**.

Lasso: shrinkage and variable selection

Shrinkage. As with ridge regression, the lasso shrinks the coefficients towards zero. However, the nature of this shrinkage is different, as we discuss further below.

Variable selection. In addition to shrinkage, the lasso also performs variable selection. With λ sufficiently large, some estimated coefficients will be exactly zero, leading to sparse models, which are easier to interpret. This is a key difference between lasso and ridge.

The Lasso

The equivalent formulation of the lasso as a constrained minimisation problem is:

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

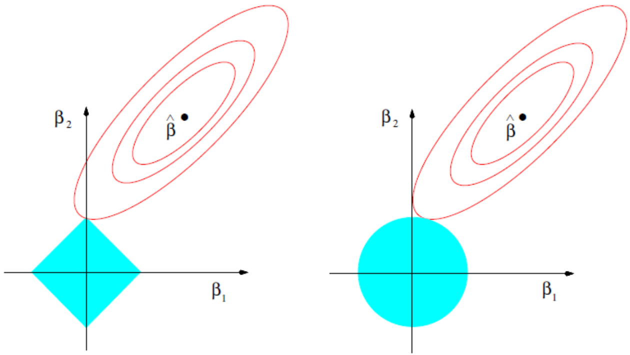
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t.$$

for some $t \ge 0$.

Lasso vs Ridge (a 2 predictor example)

Estimation picture for the constrained formulation of the lasso (left) and ridge regression (right).

The contours of the sum of squares function (which we are minimizing) are in red, and the constraint regions are in blue.



Practical details

1. We select the tuning parameter λ by cross validation.

2. As with ridge, we standardise the predictors before solving the optimisation problem.

3. Unlike OLS and Ridge, there is no explicit formula for the Lasso solutions.

4. There are efficient algorithms for computing an entire path of Lasso solutions (i.e. for all values λ), with a computation cost similar to that of the OLS.

The Lasso

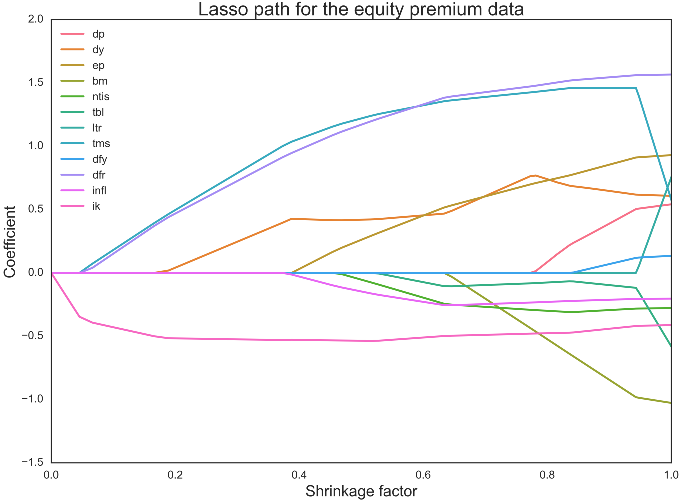
Similarly to ridge (but using the ℓ_1 norm), for each value of the tuning parameter λ we define the lasso shrinkage factor as:

$$\frac{||\widehat{\boldsymbol{\beta}}_{\text{lasso}}||_1}{||\widehat{\boldsymbol{\beta}}_{\text{ols}}||_1}$$

As λ decreases from a sufficiently large value down to zero, the lasso shrinkage factor increases from zero to one.

The next slide illustrates the effect of varying the shrinkage factor on the estimated parameters.

Lasso coefficients as functions of the shrinkage factor



Comparisons and extensions

Best subset, ridge, and lasso when predictors are orthonormal

$I(\cdot)$ denotes the indicator function. The formulas below exclude the intercept.

Estimator	Formula for $\widehat{\beta}_j$
Best subset (size k)	$\widehat{\beta}_j^{\text{ols}} \cdot I\left(\lvert\widehat{\beta}_j^{\text{ols}}\rvert \text{ is one of the } k \text{ largest } \lvert\widehat{\beta}_m^{\text{ols}}\rvert\right)$
Ridge	$\widehat{\beta}_j^{\text{ols}} \cdot \frac{1}{1+\lambda}$
Lasso	$\text{sign}(\widehat{\beta}_j^{\text{ols}})(\lvert\widehat{\beta}_j^{\text{ols}}\rvert - \frac{\lambda}{2}) I\left(\lvert\widehat{\beta}_j^{\text{ols}}\rvert > \frac{\lambda}{2}\right)$

In particular:

if $\widehat{\beta}_j^{\text{ols}} > \frac{\lambda}{2}$ then $\widehat{\beta}_j^{\text{lasso}} = \widehat{\beta}_j^{\text{ols}} - \frac{\lambda}{2}$

if $\widehat{\beta}_j^{\text{ols}} < -\frac{\lambda}{2}$ then $\widehat{\beta}_j^{\text{lasso}} = \widehat{\beta}_j^{\text{ols}} + \frac{\lambda}{2}$

and if $-\frac{\lambda}{2} \leq \widehat{\beta}_j^{\text{ols}} \leq \frac{\lambda}{2}$ then $\widehat{\beta}_j^{\text{lasso}} = 0$

Which method to use?

- Recall the no free lunch theorem: neither ridge regression nor the lasso universally outperforms the other. The choice of the method should be data driven.
- Generally, the lasso should perform better when a few predictors have large coefficients, while the remaining predictors zero or small.
- Ridge regression tends to perform better when all predictors have coefficients of similar size.
- The lasso has better interpretability as it produces sparse solutions (i.e. with some coefficients set to zero).

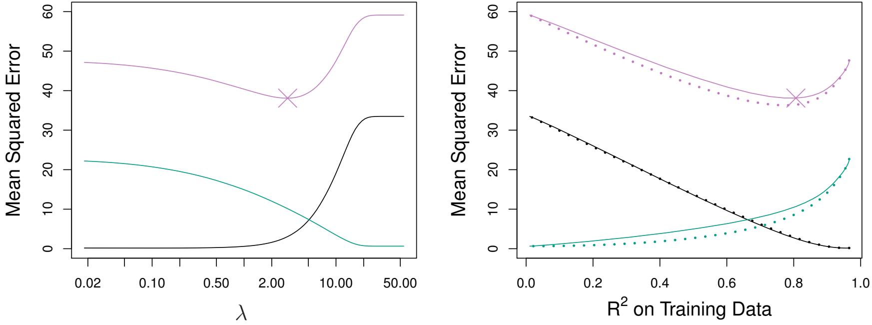
Bias-Variance trade-off for Ridge regression

Simulated data with $n = 50$ and $p = 45$, all true coefficients non-zero

Left plot: Lasso

Right plot: Lasso solid and Ridge dotted

Bias-squared **Variance** **Test MSE**



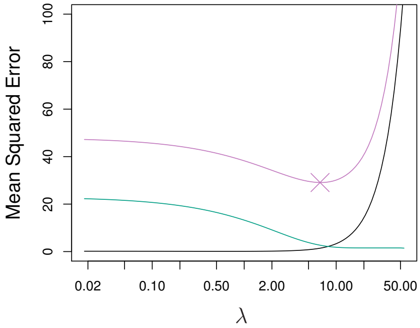
Bias-Variance trade-off for Ridge regression

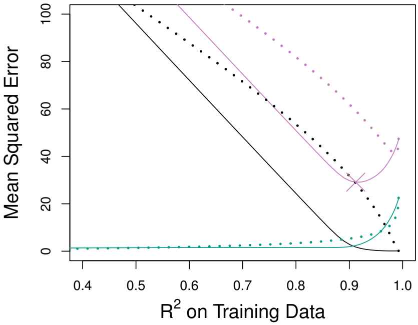
Simulated data with $n = 50$ and $p = 45$, only two true coefficients non-zero

Left plot: Lasso

Right plot: Lasso solid and Ridge dotted

Bias-squared **Variance** **Test MSE**





Elastic Net

The **Elastic Net** is a compromise between ridge and the lasso (and contains both methods as special cases):

$$\widehat{\beta}_{\mathsf{EN}} = \underset{\beta}{\mathsf{argmin}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} \Big(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\Big)$$

for $\lambda \geq 0$ and $0 \leq \alpha \leq 1$.

Penalty term involving $|\beta_j|$ encourages sparse solutions (like the lasso)

Penalty term involving β_j^2 improves the handling of highly correlated predictors (same as in ridge regression, this term encourages the coefficients of highly correlated predictors to be similar).

Illustration: equity premium data

Estimated coefficients (tuning parameters selected by CV)

	OLS	Ridge	Lasso	EN
dp	0.566	0.159	0.000	0.111
dy	0.602	0.197	0.000	0.153
ep	0.942	0.116	0.000	0.048
bm	-1.055	0.033	0.000	0.000
ntis	-0.276	-0.067	-0.000	-0.000
tbl	-0.489	-0.248	-0.000	-0.178
ltr	0.597	0.186	0.000	0.124
tms	0.762	0.286	0.161	0.239
dfy	0.145	0.031	0.000	0.000
dfr	1.570	0.377	0.131	0.294
infl	-0.202	-0.214	-0.000	-0.150
ik	-0.408	-0.318	-0.422	-0.282

Illustration: equity premium data

Prediction results

	Train R^2	Test R^2
OLS	0.108	0.014
Ridge	0.054	0.033
Lasso	0.033	0.011
Elastic Net	0.050	0.029

Comparison with subset selection

- Regularisation methods generally have lower variance than the subset selection methods.

- The computational cost of the regularisation procedures is similar to that of OLS (i.e. much lower than best subsets, for example).

Review questions

- What is best subset selection?
- What are stepwise methods?
- What are some advantages and disadvantages of subset selection methods?
- What are the penalty terms in the ridge and lasso methods?
- What are the key differences between the ridge and the lasso methods?