

# High-Dimensional Data

Aldo Solari



# Outline

## ① High-Dimensional Data

## ② Ridge Regression



# High-dimensional data

- In the past 20 years, new technologies have changed the way that data are collected in fields as diverse as finance, marketing, and medicine
- It is now commonplace to collect a large number  $p$  of predictors
- While  $p$  can be extremely large, the number of observations  $n$  is often limited due to cost, sample availability, or other considerations
- Data sets containing more predictors than observations, i.e.

$$p > n$$

are often referred to as [high-dimensional](#)

- Classical approaches such as least squares linear regression are not appropriate in this setting
- Slides based on *Lecture notes on ridge regression* by van Wieringen (2015)



# Linear regression

- Training data:  $\mathbf{y}_{n \times 1}, \mathbf{X}_{n \times p}$
- OLS estimator:  $\hat{\boldsymbol{\beta}}_{p \times 1} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$
- Fit on training data:  $\hat{\mathbf{y}}_{n \times 1} = \mathbf{X} \hat{\boldsymbol{\beta}}$
- Test data:  $\mathbf{y}^*_{m \times 1}, \mathbf{X}^*_{m \times p}$
- Prediction on test data:  $\hat{\mathbf{y}}^*_{m \times 1} = \mathbf{X}^* \hat{\boldsymbol{\beta}}$



# High-dimensional regression

- **Case**  $p < n$

$\text{rank}(\mathbf{X}) = p$  : it exists an unique solution  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

- **Case**  $p \rightarrow n$

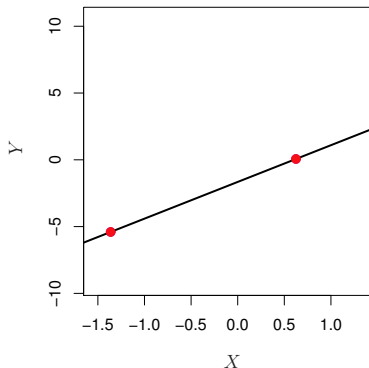
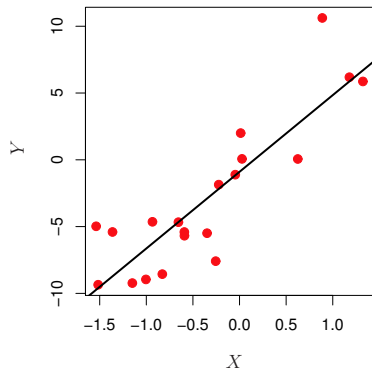
Even if  $\mathbf{X}^T \mathbf{X}$  can be inverted,  $\mathbf{X}^T \mathbf{X}$  approaches singularity and  $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$  increases without bound

- **Case**  $p > n$

$\text{rank}(\mathbf{X}) < p$  :  $\mathbf{X}^T \mathbf{X}$  is singular and does not have an inverse.  $\hat{\boldsymbol{\beta}}$  is not defined since there are infinitely many solutions  $\hat{\boldsymbol{\beta}}$



$$p = n$$

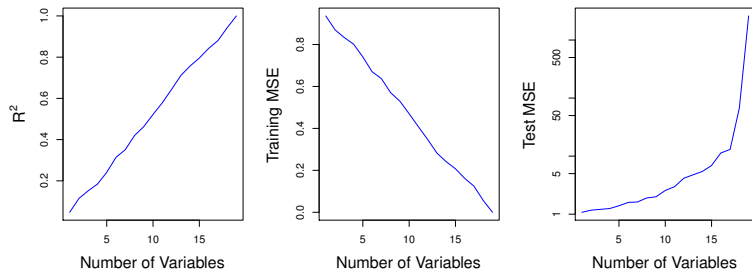


- Left: Least squares regression in the low-dimensional setting
- Right: Least squares regression with  $n = 2$  observations and two parameters to be estimated (intercept+slope)

Source: ISL p. 240 Figure 6.22



$$p \rightarrow n$$

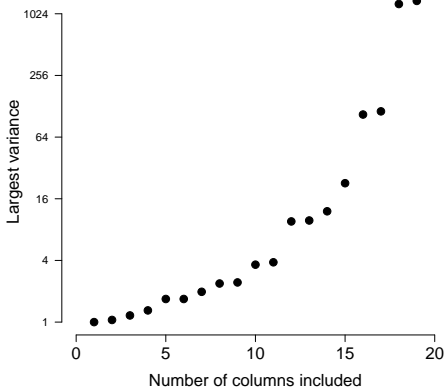


- Simulated example with  $n = 20$  and predictors that are completely unrelated to the response are added to the model
- Left: The  $R^2$  increases to 1 as more predictors are included
- Center:  $\text{MSE}_{\text{Tr}}$  decreases to 0 as more predictors are included
- Right:  $\text{MSE}_{\text{Te}}$  increases as more predictors are included

Source: ISL p. 241 Figure 6.23



Consider  $\mathbf{X}$  with  $n = 20$  and whose elements consist of independent, normally distributed random numbers; the figure below plots the largest variance of the  $\hat{\beta}_j$  estimates as we increase the number of columns in  $\mathbf{X}$



Source: P. Breheny





# Collinearity

- Collinearity in regression refers to the event of two (or multiple) predictors in  $\mathbf{X}$  being highly linearly related
- Then, the subspace spanned by the columns of  $\mathbf{X}$  is close to not being of full rank
- When the subspace onto which  $\mathbf{y}$  is projected, is close to rank deficient, it is almost impossible to separate the contribution of the individual predictors
- The uncertainty with respect to the predictors responsible for the variation explained in  $\mathbf{y}$  is often reflected in the fit of the linear regression model to data by a large error of the estimates of the regression parameters corresponding to the collinear predictors



```
n = 20
set.seed(123)
x1 <- rnorm(n)
x2 <- rnorm(n)
x3 <- rnorm(n,mean=x2,sd=.01)
y <- 1 + 1*x1 + 3*x2 + 3*x3 + rnorm(n)
summary( lm(y ~ x1 + x2 + x3 ) )
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.8949	0.2301	3.889	0.001305	**
x1	1.1254	0.2402	4.685	0.000248	***
x2	19.3422	24.4470	0.791	0.440405	
x3	-13.0406	24.4209	-0.534	0.600686	



# Super-collinearity

- The case of two (or multiple) predictors in  $\mathbf{X}$  being perfectly linearly dependent is referred to as [super-collinearity](#)
- An high-dimensional  $\mathbf{X}$  suffers from super-collinearity
- If  $p > n$ , then  $\text{rank}(\mathbf{X}) \leq n < p$ . This implies that the columns of  $\mathbf{X}$  are linearly dependent

$$\exists \mathbf{v} \in \mathbb{R}^p, \mathbf{v} \neq \mathbf{0} : \mathbf{X}\mathbf{v} = \mathbf{0}$$

- If  $p > n$ , then  $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^T \mathbf{X}) < p$ . This implies that the square matrix  $\mathbf{X}^T \mathbf{X}$  is singular with  $\det(\mathbf{X}^T \mathbf{X}) = 0$  and it does not have an inverse  $(\mathbf{X}^T \mathbf{X})^{-1}$



# Outline

① High-Dimensional Data

② Ridge Regression



# Ridge regression

- To obtain an estimate of the regression parameter  $\beta$  when  $\mathbf{X}$  is super-collinear, Hoerl and Kennard (1970) proposed an ad-hoc fix to resolve the singularity of  $\mathbf{X}^T\mathbf{X}$
- Simply replace  $\mathbf{X}^T\mathbf{X}$  by  $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p \times p}$  with  $\lambda > 0$
- The scalar  $\lambda$  is a tuning parameter, henceforth called the **penalty** parameter
- The **ridge regression estimator** is defined as

$$\hat{\beta}^\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p \times p})^{-1}\mathbf{X}^T\mathbf{y}$$

- The set of all ridge regression estimates  $\{\hat{\beta}^\lambda : \lambda \in [0, \infty)\}$  is called the **solution path** of the ridge estimator



# Ridge regression fit

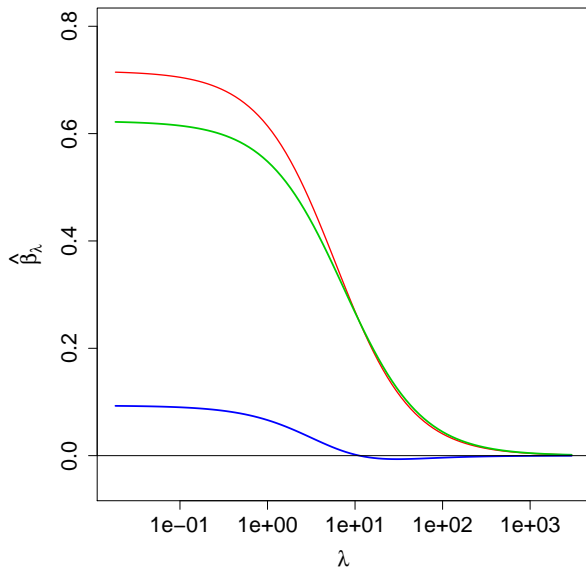
- Ridge regression fit on the training data  $\hat{\mathbf{y}}^\lambda = \mathbf{H}^\lambda \mathbf{y}$  with

$$\mathbf{H}^\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_{p \times p})^{-1} \mathbf{X}^\top$$

- If  $\text{rank}(\mathbf{X}) = p$  and  $\lambda = 0$ , the OSL fit  $\hat{\mathbf{y}}^0 = \hat{\mathbf{y}}$  is the projection of  $\mathbf{y}$  onto the subspace spanned by the columns of  $\mathbf{X}$
- With  $\lambda > 0$ , the fit  $\hat{\mathbf{y}}^\lambda$  is not the projection of  $\mathbf{y}$  onto the subspace spanned by the columns of  $\mathbf{X}$ , i.e.  $\mathbf{H}^\lambda$  is not a projection matrix. Consequently, the ridge residuals  $\mathbf{y} - \hat{\mathbf{y}}^\lambda$  are not orthogonal to the fit  $\hat{\mathbf{y}}^\lambda$



# Solution path



```
X = matrix(c(1,1,1,1,-1,0,2,1,2,1,-1,0), byrow=F, ncol=3)
p = ncol(X)
qr(X)$rank
y = c(1.3,-.5,2.6,.9)
```

```
lambdas = exp(seq(-4,8,length.out = 100))
hatbetas =sapply(lambdas, function(lambda)
solve(t(X)%*%X + lambda*diag(p)) %*% t(X) %*%y
)
plot(lambdas, hatbetas[1,], type="l", log="x")
lines(lambdas, hatbetas[2,])
lines(lambdas, hatbetas[3,])
```





# Expectation

- Assume the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n})$$

- The expectation of the ridge estimator

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^\lambda) = \mathbf{X}^\top \mathbf{X} (\lambda \mathbf{I}_{p \times p} + \mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{\beta}$$

- Clearly,  $\mathbb{E}(\hat{\boldsymbol{\beta}}^\lambda) \neq \boldsymbol{\beta}$  for any  $\lambda \neq 0$ . Hence the ridge estimator is biased



# Shrinkage

- The expectation of the ridge estimator vanishes as  $\lambda$  tends to infinity

$$\lim_{\lambda \rightarrow \infty} \mathbb{E}(\hat{\beta}^\lambda) = \mathbf{0}$$

- All regression coefficients are shrunk towards zero as the penalty parameter  $\lambda$  increases
- This behaviour is not strictly monotone in  $\lambda$ :  $\lambda_a > \lambda_b$  does not necessarily imply  $|\hat{\beta}_j^{\lambda_a}| < |\hat{\beta}_j^{\lambda_b}|$
- If  $\mathbf{X}$  is an orthonormal matrix, i.e.  $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_{p \times p} = (\mathbf{X}^\top \mathbf{X})^{-1}$ , the relation between OLS  $\hat{\beta} = \hat{\beta}^0$  and ridge estimator  $\hat{\beta}^\lambda$  is

$$\hat{\beta}^\lambda = \frac{1}{1 + \lambda} \hat{\beta}$$



# Variance

- The linear operator

$$\mathbf{W}^\lambda = [\mathbf{I}_{p \times p} + \lambda(\mathbf{X}^\top \mathbf{X})^{-1}]^{-1}$$

transforms the OLS estimator into the ridge estimator

$$\mathbf{W}^\lambda \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\lambda$$

- The variance of the ridge estimator

$$\mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}^\lambda) = \mathbf{W}^\lambda \mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}) (\mathbf{W}^\lambda)^\top = \sigma^2 \mathbf{W}^\lambda (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{W}^\lambda)^\top$$

- Like the expectation the variance of the ridge estimator vanishes as  $\lambda$  tends to infinity

$$\lim_{\lambda \rightarrow \infty} \mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}^\lambda) = \mathbf{0}$$

- If  $\mathbf{X}$  is an orthonormal matrix, then  $\mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{I}_{p \times p}$  and

$$\mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}}^\lambda) = \frac{\sigma^2}{(1 + \lambda)^2} \mathbf{I}_{p \times p}$$



# Variance comparison

- The difference

$$\mathbb{V}\text{ar}(\hat{\beta}) - \mathbb{V}\text{ar}(\hat{\beta}^\lambda)$$

is non-negative definite

- In words, the variance of the OLS estimator is larger than of the ridge estimator (in the sense that their difference is not-negative definite)
- The variance inequality

$$\mathbb{V}\text{ar}(\hat{\beta}) \succeq \mathbb{V}\text{ar}(\hat{\beta}^\lambda)$$

can be interpreted in terms of the uncertainty of the estimate



### Example 1.5 Variance comparison

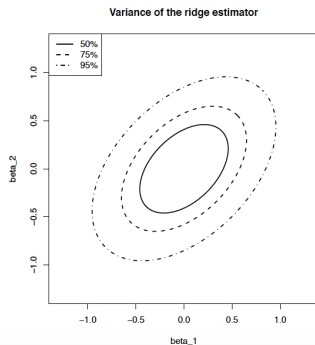
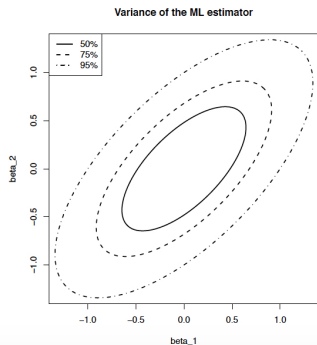
Consider the design matrix:

$$\mathbf{X} = \begin{pmatrix} -1 & 2 \\ 0 & 1 \\ 2 & -1 \\ 1 & 0 \end{pmatrix}.$$

The variances of the ML and ridge (with  $\lambda = 1$ ) estimates of the regression coefficients then are:

$$\text{Var}(\hat{\beta}) = \sigma^2 \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix} \quad \text{and} \quad \text{Var}(\hat{\beta}) = \sigma^2 \begin{pmatrix} 0.1524 & 0.0698 \\ 0.0698 & 0.1524 \end{pmatrix}.$$

These variance can be used to construct confidence intervals of the estimates. The 50%, 75% and 95% confidence intervals for the ML and ridge estimates are plotted in Figure 1.2. In line with inequality (1.3) the confidence intervals of the ridge estimate are smaller than that of the ML estimate.  $\square$



# Penalized estimation

- The ad-hoc fix of Hoerl and Kennard (1970) to super-collinearity of the design matrix (and, consequently the singularity of  $\mathbf{X}^T \mathbf{X}$ ) has been motivated post-hoc
- The ridge estimator minimizes the [ridge loss function](#)

$$\begin{aligned} & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_2}^2 \\ &= \sum_{i=1}^n (y_i - x_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2 \end{aligned}$$

where  $\|\mathbf{v}\|_{\ell_2} = \sqrt{\sum_{j=1}^p v_j^2}$  denotes the  $\ell_2$  norm

- This loss function is the traditional residual sum-of-squares (RSS)  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2$  augmented with a penalty  $\lambda \sum_{j=1}^p \beta_j^2$



# Penalized estimation

- The minimum of the RSS is attained at  $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$
- The minimum of the ridge penalty is attained at  $\boldsymbol{\beta} = \mathbf{0}$
- The  $\boldsymbol{\beta}$  that minimizes the ridge loss function balances the RSS and the penalty:

$$\hat{\boldsymbol{\beta}}^\lambda = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_2}^2$$

- The effect of the penalty in this balancing act is to shrink the regression coefficients towards zero, its minimum



# Constrained estimation

- The penalized minimization problem

$$\hat{\boldsymbol{\beta}}^\lambda = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_2}^2$$

can be reformulated into the following constrained optimization problem

$$\hat{\boldsymbol{\beta}}^\lambda = \arg \min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_{\ell_2}^2 \leq s} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2$$

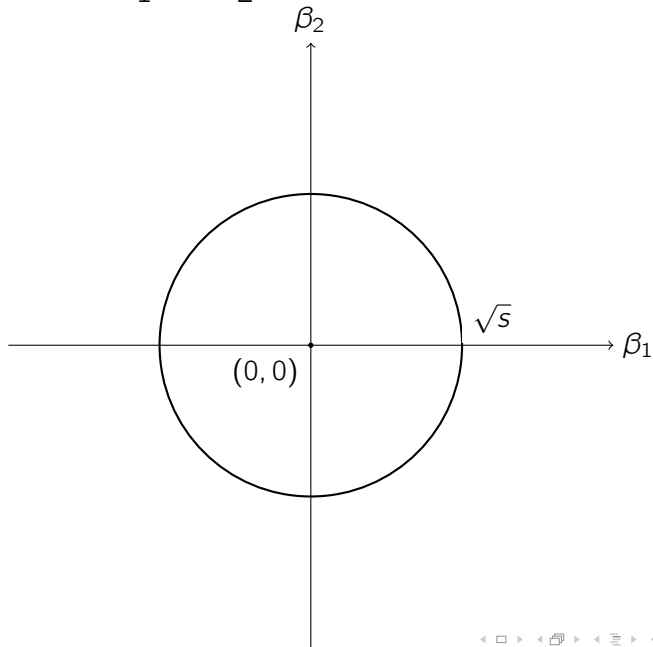
for some suitable  $s > 0$

- The constrained problem can be solved by means of the Karuch-Kuhn-Tucker (KTT) multiplier method, which minimizes a function subject to inequality constraint

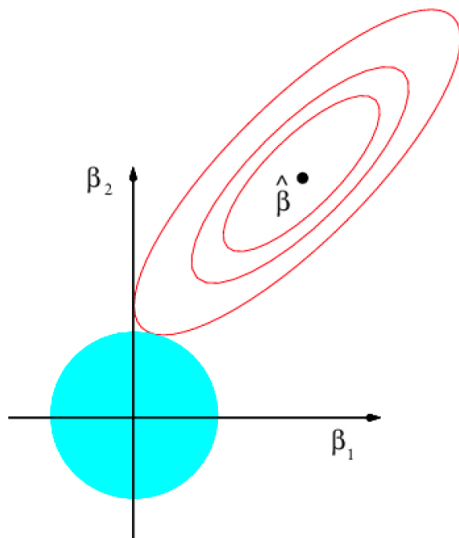




**Constraint**  $\beta_1^2 + \beta_2^2 \leq s$



# Constrained estimation



Source: Hastie et al. (2009) p. 71



# Penalized and constrained problems

- There is a one-to-one correspondence between the penalized problem and the constrained problem
- For each value  $s$ , there is a corresponding value  $\lambda$  that yields the same solution
- Conversely, the solution  $\hat{\beta}^\lambda$  to the penalized problem solves the constrained problem with  $s = \|\hat{\beta}^\lambda\|_{\ell_2}^2$



# Overfitting in high-dimensions

- The relevance of viewing the ridge regression estimator as the solution to a constrained estimation problem becomes obvious when considering a typical threat to high-dimensional data analysis: overfitting
- Overfitting refers to the phenomenon of modelling the noise rather than the signal
- In high-dimensional settings overfitting is a real threat. The number of predictors exceeds the number of observations. It is thus possible to form a linear combination of the predictors that perfectly explains the response, including the noise
- Large estimates of regression coefficients are often an indication of overfitting
- The estimation procedure with a constraint on the regression coefficients is a simple remedy to large parameter estimates. As a consequence it decreases the probability of overfitting



**Example 1.6** (*Overfitting*)

Consider an artificial data set comprising of ten observations on a response  $Y_i$  and nine covariates  $X_{i,j}$ . All covariate data are sampled from the standard normal distribution:  $X_{i,j} \sim \mathcal{N}(0,1)$ . The response is generated by  $Y_i = X_{i,1} + \varepsilon_i$  with  $\varepsilon_i \sim \mathcal{N}(0,1/4)$ . Hence, only the first covariate contributes to the response.

The regression model :

$$Y_i = \sum_{j=1}^9 X_{i,j} \beta_j + \varepsilon_i$$

is fitted to the artificial data using R. This yields the regression parameter estimates:

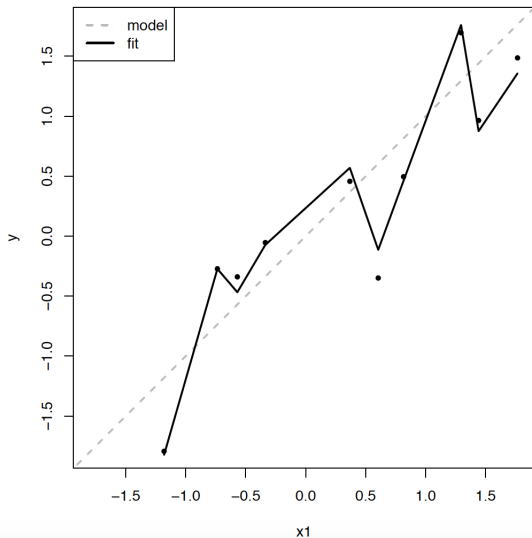
$$\hat{\beta}^\top = (0.048, -2.386, -5.528, 6.243, -4.819, 0.760, -3.345, -4.748, 2.136).$$

As  $\beta^\top = (1, 0, \dots, 0)$ , many regression coefficient are clearly over-estimated.

Source: van Wieringen (2015) p. 11



## Overfitting



Source: van Wieringen (2015) p. 10 Figure 1.3



# Perfect overfit: $p = n$

```
y = c(-1,0); x1 = c(.75,.5); x2 = c(1,.5)
summary(lm(y~0+x1+x2))
```

Residuals:

ALL 2 residuals are 0: no residual degrees of freedom!

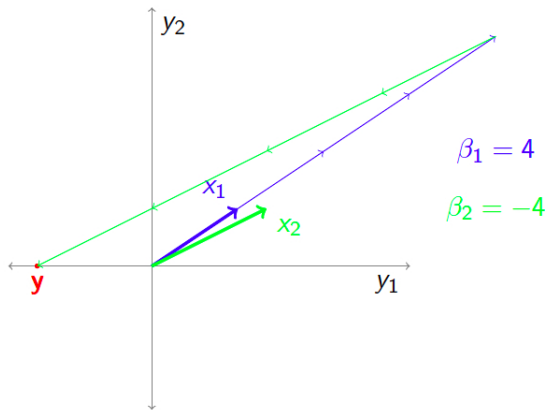
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
x1	4	NA	NA	NA
x2	-4	NA	NA	NA

Residual standard error: NaN on 0 degrees of freedom  
Multiple R-squared: 1, Adjusted R-squared: NaN  
F-statistic: NaN on 2 and 0 DF, p-value: NA



# Perfect overfit: $p = n$



Source: J. Goeman *Lasso, Ridge and Cross-validation* slides





# Shrinkage

Estimates are shrunk towards zero

## Consequence

- Introduces bias
- Reduces variance



# Predicting a Gaussian random variable

- $Y = \mu + \varepsilon$  with  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Training data:  $Y_1, \dots, Y_n$
- Sample mean:  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$
- Bias-Variance decomposition

$$\begin{aligned}\mathbb{E}\{[Y - \bar{Y}]^2\} &= \sigma^2 + [\text{Bias}(\bar{Y})]^2 + \text{Var}(\bar{Y}) \\ &= \sigma^2 + 0 + \sigma^2/n\end{aligned}$$

## Question

Rao-Blackwell theorem says that  $\bar{Y}$  has lower variance than any other unbiased estimator. Is  $\bar{Y}$  the optimal prediction for  $Y$  ?



# Shrunk sample mean

- Shrunk sample mean:  $\hat{Y} = \lambda \bar{Y}$  with  $\lambda \in [0, 1]$
- Bias-Variance decomposition

$$\begin{aligned}\mathbb{E}\{[Y - \hat{Y}]^2\} &= \sigma^2 + [\text{Bias}(\hat{Y})]^2 + \text{Var}(\hat{Y}) \\ &= \sigma^2 + [\mu - \lambda\mu]^2 + \lambda^2(\sigma^2/n)\end{aligned}$$

- Solving  $\frac{\partial}{\partial \lambda} \mathbb{E}\{[Y - \hat{Y}]^2\} = 0$  gives the value

$$\lambda = \frac{\mu^2}{\mu^2 + \sigma^2/n}$$

that minimizes  $\mathbb{E}\{[Y - \hat{Y}]^2\}$

- However, the optimal  $\lambda$  depends on unknown parameters



# Theorem

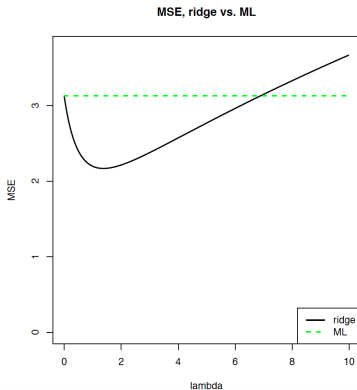
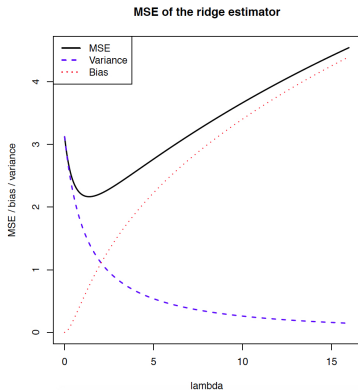
- There exists  $\lambda > 0$  such that  $(\text{Bias})^2 + \text{Var}$  of the ridge regression estimator is lower than the OLS estimator (Theorem 2 of Theobald, 1974)
- For large  $\lambda$ , the variance vanishes and the bias dominates
- For small enough values of  $\lambda$ , the decrease in variance of the ridge regression estimator exceeds the increase in its bias
- The optimal choice of  $\lambda$  depends on the quantities  $\boldsymbol{\beta}$  and  $\sigma^2$ . These are unknown in practice
- If  $\mathbf{X}$  is an orthonormal matrix, the bias-variance decomposition is

$$(\text{Bias})^2 + \text{Var} = \frac{\lambda^2}{(1 + \lambda)^2} \boldsymbol{\beta}^\top \boldsymbol{\beta} + \frac{p\sigma^2}{1 + \lambda}$$

and the optimal choice of  $\lambda$  is  $p\sigma^2/\boldsymbol{\beta}^\top \boldsymbol{\beta}$



# Bias-variance trade-off



Source: van Wieringen (2015) p. 14 Figure 1.4



# Degrees of freedom

- In ordinary regression, model complexity can be measured by the degrees of freedom consumed by the model
- In ordinary regression, the degrees of freedom consumed by the model is  $\text{trace}(\mathbf{H}) = p$ , the trace of the projection matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- By analogy, the ridge version of  $\mathbf{H}$  (which is not a projection matrix) is

$$\mathbf{H}^\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p \times p})^{-1}\mathbf{X}^T$$

- The degrees of freedom consumed by ridge regression (effective degrees of freedom) is

$$\text{trace}(\mathbf{H}^\lambda) = q$$

- The degrees of freedom consumed by ridge regression is monotone in  $\lambda$ . In particular

$$\lim_{\lambda \rightarrow \infty} \text{trace}(\mathbf{H}^\lambda) = 0$$



# Choice of the penalty parameter

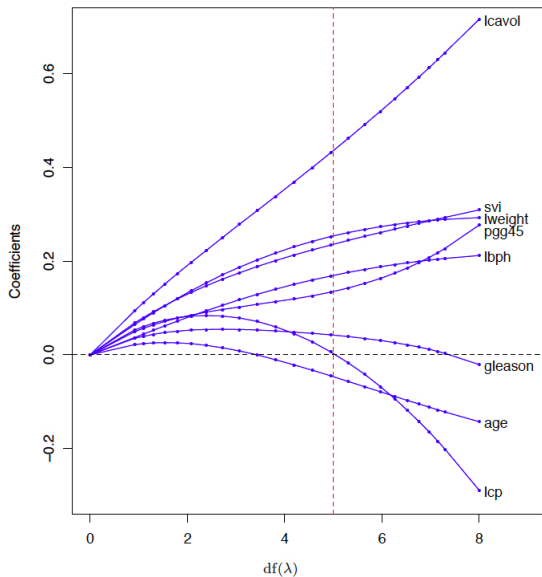
- Throughout the introduction of ridge regression and the subsequent discussion of its properties the penalty parameter  $\lambda$  is considered known or 'given'
- In practice, it is unknown and the user needs to make an informed decision on its value
- Information criteria (AIC, BIC, .. ) measure the balance between model fit and model complexity (degrees of freedom)

$$\text{AIC}(\lambda) = -2\ell(\hat{\boldsymbol{\beta}}(\lambda), \hat{\sigma}^2(\lambda)) + 2q$$

- Generalized cross-validation =  $\frac{\text{MSE}_{\text{Tr}}}{(1 - q/n)^2}$
- Cross-validation
- Cross-validated log-likelihood (R package `penalized`)



# Prostate Cancer Data

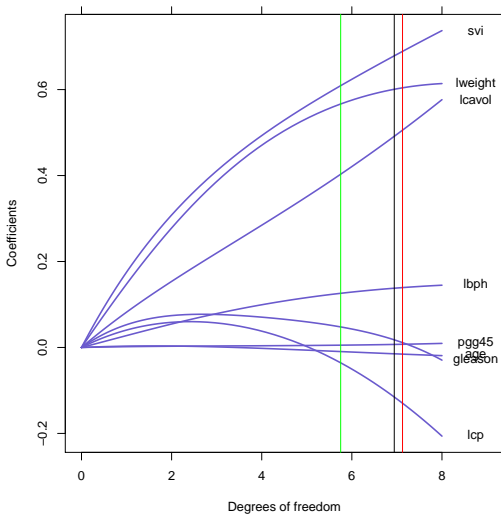


Source: Hastie et al (2009) p. 65





# Prostate Cancer Data

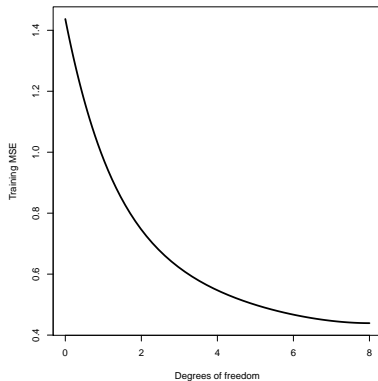
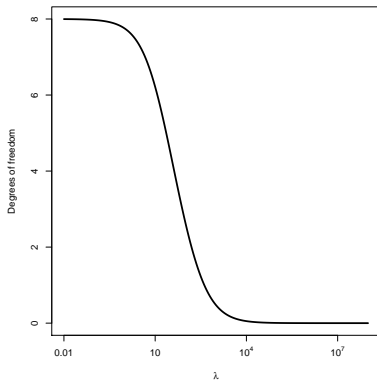


Red=AIC, black=GCV, green=BIC

Source: P. Breheny



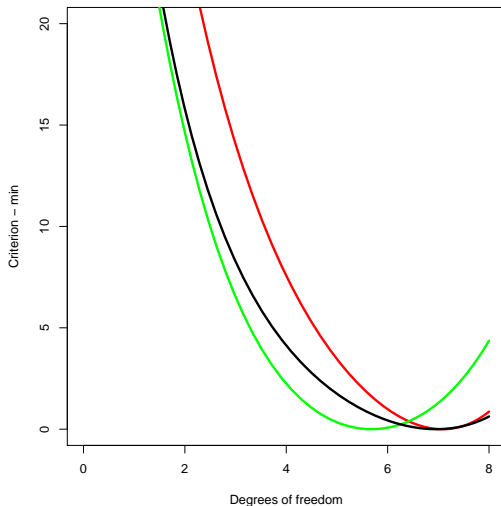
# Prostate Cancer Data



Source: P. Breheny



# Prostate Cancer Data



Red=AIC, black=GCV, green=BIC

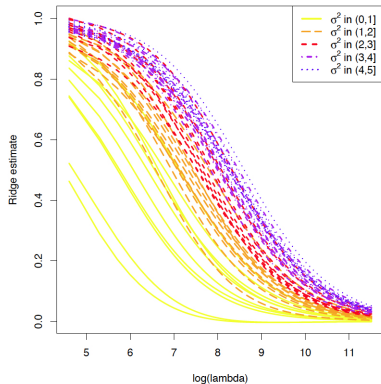


# Role of the variance of predictors

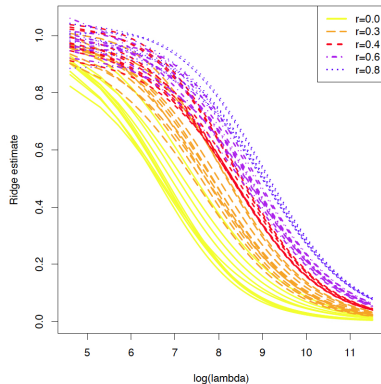
- Usually the intercept is not included in the penalty, unless you have some reason to think that the mean of  $y$  should be 0
- Usually the predictors are standardized to have mean 0 and standard deviation 1 prior to the execution of the ridge regression
- This is because ridge regularization path of coefficients corresponding to predictors with a large variance dominate those with low variances
- Ridge regression is not invariant to scaling: ridge estimates prior or posterior to scaling of predictors do not simply differ by a factor



Solution paths of covariates with distinct variance



Solution paths of correlated covariates



Source: van Wieringen (2015) p. 19 Figure 1.6



# Ridge regression and collinearity

- Suppose that  $X_1$  and  $X_2$  are strongly positively correlated
- We can write the linear model as

$$\mathbb{E}(Y|X_1, X_2) = \beta_1 X_1 + \beta_2 X_2$$

$$\mathbb{E}(Y|X_1, X_2) + \gamma(X_1 - X_2) = (\beta_1 + \gamma)X_1 + (\beta_2 - \gamma)X_2$$

where  $X_1 - X_2 \approx 0$  because  $X_1 \approx X_2$

- Then  $3X_1 + 3X_2$ ,  $4X_1 + 2X_2$ ,  $5X_1 + X_2$  etc. all have very similar fit, and least squares estimation can't easily distinguish them



# Ridge regression and collinearity

- For large enough  $\lambda$ , ridge regression favors the fits that minimize

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2$$

- This expression is minimized at  $\gamma = (\beta_2 - \beta_1)/2$ , giving

$$\frac{(\beta_1 + \beta_2)}{2}X_1 + \frac{(\beta_1 + \beta_2)}{2}X_2$$

- Ridge regression favors coefficient estimates for which strongly correlated predictors have similar effect sizes



# Ridge regression and collinearity

```
n = 20
set.seed(123)
x1 <- rnorm(n)
x2 <- rnorm(n, mean=x1, sd=.01)
y <- 10 + 3*x1 + 3*x2 + rnorm(n)

lm(y~x1+x2)$coef
(Intercept)          x1          x2
  10.127965   -7.155559   13.040228

require(MASS)
lm.ridge(y~x1+x2, lambda=1)
          x1          x2
10.144123  2.858528  2.875259
```





# glmnet

Read *Glmnet Vignette* by Hastie and Qian

[https://web.stanford.edu/~hastie/glmnet/glmnet\\_alpha.html](https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html)

```
glmnet(X, y,  
family="gaussian",  
alpha = 0,  
nlambda = 100,  
standardize = TRUE,  
intercept=TRUE  
)
```

