

SF 2930 REGRESSION ANALYSIS

LECTURE 9

High-dimensional Regression & Sparsity & Lasso

Timo Koski

KTH Royal Institute of Technology

2023

LEARNING OUTCOMES

- Bias-Variance Trade-off
- Double Descent
- $k \gg n$
 - Sparse Matrices
 - Geometry of the High-Dimensional Spaces
 - Gaussian Annulus Theorem
- Lasso Regression



AUXILIARY TERMINOLOGY

Spline is a long, flexible strip of wood that can be bent to draw curves.



Data Science Decipherer: What is a Spline?

Mathematics: **Spline** is a function defined piecewise by **polynomials**. Instead of fitting a single, high-degree polynomial to all data of a training set at once, spline interpolation fits low-degree polynomials to small subsets of the training set.

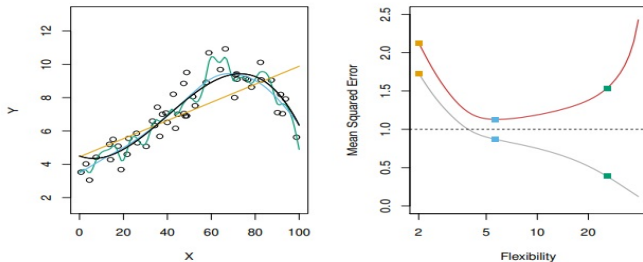


FIGURE 2.9. Left: Data simulated from f , shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

By courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert: *An introduction to statistical learning*, 2021.

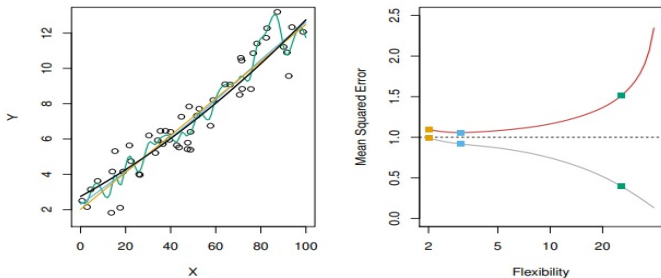


FIGURE 2.10. Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

By courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert: *An introduction to statistical learning*, 2021.

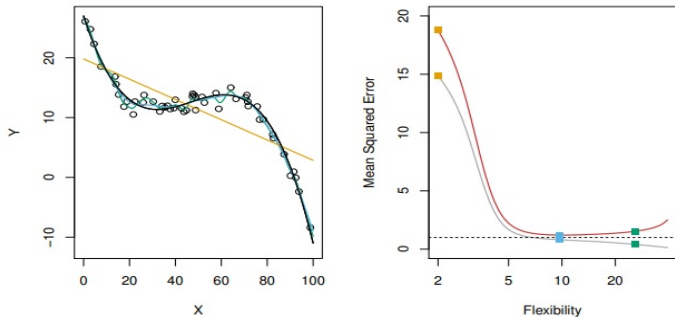
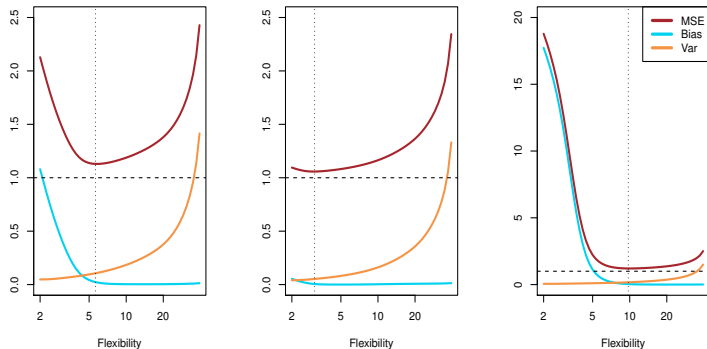


FIGURE 2.11. Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.

By courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert: *An introduction to statistical learning*, 2021.



By courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert: *An introduction to statistical learning*, 2021.

BIAS -VARIANCE TRADE-OFF: THE TRUE MODEL

The True Model

$$Y = Y(x) = E[(Y | X = x)] + \varepsilon = f(x) + \varepsilon,$$

where $E[(Y | X = x)] = f(x)$, $E[\varepsilon] = 0$,
 $\text{Var}[\varepsilon] = \sigma^2$.

The training set is $\mathcal{D}_{tr} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ are all independent samples from the same joint distribution $P(x, y)$. A learning method gives us an approximation, estimate of $f(x)$ as

$$\hat{f}(x; \mathcal{D}_{tr})$$

in some class of functions of x .

BIAS -VARIANCE TRADE-OFF: THE TRUE MODEL

We do not want to know, whether $\hat{f}(x_i; \mathcal{D}) \approx y_i$, we want to know whether $\hat{f}(x_i; \mathcal{D})$ is $\approx y$, where (x, y) is a previously unseen test observation, i.e., it has not been used to train $\hat{f}(x_i; \mathcal{D})$. We want to choose the statistical learning method that gives the lowest MSE on the test set, as opposed to the lowest training MSE.

BIAS -VARIANCE TRADE-OFF: MSE

We have $P(x, y) = P(x)P(y|x)$. We draw a new sample $X = x$ from $P(x)$ and then $Y|X = x \sim P(y|x)$. We now want to find the expected error on a new sample (x, Y)

$$\mathbb{E}_{\mathcal{D}_{tr}, \epsilon} \left[(Y - \hat{f}(x; \mathcal{D}_{tr}))^2 \right]$$

The expectation ranges over different choices of the training set $\mathcal{D}_{tr} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, all sampled from the same joint distribution $P(x, y)$.

BIAS -VARIANCE TRADE-OFF

It will be demonstrated that we can decompose its expected squared error on an another sample (x, Y) as follows

$$\mathbb{E}_{\mathcal{D}_{tr}, \epsilon} \left[(Y - \hat{f}(x; \mathcal{D}_{tr}))^2 \right] = \left(\text{Bias}_{\mathcal{D}} [\hat{f}(x; \mathcal{D}_{tr})] \right)^2 + \text{Var}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D}_{tr})] + \sigma^2 \quad (1)$$

Here

$$\text{Bias}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D}_{tr})] = \mathbb{E}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D}_{tr}) - f(x)] = \mathbb{E}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D}_{\sqcup \nabla})] - \mathbb{E} [Y(x)]$$

and

$$\text{Var}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D}_{tr})] = \mathbb{E}_{\mathcal{D}_{tr}} [(\mathbb{E}_{\mathcal{D}_{\sqcup \nabla}} [\hat{f}(x; \mathcal{D}_{tr})] - \hat{f}(x; \mathcal{D}_{tr}))^2].$$

BIAS -VARIANCE TRADE-OFF

In order to simplify writing in the subsequent check/derivation of (1), let us set

$$E_{\mathcal{D}_{tr}, \varepsilon} \rightarrow E,$$

$$\text{Var}_{\mathcal{D}_{tr}} \rightarrow \text{Var}$$

$$\hat{f}(x; \mathcal{D}_{tr}) \rightarrow \hat{f}.$$

$$\text{Bias}_{\mathcal{D}_{tr}} [\hat{f}(x; \mathcal{D})] \rightarrow \text{Bias} [\hat{f}]$$

$$Y = Y(x) = E[Y | X = x] + \varepsilon \rightarrow Y = f + \varepsilon$$

BIAS -VARIANCE TRADE-OFF

$$\begin{aligned}MSE &= E[(Y - \hat{f})^2] = E[(f + \varepsilon - \hat{f})^2] \\&= E[(f + \varepsilon - \hat{f} + E[\hat{f}] - E[\hat{f}])^2] \\&= E[(f - E[\hat{f}])^2] + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2] \\&\quad + 2E[(f - E[\hat{f}])\varepsilon] \\&\quad + 2E[\varepsilon(E[\hat{f}] - \hat{f})] \\&\quad + 2E[(E[\hat{f}] - \hat{f})(f - E[\hat{f}])]\end{aligned}$$

i.e.,

$$\begin{aligned}MSE &= (f - E[\hat{f}])^2 + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2] \\&\quad + 2(f - E[\hat{f}]) E[\varepsilon] + 2 E[\varepsilon] E[E[\hat{f}] - \hat{f}] + 2 E[E[\hat{f}] - \hat{f}](f - E[\hat{f}]) \\&= (f - E[\hat{f}])^2 + E[\varepsilon^2] + E[(E[\hat{f}] - \hat{f})^2] \\&= (f - E[\hat{f}])^2 + \text{Var}[\varepsilon] + \text{Var}[\hat{f}] \\&= \text{Bias}[\hat{f}]^2 + \text{Var}[\varepsilon] + \text{Var}[\hat{f}] \\&= \text{Bias}[\hat{f}]^2 + \sigma^2 + \text{Var}[\hat{f}].\end{aligned}$$

Finally, MSE loss function (or negative log-likelihood) is obtained by taking the expectation value over $x \sim P$.

BIAS -VARIANCE TRADE-OFF

$$\mathbb{E}_{\mathcal{D}, \varepsilon} \left[(Y - \hat{f}(x; \mathcal{D}))^2 \right] = \left(\text{Bias}_{\mathcal{D}} [\hat{f}(x; \mathcal{D})] \right)^2 + \text{Var}_{\mathcal{D}} [\hat{f}(x; \mathcal{D})] + \sigma^2. \quad (2)$$

The three terms on the right hand side represent (from left to right):

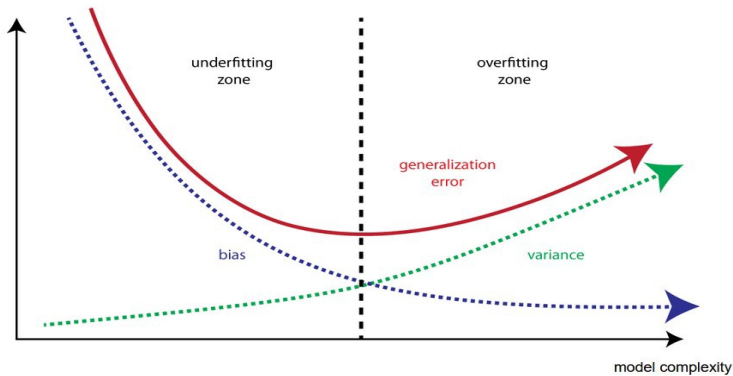
- the square of the bias of the learning method, which can be thought of as the error caused by the simplifying assumptions built into the method. E.g., when approximating a non-linear function $f(x)$ using a learning method for linear models, there will be error in the estimates $\hat{f}(x)$ due to this assumption.
- the variance of the learning method, or, intuitively, how much the learning method $\hat{f}(x)$ will move around its mean;
- the irreducible error σ^2 .

BIAS -VARIANCE TRADE-OFF

$$\mathbb{E}_{\mathcal{D}, \epsilon} \left[(Y - \hat{f}(x; \mathcal{D}))^2 \right] = \left(\text{Bias}_{\mathcal{D}} [\hat{f}(x; \mathcal{D})] \right)^2 + \text{Var}_{\mathcal{D}} [\hat{f}(x; \mathcal{D})] + \sigma^2. \quad (3)$$

This decomposition tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias. Note that variance is inherently a nonnegative quantity, and squared bias is also nonnegative. Hence, we see that the expected test MSE can never lie below $\text{Var}(\epsilon)$, variance of the unobservable irreducible error from the true model.

the bias vs. variance trade-off



BIAS -VARIANCE TRADE-OFF

What do we mean by the variance and bias of a statistical learning method? Variance refers to the amount by which \hat{f} would change if we estimated it using a different training data set.

Since the training data are used to fit the statistical learning method, different training data sets will result in a different \hat{f} . But ideally the estimate for f should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in \hat{f} . In general, more flexible statistical methods have higher variance.

QUOTE OF COUILLET, R. AND LIAO, Z.: *Random Matrix Methods for Machine Learning*, 2022, CAMBRIDGE UNIVERSITY PRESS.

Modern deep neural networks often have a huge number (billions) of parameters and are routinely trained to fit the training data almost perfectly, while still yielding remarkably good test performance in many cases. This means particularly that, in some scenarios it is possible to have good or even optimal models which contain more free parameters than intuitively needed.



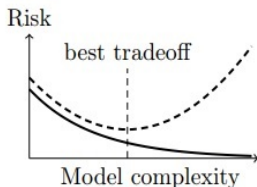
BIAS -VARIANCE TRADE-OFF

Recently, there has been observed the empirical trend where, for methods like neural networks and random forests, one sees a second bias-variance tradeoff in the out-of-sample prediction risk beyond the interpolation limit. The risk curve here resembles a traditional U-shape curve before the interpolation limit, and then descends again beyond the interpolation limit, which is known as “double descent”.

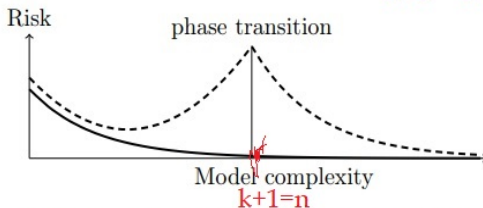
DOUBLE-DESCENT OF DEEP NEURAL NETWORKS

some scenarios, it is possible to have good or even optimal models which contain much more free parameters than intuitively needed (with typically $N > n$).

$$k+1 \gg n$$



(a) Classical U-shaped curve



(b) Modern double descent "UL"-shaped test curve

Figure 5.4: Comparison between training risk (solid lines) and true/test risk (dashed lines).

HIGH-DIMENSIONAL DATA & HIGH-DIMENSIONAL MODELS,

Multiple regression with $k \gg n$ studied.

Q: Is double descent possible here ?

A: No answer today.

HIGH-DIMENSIONAL VECTORS

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

$p=k+1$ $k \gg n$

high-dimensional vector

HIGH-DIMENSIONAL DATA

Regression analysis with $k \gg n$. David Donoho (Stanford):

The $k > n$ case is not anomalous; it is in some sense (nowadays) the generic case. For many types of event we can think of, we have the potential of a very large number of measurables quantifying that event, and a relatively few instances of that event.

GeneSPIDER - Generation and Simulation Package for Informative Data ExploRation

Andreas Tjärnberg^{1,2,†}, Torbjörn E. M. Nordling^{1,3,5}, Daniel Morgan^{1,2}, Matthew Studham^{1,2}, and Erik L.L. Sonnhammer^{1,2,4}

¹Stockholm Bioinformatics Center, Science for Life Laboratory, Sweden

²Department of Biochemistry and Biophysics, Stockholm University, Sweden

³Department of Immunology, Genetics and Pathology, Uppsala University, Rudbeck laboratory, 75185 Uppsala,

GENESPIDER: A QUOTE FROM INTRODUCTION

The primary objective in network inference is to obtain a network where each link corresponds to a real influence of importance in the biological system

Based on the [principle of parsimony](#), we recommend the use of [linear models](#) (= regression analysis to find the adjacency matrix of genes T.K.) unless a hypothesis that requires nonlinearities is tested or the class of linear models is rejected based on data. Publicly available gene expression datasets typically suffer from [few data points compared to the high number of genes and possible interactions](#), ... and redundant nearly collinear variables, i.e. ill-conditioned data matrices.

GEOMETRY OF HIGH-DIMENSIONAL SPACES

High-dimensional spaces are very strange for our intuition about geometry and convex bodies. This has important implications for high-dimensional regression and data-analysis.

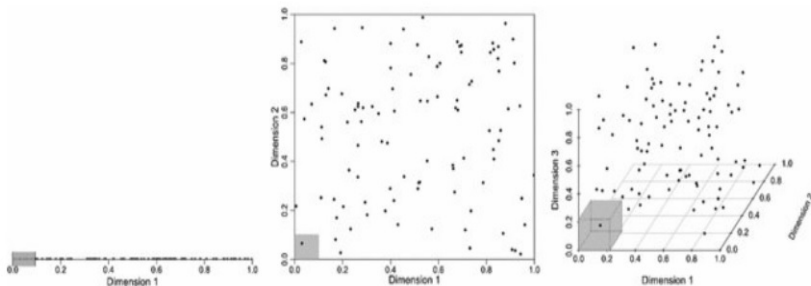
WE SHALL ARGUE THAT HIGH DIMENSION WILL LEAD TO SPARSE X

The data you have collected is as follows:

| Sparse Data | | | | | | | | | | | | |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|-----------|-----------|
| Day | Sensor 1 | Sensor 2 | Sensor 3 | Sensor 4 | Sensor 5 | Sensor 6 | Sensor 7 | Sensor 8 | Sensor 9 | Sensor 10 | Sensor 11 | Sensor 12 |
| 1-Jan | 0 | 0.89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.911 |
| 2-Jan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.931 |
| 3-Jan | 0 | 0.951 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.951 |
| 4-Jan | 0.954 | 0.911 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.899 |
| 5-Jan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.897 |
| 6-Jan | 0 | 0.899 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.968 |
| 7-Jan | 0.895 | 0.911 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.991 |
| 8-Jan | 0.911 | 0.962 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.951 |
| 9-Jan | 0 | 0.954 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.898 |
| 10-Jan | 0.898 | 0.934 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.962 |

A sparse matrix is a matrix in which most of the elements are zero. There is no strict quantitative definition regarding the proportion of zero-value elements for a matrix to qualify as sparse.

WHY IS THIS?



--

The Figure makes the point that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

GEOMETRY OF HIGH-DIMENSIONAL SPACES

Material taken from Chapter 2 in

*Avrim Blum, John Hopcroft, and Ravindran Kannan:
Foundations of Data Science, 2018*

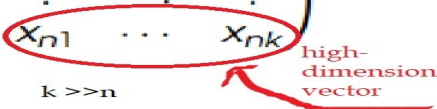
*Summary: The higher the dimension of your space,
the more likely the points are to lie near the edges
of the space rather than the center.*

GEOMETRY OF HIGH-DIMENSIONAL SPACES

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

$p=k+1$ $k \gg n$

high-dimensional vector



GEOMETRY OF HIGH-DIMENSIONAL SPACES

An important property of high-dimensional objects is that most of their volume is near the surface.

Consider any object A in \mathbb{R}^d . Next shrink A by a small amount to produce a new object: $(1 - \epsilon)A$ defined by

$$(1 - \epsilon)A := \{(1 - \epsilon)\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} \in A\}$$

Then the following equality holds (proof omitted) for volumes

$$V((1 - \epsilon)A) = (1 - \epsilon)^d V(A)$$

GEOMETRY OF HIGH-DIMENSIONAL SPACES

An important property of high-dimensional objects is that most of their volume is near the surface.

$$V((1 - \epsilon)A) = (1 - \epsilon)^d V(A)$$

Then

$$\frac{V((1 - \epsilon)A)}{V(A)} = (1 - \epsilon)^d \leq e^{-d\epsilon}$$

where we used $1 - x \leq e^{-x}$. This means that nearly all of the volume of A must be in the portion of A that does not belong to $(1 - \epsilon)A$.

VOLUME NEAR THE SURFACE

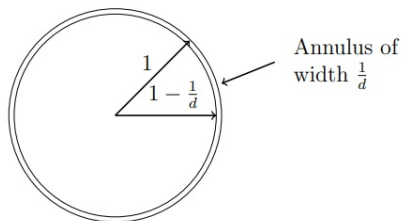


Figure 2.2: Most of the volume of the d -dimensional ball of radius r is contained in an annulus of width $O(r/d)$ near the boundary.

GEOMETRY OF HIGH-DIMENSIONAL SPACES: NORMS

$$\mathbf{x} \in \mathbb{R}^d$$

- $\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i|$
- $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$
- $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|$

GEOMETRY OF HIGH-DIMENSIONAL SPACES

DEFINITION

d-hypersphere in with radius r centered at origin of

$$B_r^d := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq r\}$$

DEFINITION

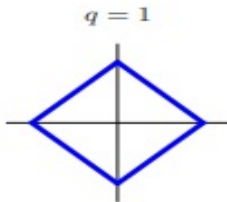
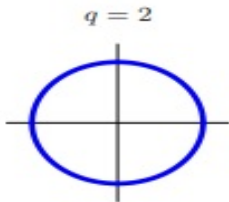
d-hypercube of side $2r$ centered at origin

$$C_r^d := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_\infty = \max_{1 \leq j \leq d} |x_j| \leq r\} = [-r, r]^d$$

GEOMETRY OF HIGH-DIMENSIONAL SPACES

On the left hand, B_1^2 , on the right hand the $\|\mathbf{x}\|_1$ unit sphere

$$\{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\|_1 \leq 1\}$$



A FIRST STRANGE PROPERTY

Let \mathbf{v} be a diagonal vector from the center to a corner of $C_r^d = [-r, r]^d$. Then $\mathbf{v} = \pm 1 \cdot \mathbf{1}_d$. Let $\mathbf{e}_j = (0, \dots, 1, \dots, 0)^T$. Then

$$\cos(\theta) = \frac{\mathbf{v}^T \mathbf{e}_j}{\|\mathbf{v}\|_2 \|\mathbf{e}_j\|_2} = \frac{\pm 1}{\sqrt{d}}$$

Hence for large d , all diagonals are almost orthogonal to the coordinate axes.

The hypervolume of B_r^d is given by

$$V(B_1^d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2} + 1)}, \quad V(B_r^d) = V(B_1^d) r^d.$$

The hypervolume of C_r^d is

$$V(C_1^d) = 2^d, \quad V(C_r^d) = V(C_1^d) r^d = 2^d r^d.$$

We have $B_1^d \subset C_1^d$: take any $\mathbf{x} \in B_1^d$, then $\|\mathbf{x}\|^2 = \sum_{i=1}^d x_i^2 \leq 1$. Hence $x_i^2 \leq 1$ holds for every i (equality when one $x_i = 1$ and the rest zeros), which means that $|x_i| = \sqrt{x_i^2} \leq \sqrt{1} = 1$ for every i . Thus $\max_{1 \leq i \leq d} |x_i| < 1$, which by definition of C_1^d means that $\mathbf{x} \in C_1^d$.

$B_1^d \subset C_1^d$. With increasing dimension the volume of the hypercube concentrates in its corners and the centre becomes less important:

$$\frac{V(B_1^d)}{V(C_1^d)} = \frac{\pi^{d/2}}{2^d \Gamma\left(\frac{d}{2} + 1\right)} \rightarrow 0, \quad \text{as } d \rightarrow \infty.$$

| | | | | | |
|---|-----------------|------------------|--------------------|-----|---------------------------------------|
| 1 | 2 | 3 | 4 | ... | 10 |
| 2 | π | $\frac{4}{3}\pi$ | $\frac{\pi^2}{2}$ | ... | $\frac{\pi^5}{120} \approx 2.55$ |
| 2 | 4 | 8 | 16 | ... | 1024 |
| 1 | $\frac{\pi}{4}$ | $\frac{\pi}{6}$ | $\frac{\pi^2}{32}$ | ... | $\frac{\pi^5}{122880} \approx 0.0025$ |

d

Volume of sphere

Volume of cube

Ratio

$$\frac{V(B_1^d)}{V(C_1^d)} = \frac{\pi^{d/2}}{2^d \Gamma\left(\frac{d}{2} + 1\right)} \rightarrow 0, \quad \text{as } d \rightarrow \infty.$$

$\Gamma\left(\frac{d}{2} + 1\right) = \frac{d}{2} \Gamma\left(\frac{d}{2}\right)$ (BETA 12.5). Then

$$\frac{V(B_1^d)}{V(C_1^d)} = \left(\frac{\pi^{1/2}}{2}\right)^d \frac{2}{d \Gamma\left(\frac{d}{2}\right)}$$

Here $0 < \frac{\pi^{1/2}}{2} < 1$. Hence $\left(\frac{\pi^{1/2}}{2}\right)^d \rightarrow 0$, as $d \rightarrow +\infty$.

$d \Gamma\left(\frac{d}{2}\right) \rightarrow +\infty$, as $d \rightarrow +\infty$.

$$\mathbf{X} \sim N_d(\mathbf{0}, \mathbb{I}_d),$$

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|\mathbf{x}\|_2^2}{2}}$$

The contour lines are for $\epsilon \in (0, 1)$

$$L = \{\mathbf{x} \in \mathbb{R}^d \mid e^{-\frac{\|\mathbf{x}\|_2^2}{2}} = \epsilon\}$$

It follows that

$$L = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2^2 = -2 \ln \epsilon\}$$

Note $-2 \ln \epsilon > 0$. Hence the probability of hitting the d -sphere inscribed in the contour line is

$$P(\mathbf{X} \in B_{2\ln(1/\epsilon)}^d) = P(\|\mathbf{X}\|_2^2 \leq 2\ln(1/\epsilon)) = P\left(\sum_{i=1}^d X_i^2 \leq 2\ln(1/\epsilon)\right).$$

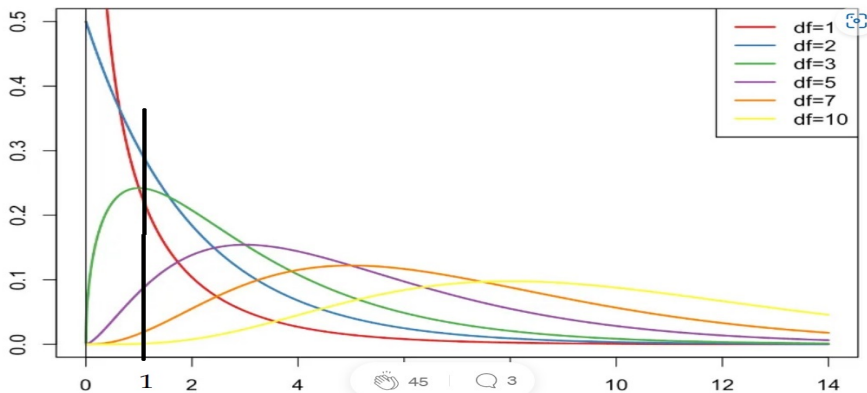
But $Z := \sum_{i=1}^d X_i^2 \sim \chi^2(d)$ so that

$$P(\mathbf{X} \in B_{2\ln(1/\epsilon)}^d) = P(Z \leq 2\ln(1/\epsilon))$$

Take $\epsilon = e^{-1/2}$. Then from the above

$$P(\mathbf{X} \in B_1^d) = P(Z \leq 1)$$

When $d \rightarrow \infty$, $P(Z \leq 1) \rightarrow 0$. This is seen graphically in the next Figure, with the vertical black line at 1. For increasing degrees of freedom the area under the pdf of $\chi^2(d)$ to the left of 1 becomes smaller and smaller.

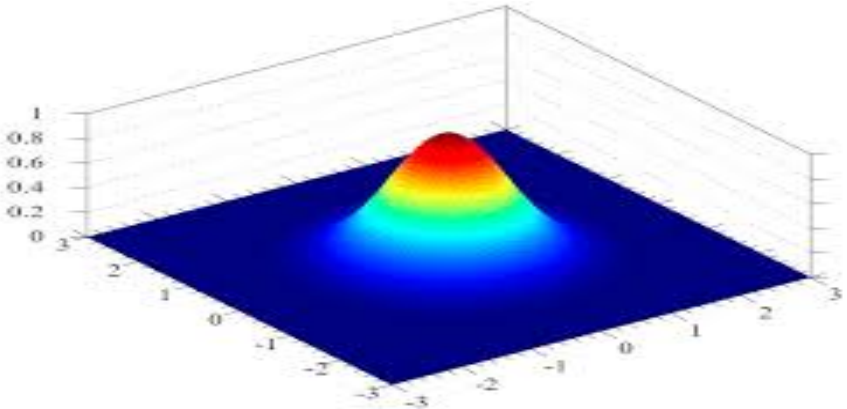


The Figure above is a slightly edited version of a Figure provided by

Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universität Berlin.

$$P(\mathbf{X} \in B_1^d) \rightarrow 0, \quad \text{as } d \rightarrow \infty$$

Hence the normal probability mass vanishes from the unit hypersphere, when dimension increases. We are used to thinking with the image of $d = 1$ and $d = 2$ that the normal distribution has most of its mass in the center.

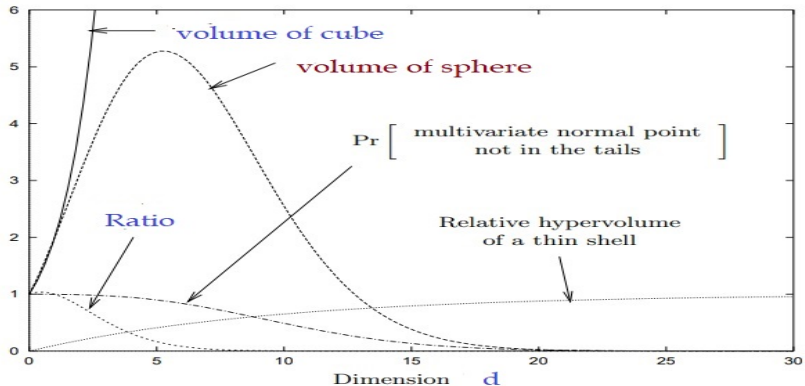


Or, we have also

$$P(\mathbf{X} \in B_1^d) = \int_{B_1^d} \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|\mathbf{x}\|_2^2}{2}} d\mathbf{x} \leq \int_{B_1^d} d\mathbf{x} = V(B_1^d)$$

and the volume of the unit hypersphere becomes vanishingly small as d increases, as found above.

M. A. Carreira-Perpinán: A Review of Dimension Reduction Techniques Technical Report CS-96-09 Dept. of Computer Science University of Sheffield, 1997.



In addition, since $\mathbf{X} = (X_1 \dots, X_d)$, each $X_i \sim N(0, 1)$,

$$E \left[\|\mathbf{X}\|_2^2 \right] = \sum_{i=1}^d \left[X_i^2 \right] = d$$

In words, the mean squared distance of a point from the center is d . The following theorem tells us, where the probability mass lies in high dimensions. This is the **Gaussian annulus theorem** found on p. 24 of Blum, Hopcroft, and Kannan.

PROPOSITION

$\mathbf{X} \sim N_d(\mathbf{0}, \mathbb{I}_d)$ For any $b \leq \sqrt{d}$

$$P \left(\sqrt{d} - b \leq \|\mathbf{X}\|_2 \leq \sqrt{d} + b \right) \geq 1 - 3e^{-cb^2},$$

where c is a fixed positive constant.

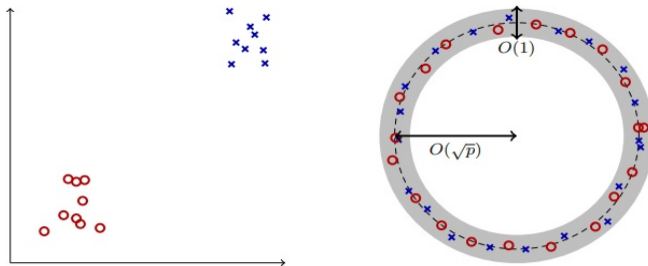


Figure 1.5: Visual representation of classification in **(left)** small and **(right)** large dimensions.

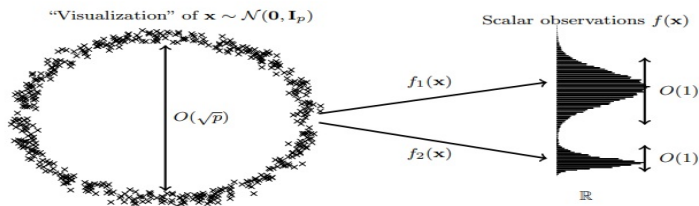


Figure 1.6: Multivariate Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, a fundamental example of concentrated random vectors. **(Left)** A visual “interpretation” of 500 independent drawings of $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. **(Right)** Concentration of observations for linear ($f_1(\mathbf{x}) = \mathbf{x}^\top \mathbf{1}_p / \sqrt{p}$) and Lipschitz ($f_2(\mathbf{x}) = \|\mathbf{x}\|_\infty$) maps.

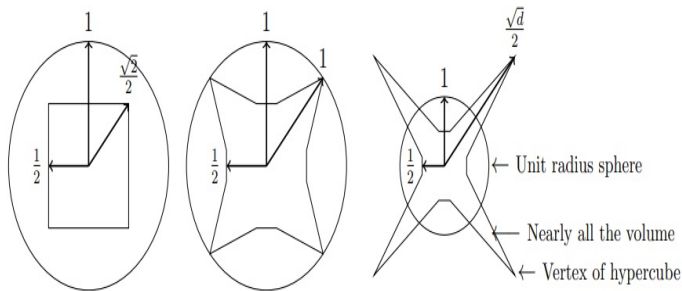
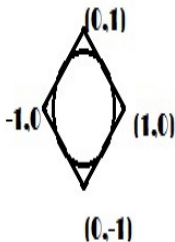
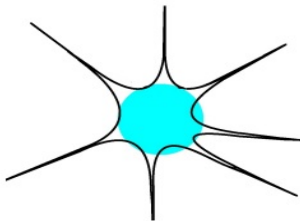


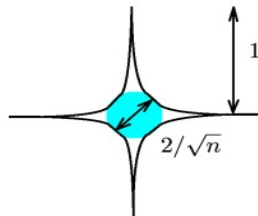
Figure 2.4: Illustration of the relationship between the sphere and the cube in 2, 4, and d -dimensions.



**Radius of
the
inscribed
sphere = $2/2^{1/2}$**



(a) A general convex set



(b) The ℓ_1 ball

Figure 2. V. Milman's “hyperbolic” drawings of high dimensional convex sets

Brin, Sergey: *Near neighbor search in large metric spaces*, VLDB, 95, 58, pp. 574–4584, 1995.

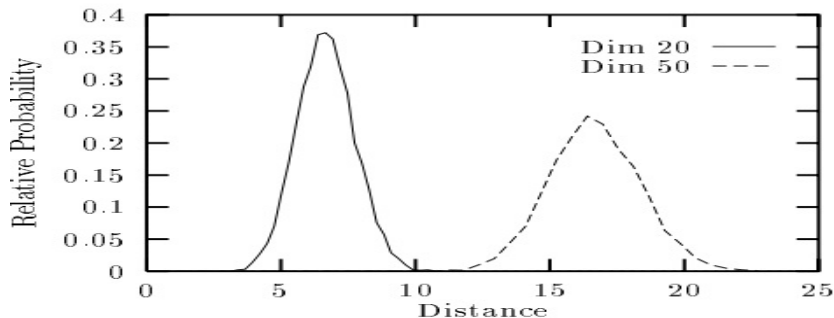


Figure 1: Distribution of distances between vectors chosen uniformly from unit cubes under the L_1 metric in 20 and 50 dimensions.

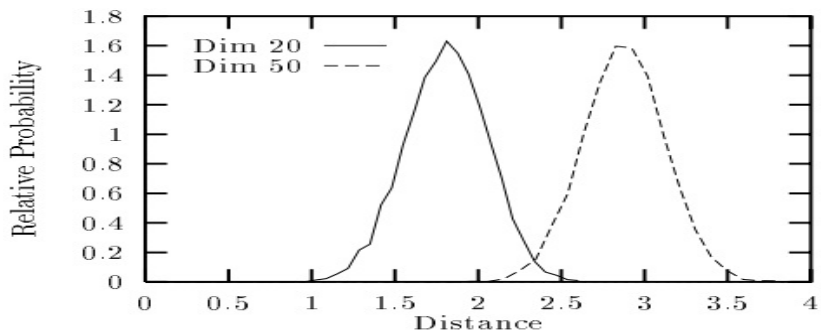


Figure 2: Distribution of distances between vectors chosen uniformly from unit cubes under the L_2 metric in 20 and 50 dimensions.

The distributions of distances between random, uniformly chosen vectors in 20 and 50 dimensional hypercubes of side 1 under the L_1 metric are very closely approximated by Gaussian distributions because of the Central Limit Theorem (Figure 1). For the L_2 metric, we obtain a Gaussian-like (though not exactly Gaussian) distribution (Figure 2). Note that the distributions for 50 dimensions should be viewed in relation to their larger ranges and hence are really quite narrow. The fact that the peaks are narrow indicate that the distance function has low entropy and that it may be difficult to index the data since arbitrary distance measurements will provide us with little information (by contrast a uniform distribution would have high entropy and distance measurements would give us a lot of information).

Lasso= Least Absolute Selection and Shrinkage Operator

LASSO

X is $k \times k$, the Lasso finds the regression coefficients by

$$\min_{\beta \in \mathbb{R}^k} \frac{1}{2n} \| \mathbf{y} - \beta_0 \mathbb{I}_n - X\beta \|_2^2$$

subject to

$$\| \beta \|_1 \leq t.$$

This is a piece of optimization theory known as convex programming. It forces sparsity, i.e., some or many $\hat{\beta}_j$ will be zeros.

LASSO'S EQUIVALENT LAGRANGE DUAL

X is $k \times k$, center the model, $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^k \hat{\beta}_j \bar{x}_j$.

$$\min_{\beta \in \mathbb{R}^k} \frac{1}{2n} \| \mathbf{y} - X\beta \|_2^2 + \lambda \| \beta \|_1$$

RIDGE REGRESSION

X is $k \times k$, ridge regression finds the regression coefficients by

$$\min_{\beta \in \mathbb{R}^k} \frac{1}{2n} \| \mathbf{y} - \beta_0 \mathbb{I}_n - X\beta \|_2^2$$

subject to

$$\| \beta \|_2 \leq t.$$

Ridge regression strives at sparsity, i.e., some or many $\hat{\beta}_j$ approach zero values.

RIDGE REGRESSION EQUIVALENT LAGRANGE DUAL

$$\min_{\beta \in \mathbb{R}^k} \frac{1}{2n} \| \mathbf{y} - X\beta \|_2^2 + \lambda \| \beta \|_2$$

T.HASTIE, R. TIBSHIRANI AND M. WAINWRIGHT: STATISTICAL LEARNING WITH SPARSITY. THE LASSO AND GENERALIZATIONS, p. 11

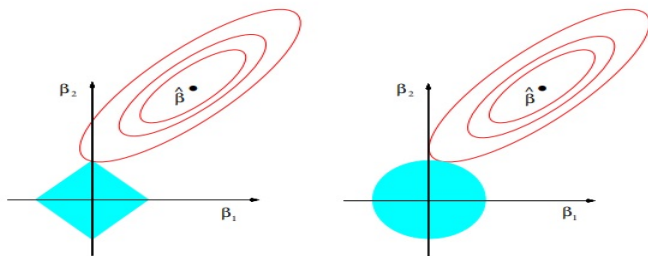


Figure 2.2 Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\hat{\beta}$ depicts the usual (unconstrained) least-squares estimate.

LAGRANGE MULTIPLIERS

$$\begin{aligned} &\text{Minimize } f_0(x) \\ &\text{subject to } f_i(x) \leq 0, \quad i \in \{1, \dots, m\} \end{aligned}$$

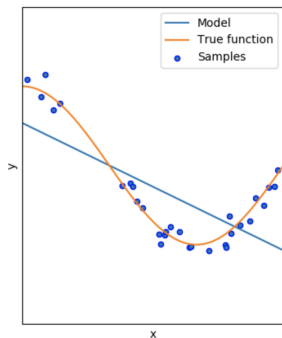
with the domain $\mathcal{D} \subset \mathbb{R}^n$ having non-empty interior. The dual Lagrangian function

$$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$$

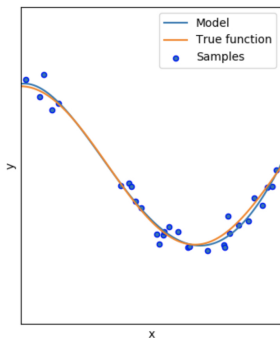
is gives the solution by

$$\text{minimize } \mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x).$$

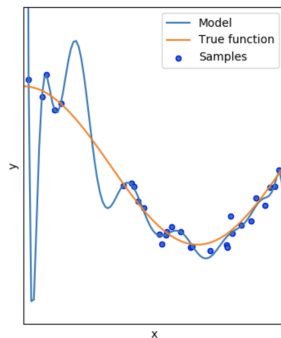
REGULARIZATION – RIDGE REGRESSION AND LASSO



High bias (underfit)



Good fit

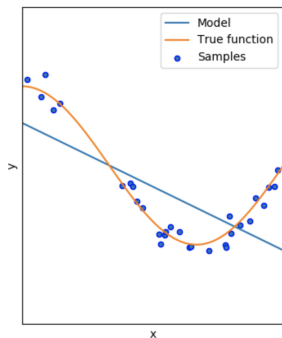


High variance

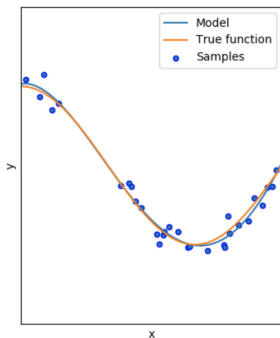
(overfit)

To overcome underfitting or high bias \Leftrightarrow Add more parameters to the model \Leftrightarrow The model complexity increases.

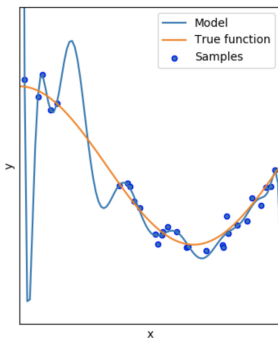
REGULARIZATION – RIDGE REGRESSION AND LASSO



High bias (underfit)



Good fit



High variance

(overfit)

To overcome underfitting or high bias \Leftrightarrow Add more parameters to the model \Leftrightarrow The model complexity increases.

REGULARIZATION – RIDGE REGRESSION AND LASSO

How can we overcome overfitting for a regression model?

- Reduce the model complexity
- Regularization – Ridge regression and Lasso

SHRINKAGE METHODS

As an alternative, to the subset selection methods discussed above, we can fit a model containing all p predictors using a technique that

- constrains or regularizes the coefficient estimates, or equivalently,
- shrinks the coefficient estimates 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

As an alternative, to the subset selection methods discussed above, we can fit a model containing all p predictors using a technique that

- constrains or regularizes the coefficient estimates, or equivalently,
- shrinks the coefficient estimates 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.

The two best-known techniques for shrinking the regression coefficients towards zero are

- ridge regression,
- the lasso.

RIDGE REGRESSION

Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity.

Ridge regression's advantage over least squares is the bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

TUNING PARAMETER

The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.

TUNING PARAMETER

The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.

TUNING PARAMETER

The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Selecting a good value for λ is critical \Rightarrow Cross-validation!

TUNING PARAMETER

The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Selecting a good value for λ is critical \Rightarrow Cross-validation!
- Choose a grid of λ values, and compute the cross-validation error for each value of λ . Then select the tuning parameter value for which the cross-validation error is smallest.

TUNING PARAMETER

The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Selecting a good value for λ is critical \Rightarrow Cross-validation!
- Choose a grid of λ values, and compute the cross-validation error for each value of λ . 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.

THE LASSO

Ridge regression does have one obvious disadvantage. Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.

The *shrinkage penalty* will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).

THE LASSO

Ridge regression does have one obvious disadvantage. Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.

The *shrinkage penalty* will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). The lasso is an alternative to ridge regression that overcomes this disadvantage.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

However,

- in the case of the lasso, the 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.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

However,

- in the case of the lasso, the 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.
- Much like best subset selection, the lasso performs variable selection.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero.

However,

- in the case of the lasso, the 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.
- Much like best subset selection, the lasso performs variable selection.
- As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.
- However, which method leads to better prediction accuracy?

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.
- However, which method leads to better prediction accuracy?
- One can see that the lasso leads to qualitatively similar behavior to ridge, in that as λ increases, the variance decreases and the bias increases.

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.
- However, which method leads to better prediction accuracy?
- One can see that the lasso leads to qualitatively similar behavior to ridge, in that as λ increases, the variance decreases and the bias increases.
- Often, the lasso and ridge regression result in almost identical biases, but the variance of ridge is slightly lower than the variance of the lasso.

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.
- However, which method leads to better prediction accuracy?
- One can see that the lasso leads to qualitatively similar behavior to ridge, in that as λ increases, the variance decreases and the bias increases.
- Often, the lasso and ridge regression result in almost identical biases, but the variance of ridge is slightly lower than the variance of the lasso.
- Consequently, the minimum MSE of ridge regression is slightly smaller than that of the lasso.

RIDGE VS LASSO

- It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.
- However, which method leads to better prediction accuracy?
- One can see that the lasso leads to qualitatively similar behavior to ridge, in that as λ increases, the variance decreases and the bias increases.
- Often, the lasso and ridge regression result in almost identical biases, but the variance of ridge is slightly lower than the variance of the lasso.
- Consequently, the minimum MSE of ridge regression is slightly smaller than that of the lasso.

- One can conclude that neither ridge regression nor the lasso will universally dominate the other.

- One can conclude that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.

- One can conclude that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.

- One can conclude that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.
- However, the number of predictors that is related to the response is never known a priori for real data sets.

- One can conclude that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.
- 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.