Ridge Regression

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Regression

Goal: Estimate response from features

Linear regression

Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Ordinary-least-squares estimator from dataset
(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

$$\ell_{\text{OLS}}(x_i) = \Sigma_{XY}^T \Sigma_X^{-1} (x_i - m(X)) + m(Y)$$

For simplicity, from now on everything centered to have zero mean

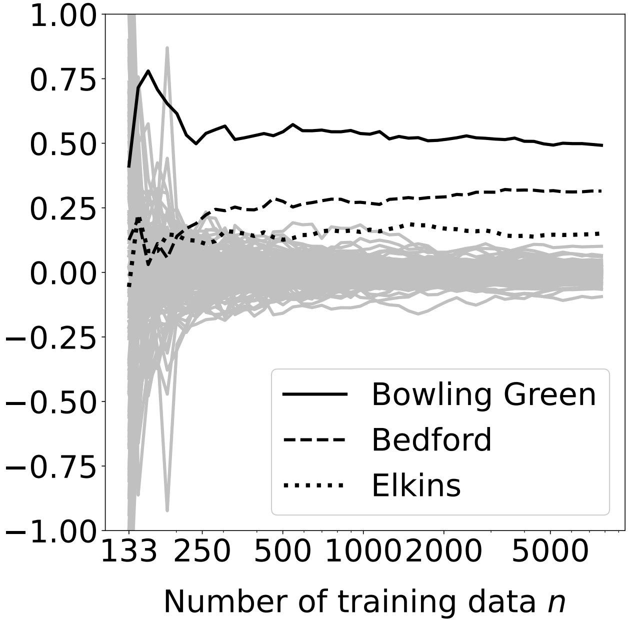
Temperature prediction

Response: Temperature in Versailles (Kentucky)

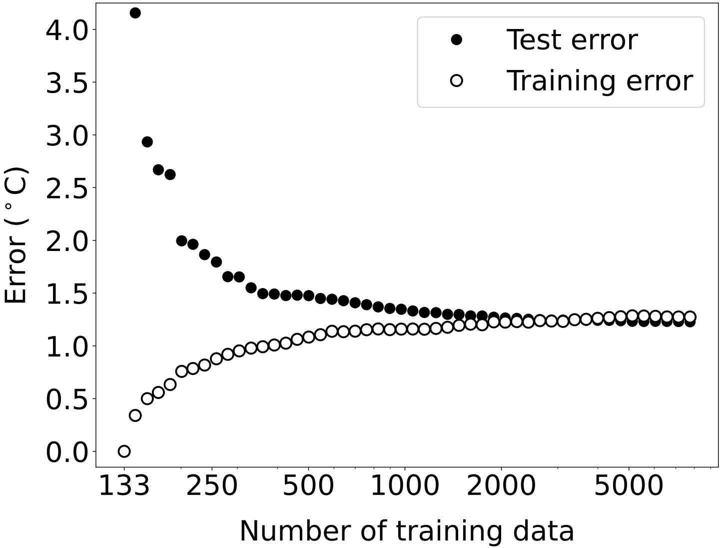
Features: Temperatures at 133 other locations

OLS coefficients (large n)

OLS coefficients



Training and test error



Ridge regression

Problem: For small n, large coefficients overfit the training data

$$\beta_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2$$

Solution: Regularization, penalize the norm of the coefficients

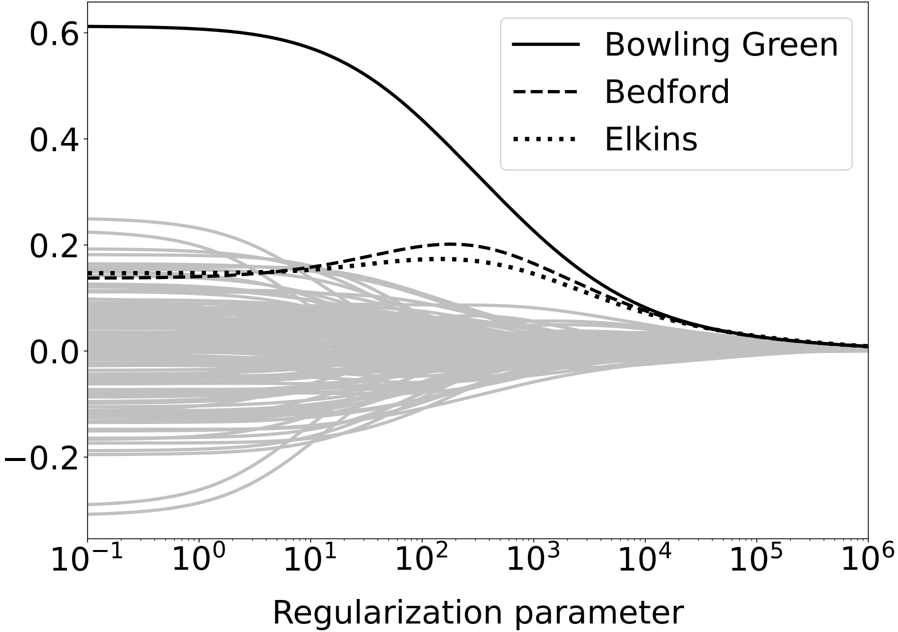
$$\beta_{\text{RR}} := \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2$$

$\lambda > 0$ is a regularization parameter

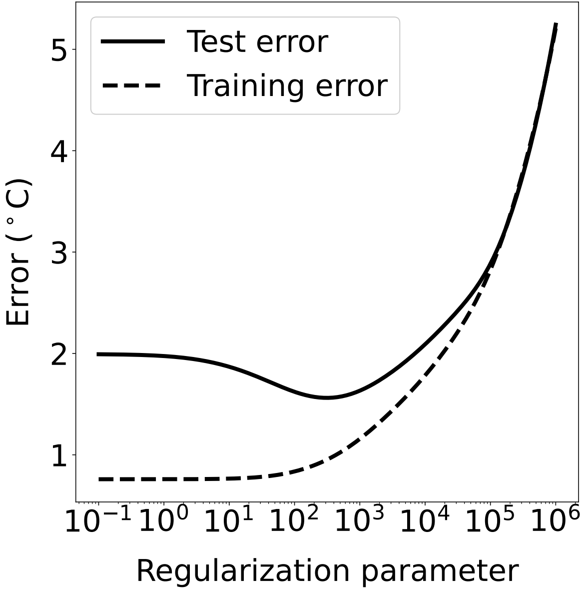
When $\lambda \to 0$? $\beta_{\text{RR}} \to \beta_{\text{OLS}}$

When $\lambda \to \infty$? $\beta_{\text{RR}} \to 0$

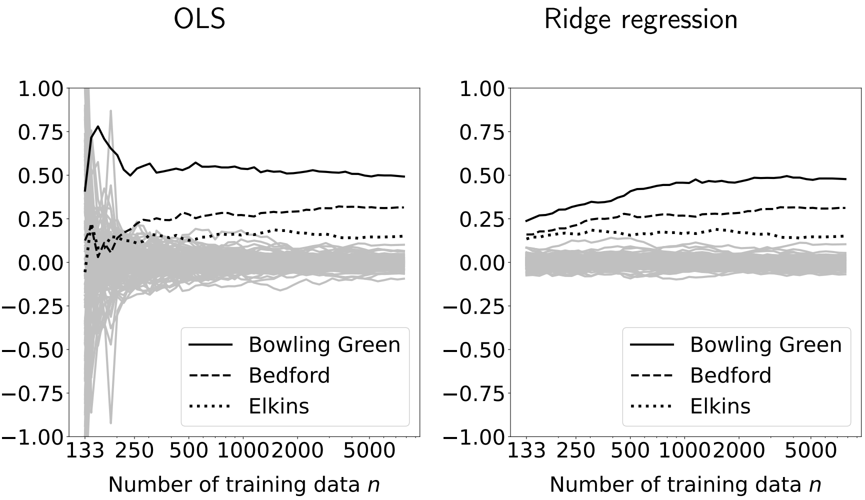
Temperature prediction via ridge regression ($n = 200$)



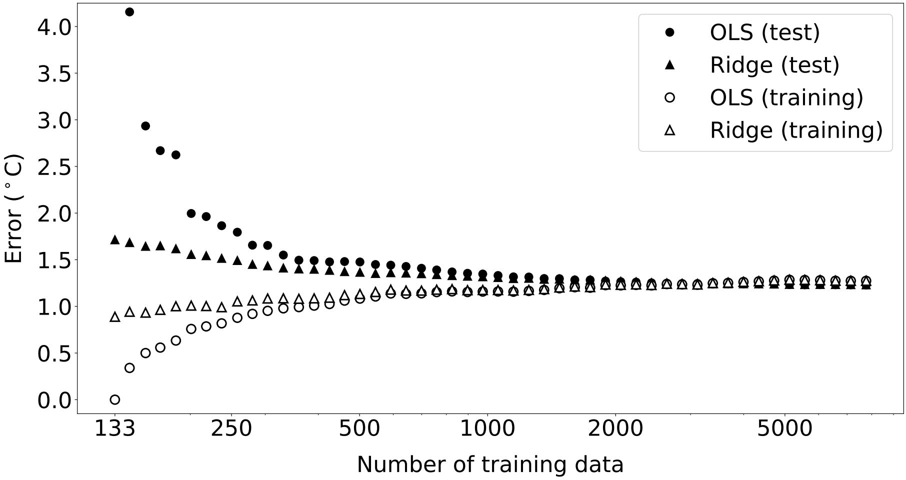
How should we choose λ?



Coefficients



Error



OLS (test)
Ridge (test)
OLS (training)
Ridge (training)
Error (°C)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
133
250
500
1000
2000
5000
Number of training data

Goal

Understand why ridge regression works

Plan:

1. Relationship between OLS and ridge-regression coefficients
2. Simple example: linear response with additive noise
3. Bias-variance comparison with OLS

1. Relationship between OLS and ridge-regression coeffs

OLS cost function in matrix vector form

$$\sum_{i=1}^{n} \left(y_i - \beta^T x_i \right)^2 = ||y_{\text{train}} - X_{\text{train}}\beta||_2^2$$

$$y_{\text{train}} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{bmatrix}$$

Ridge regression cost function

$$
\begin{aligned}
&\sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2 + \lambda \sum_{j=1}^{d} \beta_i^2 \\
&= ||y_{\text{train}} - X_{\text{train}}\beta||_2^2 + \left|\left|0 - \sqrt{\lambda} I \beta\right|\right|_2^2 \\
&= \left|\left| \begin{bmatrix} y_{\text{train}} \\ 0 \end{bmatrix} - \begin{bmatrix} X_{\text{train}} \\ \sqrt{\lambda} I \end{bmatrix} \beta \right|\right|_2^2 \\
&= ||y_{\text{RR}} - X_{\text{RR}}\beta||_2^2
\end{aligned}
$$

Equivalent to OLS cost function with

$$
X_{\text{RR}} := \begin{bmatrix} X_{\text{train}} \\ \sqrt{\lambda} I \end{bmatrix} \qquad y_{\text{RR}} := \begin{bmatrix} y_{\text{train}} \\ 0 \end{bmatrix}
$$

OLS estimator in matrix-vector form

$$
\begin{aligned}
\beta_{\text{OLS}} &= \arg\min_{\beta} ||X_{\text{train}}\beta - y_{\text{train}}||_2^2 \\
&= \Sigma_X^{-1}\Sigma_{XY} \\
&= \left(X_{\text{train}}^T X_{\text{train}}\right)^{-1} X_{\text{train}}^T y_{\text{train}}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_X &= \frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n-1} X_{\text{train}}^T X_{\text{train}} \\
\Sigma_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i = \frac{1}{n-1} X_{\text{train}}^T y_{\text{train}}
\end{aligned}
$$

Ridge-regression estimator

$$
\begin{aligned}
\beta_{\text{RR}} &= \arg\min_{\beta} ||y_{\text{RR}} - X_{\text{RR}}\beta||_2^2 \\
&= \left(X_{\text{RR}}^T X_{\text{RR}}\right)^{-1} X_{\text{RR}}^T y_{\text{RR}} \\
&= \left(\begin{bmatrix} X_{\text{train}}^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X_{\text{train}} \\ \sqrt{\lambda} I \end{bmatrix}\right)^{-1} \begin{bmatrix} X_{\text{train}}^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} y_{\text{train}} \\ 0 \end{bmatrix} \\
&= \left(X_{\text{train}}^T X_{\text{train}} + \lambda I\right)^{-1} X_{\text{train}}^T y_{\text{train}} \\
&= \left(\Sigma_X + \frac{\lambda}{n-1} I\right)^{-1} \Sigma_{XY}
\end{aligned}
$$

PCA perspective: OLS coefficients

$$\begin{aligned}
\Sigma_X &= U \Lambda U^T \\
&= \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \xi_1 & 0 & \cdots & 0 \\ 0 & \xi_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \xi_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T
\end{aligned}$$

$$\begin{aligned}
\beta_{\mathrm{OLS}} &= \Sigma_X^{-1} \Sigma_{XY} \\
&= \sum_{j=1}^{d} \frac{1}{\xi_j} u_j u_j^T \Sigma_{XY} \\
&= \sum_{j=1}^{d} c_{\mathrm{OLS}}[j] u_j \qquad c_{\mathrm{OLS}}[j] = \frac{u_j^T \Sigma_{XY}}{\xi_j}
\end{aligned}$$

PCA perspective: Ridge-regression coefficients

$\lambda_n := \lambda/(n-1)$

$$
\begin{aligned}
\Sigma_X + \lambda_n I &= U \Lambda U^T + \lambda_n U U^T \\
&= U \left(\Lambda + \lambda_n I \right) U^T \\
\left(\Sigma_X + \lambda_n I \right)^{-1} &= U \left(\Lambda + \lambda_n I \right)^{-1} U^T \\
&= \sum_{j=1}^{d} \frac{1}{\xi_j + \lambda_n} u_j u_j^T \\
\beta_{\mathrm{RR}} &= \left(\Sigma_X + \lambda_n I \right)^{-1} \Sigma_{XY} \\
&= \sum_{j=1}^{d} \frac{1}{\xi_j + \lambda_n} u_j u_j^T \Sigma_{XY} \\
&= \sum_{j=1}^{d} c_{\mathrm{RR}}[j] u_j \qquad c_{\mathrm{RR}}[j] = \frac{u_j^T \Sigma_{XY}}{\xi_j + \lambda_n}
\end{aligned}
$$

PCA perspective: OLS vs ridge regression

$$c_{\text{OLS}}[j] = \frac{u_j^T \Sigma_{XY}}{\xi_j} \qquad c_{\text{RR}}[j] = \frac{u_j^T \Sigma_{XY}}{\xi_j + \lambda_n}$$

$$= \frac{\xi_j}{\xi_j + \lambda_n} \frac{u_j^T \Sigma_{XY}}{\xi_j}$$

$$= \frac{c_{\text{OLS}}[j]}{1 + \lambda_n / \xi_j}$$

Ridge regression shrinks the OLS coefficients *in the principal directions of the features*

Key insight: The shrinkage is selective

Selective shrinkage

$$c_{\mathrm{RR}}[j] = \frac{c_{\mathrm{OLS}}[j]}{1 + \lambda_n/\xi_j}$$

Example: ξ_1 is large and ξ_2 is small

Consider λ_n such that $\lambda_n/\xi_1 \ll 1$ and $\lambda_n/\xi_2 \gg 1$

$$c_{\mathrm{RR}}[1] = \frac{c_{\mathrm{OLS}}[1]}{1 + \lambda_n/\xi_1} \approx c_{\mathrm{OLS}}[1]$$

$$c_{\mathrm{RR}}[2] = \frac{c_{\mathrm{OLS}}[2]}{1 + \lambda_n/\xi_2} \approx 0$$

2. Simple example: linear response with additive noise

Linear response with additive noise

$$y_{\text{train}} := X_{\text{train}} \beta_{\text{true}} + z_{\text{train}}$$

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

For simplicity, everything is centered to have zero mean

$$\beta_{\text{OLS}} = \beta_{\text{true}} + \Sigma_X^{-1} \Sigma_{XZ} \qquad \Sigma_{XZ} := \frac{1}{n-1} \sum_{i=1}^{n} x_i z_{\text{train}}[i]$$

Example with independent noise samples

$$\underbrace{\begin{bmatrix} 0.33 \\ 0.91 \\ -1.51 \\ -0.10 \end{bmatrix}}_{y_{\text{train}}} := \underbrace{\begin{bmatrix} 0.46 & 0.44 \\ 0.97 & 1.03 \\ -1.52 & -1.51 \\ 0.09 & 0.04 \end{bmatrix}}_{X_{\text{train}}} \underbrace{\begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}}_{\beta_{\text{true}}} + \underbrace{\begin{bmatrix} -0.13 \\ -0.08 \\ 0.01 \\ -0.18 \end{bmatrix}}_{z_{\text{train}}}$$

$$\beta_{\text{true}} := \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} = c_{\text{true}}[1] u_1 + c_{\text{true}}[2] u_2$$

$$c_{\text{true}} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^T \beta_{\text{true}} = \begin{bmatrix} 0.71 \\ -0.36 \end{bmatrix}$$

PCA perspective: OLS coefficients

$$
\begin{aligned}
\beta_{\text{OLS}} &= \beta_{\text{true}} + \Sigma_X^{-1}\Sigma_{XZ} \\
&= \sum_{j=1}^{2} c_{\text{true}}[j]u_j + \sum_{j=1}^{2} \frac{1}{\xi_j} u_j u_j^T \Sigma_{XZ} \\
&= \sum_{j=1}^{2} \left(c_{\text{true}}[j] + \frac{u_j^T \Sigma_{XZ}}{\xi_j} \right) u_j
\end{aligned}
$$

$$
c_{\text{OLS}} = c_{\text{true}} + \begin{bmatrix} \frac{u_1^T \Sigma_{XZ}}{\xi_1} \\ \frac{u_2^T \Sigma_{XZ}}{\xi_2} \end{bmatrix} = c_{\text{true}} + \begin{bmatrix} -0.03 \\ 2.03 \end{bmatrix}
$$

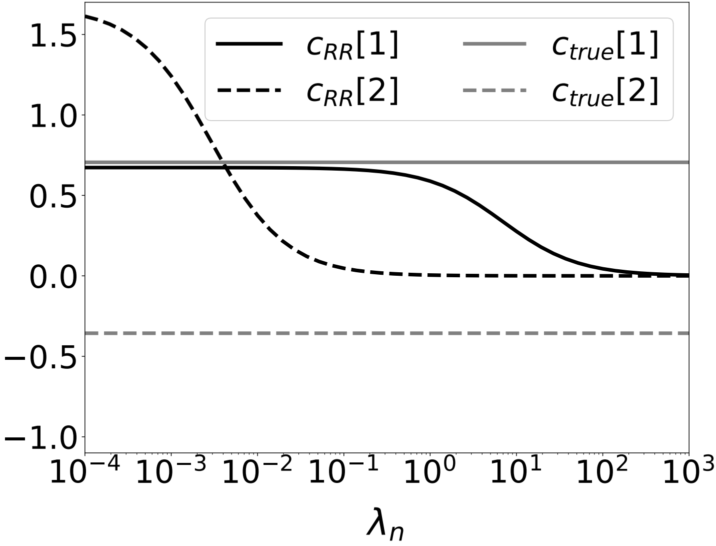
$$
\begin{aligned}
u_1^T \Sigma_{XZ} &:= -0.076 & \xi_1 &:= 2.33 \\
u_2^T \Sigma_{XZ} &:= 0.002 & \xi_2 &:= 9.68 \cdot 10^{-4}
\end{aligned}
$$

PCA perspective: Ridge-regression coefficients

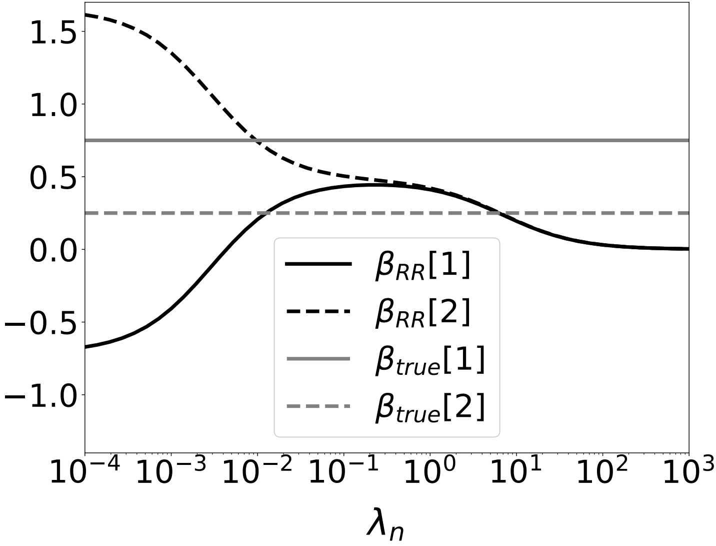
$$c_{\text{RR}}[1] = \frac{c_{\text{OLS}}[1]}{1 + \lambda_n/\xi_1}$$
$$= \frac{c_{\text{true}}[1] - 0.03}{1 + 0.43\lambda_n}$$

$$c_{\text{RR}}[2] = \frac{c_{\text{OLS}}[2]}{1 + \lambda_n/\xi_2}$$
$$= \frac{c_{\text{true}}[2] + 2.03}{1 + 1033\lambda_n}$$

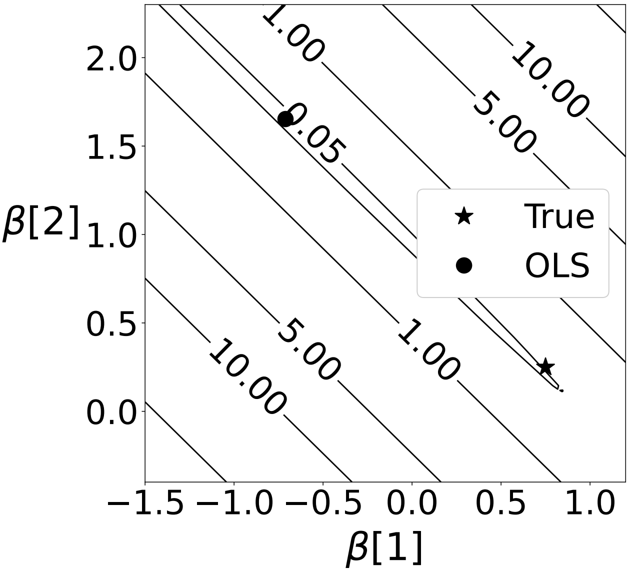
PCA perspective: Ridge-regression coefficients



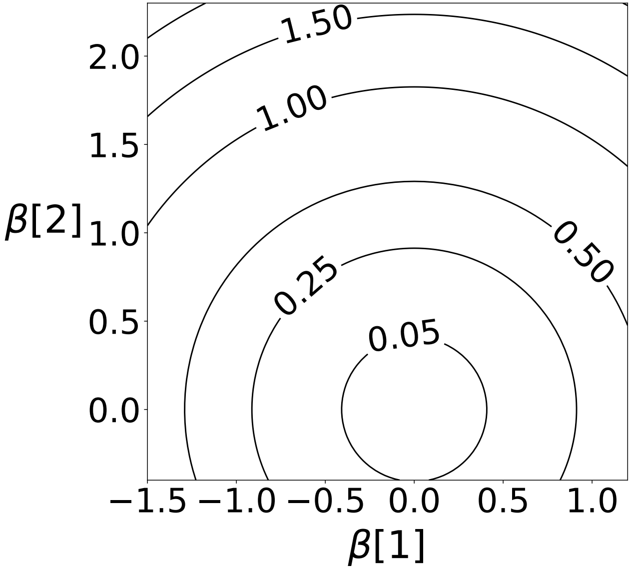
Ridge-regression coefficients



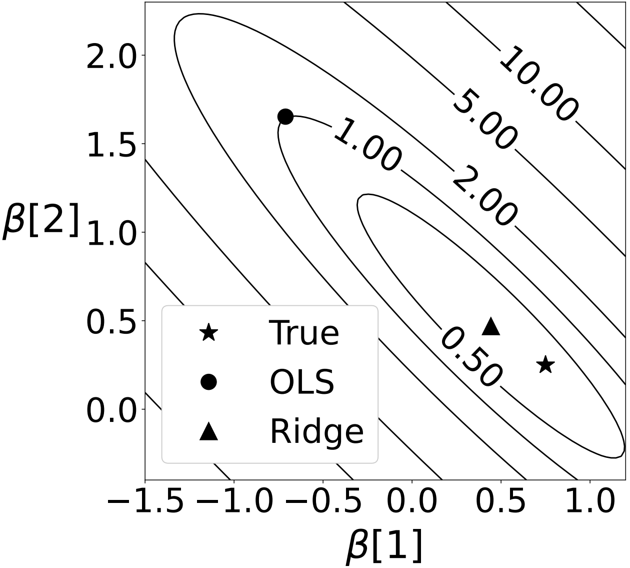
OLS cost function



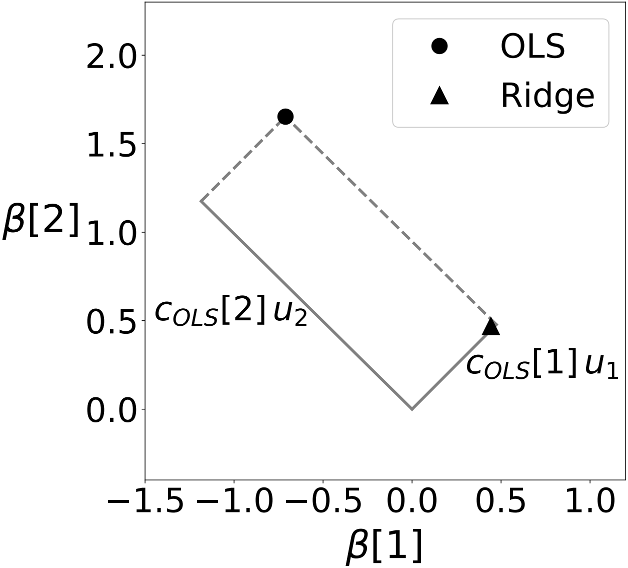
Regularization term ($\lambda_n := 0.1$)



Ridge-regression cost function ($\lambda_n := 0.1$)



Selective shrinkage



3. Bias-variance comparison with OLS

Linear response with random additive noise

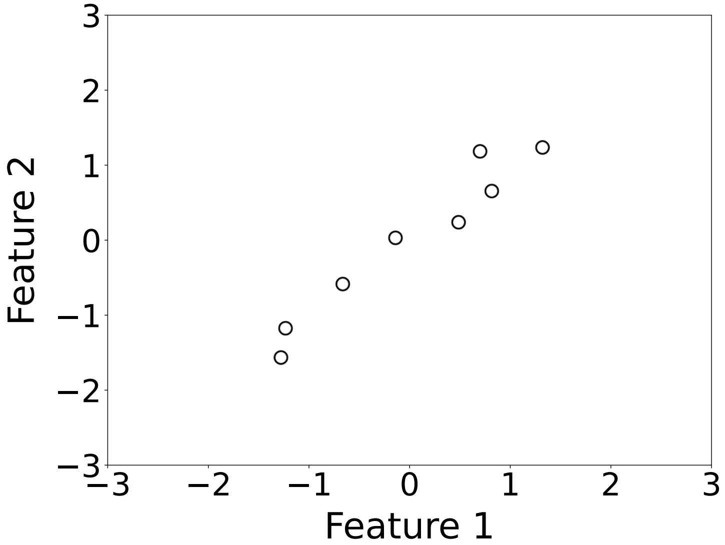
$$\tilde{y}_{\text{train}} := X_{\text{train}} \beta_{\text{true}} + \tilde{z}$$

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

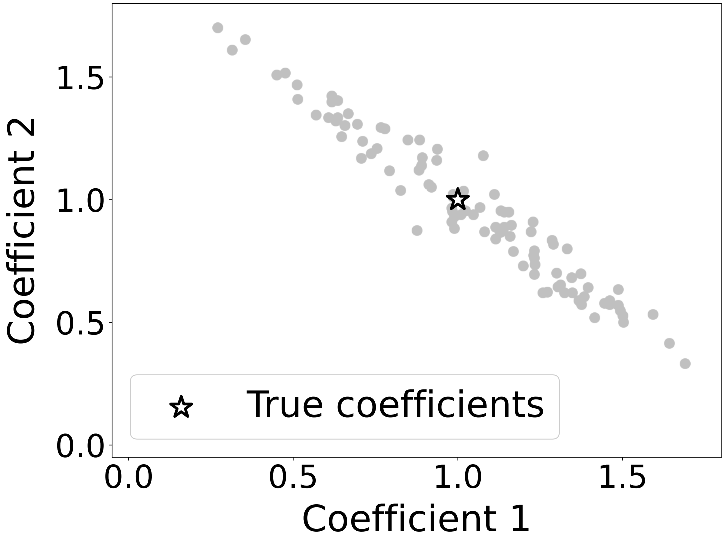
Noise $\tilde{z}$ is i.i.d. with variance σ^2

Everything is centered to have zero mean

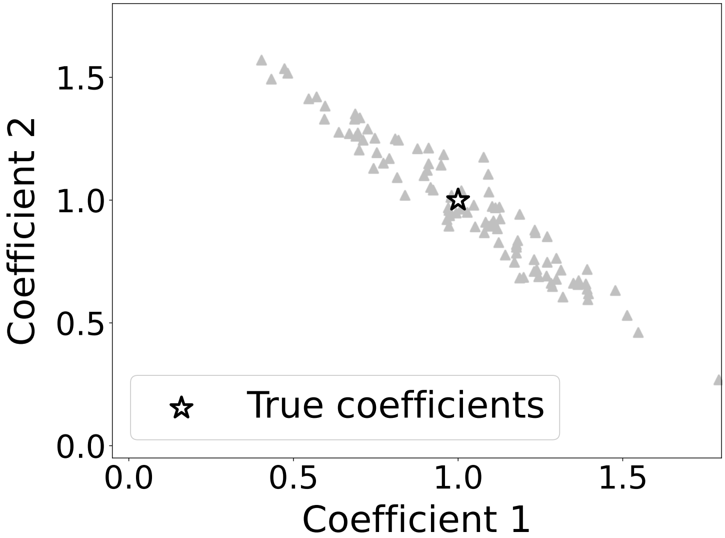
Collinear features ($n := 8$)



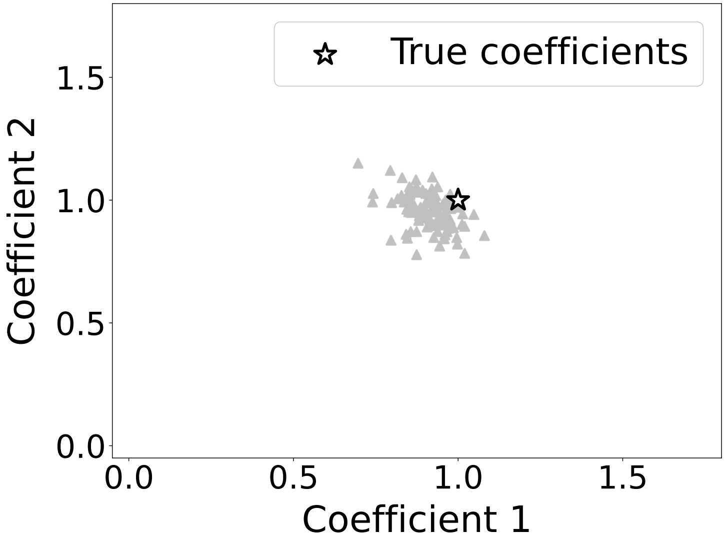
100 OLS coefficients



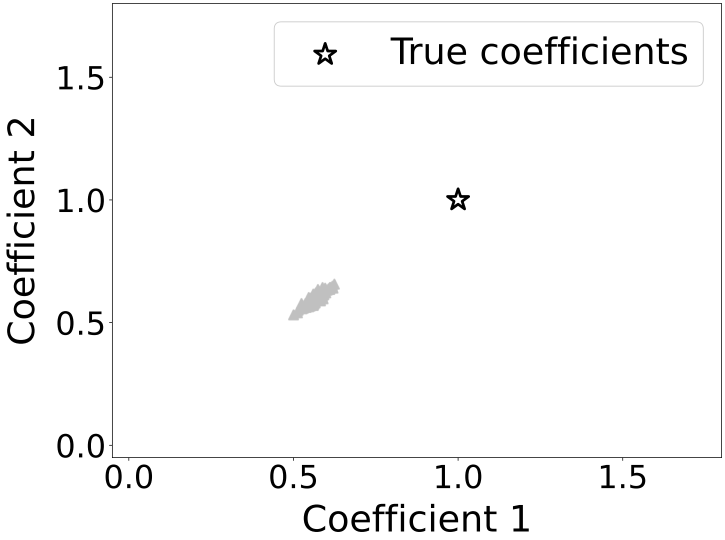
100 ridge-regression coefficients ($\lambda := 0.1$)



100 ridge-regression coefficients ($\lambda := 2$)



100 ridge-regression coefficients ($\lambda := 10$)



Empirical observations

- Unlike OLS, ridge-regression estimates are not centered at true coefficients
- As λ increases, variance decreases faster in directions of low feature variance

Bias in jth principal direction of features

$$c_{\text{true}}[j] := u_j^T \beta_{\text{true}}$$

$$\tilde{c}_{\text{OLS}}[j] := u_j^T \tilde{\beta}_{\text{OLS}}$$

$$\mathrm{E}\left[\tilde{c}_{\text{OLS}}[j]\right] = u_j^T \mathrm{E}\left[\tilde{\beta}_{\text{OLS}}\right] = u_j^T \beta_{\text{true}} = c_{\text{true}}[j]$$

$$\begin{aligned}
\mathrm{E}\left[\tilde{c}_{\text{RR}}[j]\right] &:= u_j^T \mathrm{E}\left[\tilde{\beta}_{\text{RR}}\right] \\
&= \frac{u_j^T \mathrm{E}\left[\tilde{\beta}_{\text{OLS}}\right]}{1 + \lambda_n / \xi_j} \\
&= \frac{u_j^T \beta_{\text{true}}}{1 + \lambda_n / \xi_j} = \frac{c_{\text{true}}[j]}{1 + \lambda_n / \xi_j}
\end{aligned}$$

Variance in jth principal direction of features

$$\operatorname{Var}\left[\tilde{c}_{\mathsf{OLS}}[j]\right] = \frac{\sigma^2}{(n-1)\,\xi_j}$$

$$\operatorname{Var}\left[\tilde{c}_{\mathsf{RR}}[j]\right] = \operatorname{Var}\left[\frac{\tilde{c}_{\mathsf{OLS}}[j]}{1+\lambda_n/\xi_j}\right] = \frac{\operatorname{Var}\left[\tilde{c}_{\mathsf{OLS}}[j]\right]}{(1+\lambda_n/\xi_j)^2}$$

Bias-variance tradeoff

Example: ξ_1 is large and ξ_2 is small

Consider λ_n such that $\lambda_n/\xi_1 \ll 1$ and $\lambda_n/\xi_2 \gg 1$

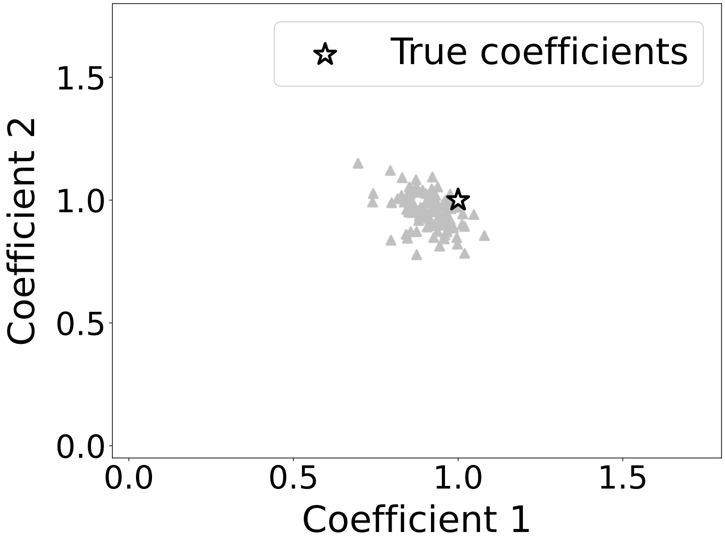
$$\mathrm{E}\left[\tilde{c}_{\text{OLS}}[1]\right] = c_{\text{true}}[1] \qquad \mathrm{Var}\left[\tilde{c}_{\text{OLS}}[1]\right] = \frac{\sigma^2}{(n-1)\,\xi_1} \quad \text{Small}$$

$$\mathrm{E}\left[\tilde{c}_{\text{OLS}}[2]\right] = c_{\text{true}}[2] \qquad \mathrm{Var}\left[\tilde{c}_{\text{OLS}}[2]\right] = \frac{\sigma^2}{(n-1)\,\xi_2} \quad \text{Large}$$

$$\mathrm{E}\left[\tilde{c}_{\text{RR}}[1]\right] = \frac{c_{\text{true}}[1]}{1+\lambda_n/\xi_1} \approx c_{\text{true}}[1] \qquad \mathrm{Var}\left[\tilde{c}_{\text{RR}}[1]\right] = \frac{\mathrm{Var}\left[\tilde{c}_{\text{OLS}}[1]\right]}{(1+\lambda_n/\xi_1)^2} \quad \text{Small}$$

$$\mathrm{E}\left[\tilde{c}_{\text{RR}}[2]\right] = \frac{c_{\text{true}}[2]}{1+\lambda_n/\xi_2} \approx 0 \qquad \mathrm{Var}\left[\tilde{c}_{\text{RR}}[2]\right] = \frac{\mathrm{Var}\left[\tilde{c}_{\text{OLS}}[2]\right]}{(1+\lambda_n/\xi_2)^2} \quad \text{Small}$$

100 ridge-regression coefficients ($\lambda := 2$)



What have we learned?

Regularization prevents overfitting

Ridge regression performs selective shrinkage of OLS coefficients

Variance in directions of low feature variance is reduced,
but bias increases