Ridge regression

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

Carlos Fernandez-Granda

Prerequisites

Ordinary least squares (OLS)

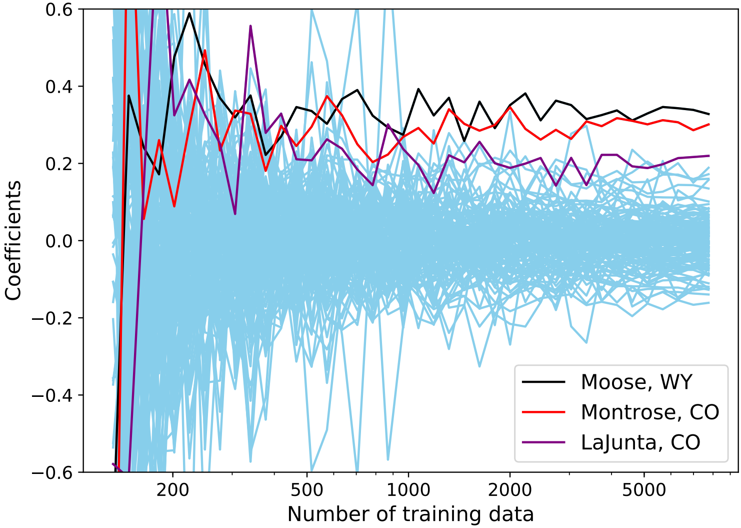
OLS coefficient analysis

OLS training and test error analysis

Temperature prediction via linear regression

- Dataset of hourly temperatures measured at weather stations all over the US
- Goal: Predict temperature in Yosemite from other temperatures
- Response: Temperature in Yosemite
- Features: Temperatures in 133 other stations ($p = 133$) in 2015
- Test set: 10^3 measurements
- Additional test set: All measurements from 2016

OLS coefficients



Motivation

Overfitting often reflected in large coefficients that cancel out to match the noise

Possible solution: Penalize large-norm solutions when fitting the model

Adding a penalty term to promote certain properties is called regularization

Ridge regression

For a fixed regularization parameter $\lambda > 0$

$$\beta_{\text{RR}} := \arg\min_{\beta} \|y - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

What happens when $\lambda \to 0$? $\beta_{\text{RR}} \to \beta_{\text{OLS}}$

What happens when $\lambda \to \infty$? $\beta_{\text{RR}} \to 0$

Ridge regression

β_{RR} is the solution to a modified least-squares problem

$$\begin{aligned}
\beta_{\text{RR}} &= \arg\min_{\beta} \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \beta \right|\right|_2^2 \\
&= \left(\begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix}^T \right)^{-1} \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} \\
&= \left(X X^T + \lambda I \right)^{-1} X y
\end{aligned}$$

Problem

How to calibrate regularization parameter

Should we choose that λ that yields the best fit? No!

Better option: Check fit on validation data

Cross validation

Given a set of examples

$$\left(y^{(1)}, x^{(1)}\right), \left(y^{(2)}, x^{(2)}\right), \ldots, \left(y^{(n)}, x^{(n)}\right),$$

1. Partition data into a training set $X_{\text{train}} \in \mathbb{R}^{n_{\text{train}} \times p}$, $y_{\text{train}} \in \mathbb{R}^{n_{\text{train}}}$ and a validation set $X_{\text{val}} \in \mathbb{R}^{n_{\text{val}} \times p}$, $y_{\text{val}} \in \mathbb{R}^{n_{\text{val}}}$

2. Fit model using the training set for every λ in a set Λ

$$\beta_{\text{RR}}(\lambda) := \arg\min_{\beta} \left|\left| y_{\text{train}} - X_{\text{train}}\beta \right|\right|_2^2 + \lambda \left|\left| \beta \right|\right|_2^2$$

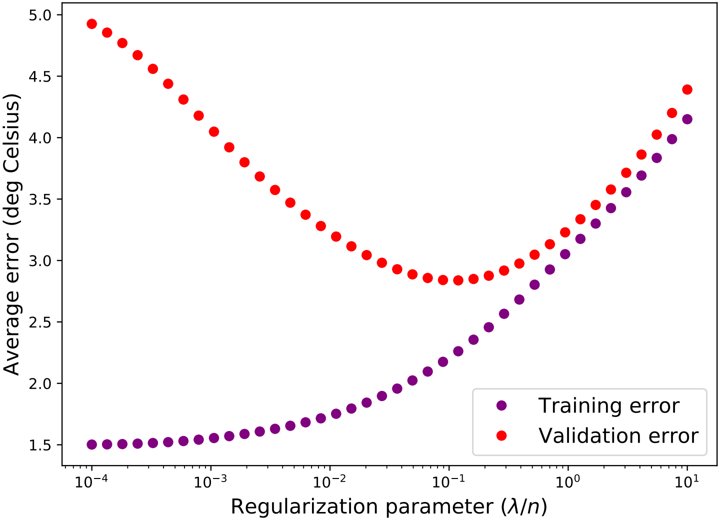
   and evaluate the fitting error on the validation set

$$\text{err}(\lambda) := \left|\left| y_{\text{val}} - X_{\text{val}}\beta_{\text{RR}}(\lambda) \right|\right|_2^2$$

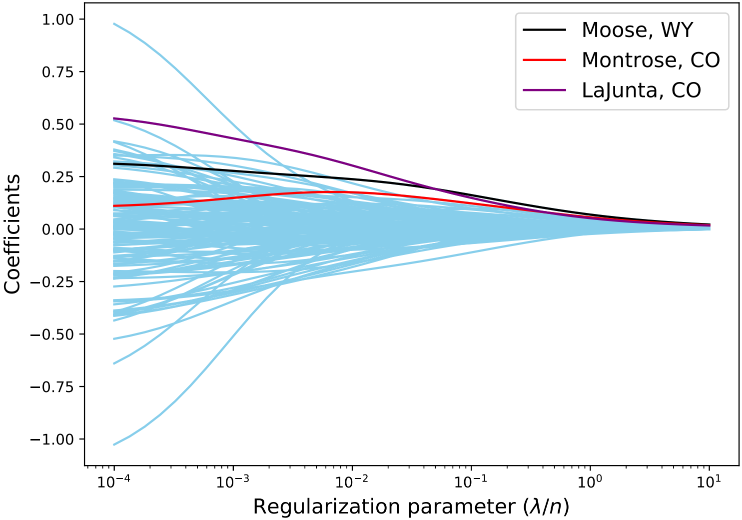
3. Choose the value of λ that minimizes the validation-set error

$$\lambda_{\text{cv}} := \arg\min_{\lambda \in \Lambda} \text{err}(\lambda)$$

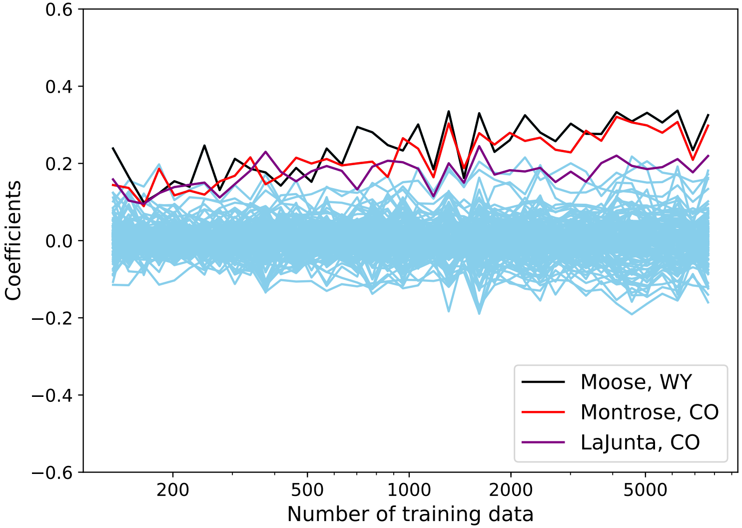
Temperature prediction via ridge regression ($n = 202$)



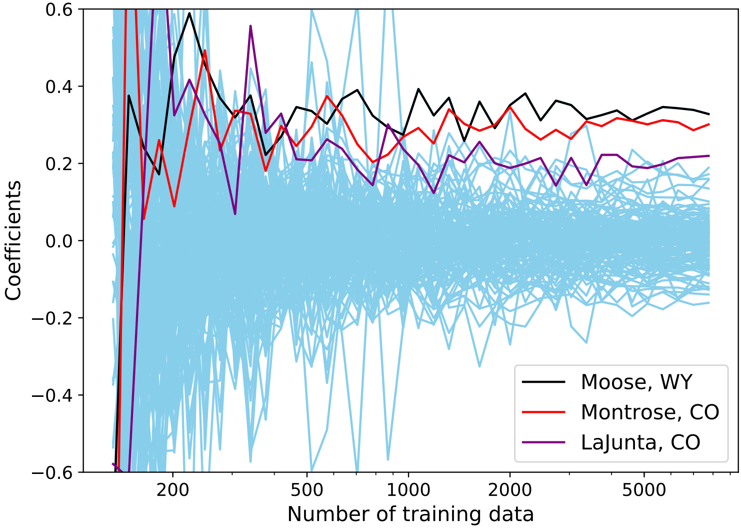
Temperature prediction via ridge regression ($n = 202$)



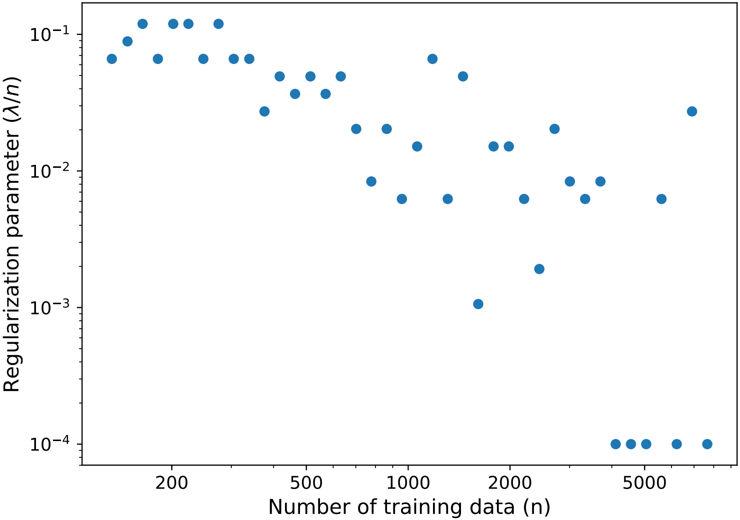
Ridge regression coefficients



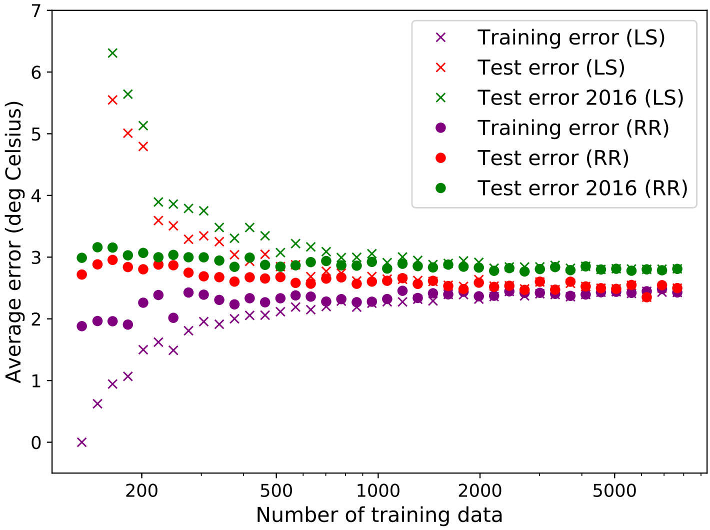
OLS coefficients



Regularization parameter



Temperature prediction via ridge regression



Additive model

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

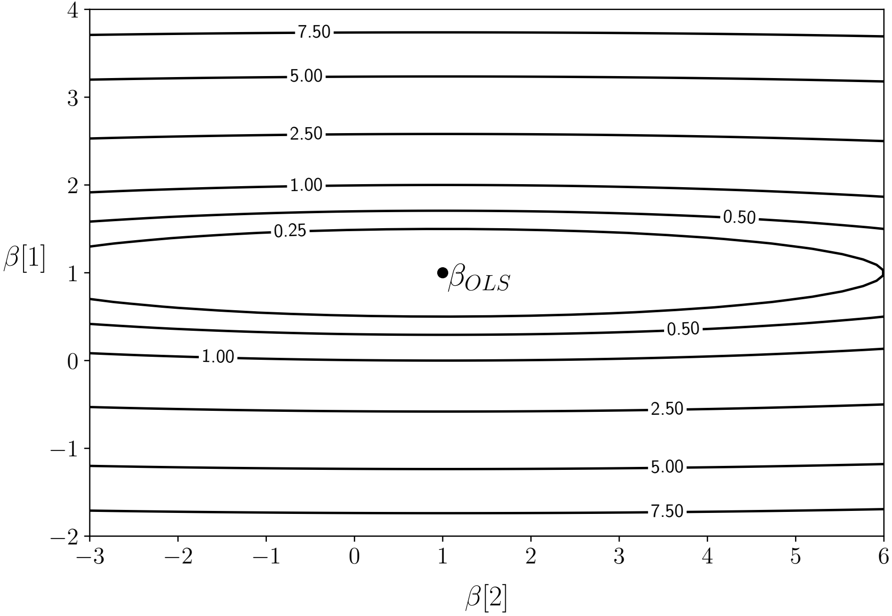
Goal: Understand how ridge regression works

Decomposition of ridge-regression cost function

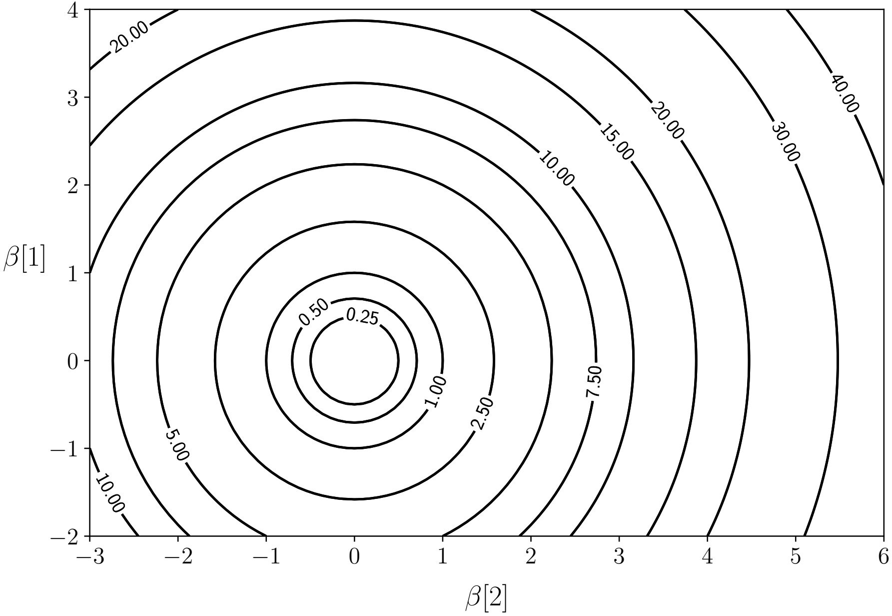
$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda\beta^T\beta - 2\tilde{z}_{\text{train}}^T X^T \beta$$

$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}})$

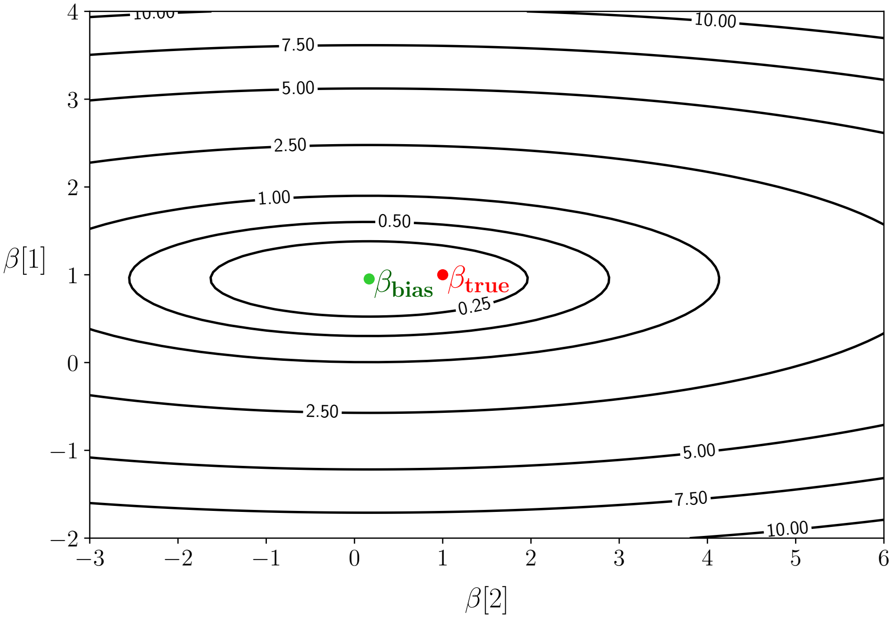


$\beta^T\beta$

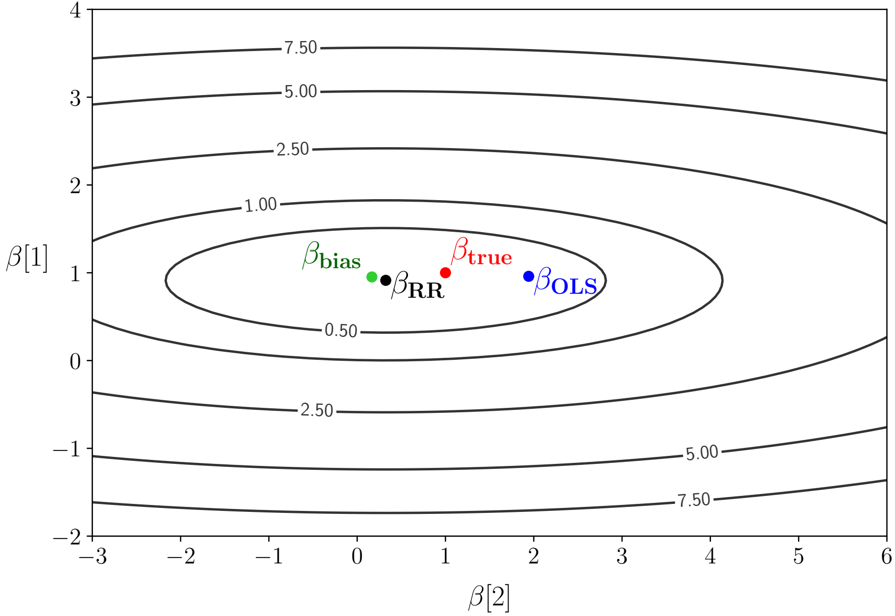


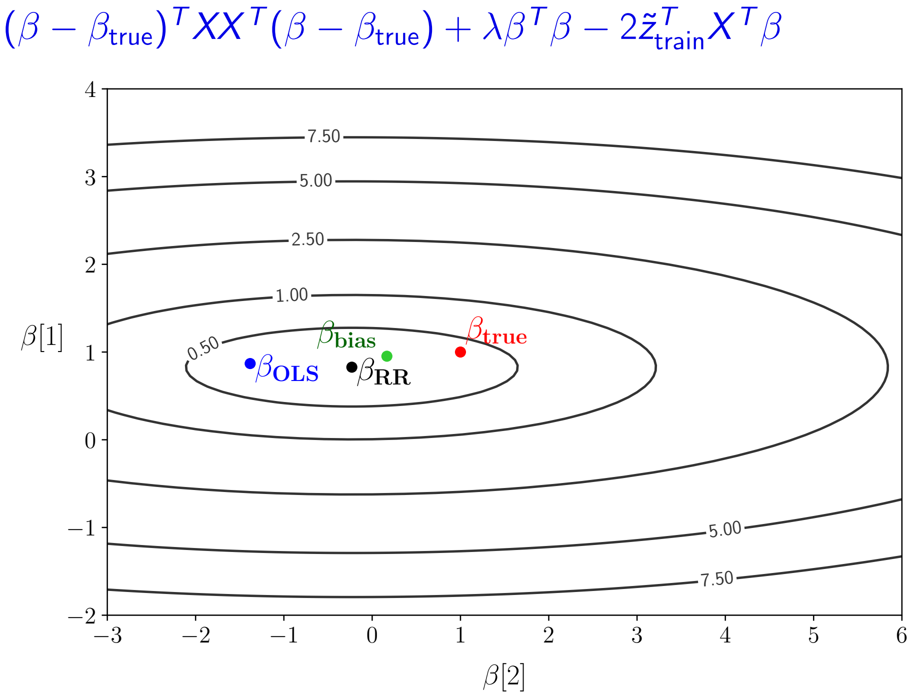
0.25
0.50
1.00
2.50
5.00
7.50
10.00
15.00
20.00
30.00
40.00
β[1]
β[2]
−3
−2
−1
0
1
2
3
4
5
6

$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta$$



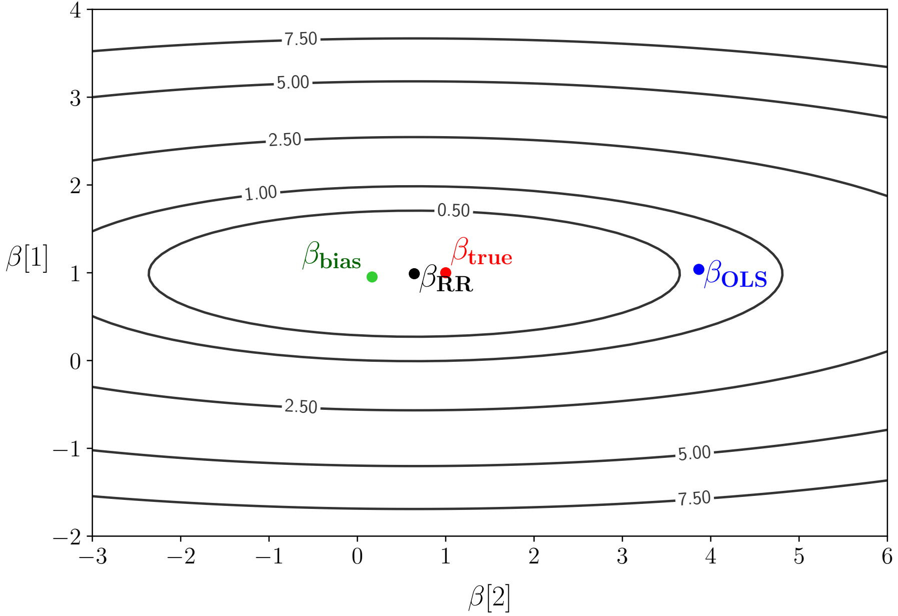
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta - 2\tilde{z}_{\text{train}}^T X^T \beta$$



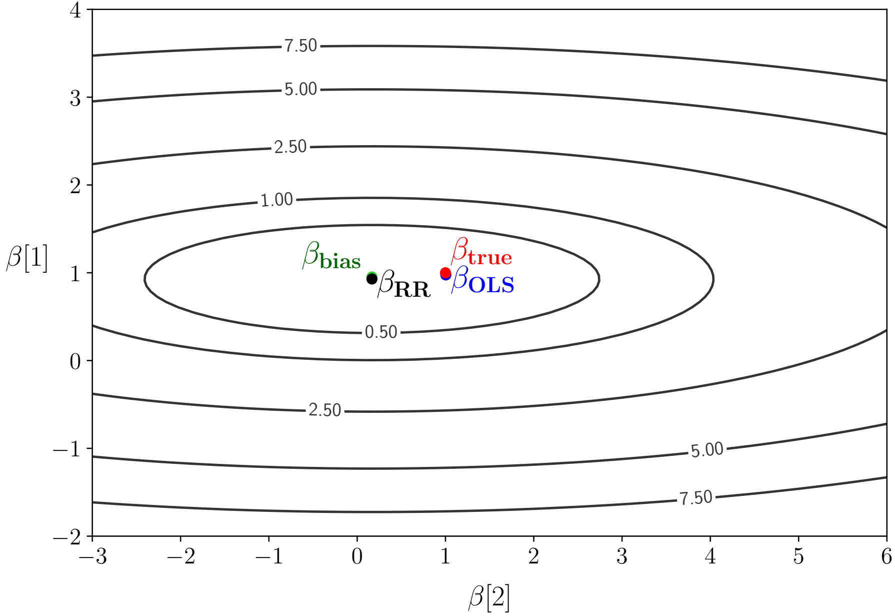


$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta - 2\tilde{z}_{\text{train}}^T X^T \beta$
7.50
5.00
2.50
1.00
0.50
β_{bias}
β_{true}
β_{OLS}
β_{RR}
5.00
7.50
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
$\beta[1]$
$\beta[2]$

$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta - 2\tilde{z}_{\text{train}}^T X^T \beta$$



$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta - 2\tilde{z}_{\text{train}}^T X^T \beta$$



Ridge-regression coefficient estimate

$$
\begin{aligned}
\tilde{\beta}_{\text{RR}} &= \left(XX^T + \lambda I\right)^{-1} X \left(X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}\right) \\
&= \left(US^2U^T + \lambda UU^T\right)^{-1} \left(US^2U^T \beta_{\text{true}} + USV^T \tilde{z}_{\text{train}}\right) \\
&= \left(U(S^2 + \lambda I)U^T\right)^{-1} \left(US^2U^T \beta_{\text{true}} + USV^T \tilde{z}_{\text{train}}\right) \\
&= U(S^2 + \lambda I)^{-1} U^T \left(US^2U^T \beta_{\text{true}} + USV^T \tilde{z}_{\text{train}}\right) \\
&= {\color{red} U(S^2 + \lambda I)^{-1} S^2 U^T \beta_{\text{true}}} + {\color{blue} U\left(S^2 + \lambda I\right)^{-1} SV^T \tilde{z}_{\text{train}}}
\end{aligned}
$$

Ridge-regression coefficient estimate

$$\tilde{\beta}_{\text{RR}} = U(S^2 + \lambda I)^{-1} S^2 U^T \beta_{\text{true}} + U \left(S^2 + \lambda I\right)^{-1} S V^T \tilde{z}_{\text{train}}$$

Distribution? Gaussian with mean

$$\beta_{\text{bias}} := \sum_{j=1}^{p} \frac{s_j^2 \left\langle u_j, \beta_{\text{true}} \right\rangle}{s_j^2 + \lambda} u_j$$

and covariance matrix

$$\Sigma_{\text{RR}} := \sigma^2 U \operatorname{diag}_{j=1}^{p} \left(\frac{s_j^2}{(s_j^2 + \lambda)^2} \right) U^T$$

Bias

In contrast to OLS, ridge regression produces systematic error

$$\mathrm{E}(\beta_{\text{true}} - \tilde{\beta}_{\text{RR}}) = \sum_{j=1}^{p} \left(\frac{\lambda \langle u_j, \beta_{\text{true}} \rangle}{s_j^2 + \lambda} - \frac{s_j \langle v_j, \mathrm{E}(\tilde{z}_{\text{train}}) \rangle}{s_j^2 + \lambda} \right) u_j$$

$$= \sum_{j=1}^{p} \frac{\lambda \langle u_j, \beta_{\text{true}} \rangle}{s_j^2 + \lambda} u_j$$

Bias grows with λ, so what's the point?

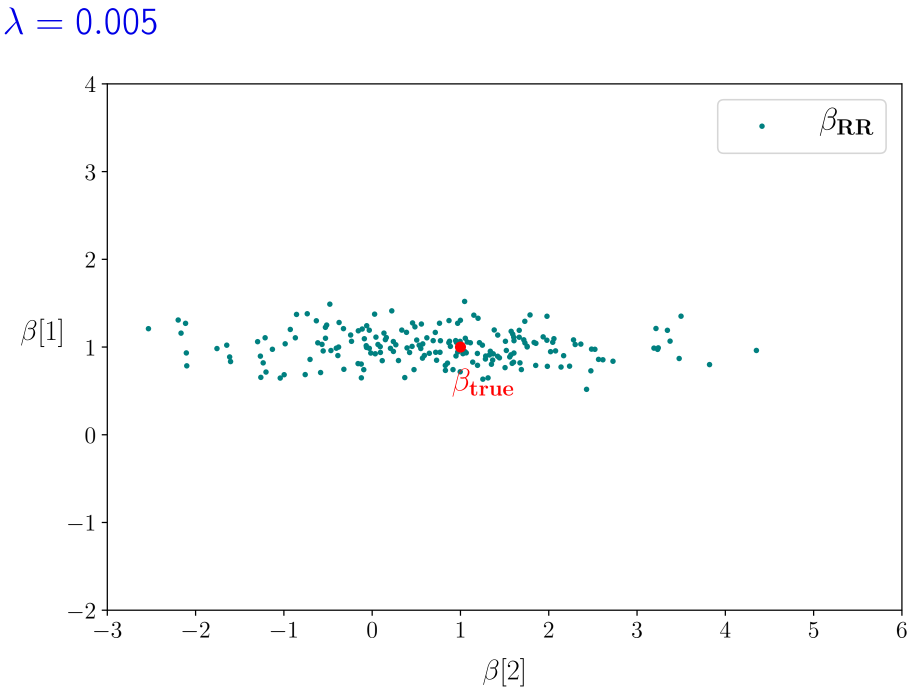
Variance

Variance in direction of u_i equals $\frac{\sigma^2 s_i^2}{(s_i^2+\lambda)^2}$

Small s_i blow up variance of OLS

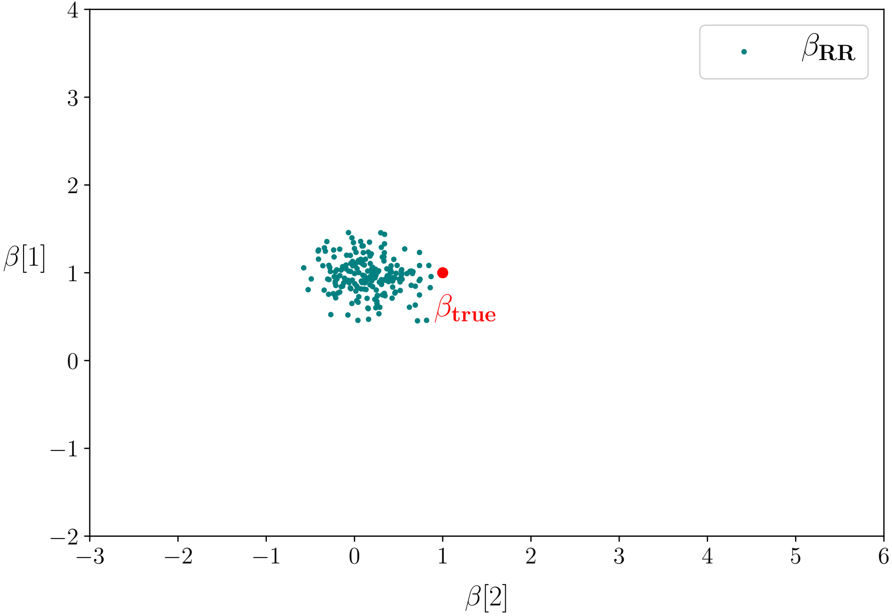
If $\lambda \gg s_i^2$, then the variance $\approx \sigma^2 s_i^2/\lambda^2 \ll \sigma^2/s_i^2$ if s_i small

Ideal λ achieves bias-variance tradeoff



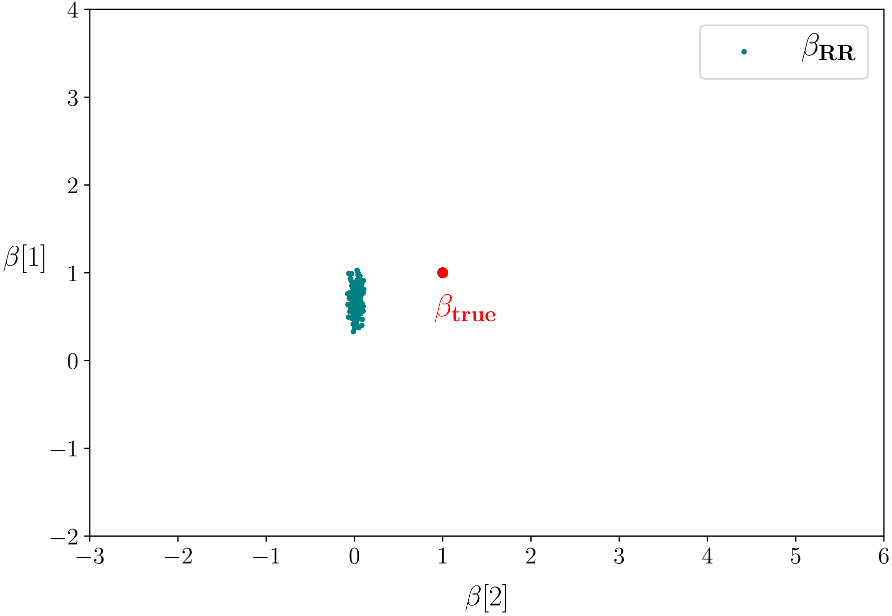
$\lambda = 0.005$
$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

$\lambda = 0.05$



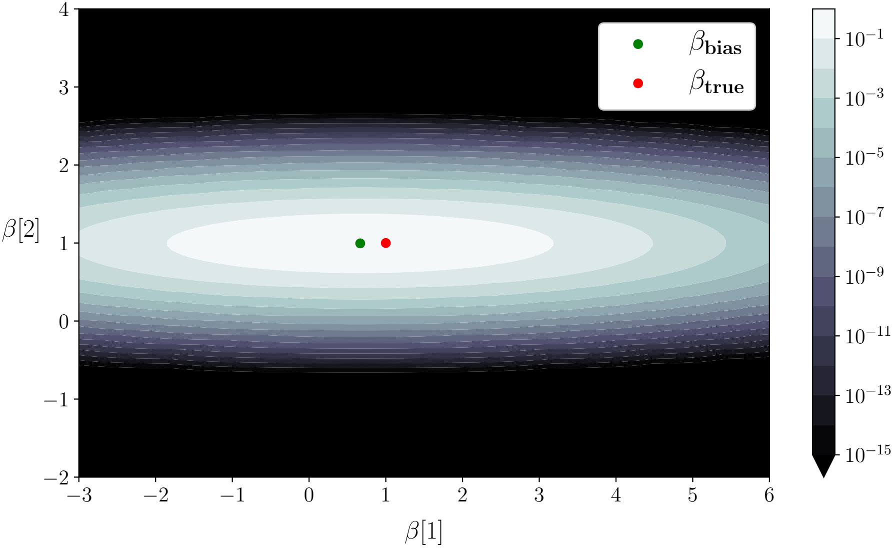
$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

$\lambda = 0.5$



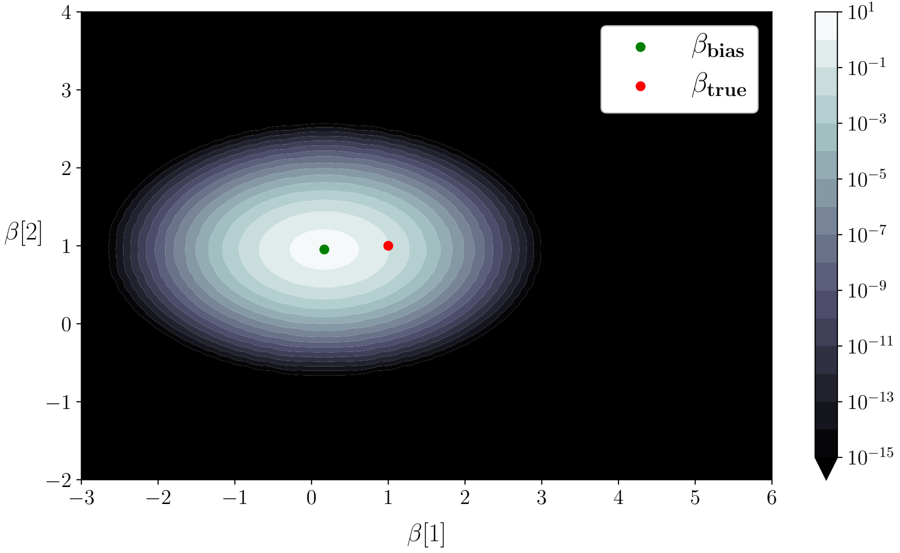
$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

$\lambda = 0.005$



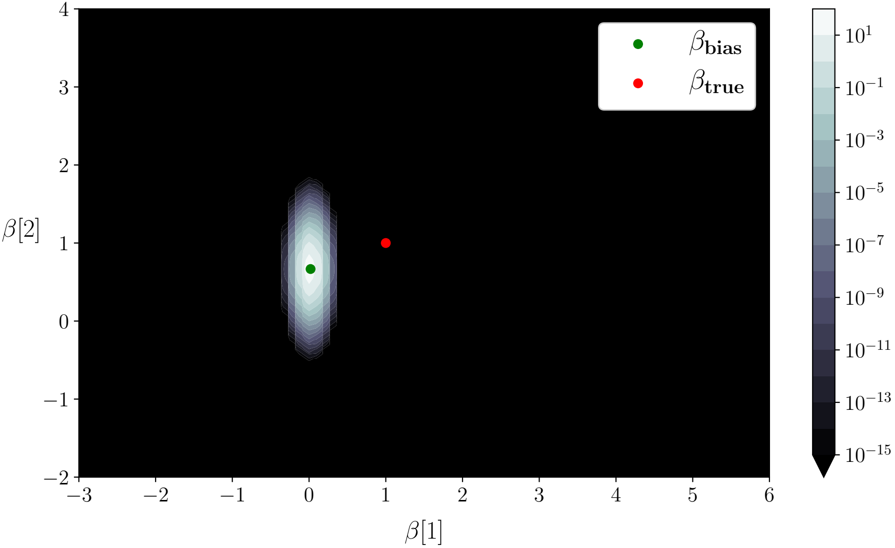
$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$\lambda = 0.05$



$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$\lambda = 0.5$



$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

What have we learned

- Ridge regression prevents overfitting by penalizing large linear coefficients
- This produces a biased estimate (under linear data model with additive noise)
- Regularization parameter balances bias and variance from small singular values of feature matrix