The Lasso:

Sparse Regression via ℓ_1-norm Regularization

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 y from d features x

Linear regression:

$$\ell(x) = \sum_{j=1}^{d} \beta[j]x[j] = \beta^T x$$

We assume features and response are centered to have zero mean

Ordinary least squares estimator

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

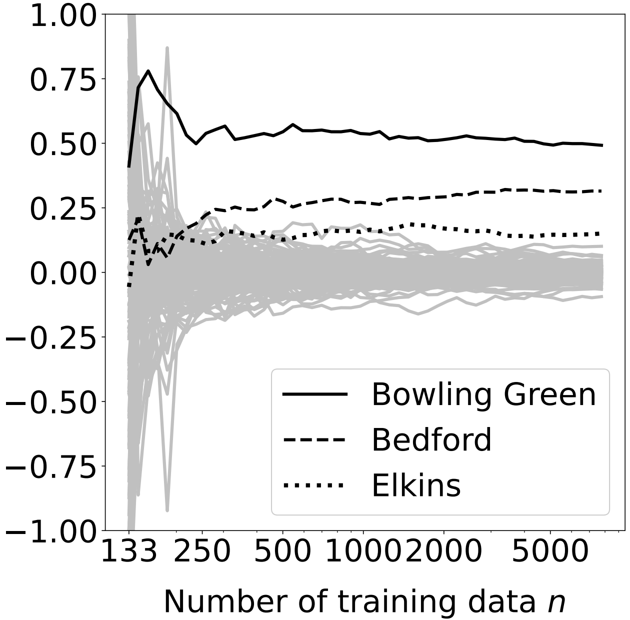
Temperature prediction

Response: Temperature in Versailles (Kentucky)

Features: Temperatures at 133 other locations

OLS coefficients (large *n*)

OLS coefficients



Sparse regression

Goal: Identify a small subset of features that provide a good fit

Equivalently, find sparse coefficients β that provide a good fit

If k out of d coefficients are nonzero

$$\sum_{j=1}^{d} \beta[j]x[j] = \sum_{j \in \mathcal{K}} \beta[j]x[j]$$

where $x \in \mathbb{R}^{d}$ is a feature vector and $\mathcal{K}$ the set of nonzero coefficients

Ridge regression

Ridge regression penalizes the ℓ_2 norm of the coefficients

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

$\lambda > 0$ is a regularization parameter

Does this produce sparse coefficients?

Linear response with random additive noise

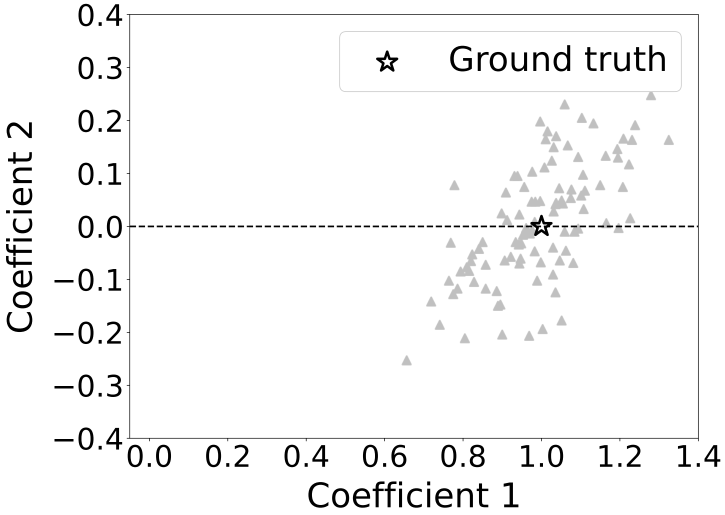
$$\tilde{y}_i := x_i[1] + \tilde{z} \quad 1 \leq i \leq n$$

$$X_{\text{train}} := \begin{bmatrix} x_1[1] & x_1[2] \\ x_2[1] & x_2[2] \\ \cdots & \\ x_n[1] & x_n[2] \end{bmatrix} \qquad \beta_{\text{true}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

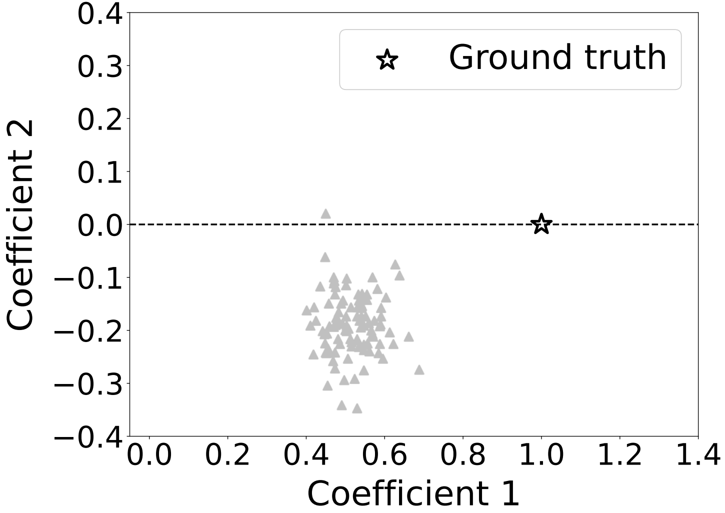
Noise $\tilde{z}$ is i.i.d. with fixed variance

Everything is centered to have zero mean

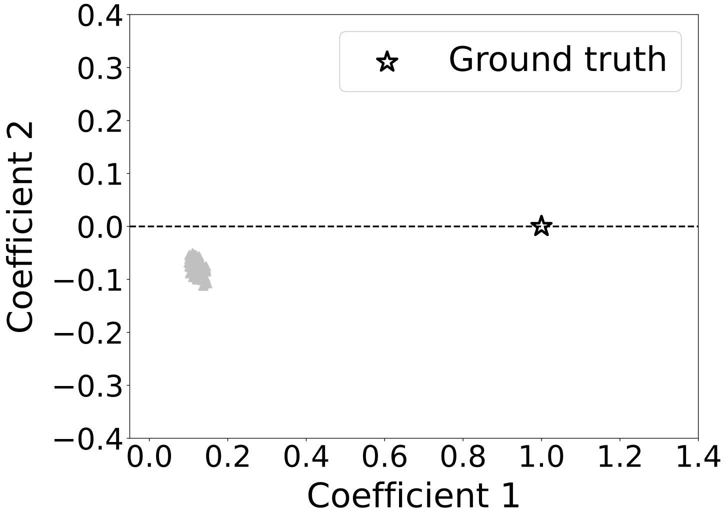
Ridge regression: Small λ



Ridge regression: Medium λ



Ridge regression: Large λ



What is going on?

Promoting sparsity requires shrinking entries

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

Principal directions are usually dense, so ridge-regression coefficients are not sparse

ℓ_2 norm vs ℓ_1 norm

$$||\beta||_2 := \sqrt{\sum_{i=1}^{d} \beta[i]^2}$$

$$||\beta||_1 := \sum_{i=1}^{d} |\beta[i]|$$

ℓ_2 norm vs ℓ_1 norm

$$\beta_{\text{dense}} := \begin{bmatrix} \frac{1}{\sqrt{d}} \\ \frac{1}{\sqrt{d}} \\ \cdots \\ \frac{1}{\sqrt{d}} \end{bmatrix} \qquad \beta_{\text{sparse}} := \begin{bmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

$$\begin{aligned} ||\beta_{\text{dense}}||_2^2 &= \sum_{i=1}^{d} \beta_{\text{dense}}[i]^2 \\ &= 1 \\ ||\beta_{\text{dense}}||_1 &= \sum_{i=1}^{d} |\beta_{\text{dense}}[i]| \\ &= \sqrt{d} \end{aligned} \qquad \begin{aligned} ||\beta_{\text{sparse}}||_2^2 &= \sum_{i=1}^{d} \beta_{\text{sparse}}[i]^2 \\ &= 1 \\ ||\beta_{\text{sparse}}||_1 &= \sum_{i=1}^{d} |\beta_{\text{sparse}}[i]| \\ &= 1 \end{aligned}$$

The lasso

Regularization penalizes the ℓ_1 norm of the coefficients (instead of ℓ_2 norm)

$$\beta_{\text{lasso}} := \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta^T x_i \right)^2 + \lambda \left|\left|\beta\right|\right|_1$$

$\lambda > 0$ is a regularization parameter

Does this produce sparse coefficients?

Linear response with random additive noise

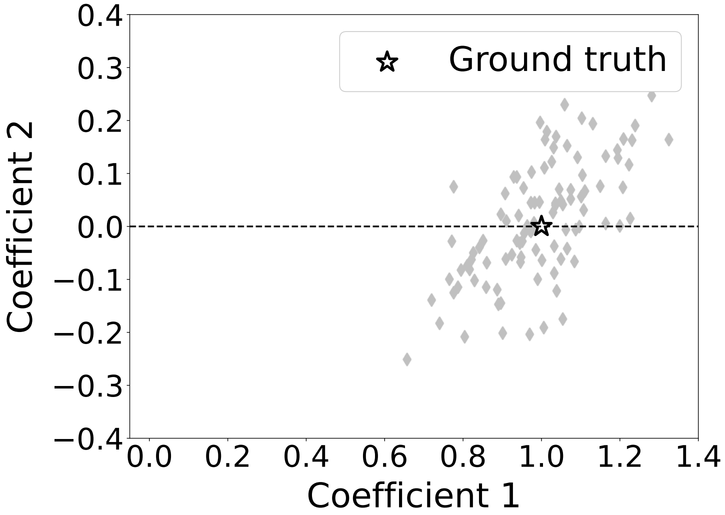
$$\tilde{y}_i := x_i[1] + \tilde{z} \quad 1 \leq i \leq n$$

$$X_{\text{train}} := \begin{bmatrix} x_1[1] & x_1[2] \\ x_2[1] & x_2[2] \\ \cdots & \\ x_n[1] & x_n[2] \end{bmatrix} \qquad \beta_{\text{true}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

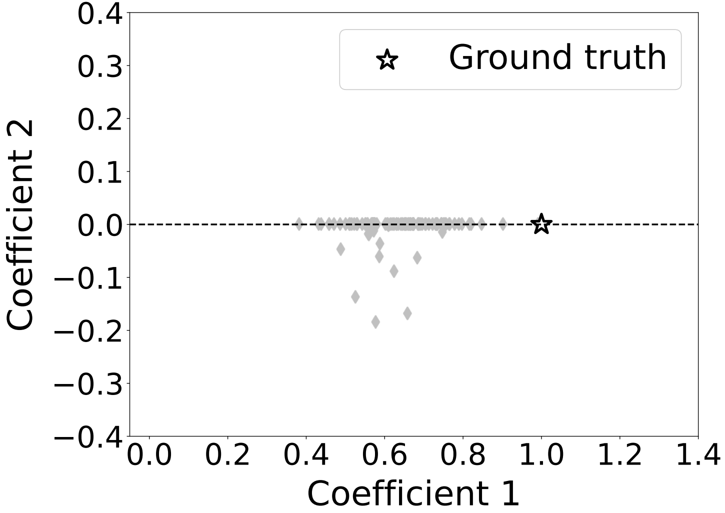
Noise $\tilde{z}$ is i.i.d. with fixed variance

Everything is centered to have zero mean

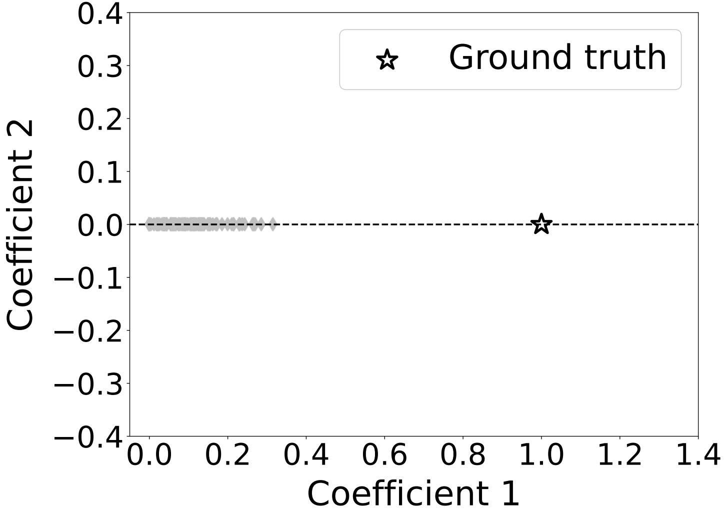
Lasso: Small λ



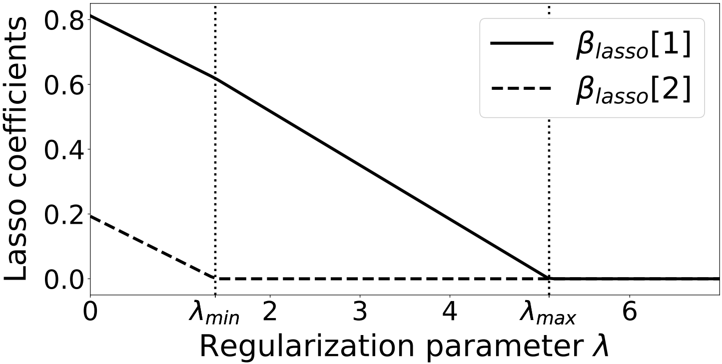
Lasso: Medium λ



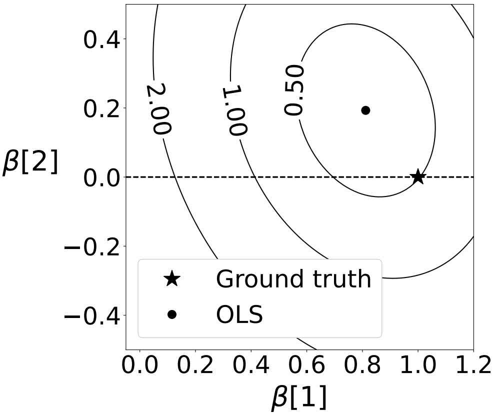
Lasso: Large λ

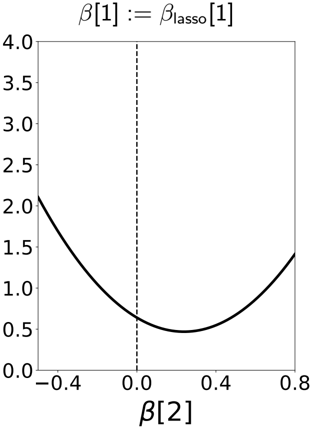


Lasso coefficients

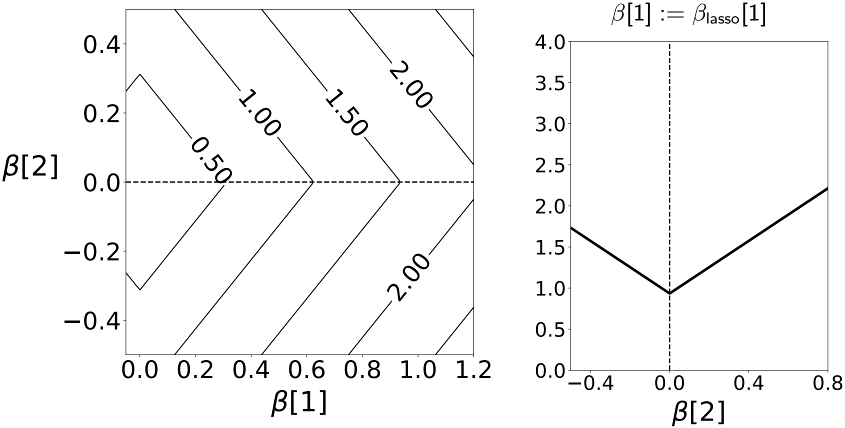


OLS cost function

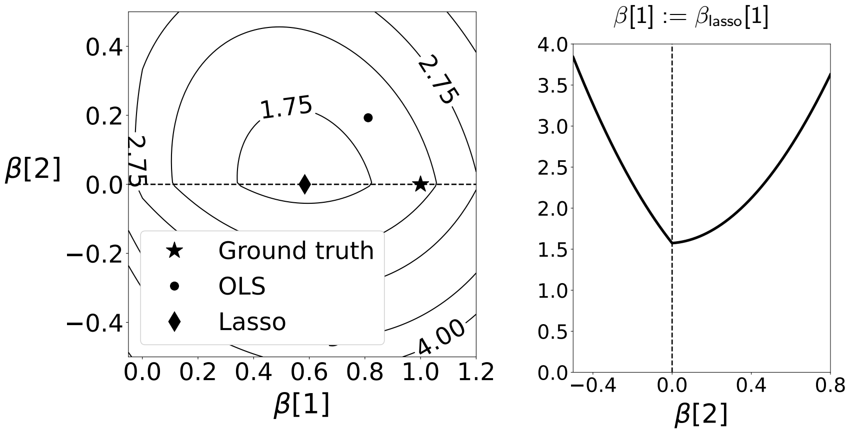




Regularization term (medium λ)



Lasso cost function (medium λ)



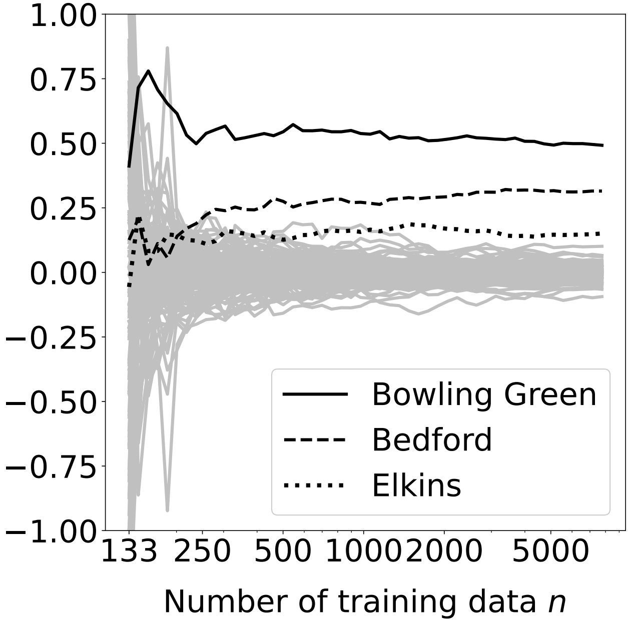
Temperature prediction

Response: Temperature in Versailles (Kentucky)

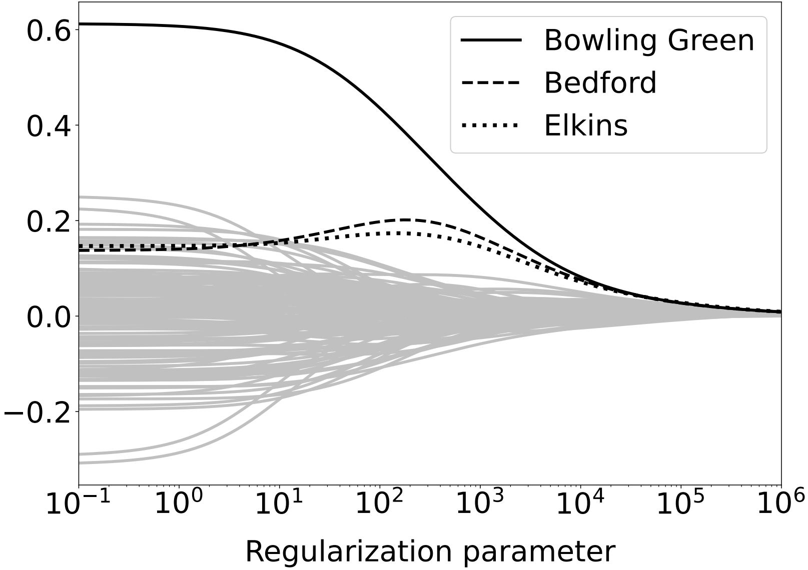
Features: Temperatures at 133 other locations

OLS coefficients (large n)

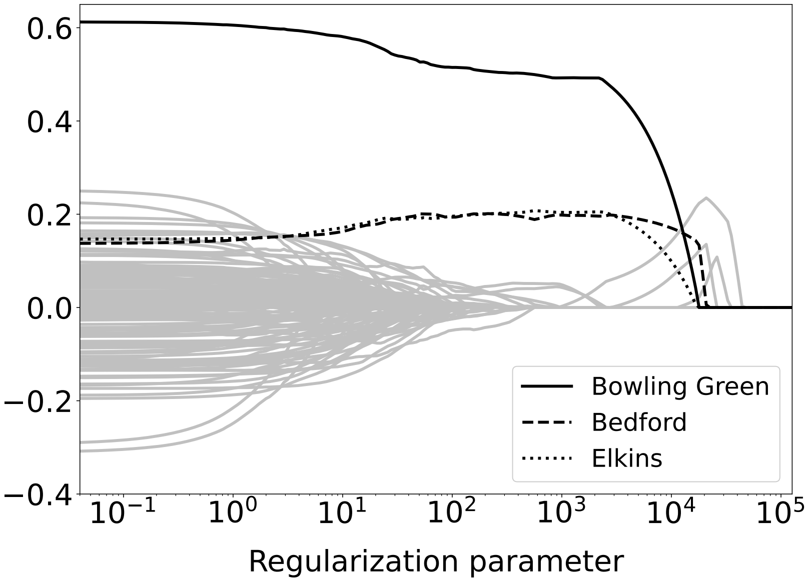
OLS coefficients



Ridge regression ($n = 200$)

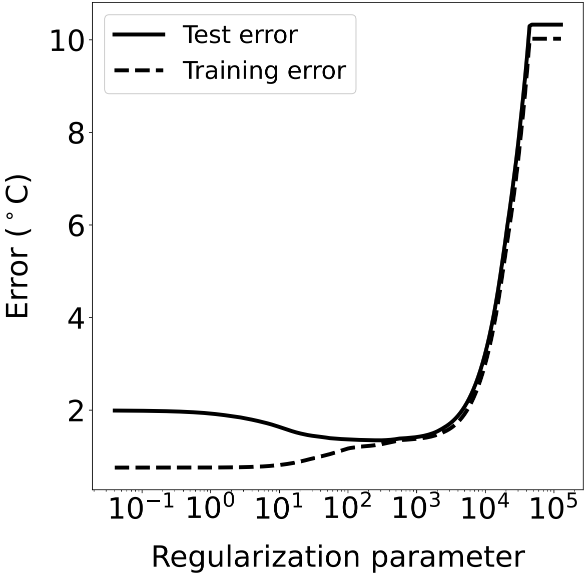


Lasso ($n = 200$)



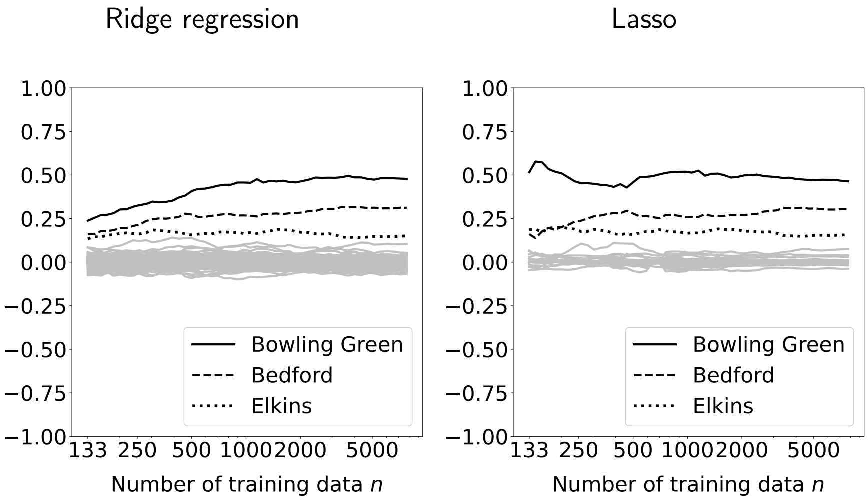
0.6
0.4
0.2
0.0
−0.2
−0.4
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
Regularization parameter
Bowling Green
Bedford
Elkins

Lasso ($n = 200$)

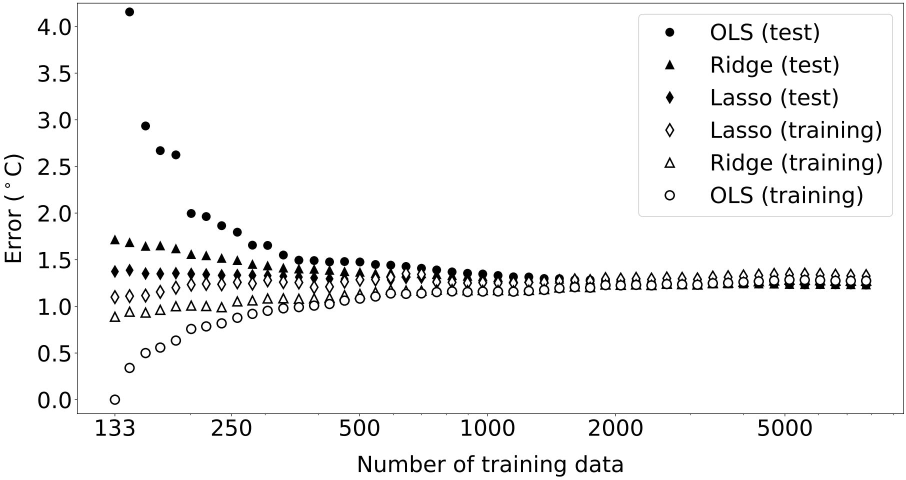


Test error
Training error
Error (°C)
Regularization parameter
10
8
6
4
2
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}

Coefficients



Error



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

What have we learned?

ℓ_1-norm regularization promotes sparsity

If there are many features, sparse regression can avoid overfitting (and provide interpretability)