Sparse regression via the lasso

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

Carlos Fernandez-Granda

Prerequisites

Ordinary least squares (OLS)

Ridge regression

Sparse regression

Linear regression is challenging when the number of features p is large

Possible solution: Select subset of relevant features $\mathcal{I} \subset \{1, \ldots, p\}$, so that

$$y \approx \sum_{i \in \mathcal{I}} \beta[i] x[i]$$

Problem: How do we find this subset?

Equivalently, how do we find a sparse coefficient vector $\beta \in \mathbb{R}^p$ such that

$$y \approx \langle x, \beta \rangle$$

Toy problem

Find t such that

$$v_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}$$

is sparse

Equivalently, find $\arg\min_t ||v_t||_0$

ℓ_0 "norm"

Number of nonzero entries in a vector

Is this a norm?

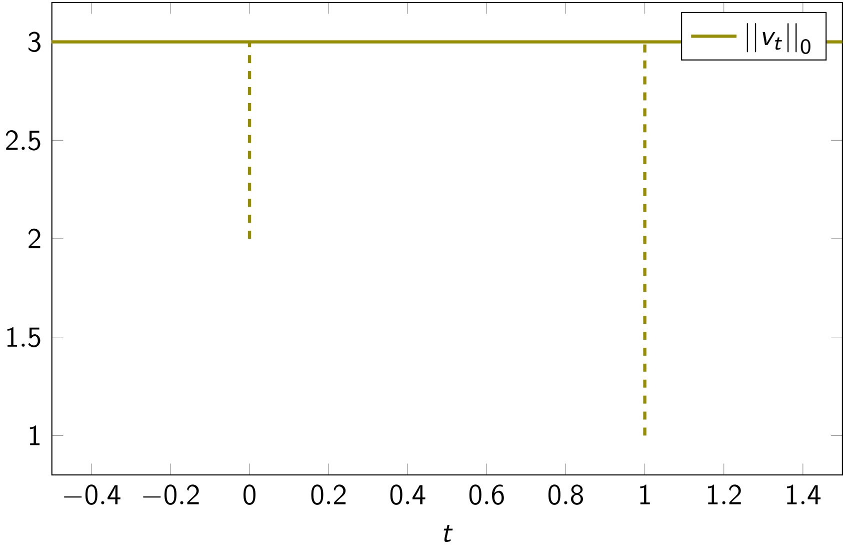
All norms are homogeneous, $||\alpha x|| = \alpha ||x||$ for any x and any scalar $\alpha > 0$

$$\begin{aligned} ||2x||_0 &= ||x||_0 \\ &\neq 2\,||x||_0 \end{aligned}$$

Toy problem: $||v_t||_0$?

$$v_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}$$

Toy problem: $||v_t||_0$?



Alternative strategy

Minimize an actual norm

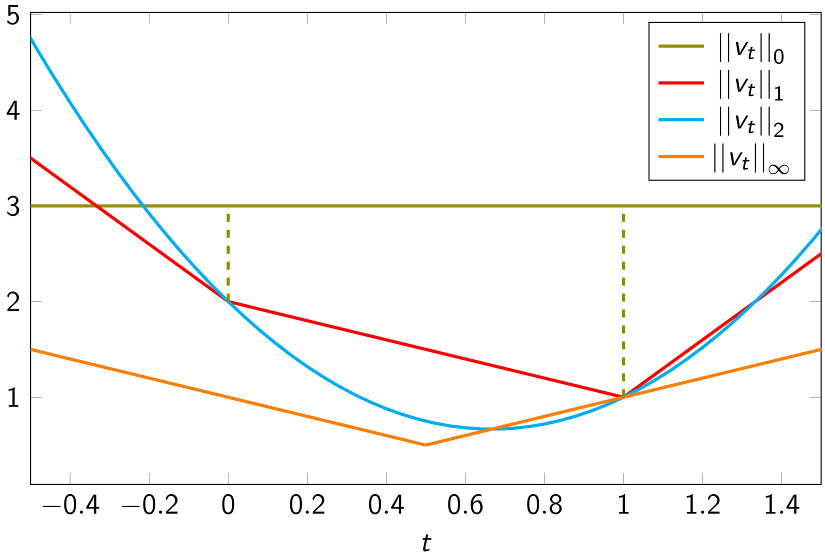
$$f(t) := ||v_t|| :$$

$$||x||_1 := \sum_{i=1}^{d} |x_i|$$

$$||x||_2 := \sqrt{\sum_{i=1}^{d} x_i^2}$$

$$||x||_\infty := \max_{1 \leq i \leq d} |x_i|$$

Toy problem



Sparse linear regression

Find a small subset of useful features

Model selection problem

Two objectives:

- Good fit to the data; $\left|\left|X^T\beta - y\right|\right|_2^2$ should be as small as possible
- Using a small number of features; β should be as sparse as possible

The lasso

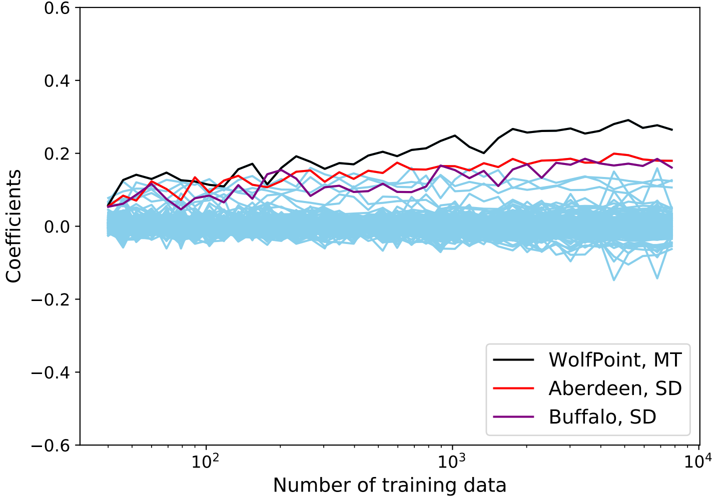
Uses ℓ_1-norm regularization to promote sparse coefficients

$$\beta_{\text{lasso}} := \arg\min_{\beta} \frac{1}{2} \left|\left| y - X^T\beta \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1$$

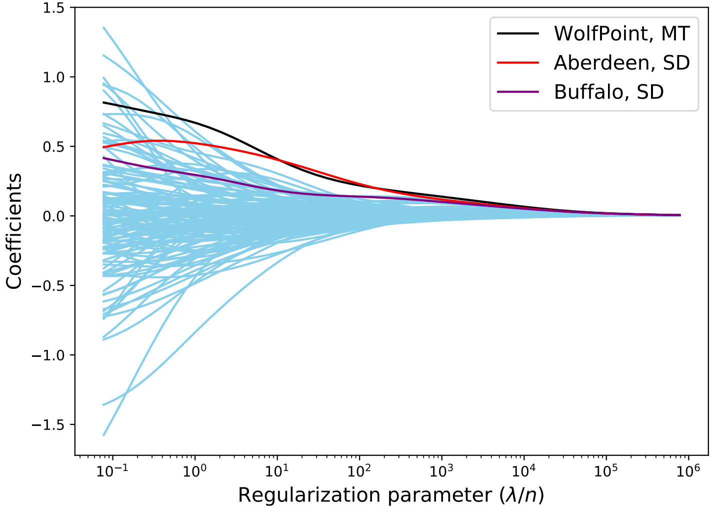
Temperature prediction via linear regression

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

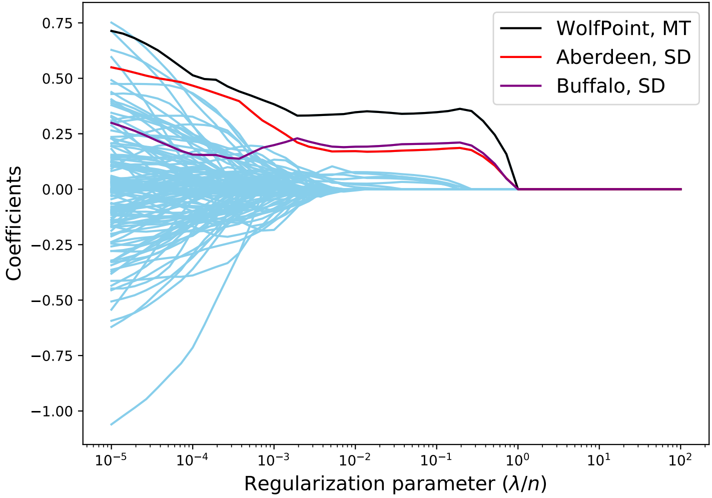
Ridge-regression coefficients



Ridge regression $n := 135$

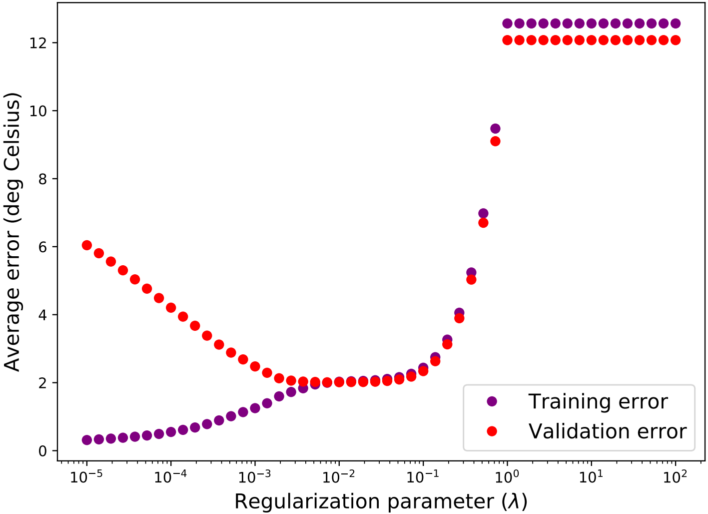


Lasso $n := 135$



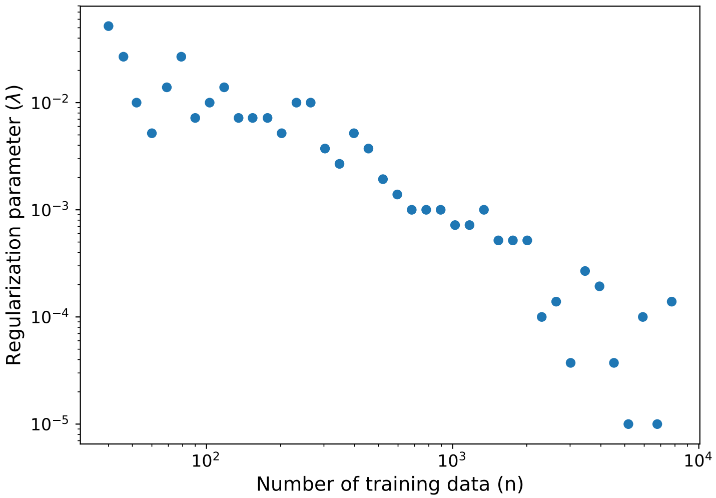
WolfPoint, MT
Aberdeen, SD
Buffalo, SD
Coefficients
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Regularization parameter (λ/n)

Lasso $n := 135$



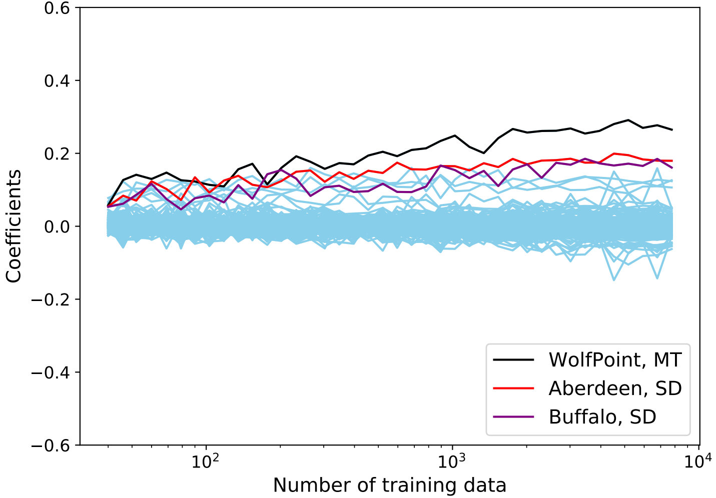
Average error (deg Celsius)
Regularization parameter (λ)
Training error
Validation error
0
2
4
6
8
10
12
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

Lasso $n := 135$

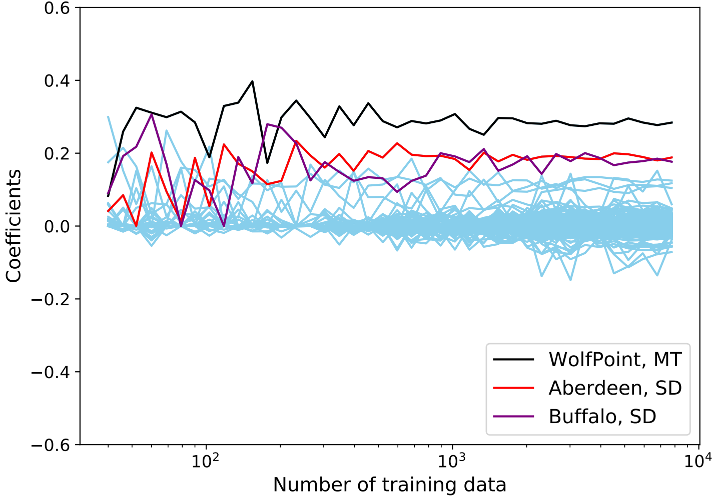


Regularization parameter (λ)
10^−2
10^−3
10^−4
10^−5
10^2
10^3
10^4
Number of training data (n)

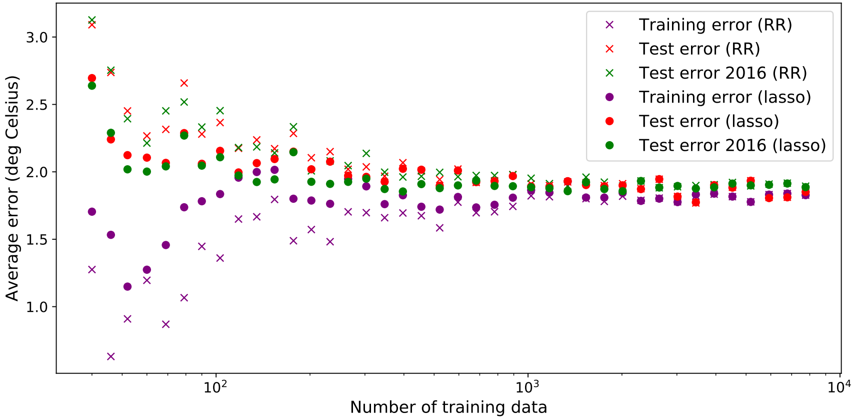
Ridge-regression coefficients



Lasso coefficients



Results



Average error (deg Celsius)
Number of training data
3.0
2.5
2.0
1.5
1.0
10^2
10^3
10^4
Training error (RR)
Test error (RR)
Test error 2016 (RR)
Training error (lasso)
Test error (lasso)
Test error 2016 (lasso)

What have we learned

- Sparse regression is the problem of fitting a linear model that includes only a subset of relevant features
- Regularization based on ℓ_1 norm makes it possible to do this automatically!