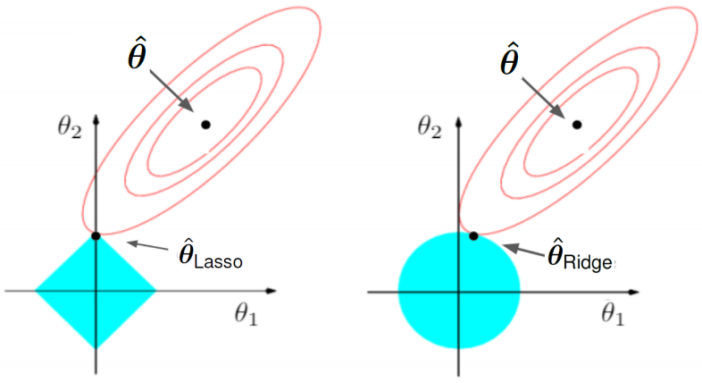
Introduction to Machine Learning

Lasso vs. Ridge Regression

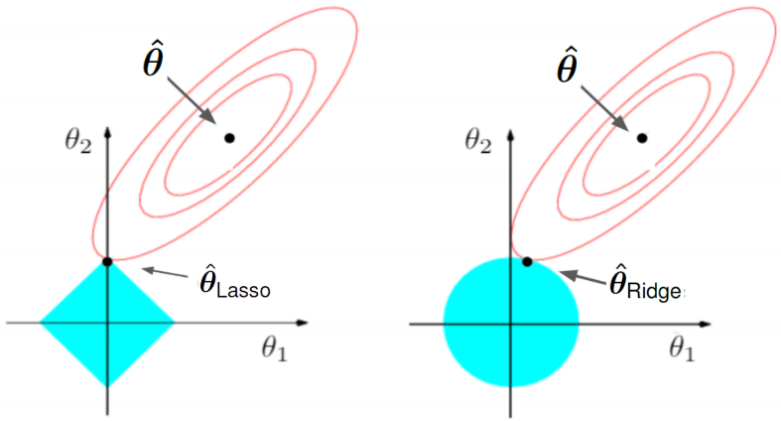


Learning goals

- Know the geometry of Ridge vs. Lasso regularization
- Understand the effects of the methods on model coefficients
- Understand that Lasso creates sparse solutions

LASSO VS. RIDGE GEOMETRY

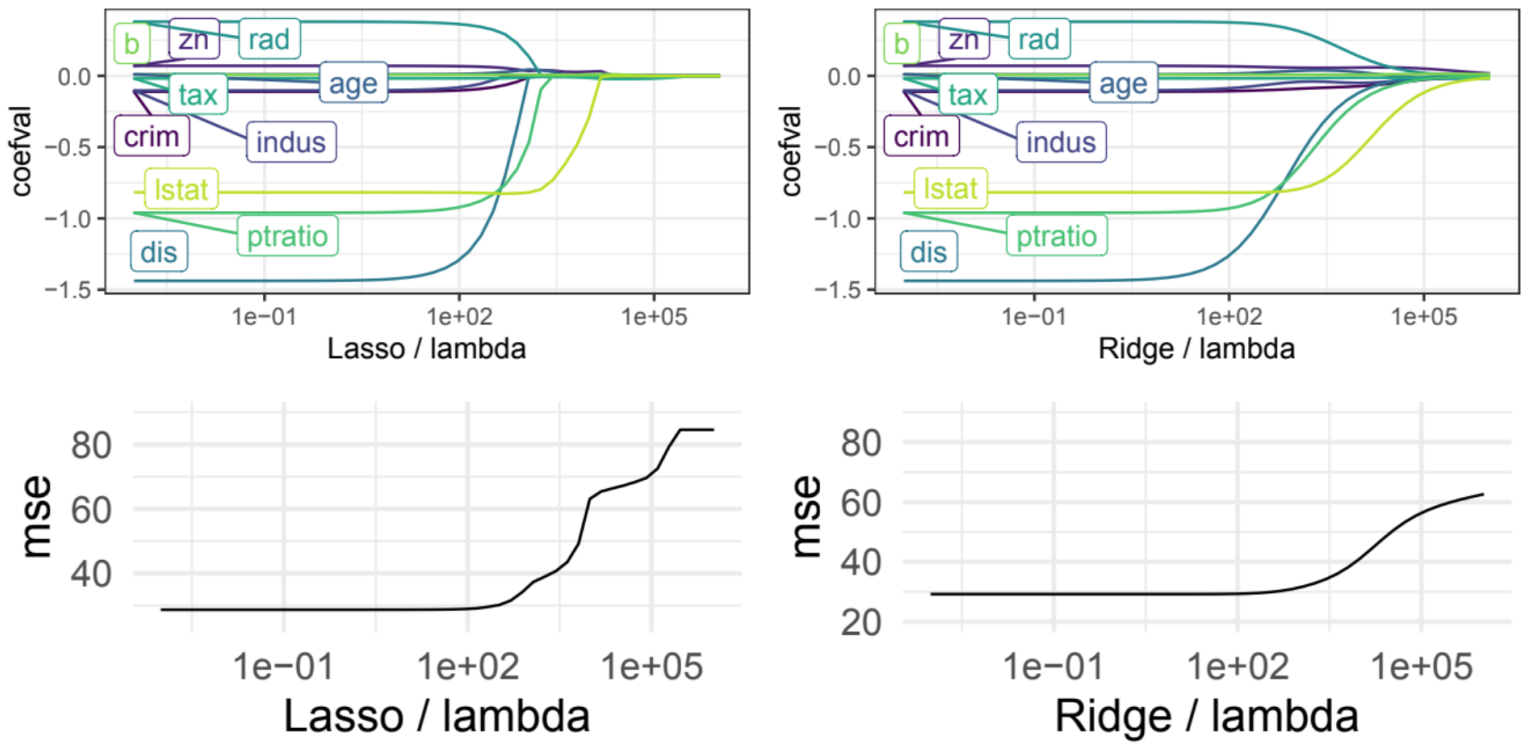
$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right)^2 \qquad \text{s.t. } \|\boldsymbol{\theta}\|_p^p \leq t$$



- In both cases, the solution which minimizes $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is always a point on the boundary of the feasible region (for sufficiently large λ).
- As expected, $\hat{\boldsymbol{\theta}}_{\text{Lasso}}$ and $\hat{\boldsymbol{\theta}}_{\text{Ridge}}$ have smaller parameter norms than $\hat{\boldsymbol{\theta}}$.
- For Lasso, the solution likely touches vertices of the constraint region. This induces sparsity and is a form of variable selection.
- In the $p > n$ case, the Lasso selects at most n features (due to the nature of the convex optimization problem).

COEFFICIENT PATHS AND 0-SHRINKAGE

Example 1: Boston Housing (few features removed for readability)
We cannot overfit here with an unregularized linear model as the task is so low-dimensional. But we see how only Lasso shrinks to sparsely 0.



Coef paths and cross-val. MSE for λ values for Ridge and Lasso.

COEFFICIENT PATHS AND 0-SHRINKAGE

Example 2: High-dimensional simulated data

We simulate a continuous, correlated dataset with 50 features, 100 observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$ and

$$y = 10 \cdot (x_1 + x_2) + 5 \cdot (x_3 + x_4) + \sum_{j=5}^{14} x_j + \epsilon$$

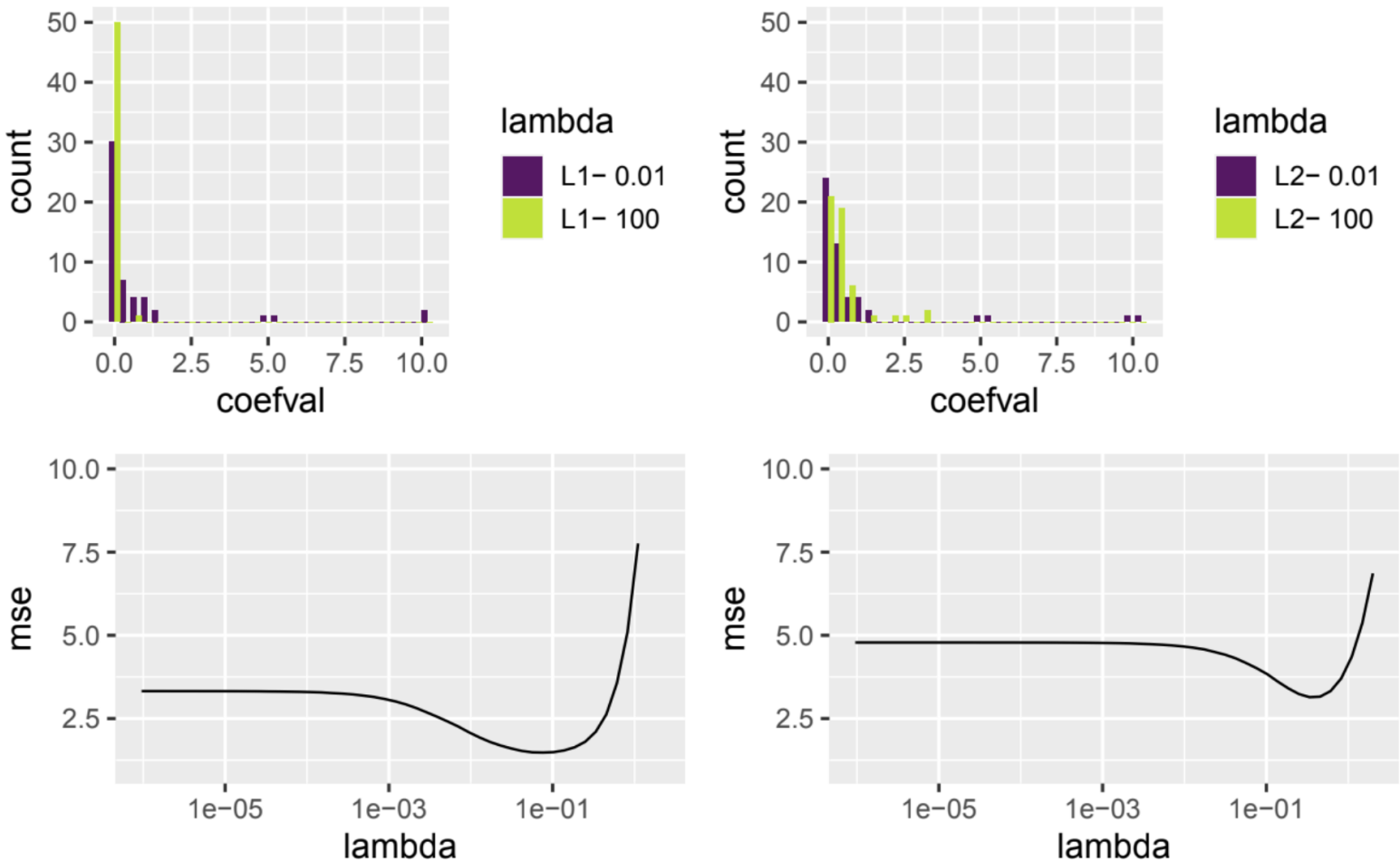
where $\epsilon \sim \mathcal{N}(0, 1)$ and $\forall k, l \in \{1, ..., 50\}$:

$$Cov(x_k, x_l) = \begin{cases} 0.7^{|k-l|} & \text{for } k \neq l \\ 1 & \textit{else} \end{cases}.$$

Note that 36 of the 50 features are noise variables.

COEFFICIENT PATHS AND 0-SHRINKAGE

Coefficient histograms for different λ values for Ridge and Lasso, on high-dimensional data along with the cross-validated MSE.



REGULARIZATION AND FEATURE SCALING

- Note that very often we do not include θ_0 in the penalty term $J(\boldsymbol{\theta})$ (but this can be implementation-dependent).
- These methods are typically not equivariant under scaling of the inputs, so one usually standardizes the features.
- Note that for a normal LM, if you scale some features, we can simply "anti-scale" the coefficients the same way. The risk does not change. For regularized models this is not so simple. If you scale features to smaller values, coefficients have to become larger to counteract. They now are penalized more heavily in $J(\boldsymbol{\theta})$. Such a scaling would make some features less attractive without changing anything relevant in the data.

REGULARIZATION AND FEATURE SCALING

Example:

- Let the true data generating process be

$$y = x_1 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1).$$

- Let there be 5 features $x_1, \ldots, x_5 \sim \mathcal{N}(0, 1)$.
- Using the Lasso (package glmnet), we get

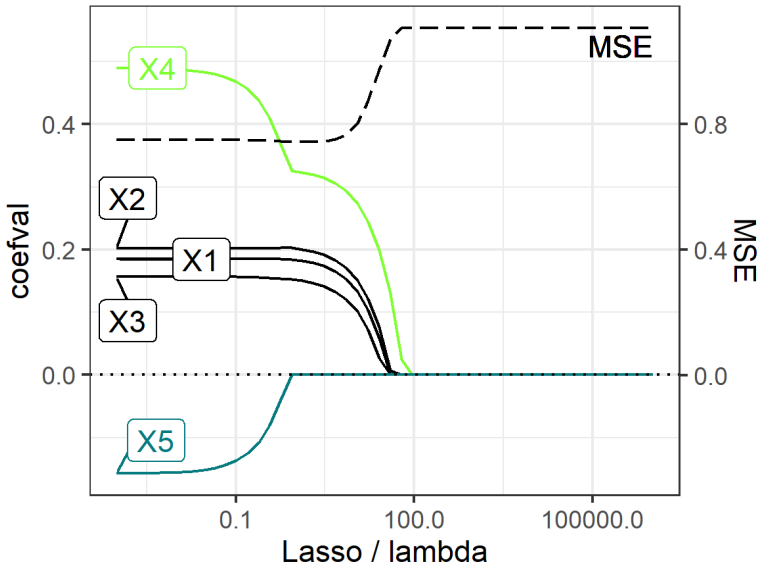
(Intercept)	x1	x2	x3	x4
-0.056	0.489	0.000	0.000	0.000

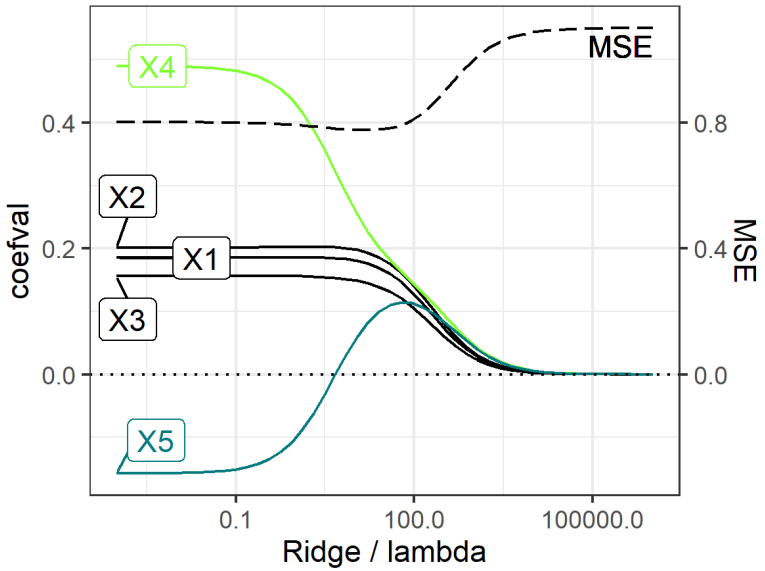
- But if we rescale any of the noise features, say $x_2 = 10000 \cdot x_2$ and don't use standardization, we get

(Intercept)	x1	x2	x3	x4
-0.106830	0.000000	-0.000013	0.000000	0.000000

- This is due to the fact, that the coefficient of x_2 will live on a very small scale as the covariate itself is large. The feature will thus get less penalized by the L_1-norm and is favored by Lasso.

CORRELATED FEATURES





Fictional example for the model $y = 0.2X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4 + 0.2X_5 + \epsilon$ of 100 observations, $\epsilon \sim \mathcal{N}(0, 1)$. X_1-X_4 are independently drawn from different normal distributions: $X_1, X_2, X_3, X_4 \sim \mathcal{N}(0, 2)$. While X_1-X_4 have pairwise correlation coefficients of 0, X_4 and X_5 are nearly perfectly correlated: $X_5 = X_4 + \delta, \delta \sim \mathcal{N}(0, 0.3), \rho(X_4, X_5) = 0.98$.

We see that Lasso shrinks the coefficient for X_5 to zero early on, while Ridge assigns similar coefficients to X_4, X_5 for larger λ.

SUMMARIZING COMMENTS

- Neither one can be classified as overall better.
- Lasso is likely better if the true underlying structure is sparse, so if only few features influence y. Ridge works well if there are many influential features.
- Lasso can set some coefficients to zero, thus performing variable selection, while Ridge regression usually leads to smaller estimated coefficients, but still dense θ vectors.
- Lasso has difficulties handling correlated predictors. For high correlation Ridge dominates Lasso in performance.
- For Lasso one of the correlated predictors will have a larger coefficient, while the rest are (nearly) zeroed. The respective feature is, however, selected randomly.
- For Ridge the coefficients of correlated features are similar.