Comparison of Ridge and Lasso

Pavlos Protopapas, Ignacio Becker


THE BEST WAY TO
EXPLAIN OVERFITTING

Ridge, LASSO - Computational complexity

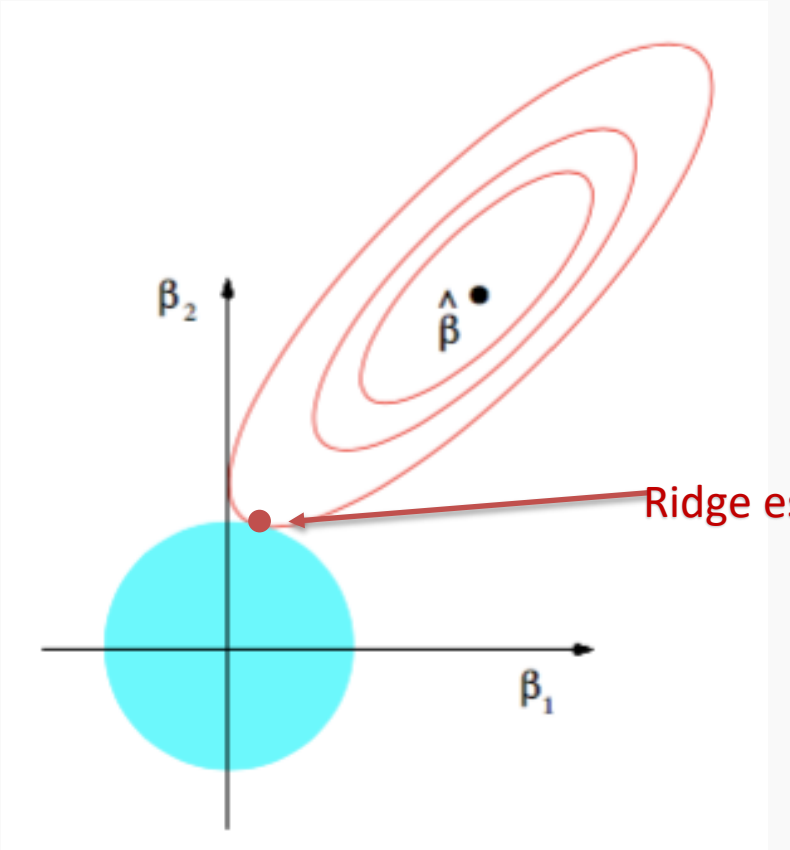
Solution to ridge regression:

$$\beta = (X^T X + \lambda I)^{-1} X^T Y$$

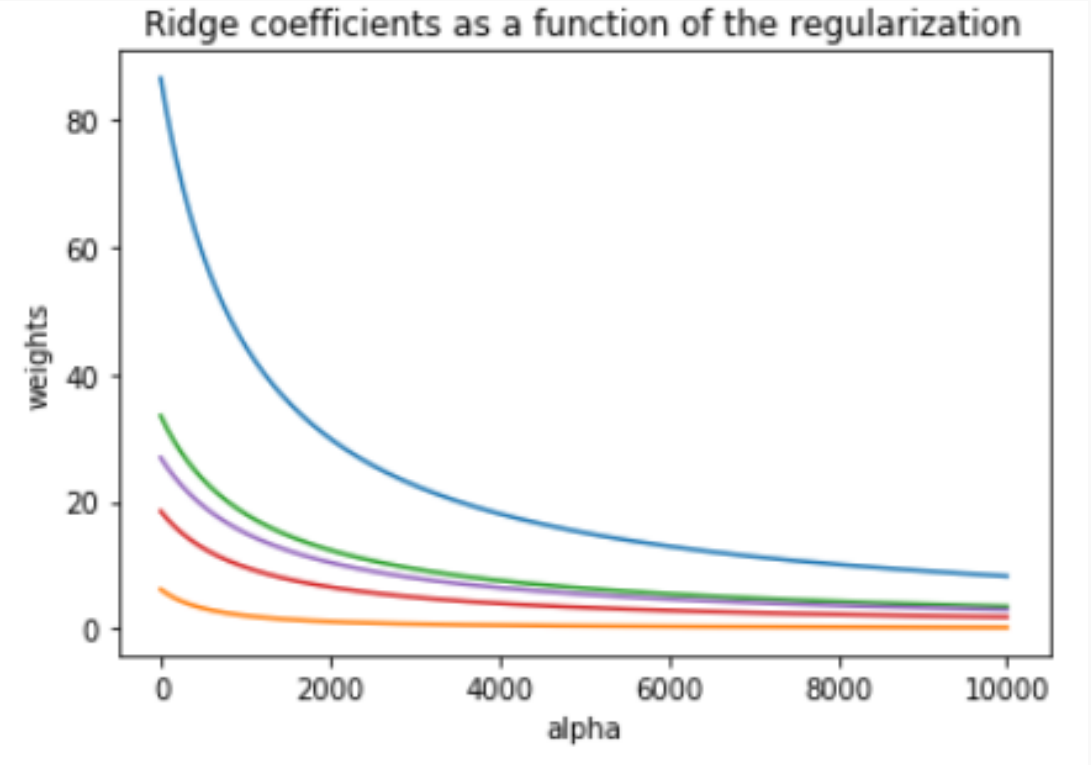
The solution to the LASSO regression:

LASSO has no conventional analytical solution, as the L1 norm has no derivative at zero. We can, however, use the concept of subdifferential or subgradient to find a manageable expression.

Ridge visualized

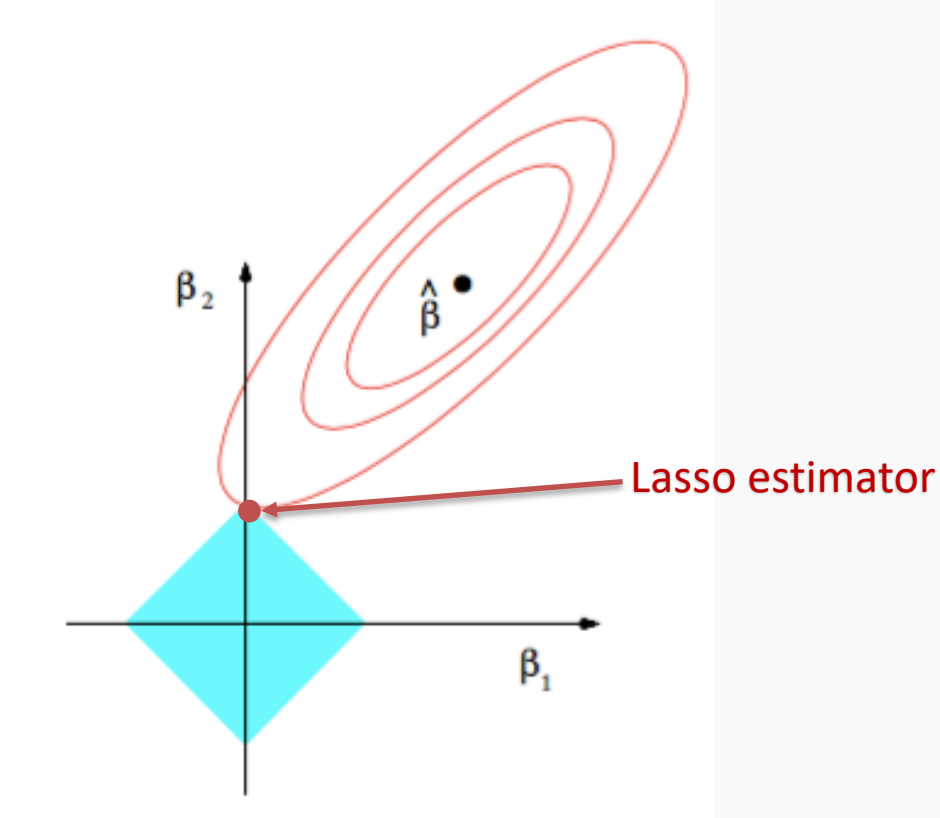


The ridge estimator is where the constraint and the loss intersect.

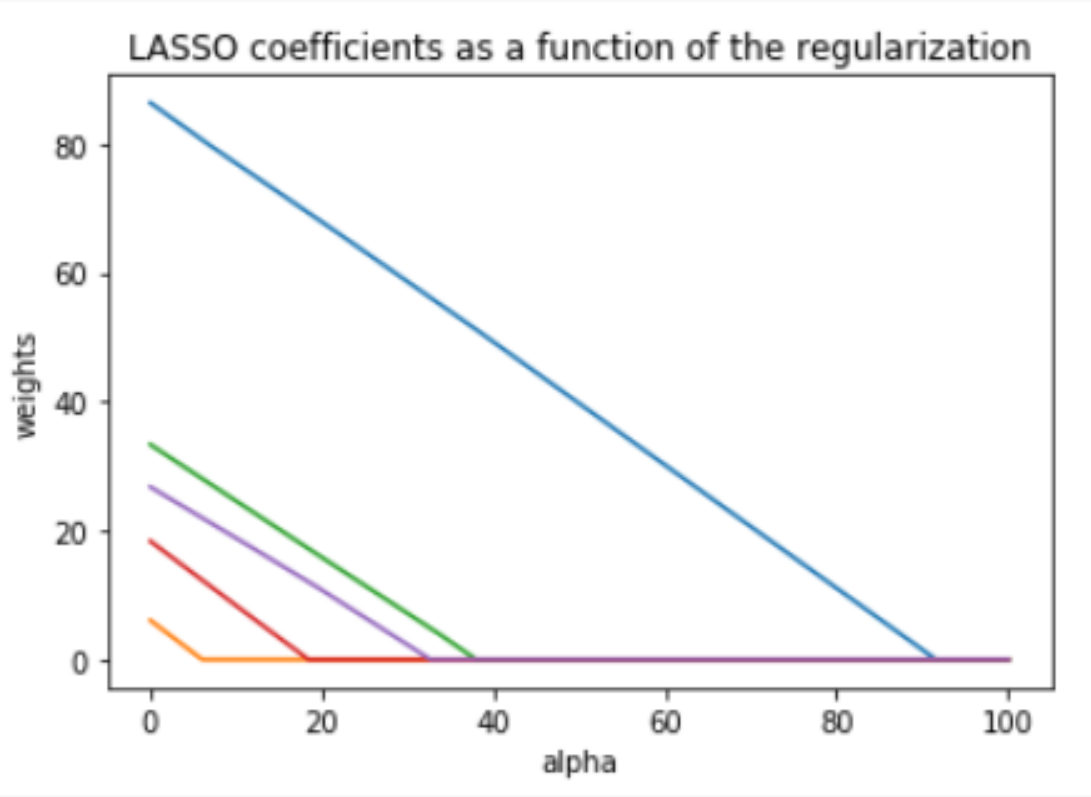


The values of the coefficients decrease as lambda increases, but they are not nullified.

LASSO visualized



The Lasso estimator tends to zero out parameters as the OLS loss can easily intersect with the constraint on one of the axis.



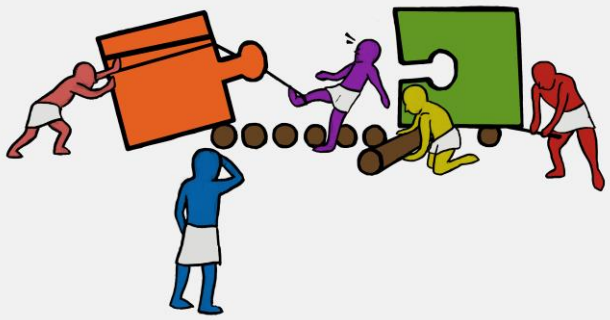
The values of the coefficients decrease as lambda increases and are nullified fast.

Variable Selection as Regularization

Question: What are the pros and cons of the two approaches?

Since LASSO regression tends to produce zero estimates for a number of model parameters - we say that LASSO solutions are **sparse** - we consider LASSO to be a variable selection method.

Ridge is faster to compute but many prefer using LASSO for variable selection (as well as for suppressing extreme parameter values) and therefore easier to interpret.

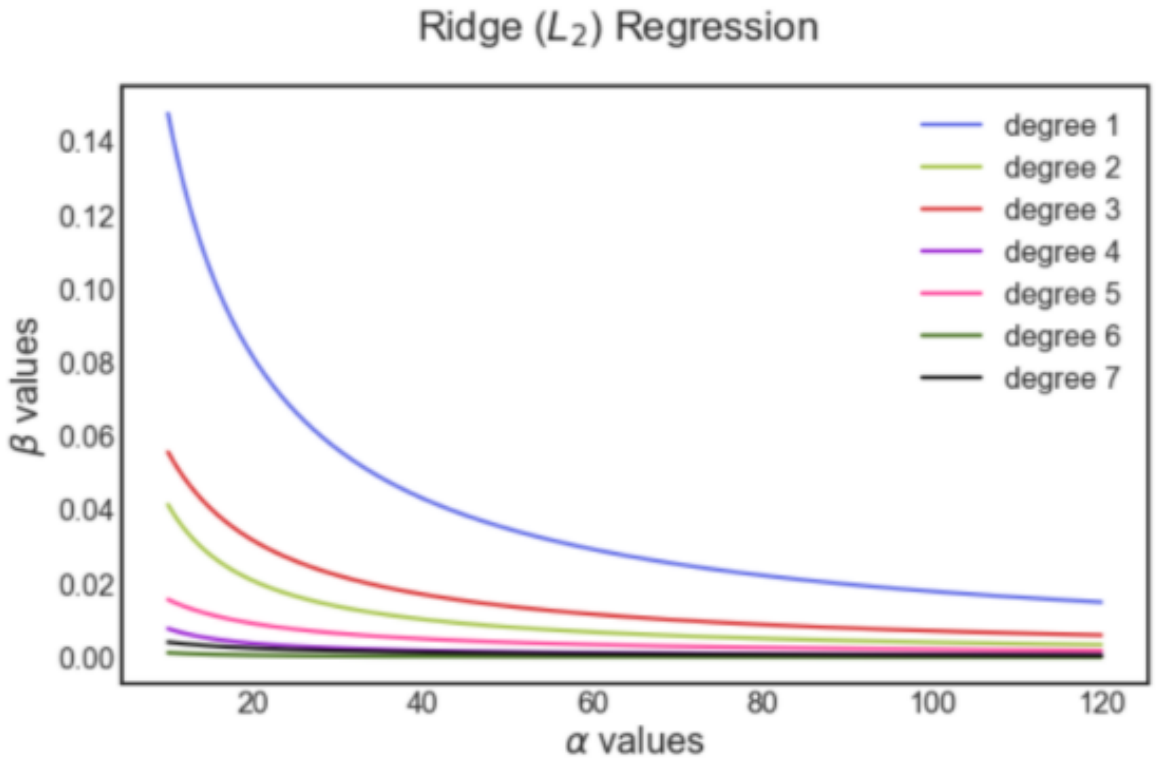


Description

🏋 Exercise: Variation of Coefficients for Lasso and Ridge Regression

The goal of this exercise is to understand the variation of the coefficients of predictors with varying values of regularization parameter in Lasso and Ridge regularization.

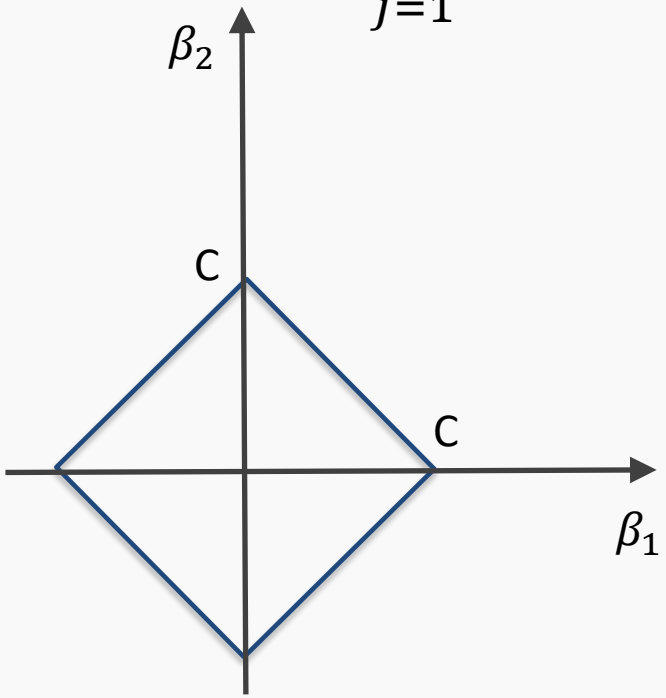
Below is a sample plot for Ridge (L_2 regularization)



The Geometry of Regularization (LASSO)

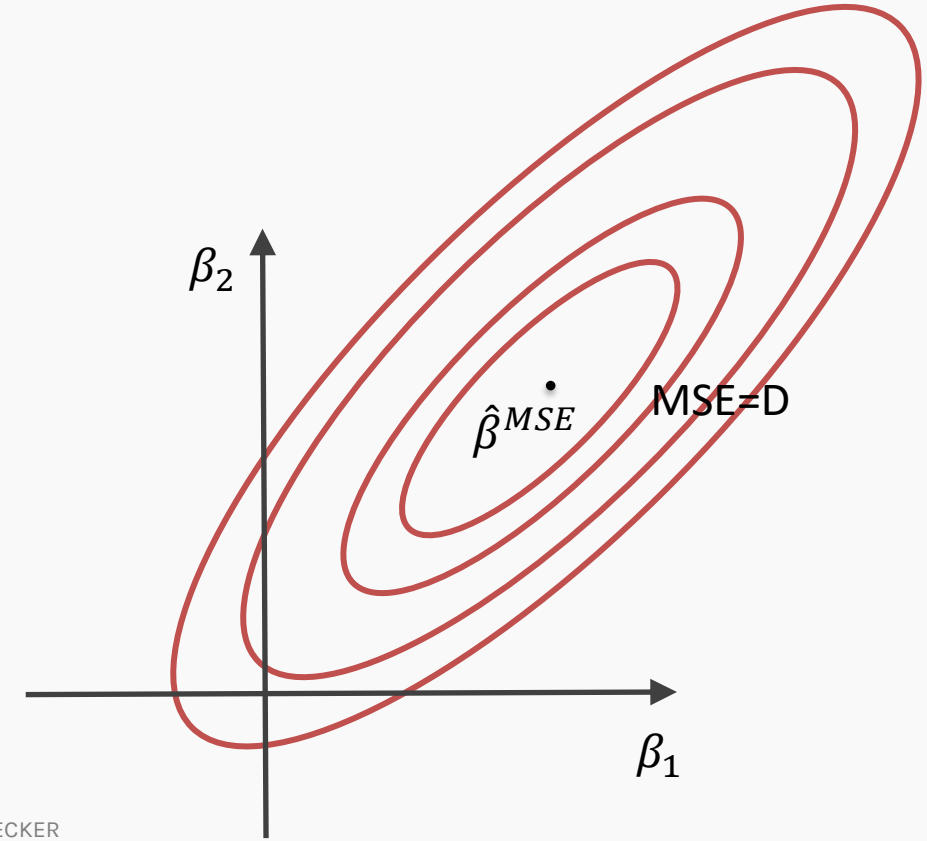
$$L_{LASSO}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^T \boldsymbol{x}|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

$$\lambda \sum_{j=1}^{J} |\hat{\beta}_j^{LASSO}| = C$$



$$\widehat{\boldsymbol{\beta}}^{LASSO} = \text{argmin}\, L_{LASSO}(\boldsymbol{\beta})$$

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\boldsymbol{\beta}}^{LASSO^T} \boldsymbol{x}|^2 = D$$

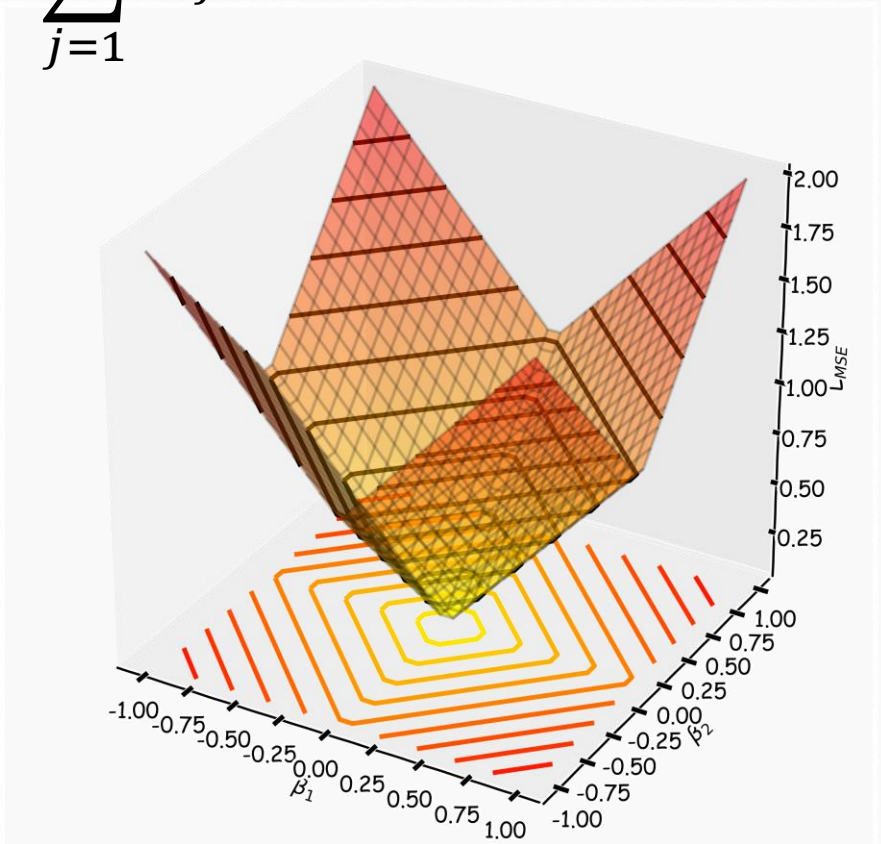


The Geometry of Regularization (LASSO)

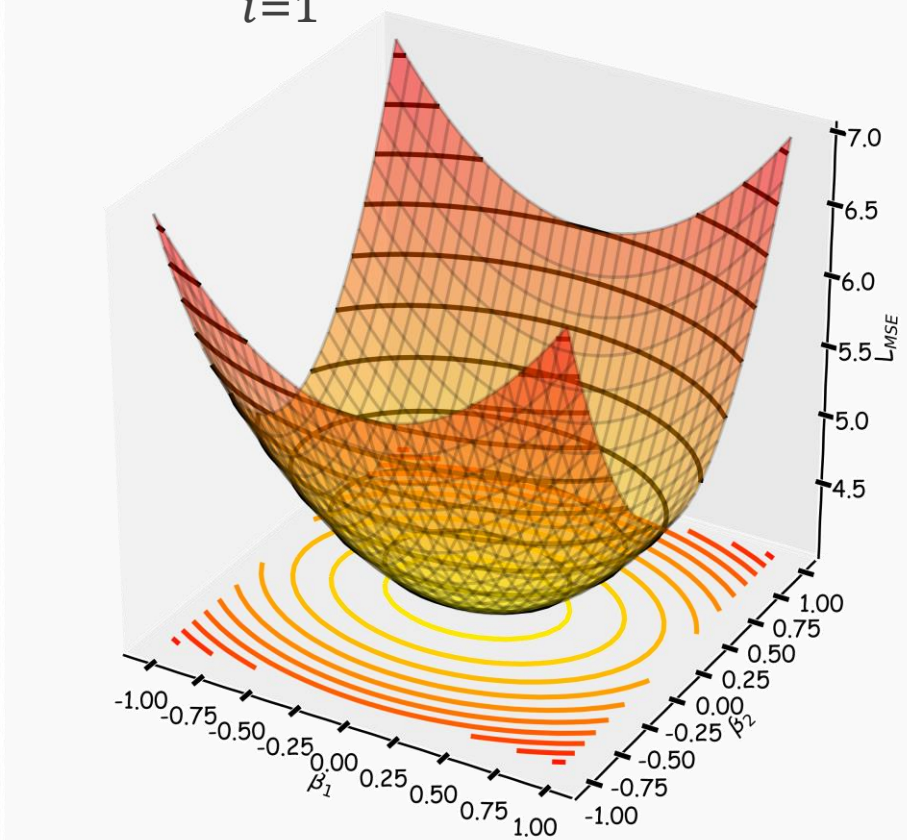
$$L_{LASSO}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^T \boldsymbol{x}\Big|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \operatorname{argmin} L_{LASSO}(\boldsymbol{\beta})$$

$$L_1 = \lambda \sum_{j=1}^{J} \Big| \hat{\beta}_j^{LASSO} \Big|$$



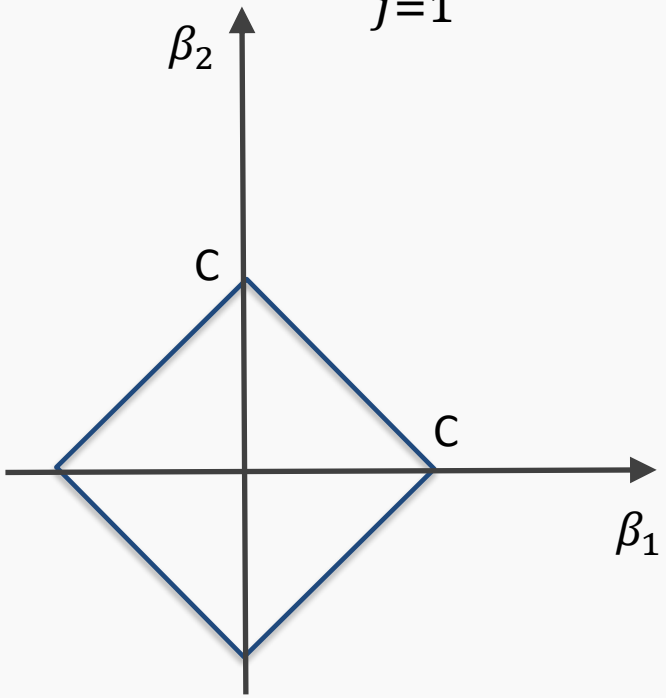
$$L_{MSE}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^T \boldsymbol{x}\Big|^2$$



The Geometry of Regularization (LASSO)

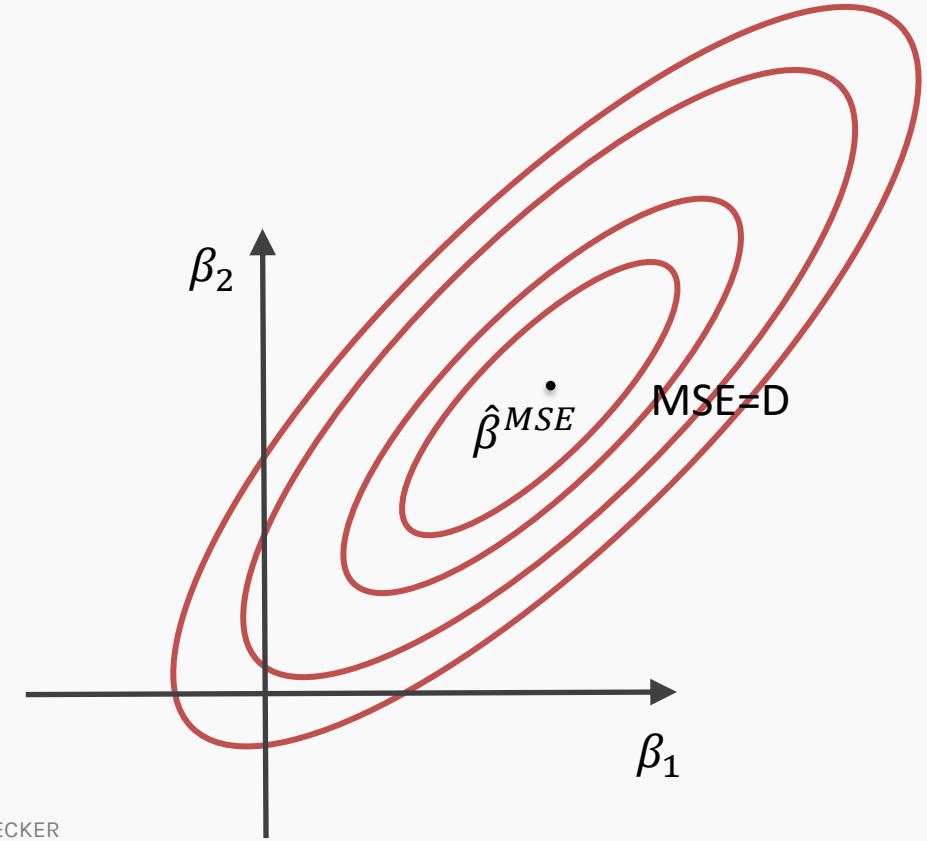
$$L_{LASSO}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^T \boldsymbol{x}|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

$$\lambda \sum_{j=1}^{J} |\hat{\beta}_j^{LASSO}| = C$$

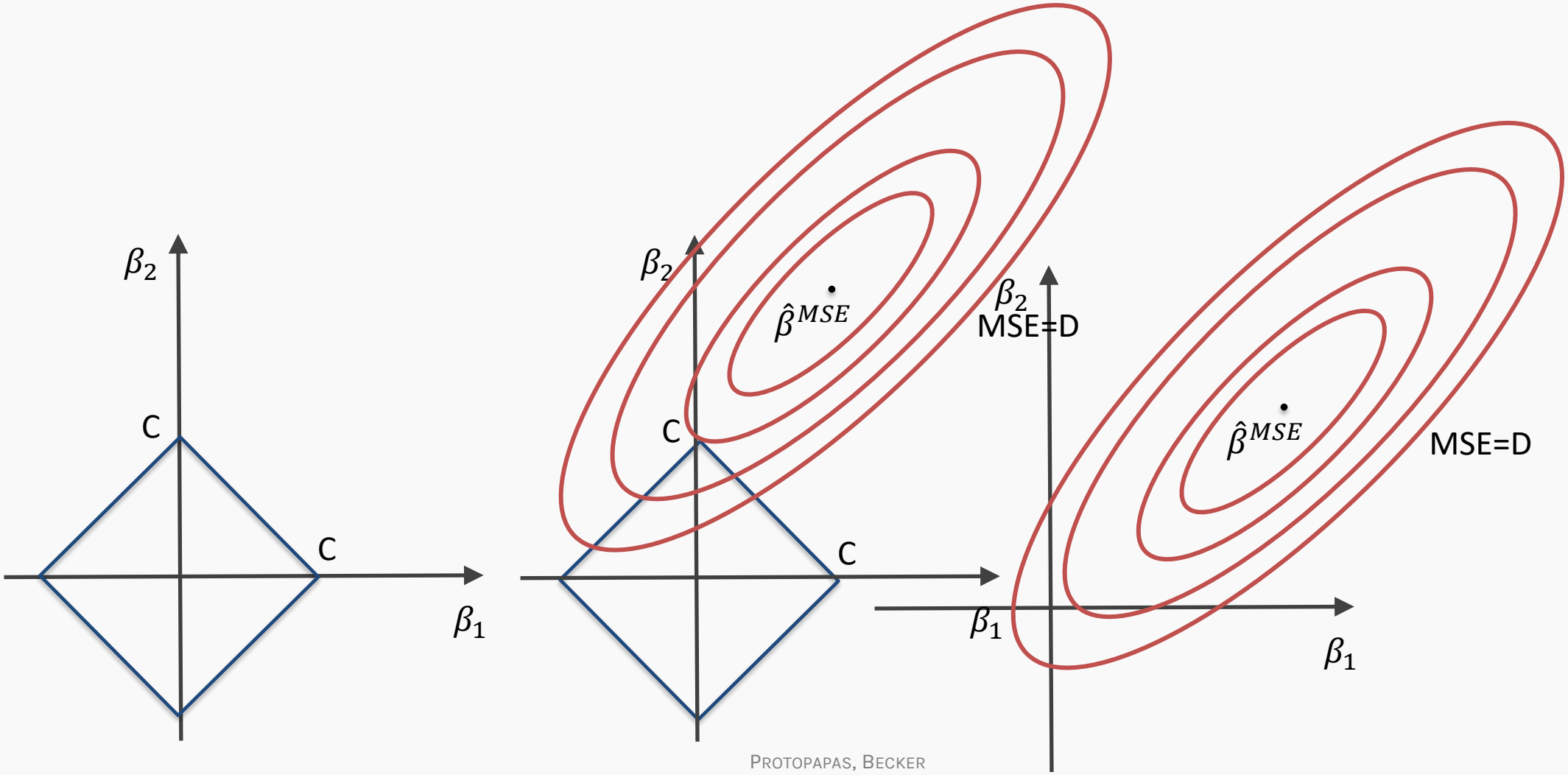


$$\widehat{\boldsymbol{\beta}}^{LASSO} = \text{argmin}\, L_{LASSO}(\boldsymbol{\beta})$$

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\boldsymbol{\beta}}^{LASSO^T} \boldsymbol{x}|^2 = D$$



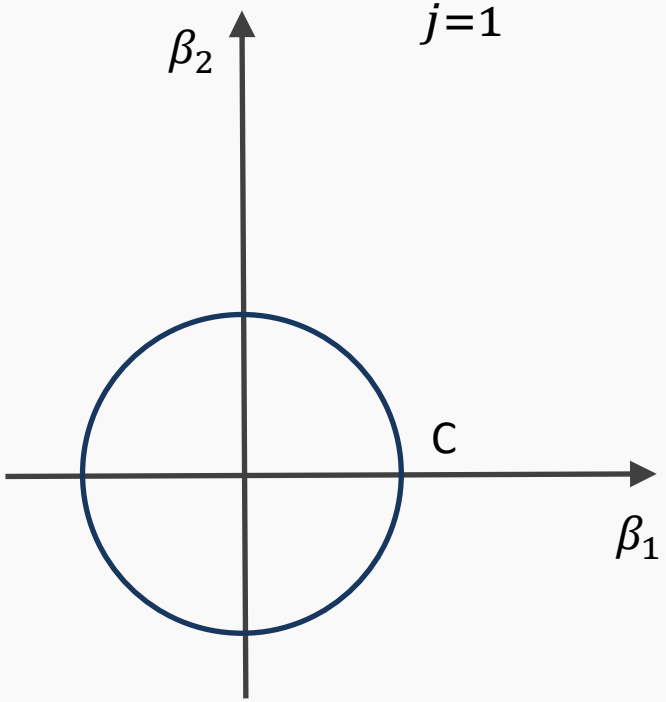
The Geometry of Regularization (LASSO)



The Geometry of Regularization (Ridge)

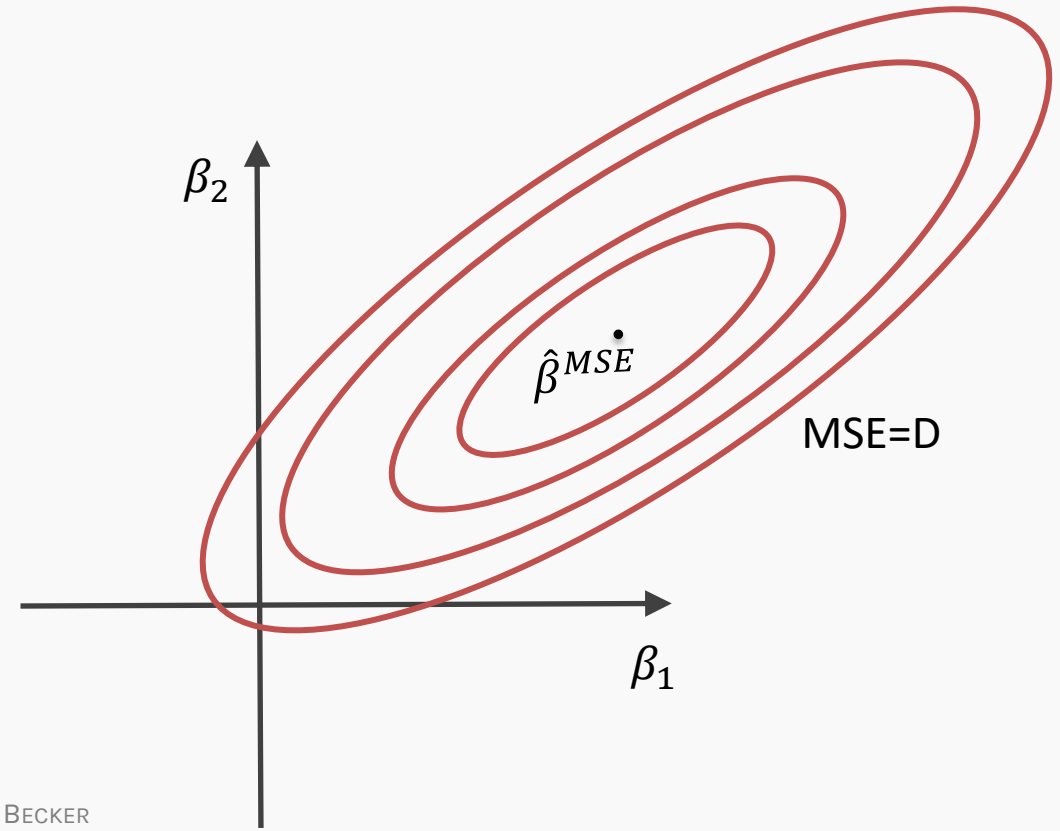
$$L_{Ridge}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{\beta}^T \boldsymbol{x}\Big|^2 + \lambda \sum_{j=1}^{J} (\beta_j)^2$$

$$\lambda \sum_{j=1}^{J} \left|\hat{\beta}_j^{Ridge}\right|^2 = C$$

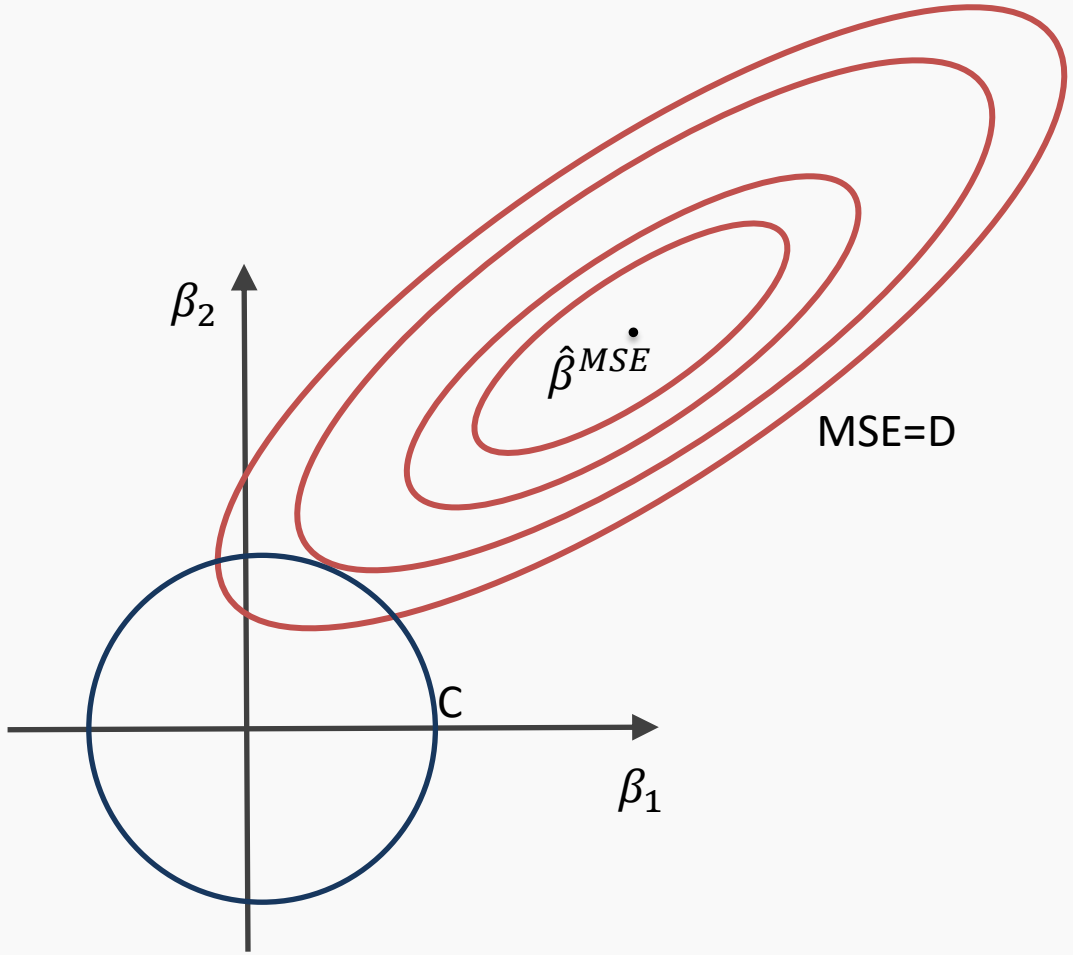


$$\hat{\boldsymbol{\beta}}^{Ridge} = \text{argmin}\, L_{Ridge}(\boldsymbol{\beta})$$

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{\boldsymbol{\beta}}^{Ridge^T} \boldsymbol{x}|^2 = D$$



The Geometry of Regularization (Ridge)



The Geometry of Regularization