Regularized Regression and Tuning

General Rule in ML

If there is a parameter or a decision to be made, we use cross-validation (or another validation technique) to optimize for predictive performance

- For regularized regression, this means choice of alpha, lambda.
- CV is done for tuning everything at once.

Survey design in CV: Can use different CV methods (such as stratified) and models.

- Most important thing is to get unbiased estimates of performance.

Steps under ML Framework

1. Split data into training and test.
2. With training, use cross-validation to tune
    a. Split into train-validation using CV folds.
    b. Fit elastic net models with various values of lambda and alpha.
    c. Find best model using validation.
3. Evaluate best model on test set.

Computation Time

Regularized regression is generally relatively fast.

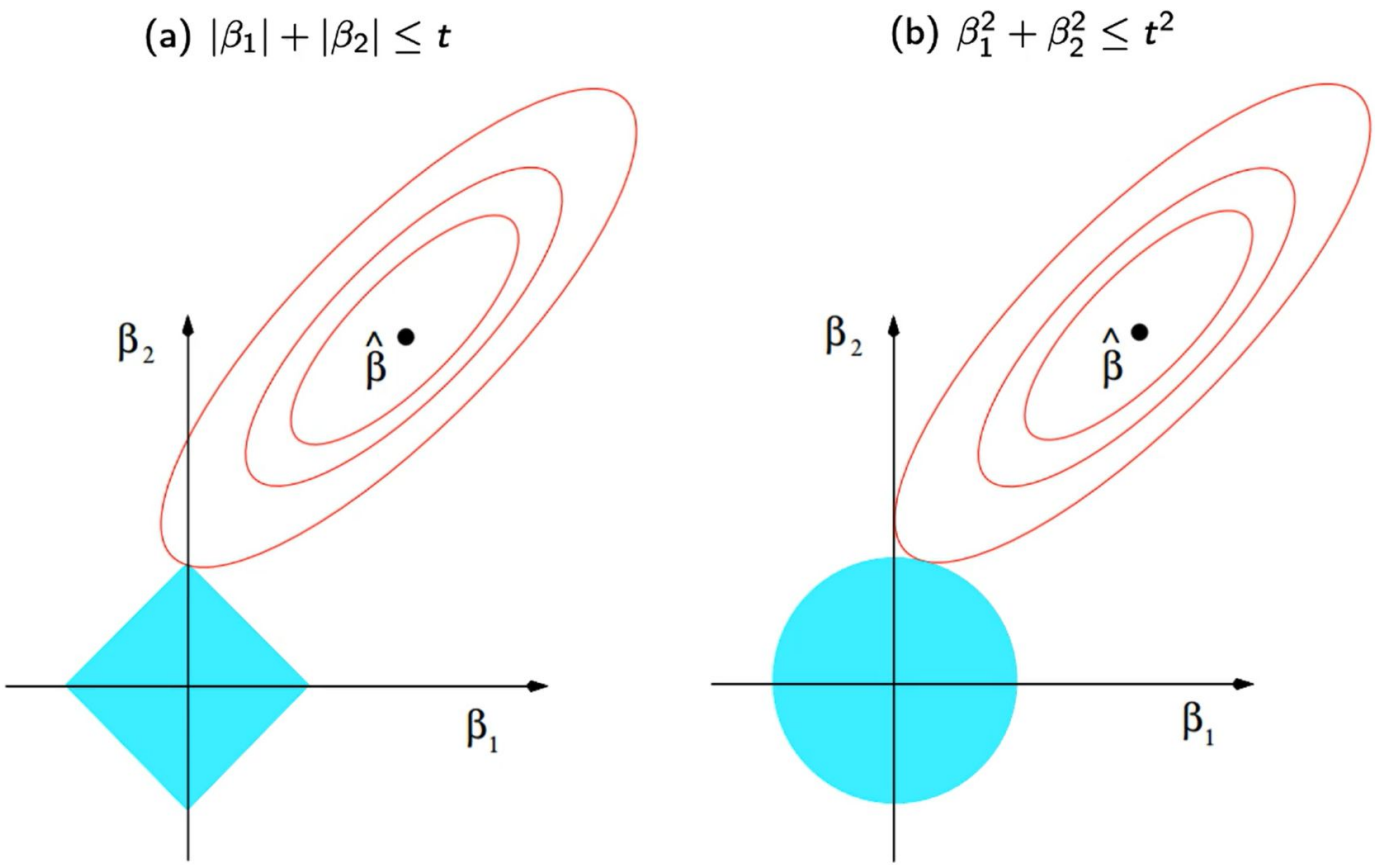
We'll really start to see it slow down with ensemble methods.

Lasso and Ridge Regression

Two ways of thinking about it:

- Shrinkage penalty - instead of finding the beta coefficients that minimize the residual sum of squares, find the beta coefficients that **minimize RSS + penalty on coefficients** (to shrink coefficients)
- Constraint regions - **minimize RSS with the constraint** that coefficients can’t be too big.

Figure: Constraint regions and RSS contours



Efron & Hastie (2016)

Features and Feature Selection

We typically do not check for significance of predictors in regularized regression under ML framework.

- Goal is prediction, so significance is not important – only performance on test set.

Usually want to include all variables and tune for best predictive ability.

- May need to use domain knowledge for generating features

Elastic Net

Note: Elastic net with alpha = 0 is Ridge Regression. Elastic Net with alpha = 1 is Lasso.

So, only need to **fit Elastic Net and tune alpha** instead of needing to do Lasso and Ridge separately.

Idea behind Group Lasso

Some variables should be considered together for variable selection purposes.

- Dummies of the same categorical variable.
- Related variables like latitude/longitude

Make sure that sending all of the related variables are sent to zero at the same time if they are set to zero.