Elastic Net

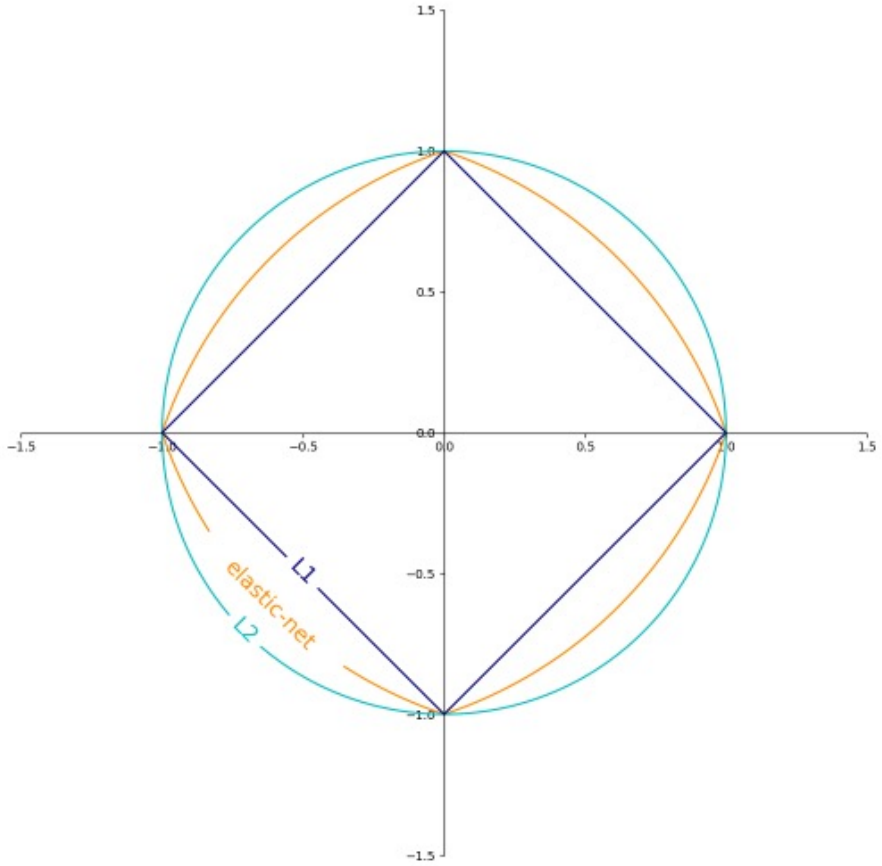
Emrys

Outline

Introduction	Use	Least Squares	Bias Variance Tradeoff
Regularization	Lasso & Ridge	Elastic Net	Implementation

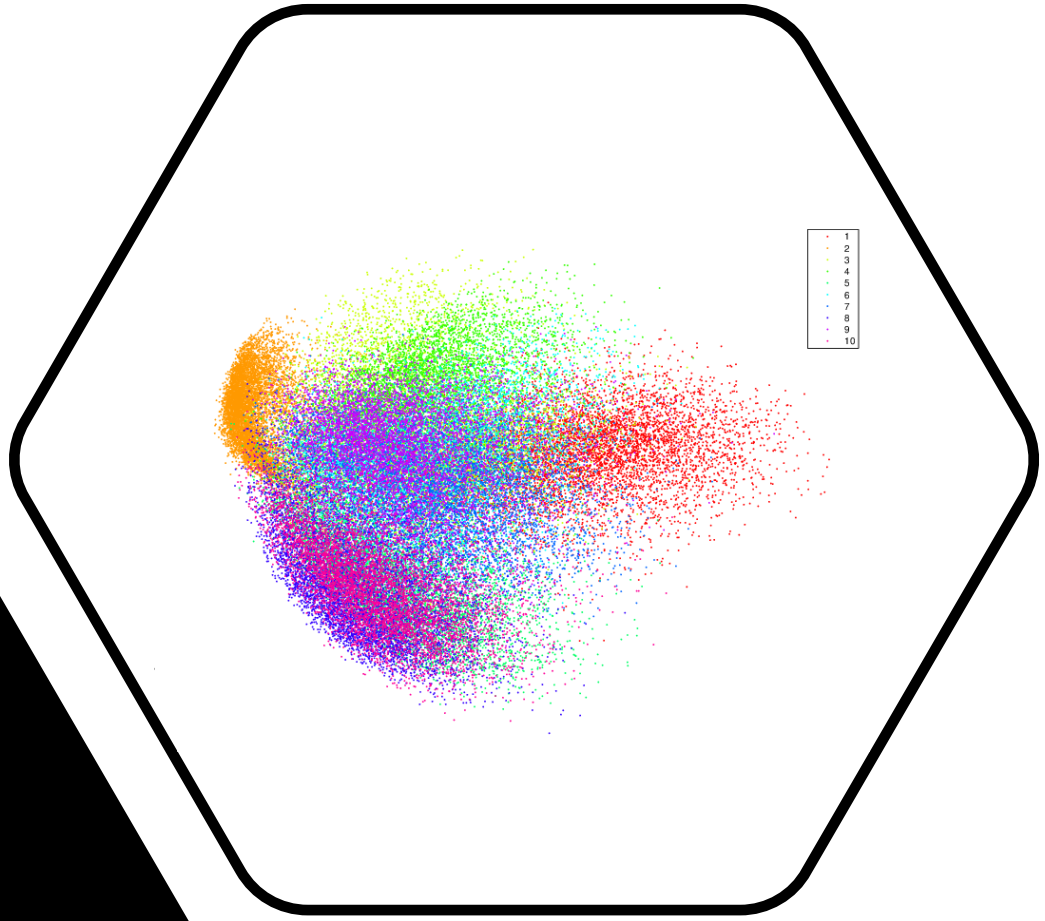
What is Elastic Net?

- A regularized regression method that simultaneously performs variable selection and regularization
- Applicable to regression and classification tasks



What makes Elastic Net useful?

- Increases interpretability of model by removing noise variables
- Provides greater accuracy for unseen data by preventing overfitting (higher generalizability)
- Appropriate for data problems with p>n (greater number of predictors than observations)



Problem Statement/Example

- Consider a standard regression problem where the response y is modeled by p predictors $x_1, x_2, \ldots, x_p$ with corresponding coefficient vector $\hat{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_p})$

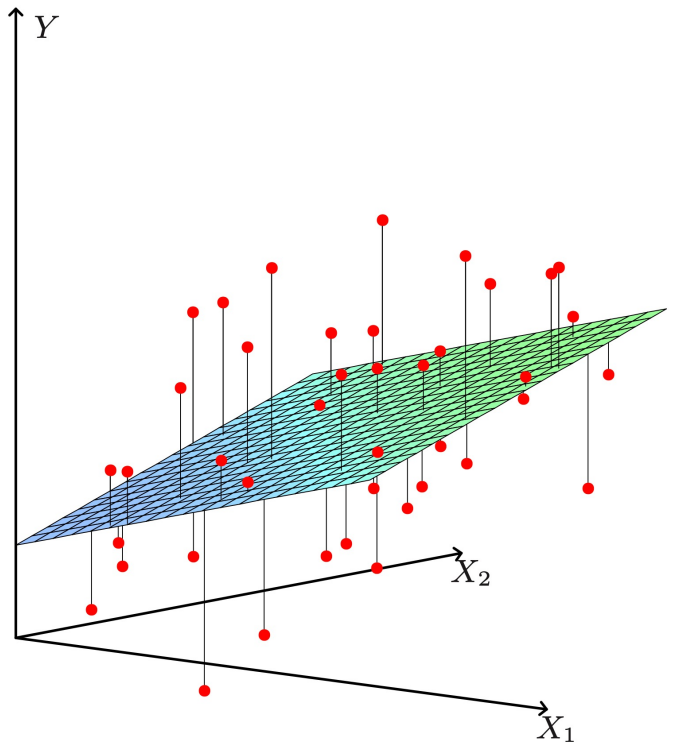
$$\hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

- e.g., Modeling DVT (a complication related to blood clots) rates post surgery for patients undergoing a knee replacement operation
- Predictors: Underlying health factors and prescribed treatment drugs

Loss Function: Ordinary Least Squares

$$L_{OLS} = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2$$

- No feature selection in the presence of too many variables
- Low prediction accuracy
- Not robust to outliers
- Low bias/High variance

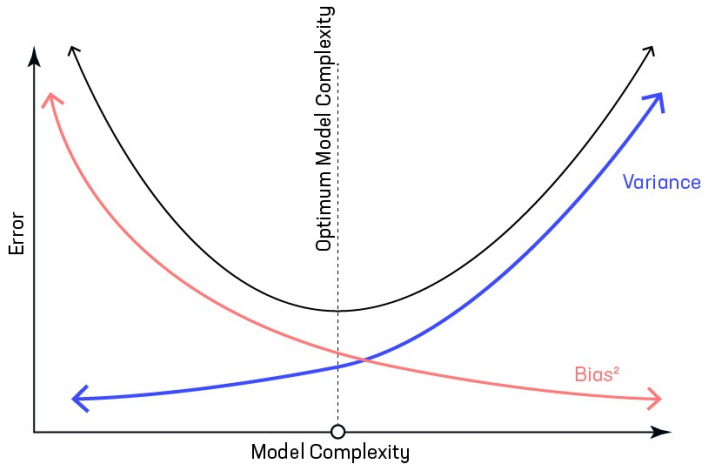


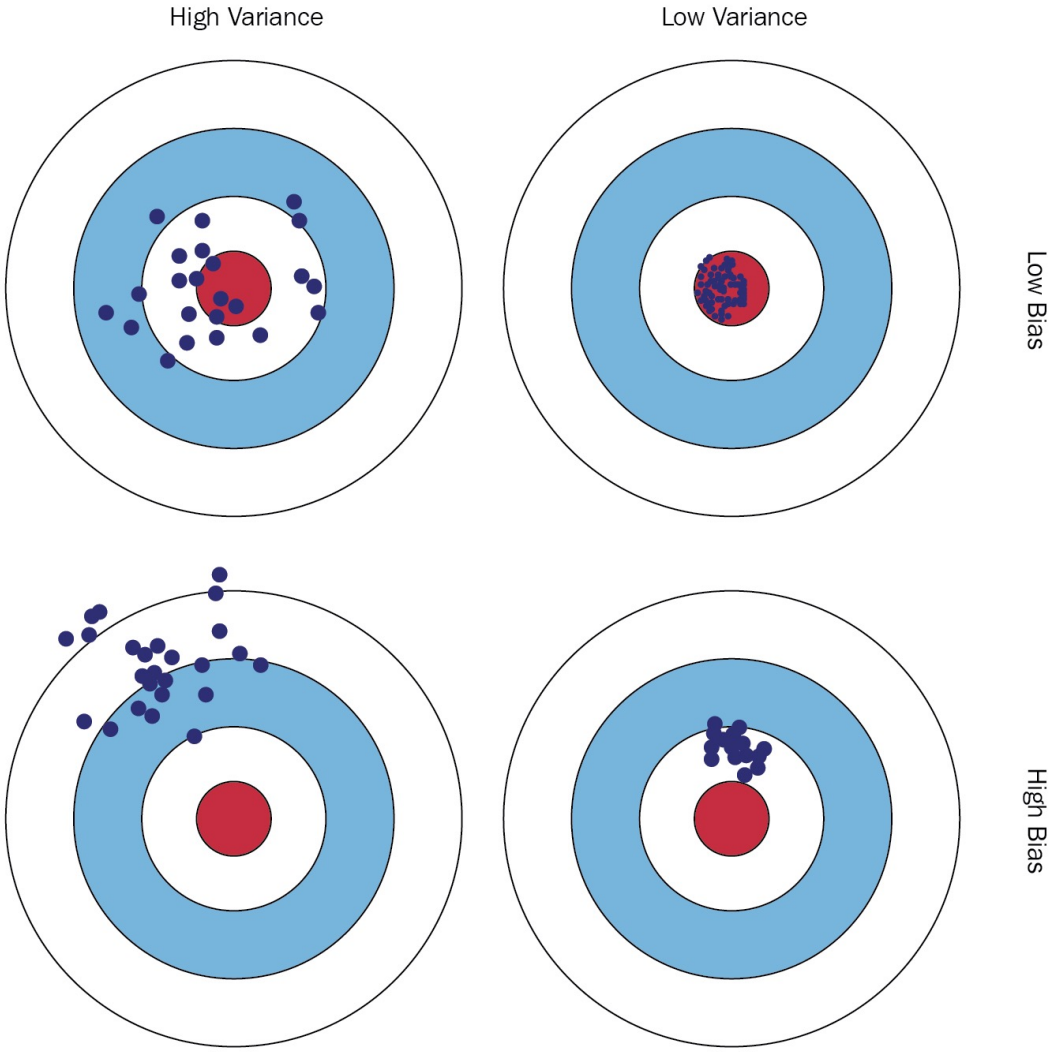
Bias Variance Tradeoff

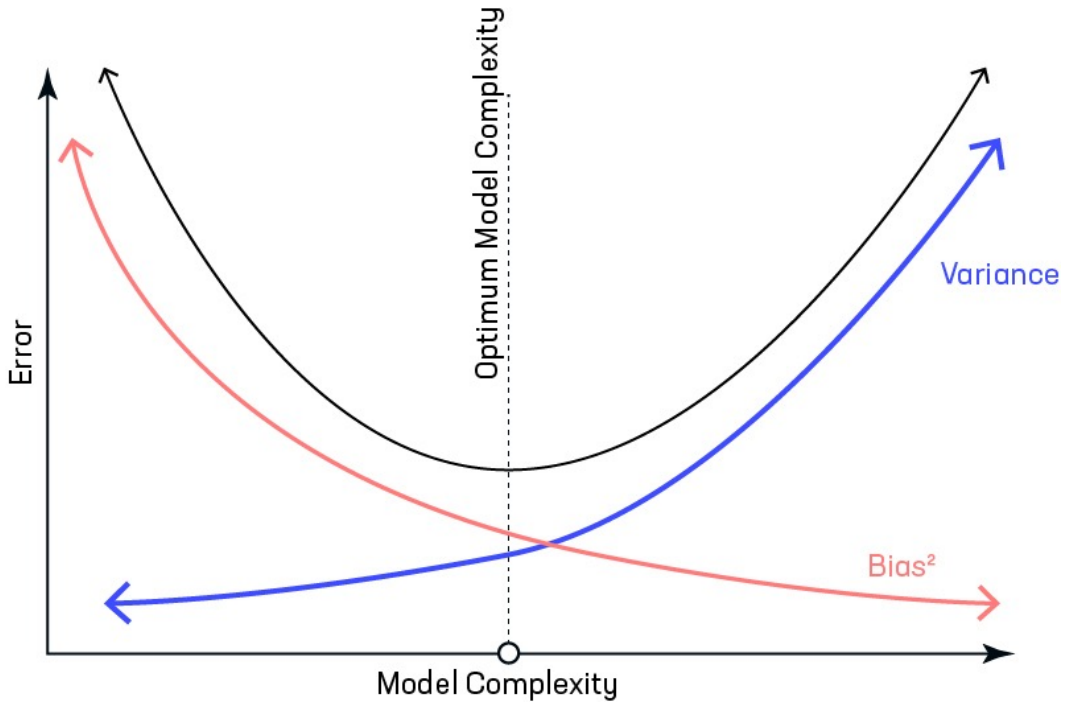
Bias: Difference between the true population parameter and the expected estimator
Variance: The spread, or uncertainty, in the estimates

$$Bias(\hat{\beta}) = E(\hat{\beta}) - \beta$$

$$\mathrm{Var}(\hat{\beta}) = E[\hat{\beta}^2] - \Big(E[\hat{\beta}]\Big)^2.$$



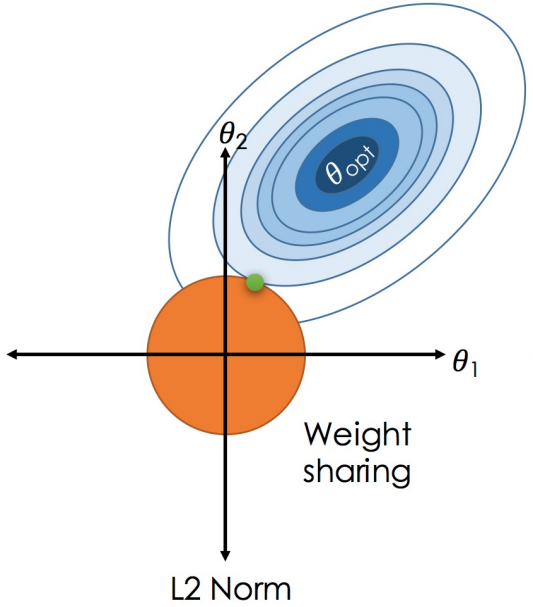


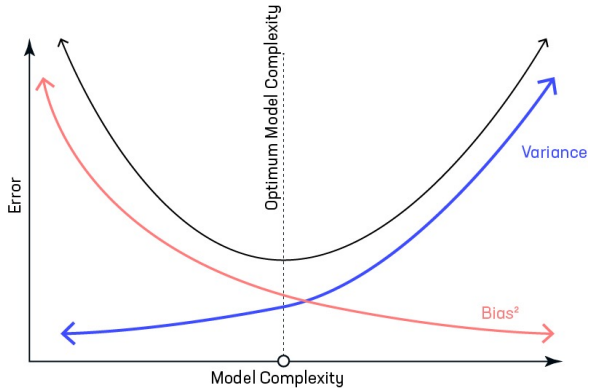


Regularization

Reducing the variance at the cost of introducing a tolerable level of bias

Ridge Regression

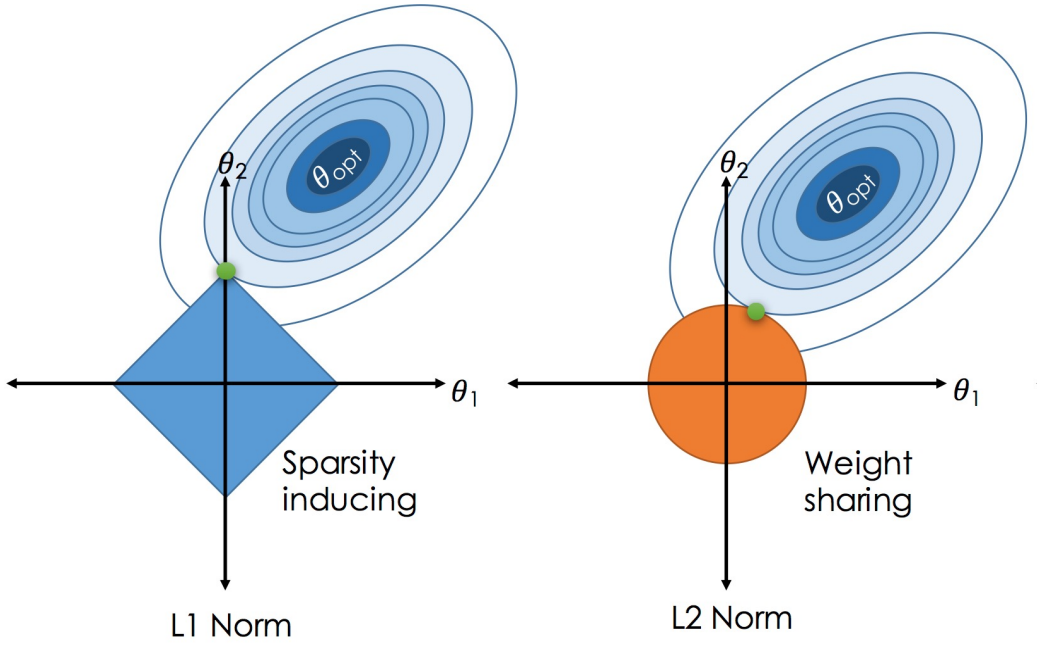




$$L_{Ridge} = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j{}^2$$

- Minimize subject to a bound on L2 norm of coefficients
- Continuous shrinkage/no variable selection

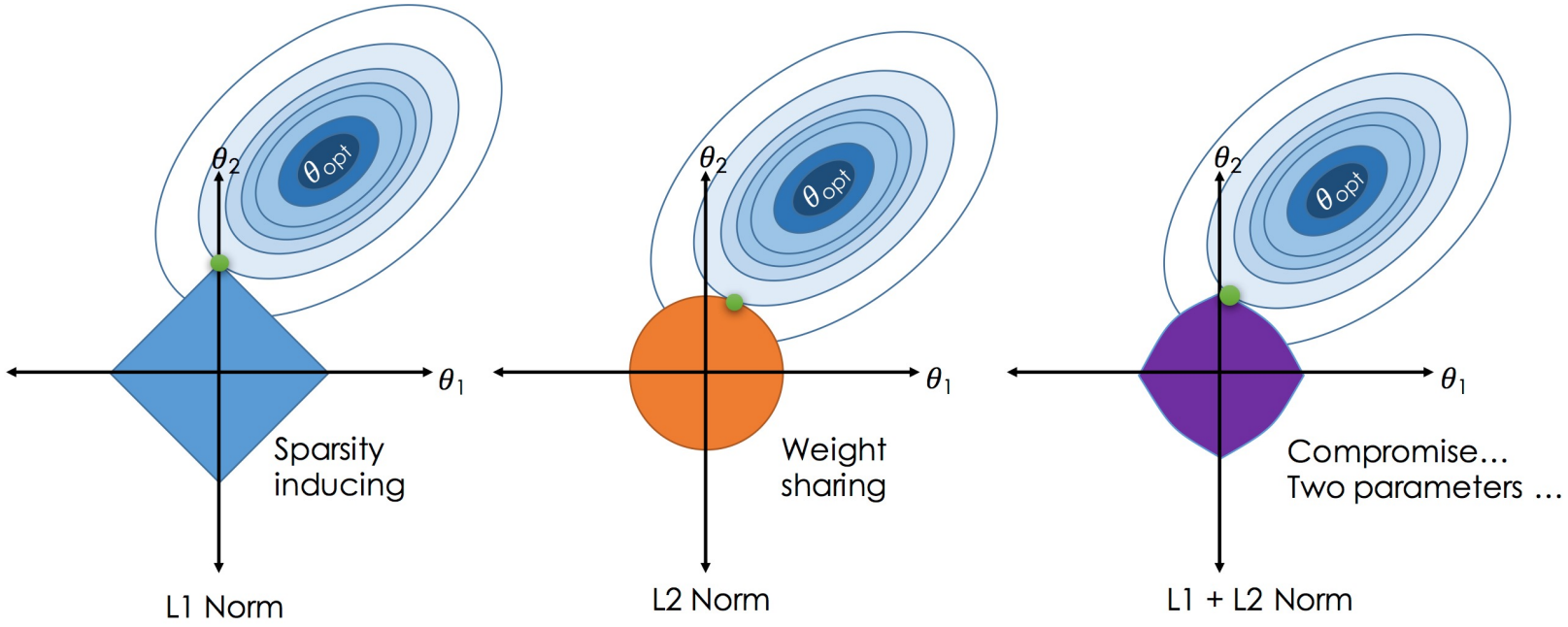
Lasso



$$L_{Lasso} = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

- Minimize subject to a bound on L1 norm of coefficients
- Sparsity inducing/removes variables

Elastic Net Penalty



$$L_{elnet} = \frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta}_i)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j{}^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

- Combination of Ridge and Lasso
- Involves a mixing parameter alpha

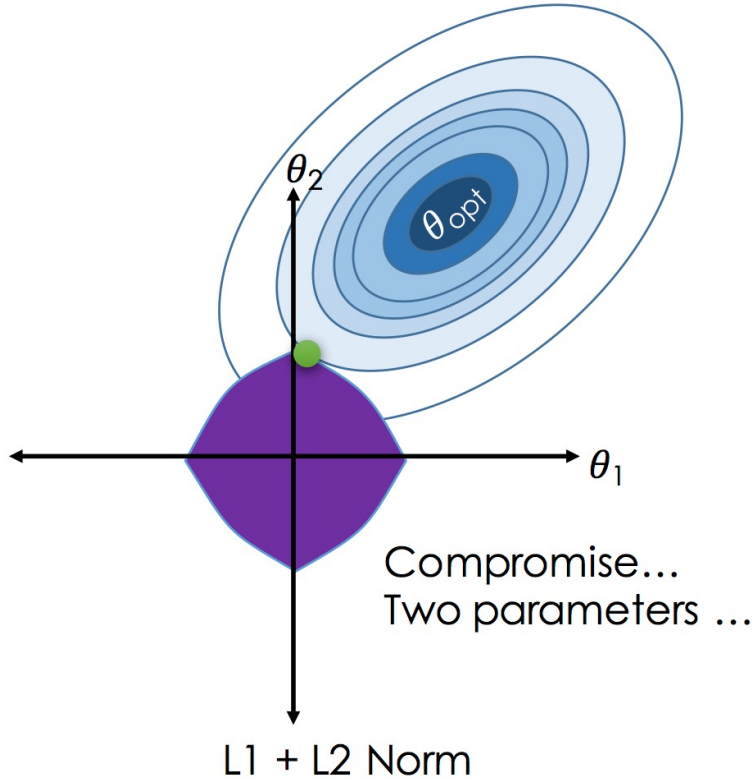
Elastic Net Advantages

<u>Compared to Ridge:</u>

- Performs feature selection

<u>Compared to Lasso:</u>

- Feature-group selection
- In the p > n case, the lasso selects at most n variables before it saturates



Downsides of Elastic Net

- Not a lot of downsides, which is good
- Computation cost
  - We have more parameters that must be tuned
- Flexibility of the estimator
  - Disadvantage? (maybe)
  - It can be hard to find the right ratio for the purpose you have
  - i.e. you might not be able to do what you're trying to do

Elastic Net Implementation w Caret in R

Variables in DVT example: Risk factors and Drugs

sex	Anesthesia	Peptic_Ulcer	dvt_hx
age	AIDS	PVD	pe_hx
race	Malignancy	CKD	pe
ethnicity	Cerebrovascular	Rheumatic	dvt
ht	COPD	htn	dvt_leg
wt	CHF	depression	dvt_leg_proximal
bmi	Dementia	obesity	dvt_leg_distal
financial_class	Diabetes_cc	alcohol	aspirin
cpt	Diabetes_no_cc	drug	coumadin
provider	Hemiplegia	tobacco	xarelto
length_of_stay	Metastatic	ulcerative_colitis	eliquis
disch_disp	Mild_Liver	lupus	heparin
length_of_surgery	Moderate_Liver	factor_v	lovenox
tourniquet_flag	MI	sickle_cell	pradaxa

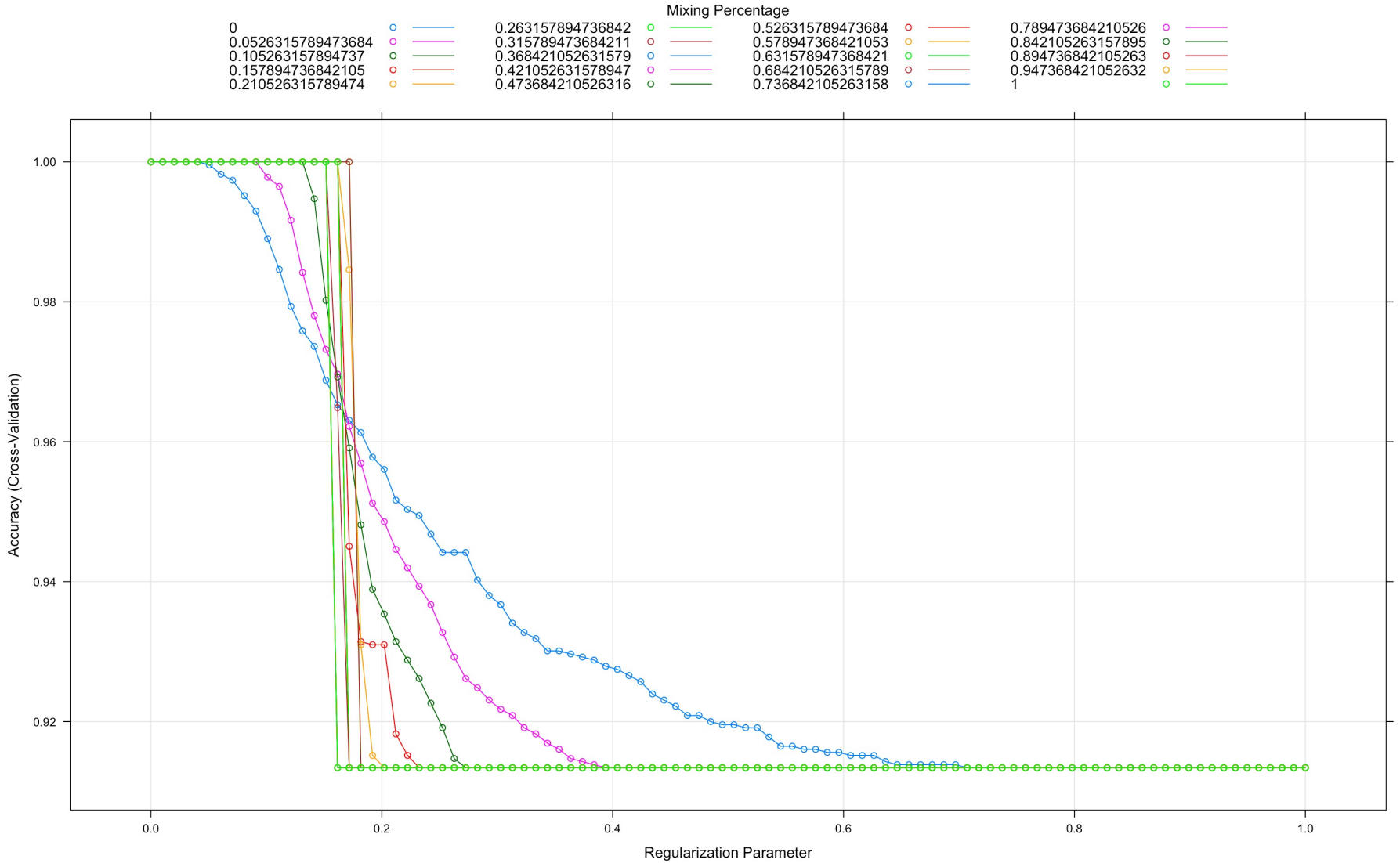
Elastic Net Implementation w Caret in R

Model DVT in Total Knee Arthroplasty

```
# Grid search & Elastic net
# Train elastic net model using caret
# while doing a parameter tuning of alpha
tka_dvt_elnet = train(
  tka_dvt_fact ~ ., data = tka_dvt_data,
  method = "glmnet",
  tuneGrid = expand.grid(alpha = seq(0, 1, length = 20),
                         lambda = seq(0.0001, 1, length = 100)),
  trControl = cv_5
)
```

- Using a grid search with library(caret)
- glmnet is the elastic net model
- Alpha: mixing parameter
- Lambda: shrinkage parameter

CV Accuracy Plot

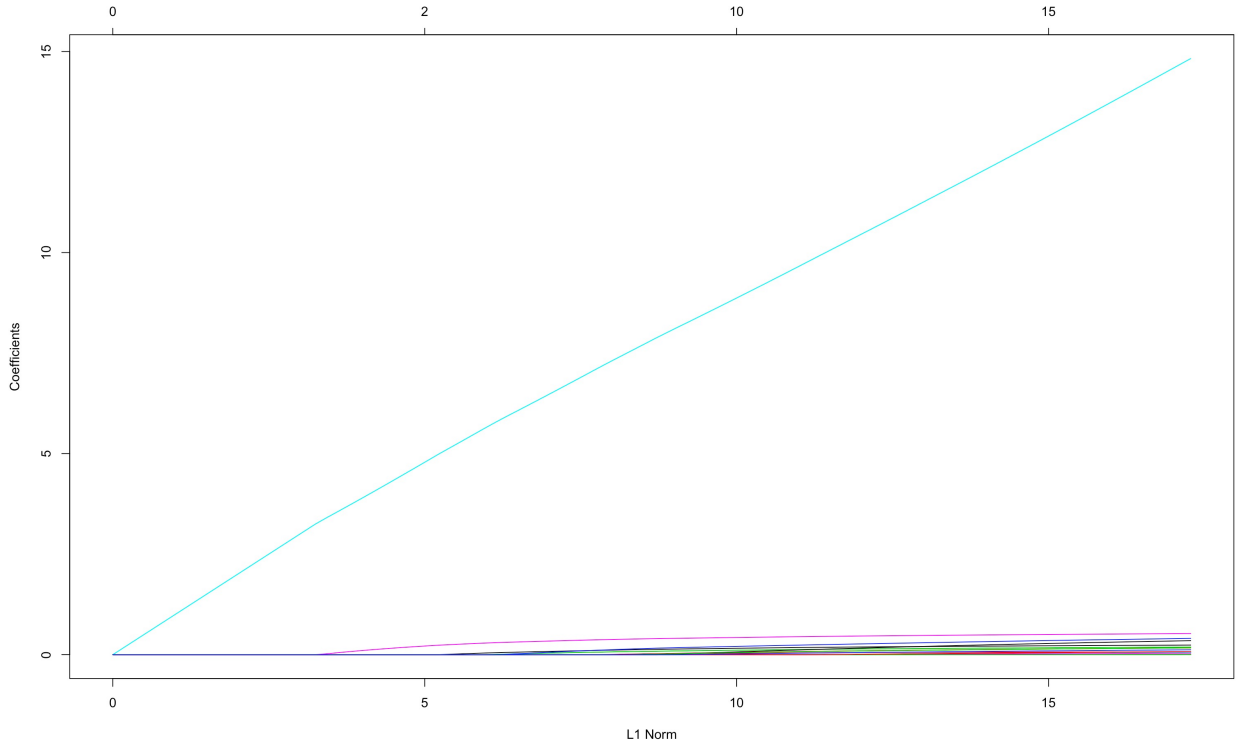


Elastic Net Implementation w Caret in R

Extract the best model

```
> get_best_result(tka_dvt_elnet)
      alpha lambda Accuracy Kappa AccuracySD KappaSD
1 0.2631579 0.1718        1     1          0       0
```

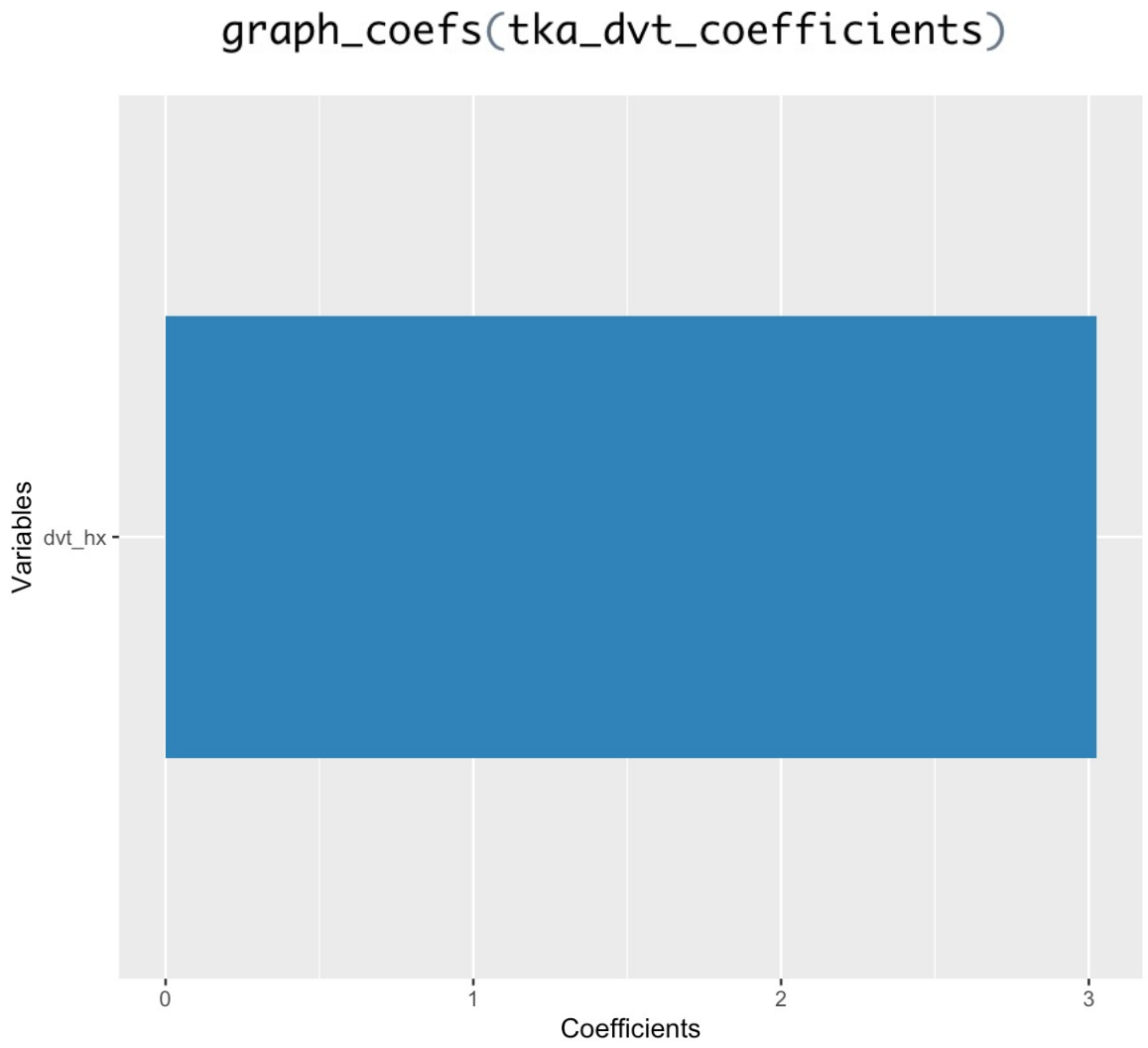
- Best values of the tuning parameters



- Coefficients of the variables against L1 norm

Elastic Net Implementation w Caret in R

Graph Non-Zero Variables



- Only a single variable dvt_hx was included in the final model out of the 56 we started with

References

- Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320. http://www.jstor.org/stable/3647580
- https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net
- https://towardsdatascience.com/bias-variance-and-regularization-in-linear-regression-lasso-ridge-and-elastic-net-8bf81991d0c5
- https://stats.stackexchange.com/questions/184029/what-is-elastic-net-regularization-and-how-does-it-solve-the-drawbacks-of-ridge
- https://corporatefinanceinstitute.com/resources/knowledge/other/elastic-net/
- https://en.wikipedia.org/wiki/Elastic_net_regularization

Questions/Comments