# Visual Explanation of Ridge Regression and LASSO

Kazuki Yoshida

2017-11-03

# OLS Problem

- The ordinary least square (OLS) problem can be described as the following optimization.

$$\arg\min_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 \right]$$

- That is, we try to find coefficient $\boldsymbol{\beta}$ that minimizes squared errors (squared distance between the outcome $Y_i$ and predicted outcome $\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}$).

# Advantages of OLS

- Provided the true relationship between the response and the predictors is approximately linear, the OLS estimates will have low bias. If the number of observations $n$ is much larger the number of variables $p$, then the OLS estimates tend to have low variance, and hence will perform well on test observations.

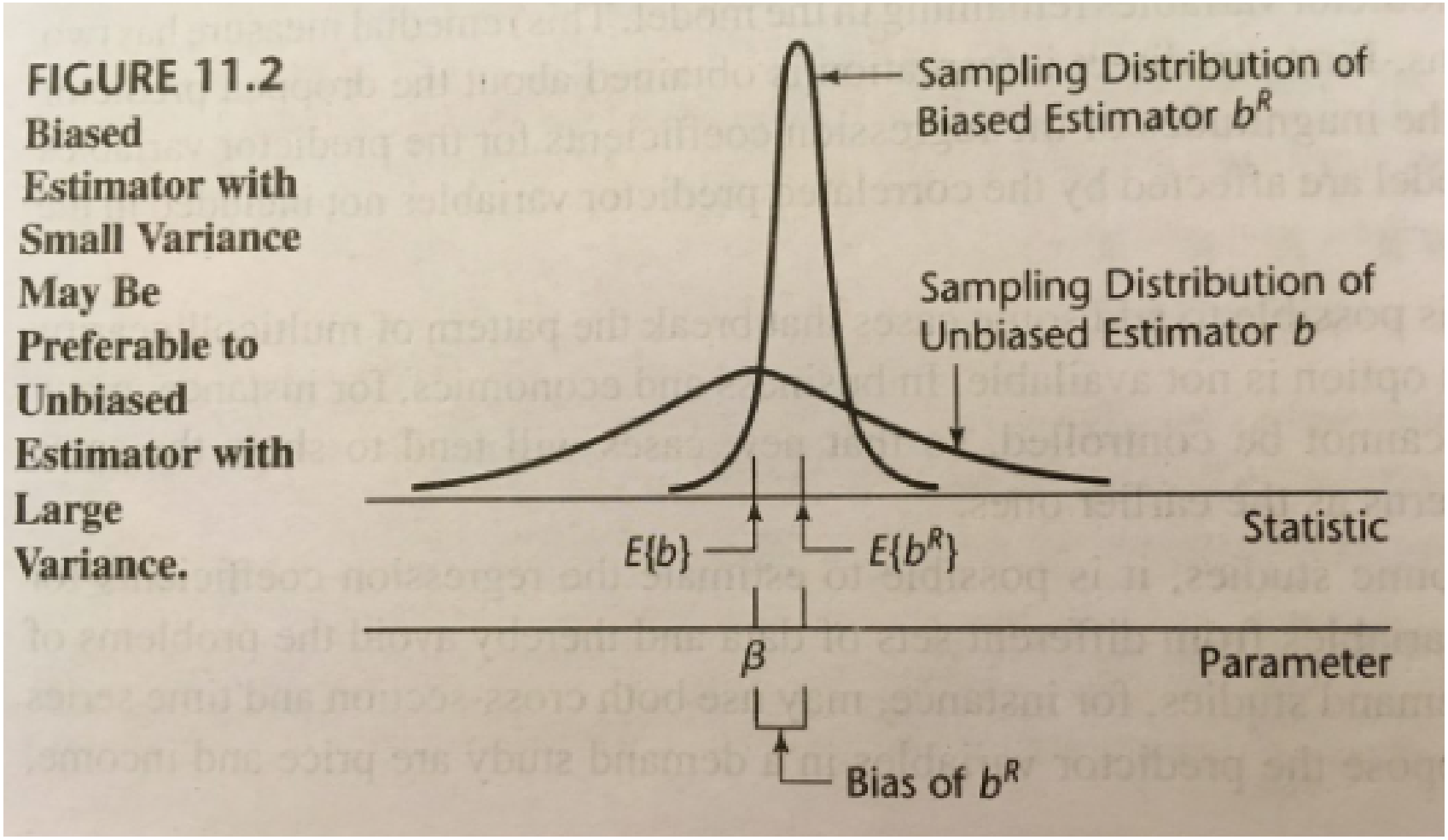
# Why OLS May Fail

- In settings with many explanatory variables, unless $n$ is extremely large, because of sampling variability the estimates $\left\{\widehat{\beta}_j\right\}$ tend to be much larger in magnitude than the true values $\{\beta_j\}$ (overfitting).
- This tendency is exacerbated when we keep only statistically significant variables in a model (typical stepwise selection). [Agresti, 2015]
- If $p > n$, there is no longer a unique OLS solution. [James et al., 2017]

# Penalized Regression Rationale

- Shrinkage (regularization) toward 0 tends to move $\left\{\widehat{\beta}_j\right\}$ closer to $\{\beta_j\}$. [Agresti, 2015]
- By shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias, substantially improving the accuracy of prediction for future observations. [James et al., 2017]

# Bias-Variance Tradeoff




[Kutner et al., 2004]

# Penalized Regression Definition

Ordinary Least Square

$$\arg\min_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 \right]$$

Ridge Regression

$$\arg\min_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right]$$

LASSO

$$\arg\min_{\boldsymbol{\beta}} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

# Penalized Regression Definition

$$\arg\min_{\beta} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]$$

- Both ridge regression and LASSO adds a penalty term for having "large" coefficients.
  - Ridge regression: Large with respect to the squared $L_2$ norm (Euclidean distance) $q = 2$
  - LASSO: Large with respect to the $L_1$ norm (Manhattan distance) $q = 1$
- Both have a tuning parameter $\lambda$, which decides how important the penalty is in relation to the squared error term.

# Ridge Regression vs OLS

- Ridge regression estimates tend to be stable in the sense that they are usually little affected by small changes in the data on which the fitted regression is based. In contrast, ordinary least squares estimates may be highly unstable under these conditions when the predictor variables are highly multicollinear.
- Predictions of new observations made from ridge estimated regression functions tend to be more precise than predictions made from OLS regression functions when the predictor variables are correlated and the new observations follow the same multicollinearity pattern. [Kutner et al., 2004]
- Unlike OLS, ridge regression will produce a different set of coefficient estimates $\widehat{\beta}_{\lambda}^{R}$ for each value of $\lambda$. Selecting a good value for $\lambda$ is critical. [James et al., 2017]

# LASSO vs Stepwise Model Selection

- Because of the nature of the constraint, making the penalty sufficiently large will cause some of the coefficients to be exactly zero.
- Thus, the lasso does a kind of *continuous* model selection [Hastie et al., 2016].
- On the other hand, the stepwise variable selection methods are *discrete* in nature because variables are either retained fully or discarded fully.

# LASSO and Coefficients

- Regarding the coefficients themselves, the LASSO shrinkage causes the estimates of the non-zero coefficients to be biased towards zero, and in general they are not consistent.
- One approach for reducing this bias is to run the LASSO to identify the set of non-zero coefficients, and then fit an unrestricted linear model to the selected set of features. This is not always feasible, if the selected set is large.
- Alternatively, one can use the LASSO again on the selected set of variable (relaxed LASSO).
  [Hastie et al., 2016] (p91)

# Ridge Regression vs LASSO

- A disadvantage of ridge regression is that it requires a separate strategy for finding a parsimonious model, because all explanatory variables remain in the model.
- When $p$ is large but only a few $\{\beta_j\}$ are practically different from 0, the LASSO tends to perform better, because many $\left\{\widehat{\beta}_j\right\}$ may equal 0.
- When $\{\beta_j\}$ do not vary dramatically in substantive size, ridge regression tends to perform better. [Agresti, 2015]
- Neither ridge regression nor the lasso will universally dominate the other. [James et al., 2017]
- Not performing variable selection may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables $p$ is quite large. The LASSO yields sparse models – that is, models that involve only a subset of the variables, which are generally much easier to interpret.

# Selecting the Tuning Parameter

- A penalty that is too large can prevent the penalized model from capturing the main signal in the data, while too small a value can lead to overfitting to the noise in the data.
- Cross-validation provides a simple way to tackle the problem of choosing a good tuning parameter $\lambda$. We choose a grid of $\lambda$ values, and compute the cross-validation prediction error for each value of $\lambda$. We then select the tuning parameter value for which the cross-validation error is smallest. Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.
  [James et al., 2017] (p227)

# Elastic Net

Elastic Net

$$\arg\min_{\beta} \left[ \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ji} \right)^2 + \lambda \sum_{j=1}^{p} \left( \alpha |\beta_j|^2 + (1-\alpha)|\beta_j| \right) \right]$$

- The elastic-net selects variables like the LASSO, and shrinks together the coefficients of correlated predictors like ridge.
- The LASSO does not handle highly correlated variables very well; the coefficient paths tend to be erratic and can sometimes show wild behaviors. If there are two identical variables, the coefficients are not uniquely estimated with the $L_1$ penalty although identical half-sized coefficients are estimated with the $L_2$ penalty.

# Set Up

$$\begin{bmatrix} X_{1i} \\ X_{2i} \end{bmatrix} \sim MVN\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$E[Y_i] = \beta_1 X_{1i} + \beta_2 X_{2i}$$
$$Y_i \sim N(\beta_1 X_{1i} + \beta_2 X_{2i}, \sigma^2)$$

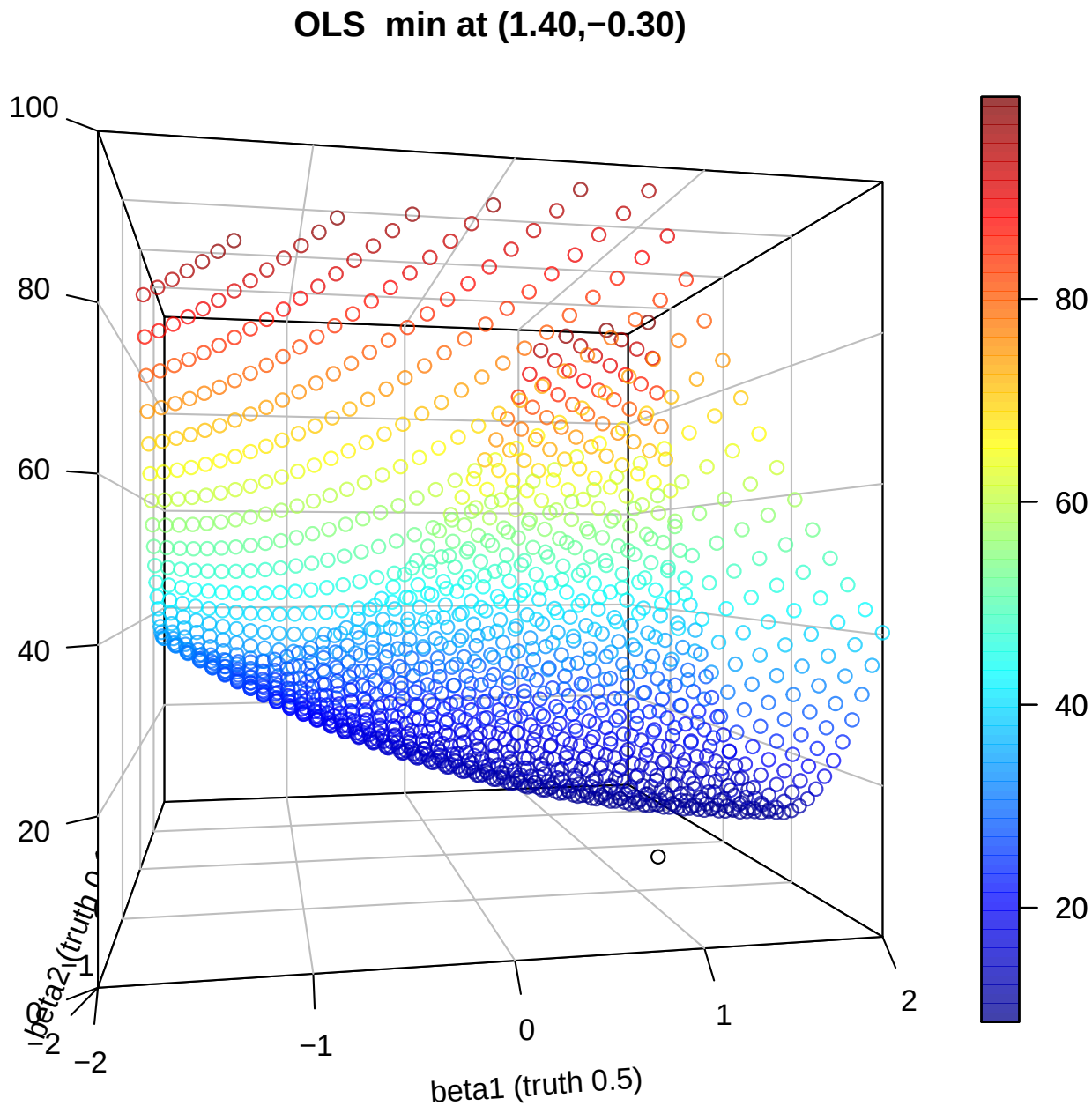
$$\rho = 0.7$$
$$\beta_1 = 0.5, \beta_2 = 0.1$$
$$\sigma^2 = 1$$

# Data

Ten data points were generated.

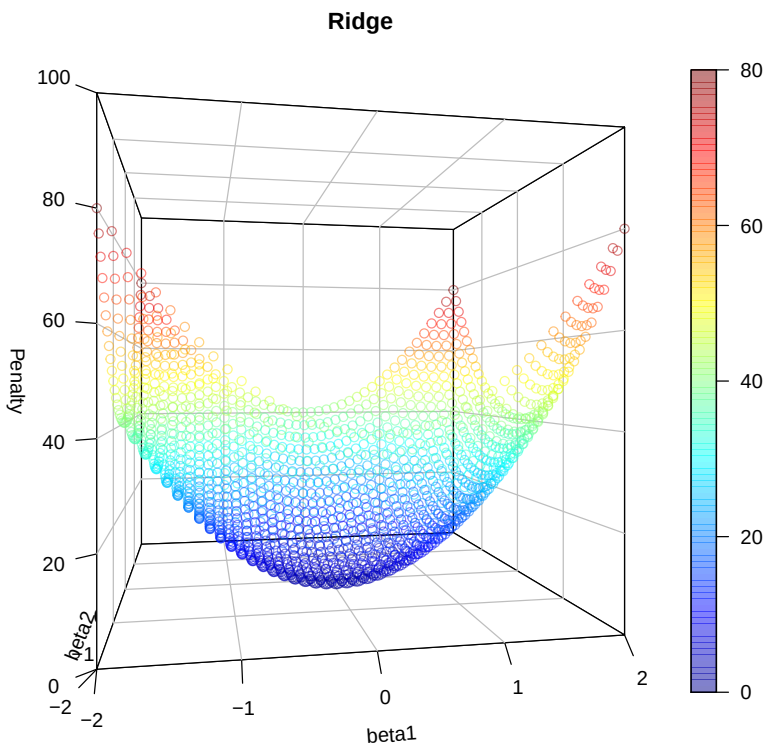
```
                Y          X1          X2
 [1,] -0.1031846  0.2975677  0.9011822
 [2,]  0.3685920 -0.8119366 -1.3164309
 [3,]  1.2895888  0.5110124 -0.8779760
 [4,] -0.9877068 -0.5490504 -0.2468024
 [5,]  1.0606268  0.8946819  0.7958214
 [6,]  3.6955798  2.6184162  3.5490718
 [7,]  0.2591585 -0.9085681 -1.1293642
 [8,] -0.4325165 -0.8695810 -0.5797419
 [9,] -0.5336736 -0.1324647  0.4957759
[10,]  1.5956205 -0.3602429 -0.1464311
```

# OLS Objective Function

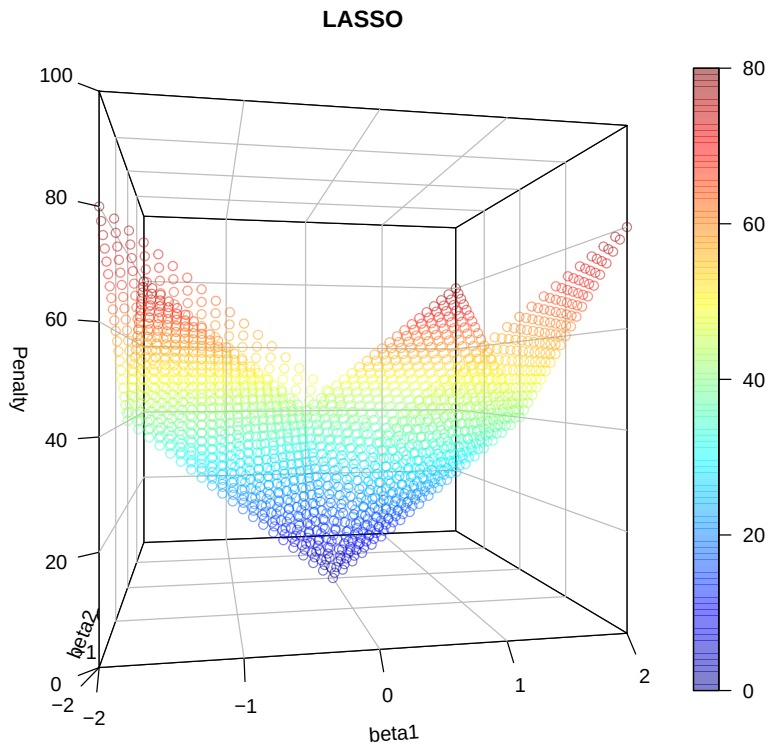


# Penalty Functions

## Ridge Regression

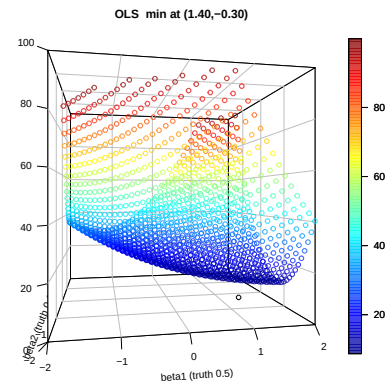



## LASSO

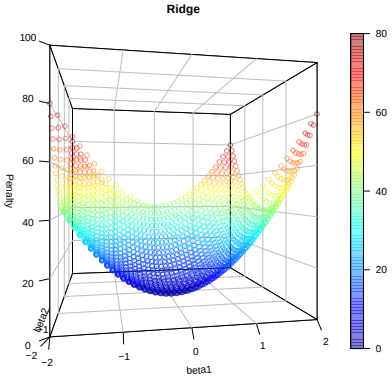


# Constructing Penalized Objective Functions

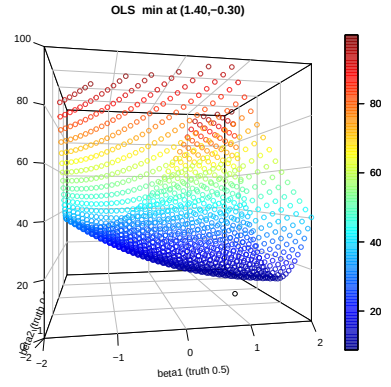
## Ridge Regression



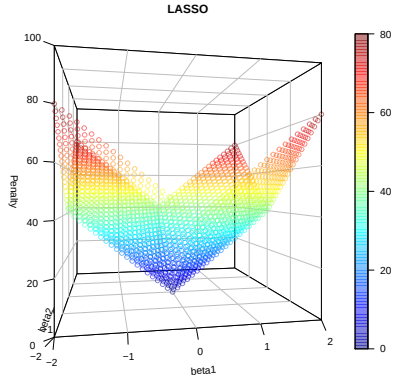

+




## LASSO

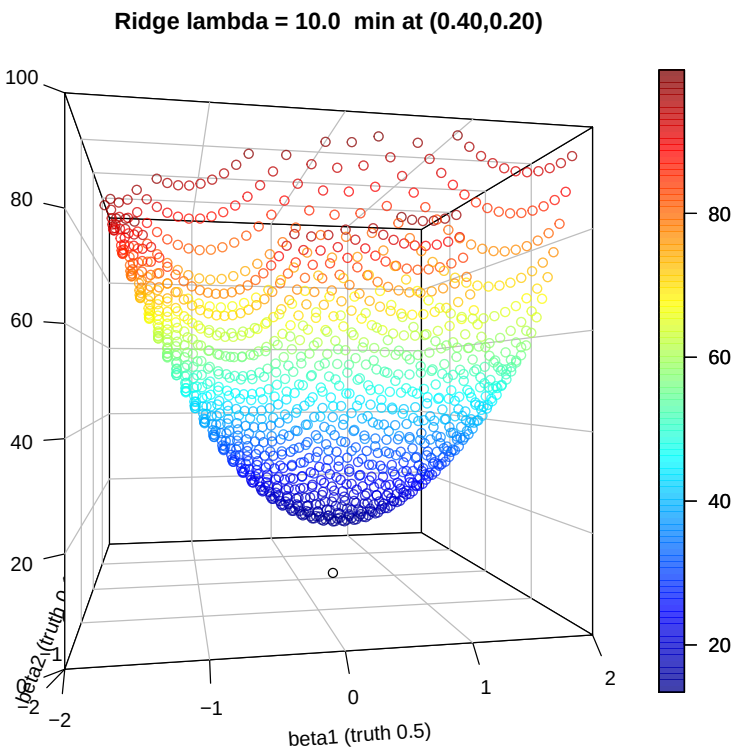



+

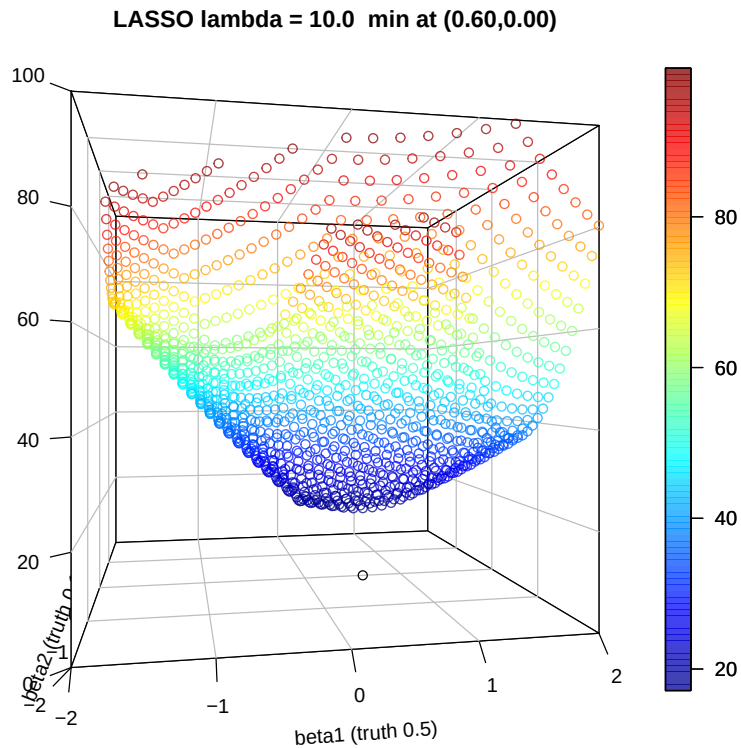


# Penalized Objective Functions

## Ridge Regression




## LASSO



# Bibliography I


[Agresti, 2015] Agresti, A. (2015).
*Foundations of Linear and Generalized Linear Models*.
Wiley, Hoboken, New Jersey, 1 edition edition.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010).
Regularization Paths for Generalized Linear Models via Coordinate Descent.
*J Stat Softw*, 33(1):1–22.

[Hastie et al., 2016] Hastie, T., Tibshirani, R., and Friedman, J. (2016).
*The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*.
Springer, New York, NY, 2nd edition edition.

[James et al., 2017] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017).
*An Introduction to Statistical Learning: With Applications in R*.
Springer, New York, 1st ed. 2013, corr. 7th printing 2017 edition edition.

[Kutner et al., 2004] Kutner, M. H., Neter, J., Nachtsheim, C. J., and Li, W. (2004).
*Applied Linear Statistical Models*.
McGraw-Hill Education, Boston, Mass., 5th international edition edition.