

MODEL FITTING

Prof. Alexander Huth

9/19/2017

RECAP

$$Y = f(X)$$

- * System identification
 - * Linear
 - * Linearized
 - * Nonlinear

RECAP

$$Y = f(X)$$

- * System identification

- * Linear

$$Y = X\beta$$

- * Linearized

$$Y = \mathbb{L}(X)\beta$$

- * Nonlinear

$$Y = \Theta(X)$$

RECAP

- * Spatiotemporal models
 - * Space-time separable
 - * Space-time inseparable
 - * Finite impulse response (FIR) model

LINEAR REGRESSION

RESPONSE

VARIABLES

WEIGHTS

NOISE

$$Y = X\beta + \epsilon$$

LINEAR REGRESSION

RESPONSE

VARIABLES

WEIGHTS

NOISE

$$Y = X\beta + \epsilon$$

Loss function:

$$||Y - X\beta||_2$$

LINEAR REGRESSION

$$\begin{array}{ccccccc} \text{RESPONSE} & & \text{VARIABLES} & & \text{WEIGHTS} & & \text{NOISE} \\ | & & | & & | & & | \\ Y & = & X & \beta & + & \epsilon \end{array}$$

- * How do we solve for beta?
- * Analytically
- * Iteratively

ANALYTIC SOLUTION TO LINEAR REGRESSION

RESPONSE

VARIABLES

WEIGHTS

NOISE

$$Y = X\beta + \epsilon$$

$$\beta = f(X, Y)$$

ITERATIVE SOLUTION TO LINEAR REGRESSION

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$\Delta\beta \propto -\frac{\partial(||Y - X\beta||_2)}{\partial\beta}$$

LINEAR REGRESSION

$$\begin{array}{ccccccc} \text{RESPONSE} & & \text{VARIABLES} & & \text{WEIGHTS} & & \text{NOISE} \\ | & & | & & | & & | \\ Y & = & X & \beta & + & \epsilon \end{array}$$

Constraining the values **beta** can take improves model performance:

REGULARIZATION

REGULARIZATION

- * Regularization can be thought of in three ways:
- * **Prior**
- * **Penalty**
- * **Geometry**

REGULARIZATION AS PRIOR

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$Y_{t,j} \sim \mathcal{N}(X\beta, \sigma^2)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} P(Y|X, \beta)$$

$Y_{t,j} \sim \mathcal{N}(X\beta, \sigma^2)$

$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} P(Y|X, \beta)$

REGULARIZATION AS PRIOR

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$Y_{t,j} \sim \mathcal{N}(X\beta, \sigma^2)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} P(Y|X, \beta)P(\beta)$$

$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} P(Y|X, \beta)P(\beta)$

REGULARIZATION AS PENALTY

RESPONSE

VARIABLES

WEIGHTS

NOISE

$$Y = X\beta + \epsilon$$

$$E(\beta) = ||Y - X\beta||_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} E(\beta)$$

$$E(\beta) = ||Y - X\beta||_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} E(\beta)$$

REGULARIZATION AS PENALTY

$$E_{\text{pen}}(\beta) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} E_{\text{pen}}(\beta)$$

RESPONSE

VARIABLES

WEIGHTS

NOISE

$$Y = X\beta + \epsilon$$

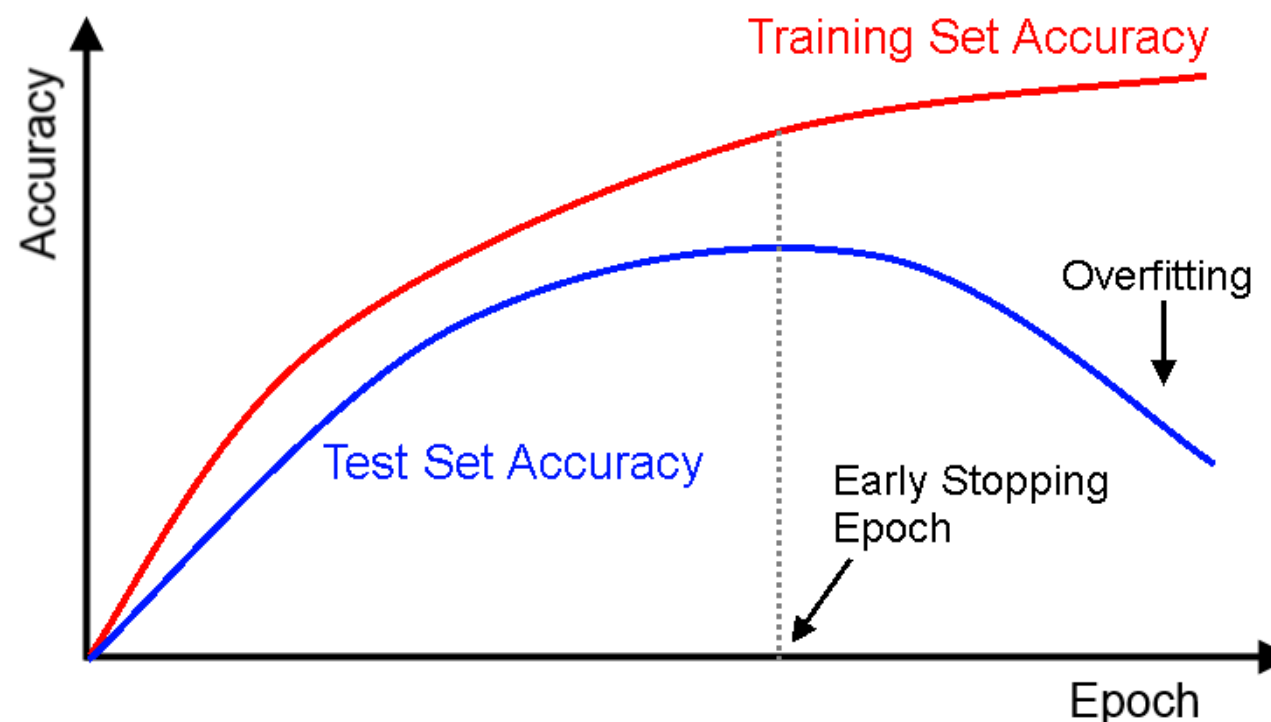
$$E_{\text{pen}}(\beta) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} E_{\text{pen}}(\beta)$$

REGULARIZATION AS GEOMETRY - EARLY STOPPING

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$



COMMON TYPES OF REGULARIZATION

- * **Beta is small (L2-sense) = ridge =**
gradient descent w/ early stopping
- * **Beta is small, sparse (L1-sense) = LASSO**
= coord. descent w/ early stopping
- * **Beta is small & sparse (L1+L2 sense) =**
elastic net
- * **Beta is sparse (L0-sense) = variable**
selection

RIDGE REGRESSION

- * Multivariate normal (MVN) prior on β
- * L2 penalty on β
- * Gradient descent w/ early stopping

RIDGE REGRESSION

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right]$

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}$

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\|Y - X\beta\|_2^2}_{\text{ERROR or LOSS}} + \underbrace{\lambda \|\beta\|_2^2}_{\text{PENALTY}} \right]$$

RIDGE REGRESSION

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

$$\hat{\beta} = X_{ridge}^{+} Y$$

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

RIDGE REGRESSION

* Efficient solution with SVD

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

$$\text{(SVD)} \quad X = U S V^{\top} \quad D = \frac{S}{S^2 + \lambda^2}$$

$$\hat{\beta} = V D U^{\top} Y$$

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

$$D = \frac{S}{S^2 + \lambda^2}$$
$$\hat{\beta} = V D U^{\top} Y$$

RIDGE REGRESSION

- * How to choose λ ?
- * GCV - Generalized Cross Validation 🙄
- * Block-wise cross-validation 👍

RIDGE REGRESSION

- * Good implementation: `scikit-learn`
- * Awesome implementation:
<http://github.com/alexhuth/ridge>

LASSO

- * Laplacian prior on β_i
- * L1 penalty on β
- * Coordinate descent w/ early stopping

LASSO

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right]$

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}$

$$Y = X\beta + \epsilon$$

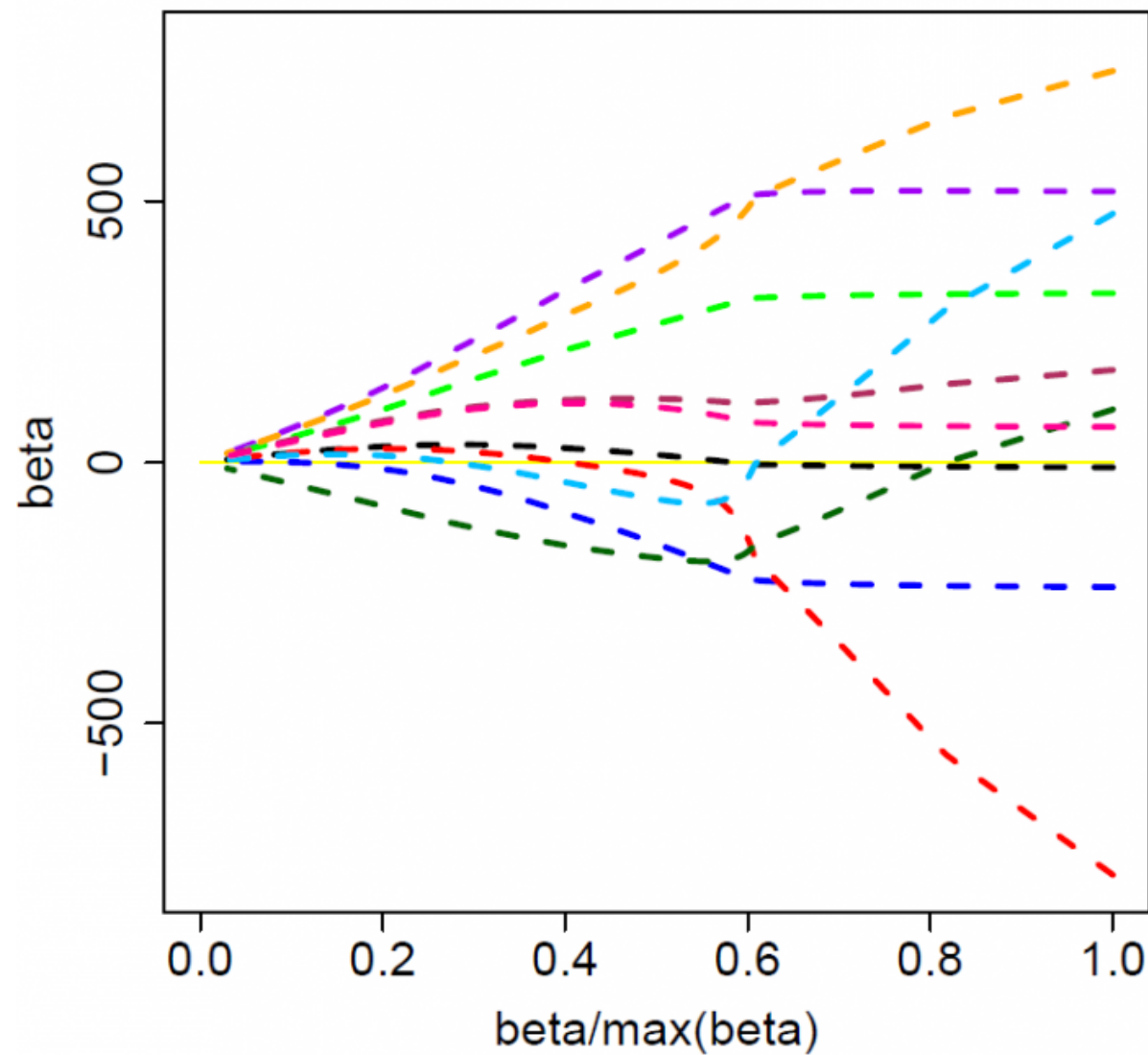
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\underbrace{\|Y - X\beta\|_2^2}_{\text{ERROR or LOSS}} + \underbrace{\lambda \|\beta\|_1}_{\text{PENALTY}} \right]$$

LASSO

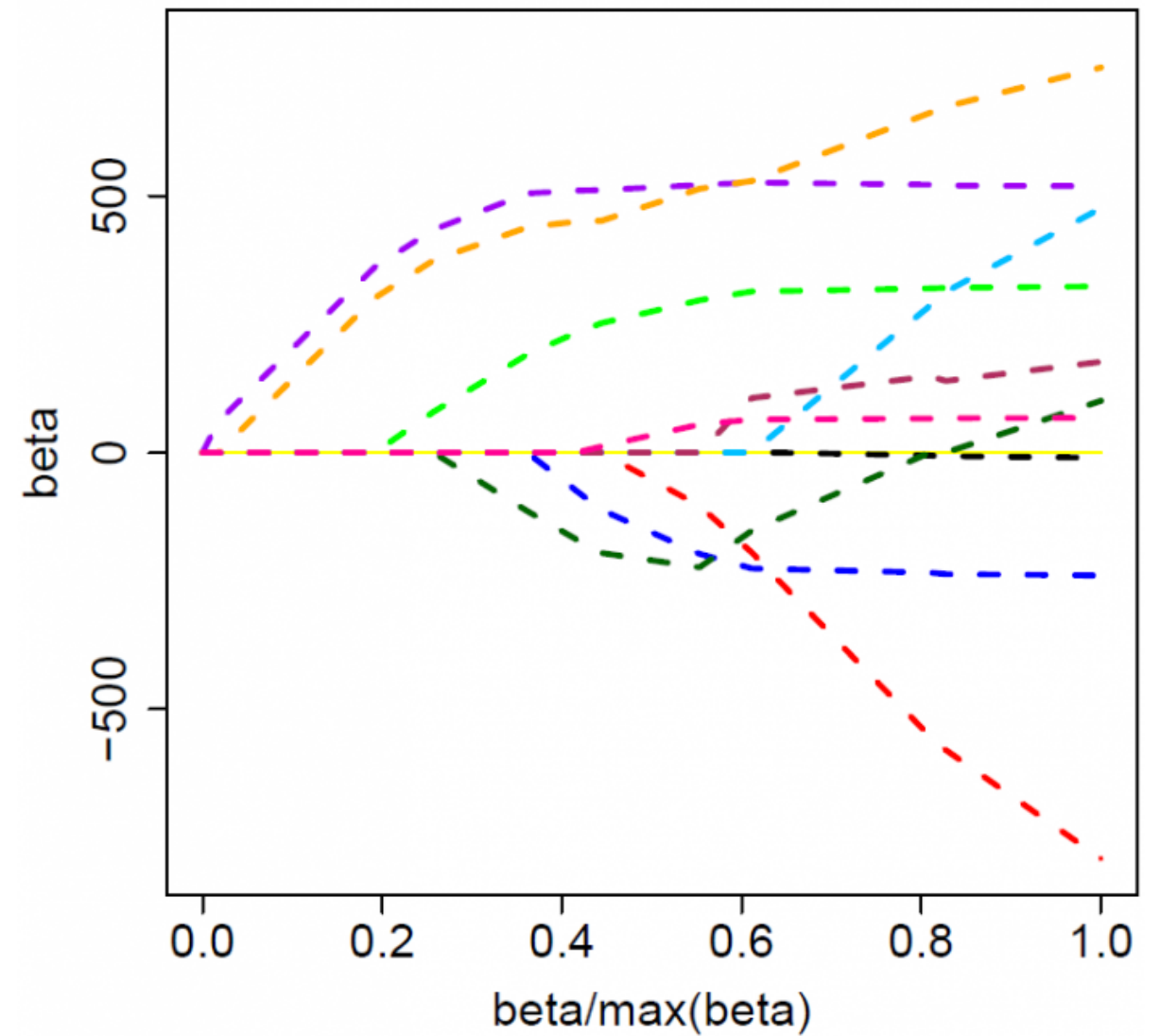
- * No closed form solution
- * Solved via coordinate descent, LARS (least-angle regression) or other methods
- * *SssssLLLLLoooooWWWW.....*

LASSO

Ridge Regression



Ordinary Lasso



OTHER METHODS

- * Neural networks
- * Random forests
- * Feature selection ($\sim L_0$ -norm)

NEXT TIME

- * Tikhonov regression