Linear Methods for Regression

Parthajit Mohapatra
and
Venugopalakrishna Y R

SPC Lab, IISc

15^{th} of February 2013

Outline

- Introduction to Regression and Linear Methods
- Linear Least Squares Problem
- Least Squares with Regularization
- Subset Selection
- Shrinkage Methods
  - Ridge Regression
  - Least Absolute Shrinkage and Selection Operator

Machine Learning

- Learning from data
- Training set: $(\mathbf{x}^{(1)}, y^{(1)}),\ (\mathbf{x}^{(2)}, y^{(2)}) \ldots (\mathbf{x}^{(N)}, y^{(N)})$
- $\mathbf{x}^{(i)} \in \mathcal{X}$ is i^{th} input feature vector
- $y^{(i)} \in \mathcal{Y}$ is i^{th} ouput measurement

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. A prediction model was desired for the response variable, a measure of disease progression one year after baseline.

- Regress: the act of reasoning backwards from effect to the cause
- Regression Analysis: Learning the model that best describes the relationship between the output measurements and the corresponding input feature vector
- $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Purpose of the model: to predict output for new inputs

Key questions in Regression (Machine Learning)

- What is the best $f : \mathcal{X} \to \mathcal{Y}$ that agrees with our data?
- What is the best f that generalizes for a new data point $\mathbf{x}^{(new)}$?
- What are the efficient algorithms to find f?

Empirical Risk function

- $C(f(\mathbf{x}), y)$ - cost of using $f(\mathbf{x})$ as an estimate of y
- Minimum expected risk

$$f^* = \arg\min_f \int_{\mathcal{X},\mathcal{Y}} C(f(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy$$

- Joint distribution $p(\mathbf{x}, y)$ is not known
- Minimum empirical risk

$$f_N^* = \arg\min_f \sum_{i=1}^{N} C(f(\mathbf{x}^{(i)}), y^{(i)})$$

Cost Function and Class of Hypothesis Functions

- $C(f(\mathbf{x}), y) = (y - f(\mathbf{x}))^2$
- Cost is high for large errors
- Linear model: $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \mathbf{x}^T \beta$
- Cost function remains convex (quadratic) in βs

Linear Methods

- Linear models are simple and interpretable (Linear algebra)
- Closed form solutions/efficient algorithms are available
- Non-linear problems can be transformed to linear problems
- Eg. $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_2 x_2 + \beta_3 x_2^2$
  $[x_1 \ x_1^2 \ x_2 \ x_2^2] \rightarrow [z_1 \ z_2 \ z_3 \ z_4] \in \mathbb{R}^4$
  $g(\mathbf{z}) = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4$ is linear in $\mathbf{z}$

Least Squares Problem

- $\hat{\beta} = \arg\min_\beta \sum_{i=1}^{N} (y^{(i)} - \mathbf{x}^{(i)^T} \beta)^2$
- $\hat{\beta} = \arg\min_\beta (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$
- $\mathbf{X}$ is $N \times p$, i^{th} row- input features of i^{th} example, j^{th} column- j^{th} input feature of all examples
- $f(\mathbf{x}) = \mathbf{x}^T \hat{\beta}$

Solution to Least Squares Problem

- $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- $N < p$, $\mathbf{X}^T\mathbf{X}$ is not invertible
- $N > p$, $\mathbf{X}^T\mathbf{X}$ should be of full rank p for it to invertible
- Cause of non-invertibility: Redundant features or too many features

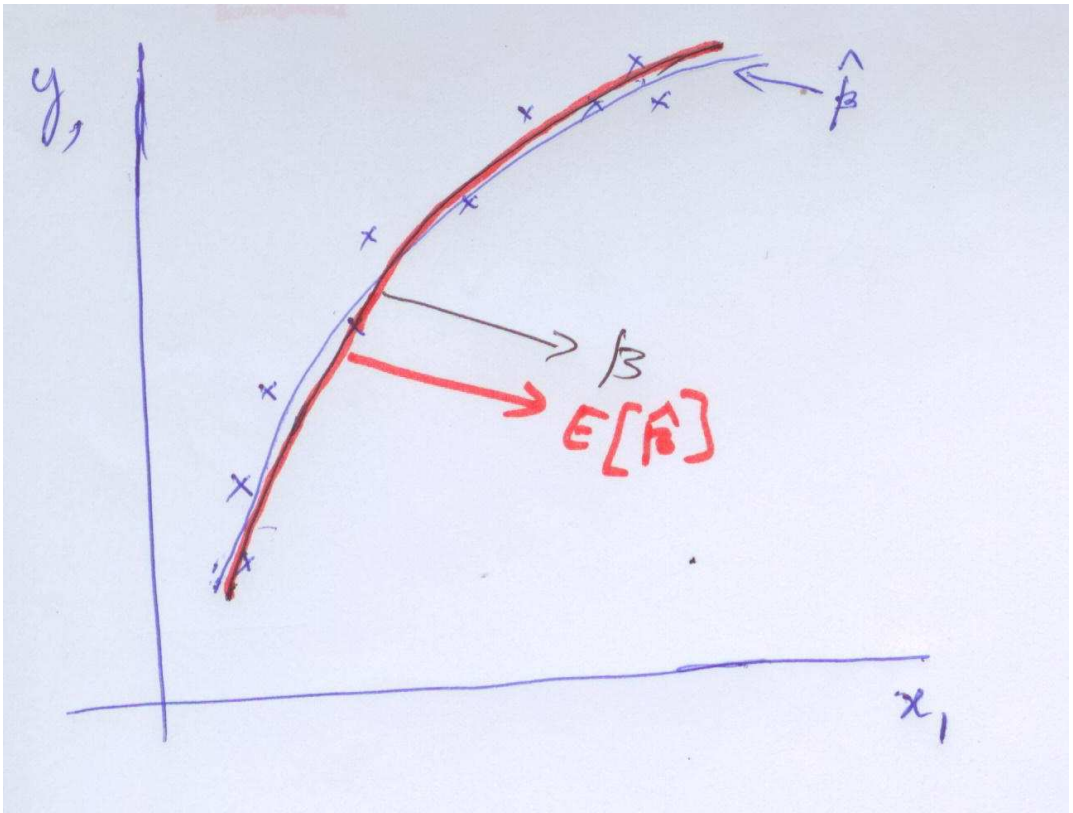
Gauss Markov Theorem

- How to test the goodness of the model? - Test data or Probabilistic Model
- $\mathbf{y} = \mathbf{X}\beta + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(0, \sigma^2)$
- Least squares estimate is MVUE ($E[\hat{\beta}] = \beta$)
- $MSE[\hat{\beta}] = Var(\hat{\beta}) + (E[\hat{\beta}] - \beta)^2$

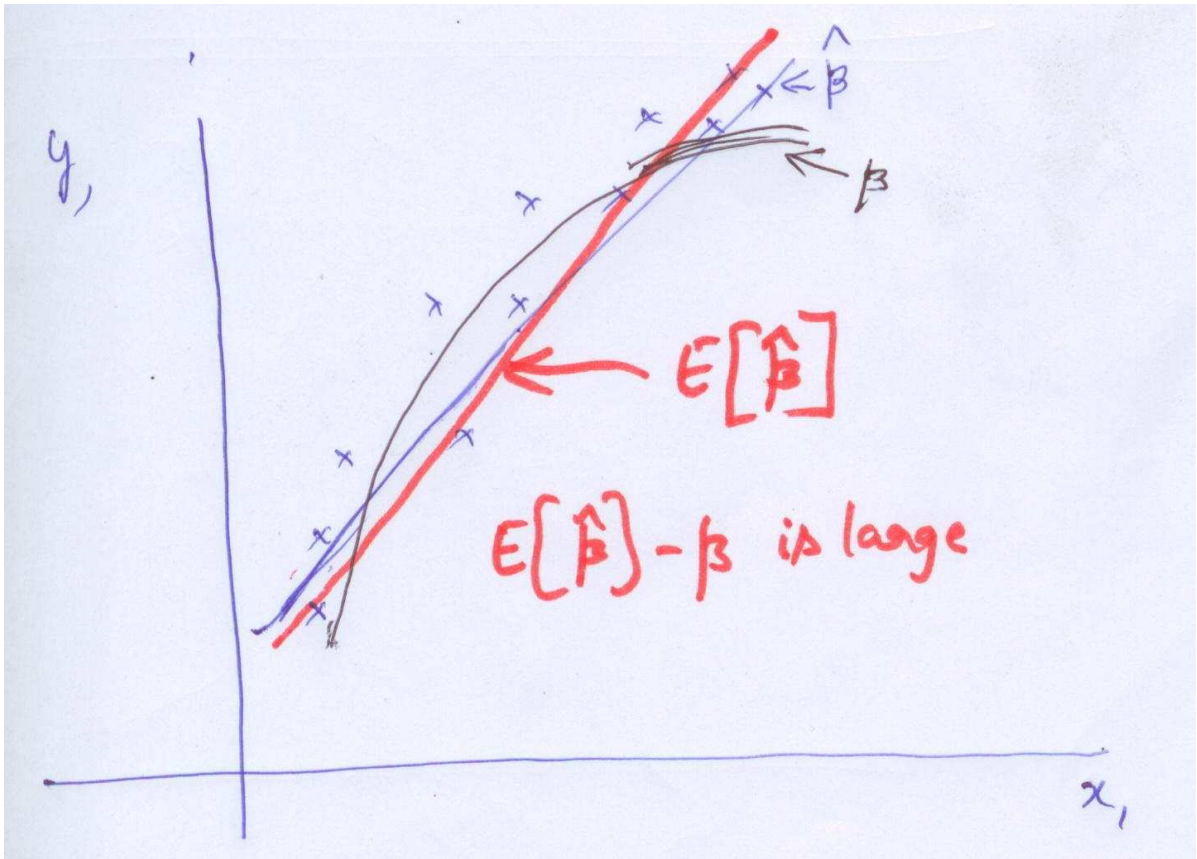
Need for Regularization

- Actual underlying model is not known
- Consider to fit a model
  $f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_2 x_1^3 + \beta_3 x_1^4$
- Let actual data be from a quadratic model
- Expected prediction error$= \sigma^2 + MSE[\hat{\beta}]$

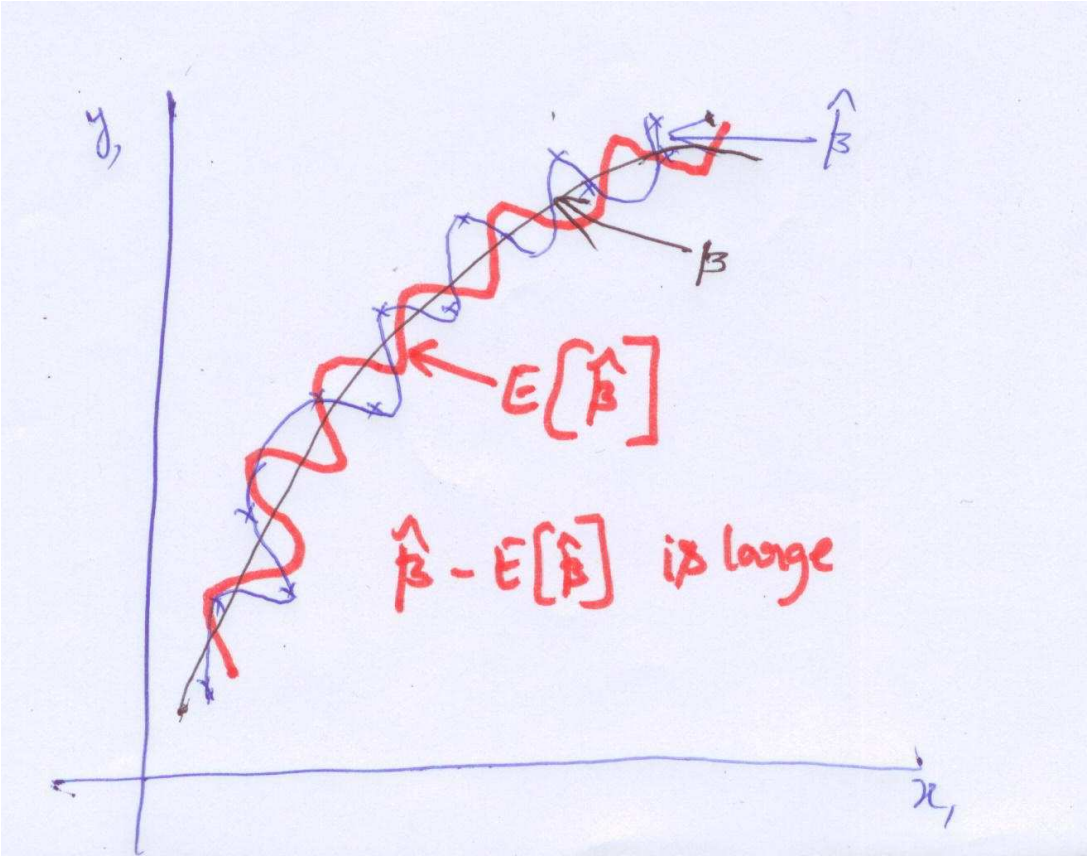
Good Fit



Under Fit



Over Fit



Subset selection

- Problem with least square estimate
  - Prediction accuracy: low bias but large variance
  - Model interpretation: large number of predictors
- How to overcome
  - Prediction accuracy: shrinking or setting some coefficients to zero
  - Determine a smaller subset that exhibit the strong effects
  - "Big picture" of the model with sacrifice in small details

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. A prediction model was desired for the response variable, a measure of disease progression one year after baseline.

Notion of subset selection

- Retain only a subset of the variables
- Least square regression: to estimate coefficients of the inputs that are retained
- Different strategies to select the subset
  - Best subset selection
  - Forward and backward-stepwise selection
  - Forward stagewise (FS) regression

Best subset selection

- Finds for each $k \in \{0, 1, 2, ..., p\}$ the subset of size k that gives smallest residual sum of squares (RSS)
- Algorithm: leaps and bounds procedure (Furnival and Wilson, 1974)
- Works for p as large as 30 or 40
- How to choose k
  - Involves the tradeoff between bias and variance
  - Can be subjective
  - Typically used: smallest model that minimizes an estimate of the expected prediction error
- Need to search all possible subsets

Forward and backward-stepwise selection

- Forward-stepwise selection
  - Starts with intercept, and then sequentially adds into the model the predictor that most improves the fit
  - Builds the model sequentially by adding one variable at a time
- Backward-stepwise selection
  - Starts with the full model, and sequentially deletes the predictor that has the least impact on the fit
  - Can be used only when $N > p$
  - Forward stepwise can always be used

Forward stagewise regression

- Cautious version of forward selection
- May take a large number of steps as it moves towards a final model
- Notation:
  $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$: n-vectors representing the features
  $\mathbf{x}_j$: represents jth feature vector
  $\mathbf{y}$: vector of responses
  $\beta = [\beta_1, \ldots, \beta_p]^T$
- Total squared error:

$$S(\beta) = \| \mathbf{y} - \mu \|^2, \text{ where } \mu = \mathbf{X}\beta$$

Algorithm

- Starts with $\mu = 0$
- Let μ current stagewise estimate
- $\hat{c} = \mathbf{X}^T(\mathbf{y} - \mu)$: vector of current correlation and c_j: proportional to the correlation between x_j and current residual vector
- Next step is taken in the direction of the greatest current correlation
- $j = \arg\max |\hat{c}_j|$ and $\mu \leftarrow \mu + \epsilon \text{sign}(\hat{c}_j) x_j$
- This is continued till none of the variables have correlation with the residuals

Shrinkage methods

- Subset selection methods: discrete in nature
- Suffer from high variance
- Shrinkage methods: continuous in nature
- Don't suffer much from high variability

Ridge regression

- Shrinks the regression coefficients by imposing a penalty in their sizes
- Minimize a penalized residual sum of squares

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{p} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where $\lambda \geq 0$: controls the amount of shrinkage

- Equivalent problem

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{p} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{sub. to} \sum_{j=1}^{p} \beta_j^2 \leq t$$

- One to one correspondence between λ and t

- Assuming the data is centered, β_0 can be removed
- Residual sum of square:

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

- $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- One can find an estimate of $\hat{\beta}$ even if $\mathbf{X}^T\mathbf{X}$ is singular

Interpretation of ridge regression

- SVD: $X = \mathbf{UDV}^T$
- Least square fitted vector:

$$\begin{aligned} \mathbf{X}\hat{\beta}_{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y} \end{aligned}$$

- $\mathbf{U}^T\mathbf{y}$: coordinates of $\mathbf{y}$ wrt orthonormal basis $\mathbf{U}$

- Ridge regression

$$\begin{aligned}\mathbf{X}\hat{\beta}_{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

  where $\mathbf{u}_j$: jth column of $\mathbf{U}$

- Computes the coordinates of $\mathbf{y}$ wrt the orthonormal basis $\mathbf{U}$ and then it shrinks the coordinates by $\frac{d_j^2}{d_j^2+\lambda}$
- $\lambda = 0$: reduces to least square solution

LASSO

- Subset selection: Can provide interpretable models but can be extremely variable
- Ridge regression: More stable but does not give an easily interpretable model
- Can we get best of both these models?

- Shrinks some coefficients and sets other to 0
- Retains good features of ridge regression and subset selection
- Hence, named as LASSO (Least absolute shrinkage and selection operator)
- In the signal processing literature, the LASSO is also known as basis pursuit

- LASSO estimate

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Equivalent problem

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{sub. to} \sum_{j=1}^{p} |\beta_j| \leq t$$

- Parameter t: controls the amount of shrinkage
- Let $\hat{\beta}_j^{\text{ls}}$: ordinary least square estimate and $t_0 = \sum |\hat{\beta}_j^{\text{ls}}|$
- If $t > t_0$, then LASSO estimates are the $\hat{\beta}_j^{\text{ls}}$
- If $t = \frac{t_0}{2}$, then the least squares coefficients are shrunk by about 50%
- Making t sufficiently small will cause some of the coefficients to be exactly zero

- No closed form expression
- Solution of LASSO: quadratic programming problem
- Efficient algorithm: LAR (Least angle regression)

Orthonormal design case

- Ridge regression: $\hat{\beta}_j^{\text{ridge}} = \frac{\hat{\beta}_j^{\text{ls}}}{1+\lambda}$
- Performs a proportional shrinkage
- LASSO: $\hat{\beta}_j^{\text{LASSO}} = \text{sgn}(\beta_j^{\text{ls}})(|\beta_j^{\text{ls}}| - \frac{\lambda}{2})^+$
- Performs soft thresholding

Summary

	LEAST SQ.	SUBSET SELECTION	RIDGE REGRESSION	LASSO
BIAS	UNBIASED	BIASED	PAY LITTLE MORE ON BIAS	PAY LITTLE MORE ON BIAS
VARIANCE	HIGH VARIANCE	HIGH	LOW VARIANCE	LOW
REMARK	DOESNOT GIVE INTERPRETABLE MODEL	GIVES INTER-PRETABLE MODEL	DOESNOT GIVE INTERPRETABLE MODEL	GIVES INTER-PRETABLE MODEL
CLOSED FORM SOLUTION	IF X^TX IS INVERTIBLE		EXISTS	DOES NOT EXIST