Predicting House Price on Ames Dataset

Evaluating Linear and Bayesian Regression

Supervisor :
Anirban Basu

Prepared By:-
Archit Vora (2015HT12480)

Data Description

- 80 features
  - 23 Nominal
    - Hair color : Black, Brown
    - Agriculture, Commercial, Ordinal
  - 23 Ordinal
    - Service satisfaction
      - Unsatisfied, neutral, satisfied
    - Bad, Good, Excellent
  - 14 Discrete
  - 20 Continuous

- Example Features
  - Nearby Area
  - No of bedrooms
  - Parking Facility
  - Kitchen Quality
  - Swimming Pool
  - No of Floors

Filling Up Missing Values

- Common
  - Encoding for Ordinal variables
  - Leveling for nominal variables
- Specific
  - Replace with most frequent value for discrete and nominal
    - When no of missing values are less
    - Mostly the case with our data
  - Sometimes null value implies feature is not available
    - Basement area null gets replaced with zero
  - Inferring from related variables
    - When garage year is null, we say house is without garage
    - Allows us to infer garage area, no of cars, garage condition
  - Linear Regression : Predict Lot Frontage from lot area (after sqrt)
- Improvements:
  - Creating Levels blindly : 288 features, 0.157122 rmse
  - After logically filling missing values : 263 features, 0.155965

Linear Regression

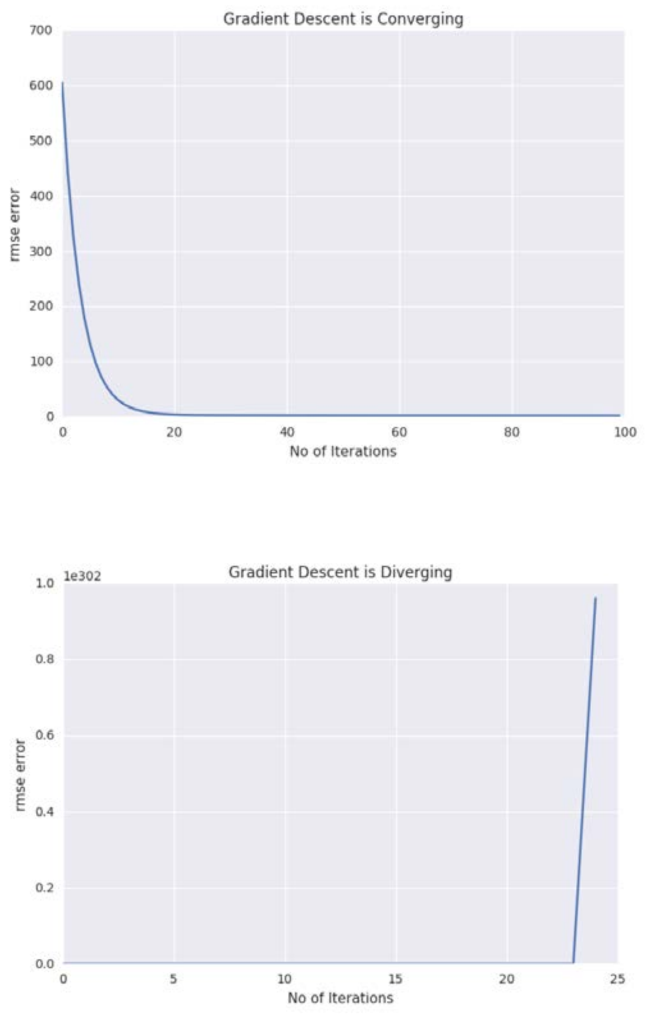
$$x^* = A^{-1}b = (UDV^T)^{-1}b,$$

$$(UDV^T)^{-1} = V^{-T}\, D^{-1}\, U^{-1}$$

$$x^* = A^{-1}b = VD^{-1}U^Tb$$

LLS with SVD

- LLS with SVD
  - A = U.D.V
  - U and V are orthogonal hence is transpose
  - D is diagonal, so inverse each of them
- Accuracy – Not possible as Rank is low
  - 0.14949 via PCR

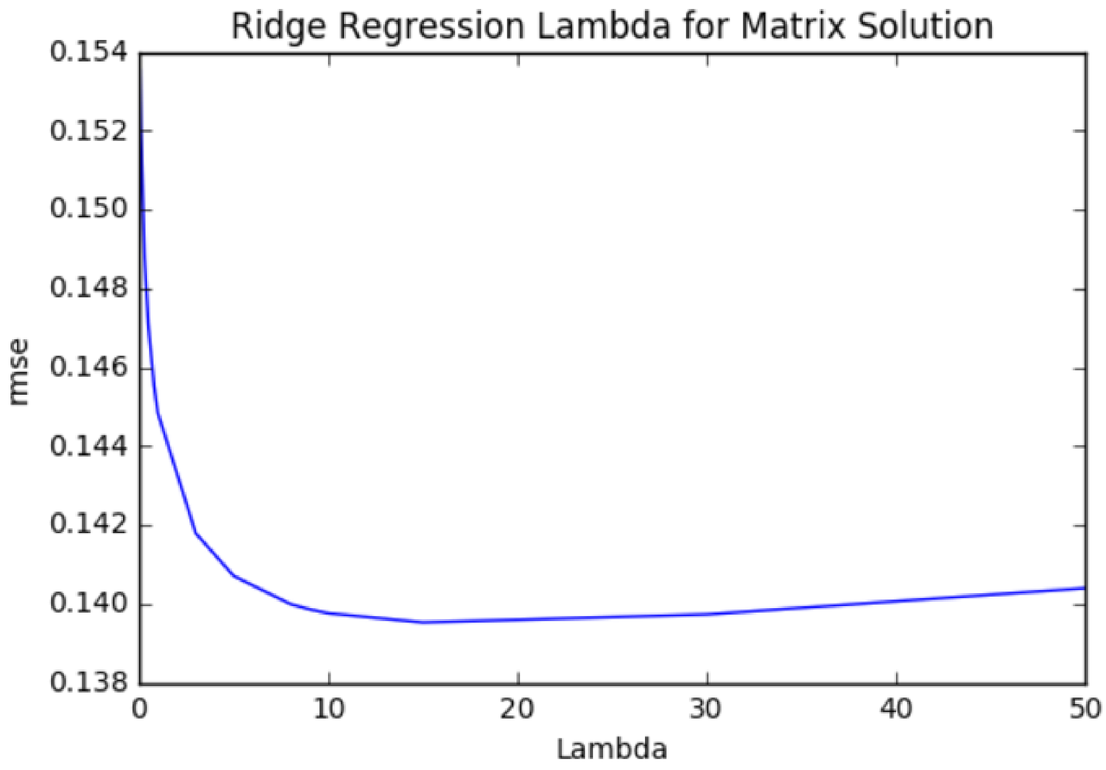


Gradient Descent

- Importance of Normalizing
- Iterations (step size 0.05)
  - 100 – rmse 1.93
  - 5k – rmse 0.48566
  - 10k – rmse 0.38121
  - 100k – rmse 0.18964

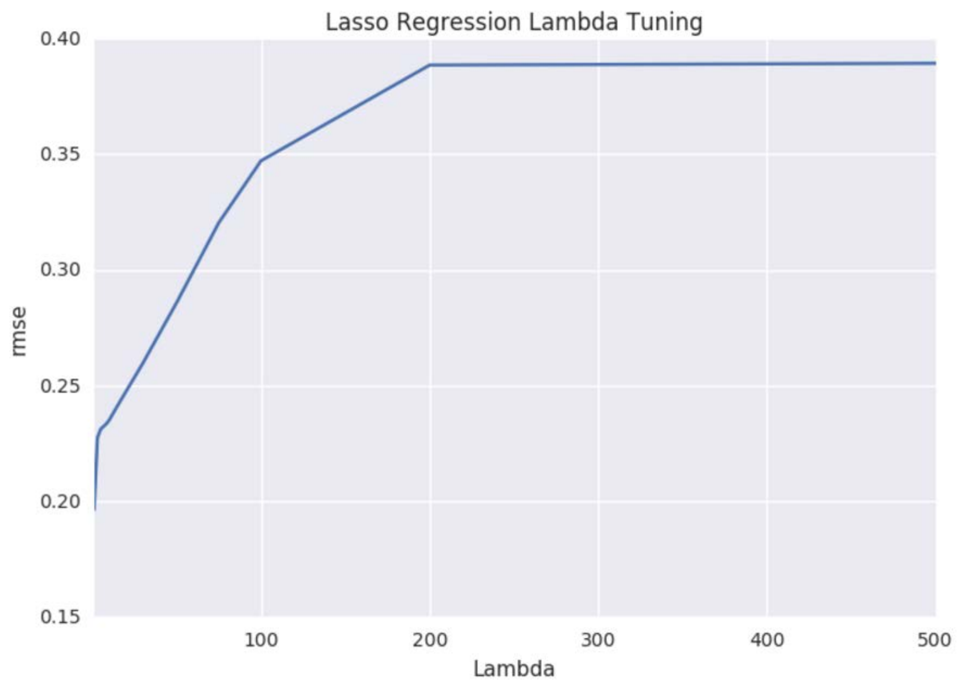
Ridge

- Matrix Inverse
  - Unique inverse Exists
  - Accuracy - 0.1395
- Matrix SVD
  - Accuracy - 0.13927
- SAG
  - Accuracy - 0.230815
- Tuning
  - Alpha = 15



Lasso

- Indicating not lo use Lasso
- Generally used when we have too many features say in the range of 1000
- Accuracy - 0.19638
  - Non zero coefficients - 13
- Tuning
  - Alpha = 1



Shrinkage Intuition

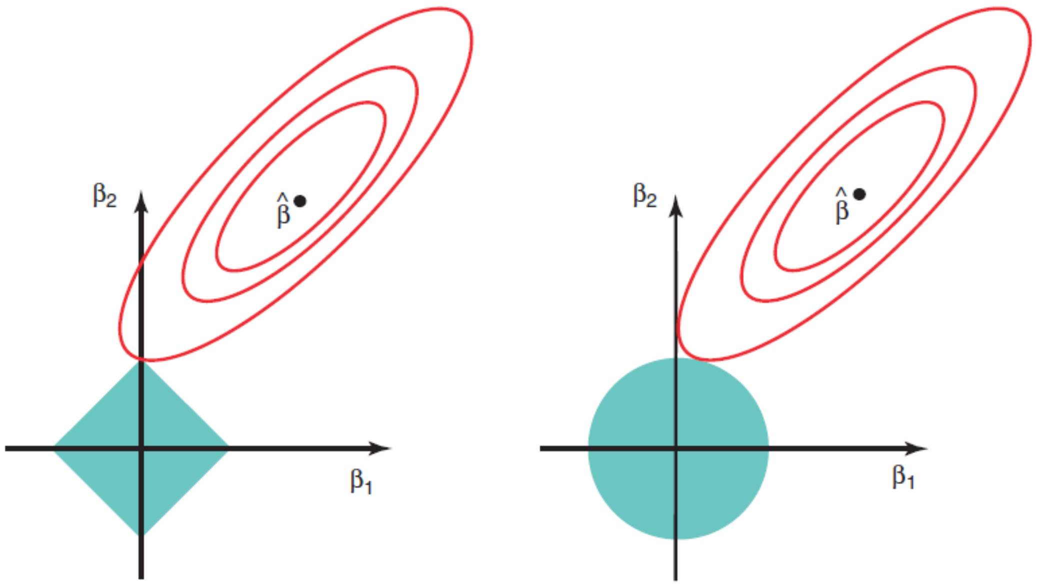
- Ridge

$$J = \sum_{i=1}^{n} (y_i - y_i')^2 + \lambda \sum_{i=1}^{p} (b_i)^2$$

- Lasso

$$J = \sum_{i=1}^{n} (y_i - y_i')^2 + \lambda \sum_{i=1}^{n} |b_i|$$

- Assumption
  - Two parameters
  - Without shrinkage entire plane is available



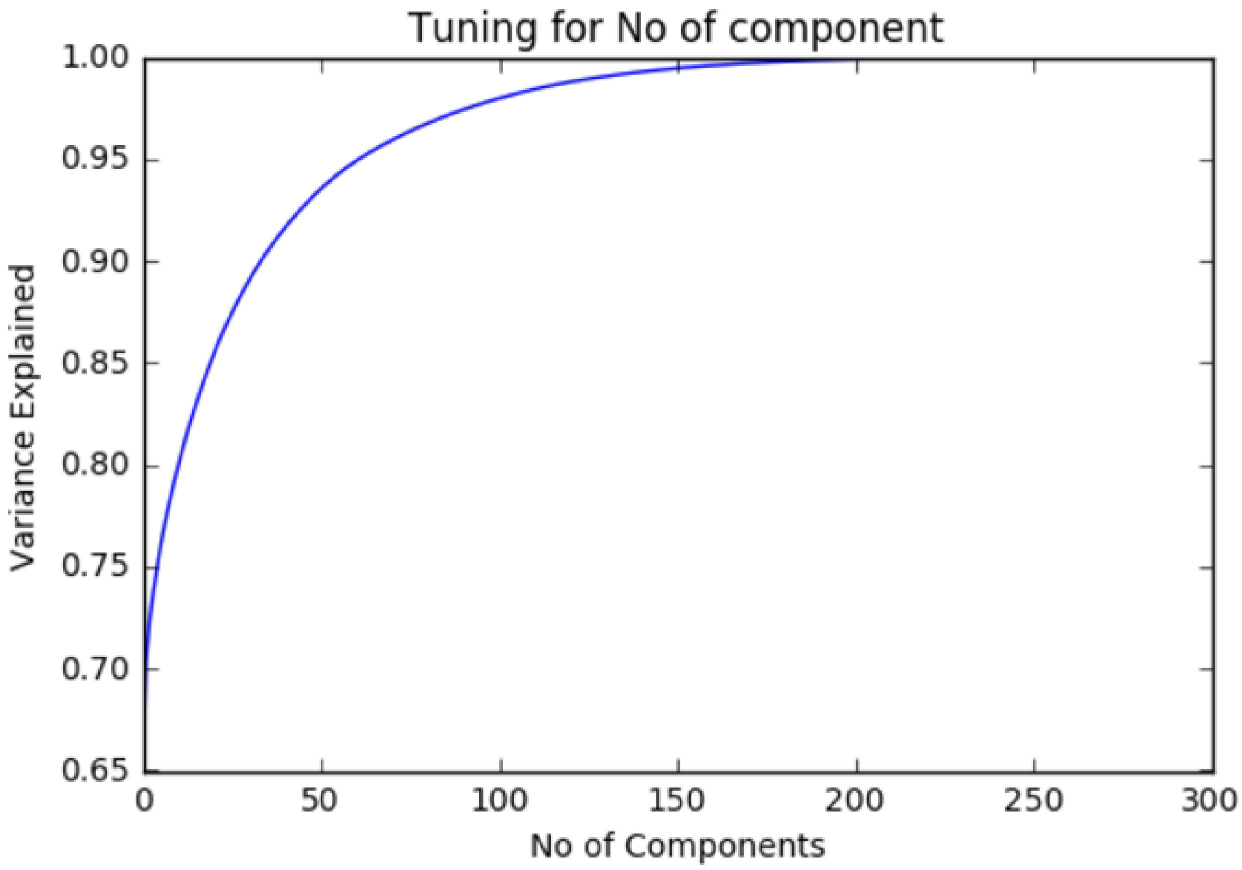
PCR

- SVD
- U reduced
- First 100 component
  - Var explained = 0.9797
  - rmse 0.14949

$$\hat{\mathbf{y}}_{\mathrm{OLS}} = \mathbf{X}\beta_{\mathrm{OLS}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}$$

$$\hat{\mathbf{y}}_{\mathrm{ridge}} = \mathbf{X}\beta_{\mathrm{ridge}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathrm{U}\,\mathrm{diag}\left\{\frac{s_i^2}{s_i^2+\lambda}\right\}\mathrm{U}^\top\mathbf{y}$$

$$\hat{\mathrm{y}}_{\mathrm{PCR}} = \mathrm{X}_{\mathrm{PCA}}\beta_{\mathrm{PCR}} = \mathrm{U}\,\mathrm{diag}\left\{1,\ldots,1,0,\ldots 0\right\}\mathrm{U}^\top\mathrm{y},$$



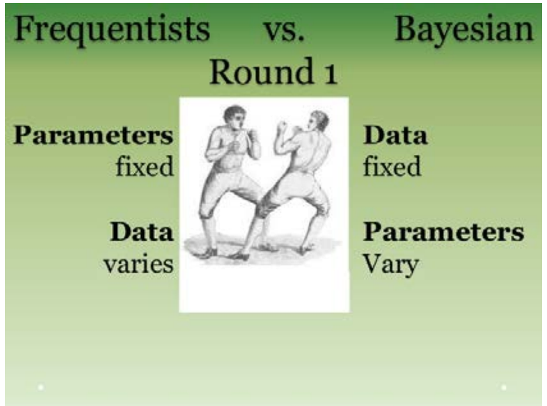
Linear Regression - Summary

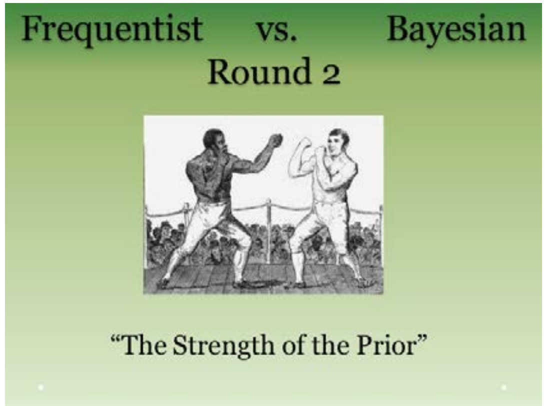
- Simple Least Square
  - LLS with SVD - rmse 0.155965
  - Gradient descent – step size 0.005 (Did not help), diverging
  - After Normalizing
    - Iter 1,00,000 rmse 0.18964
    - Iter 10, 000 rmse 0.38121
    - Iter 5000 rmse 0.48566
- PCR
  - 100 component
  - Rmse 0.14949
- Ridge
  - Matrix - alpha 15, rmse 0.1395
  - Svd instead of finding inverse – 0.13927
  - Sag – 0.230815
- Lasso
  - Alpha 1, rmse 0.19638, non zero 13

Bayesian Regression

Bayes Theorem

- Role of prior
- Confidence Interval vs Credible Interval
- What is fixed ?
    - Data
    - Parameter
- Denominator as complex multiple integration over parameters





Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

Markov Chain Monte Carlo

- Markov Chain
  - Next state depends only on current steps
  - Application
    - Hidden Markov models in speech recognition
    - Allows to draw independent samples
- Monte Carlo
  - Equation is simple mathematical equation
  - Application
    - Uncertainty in market is random, predict sales
    - Solving complex Integration

MCMC Steps – Metropolis Hasting

1. muProposal = norm(muCurrent, StdDev).rvs()
2. likelihoodCurrent = norm(muCurrent, 1).pdf(data).prod()
3. likelihoodProposal = norm(muProposal, 1).pdf(data).prod()
4. priorCurrent = norm(muPriorMu, muPriorSd).pdf(muCurrent)
5. priorProposal = norm(muPriorMu, muPriorSd).pdf(muProposal)
6. probCurrent = likelihoodCurrent * priorCurrent
7. probProposal = likelihoodProposal * priorProposal
8. probAccept = probProposal / probCurrent
9. accept = np.random.rand() < probAccept
10. if (accept == True) then muCurrent = muProposal

Pymc3 style

1. Model

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n + \epsilon$$

2. Describe

```
from pymc3 import Model, Normal, HalfNormal

regression_model = Model()
with regression_model:
    # priors for unknown model parameters
    alpha = Normal('alpha', mu=2, sd=10)
    beta = Normal('beta', mu=0, sd=15, shape=X.shape[1])
    sigma = HalfNormal('sigma', sd=10)

    # expected value of outcome
    ll = [beta[q]*X[:,q] for q in range(0, X.shape[1])]
    mu = alpha + sum(ll)

    # likelihood (sampling distribution) of observations
    y_obs = Normal('y_obs', mu=mu, sd=sigma, observed=y)
```

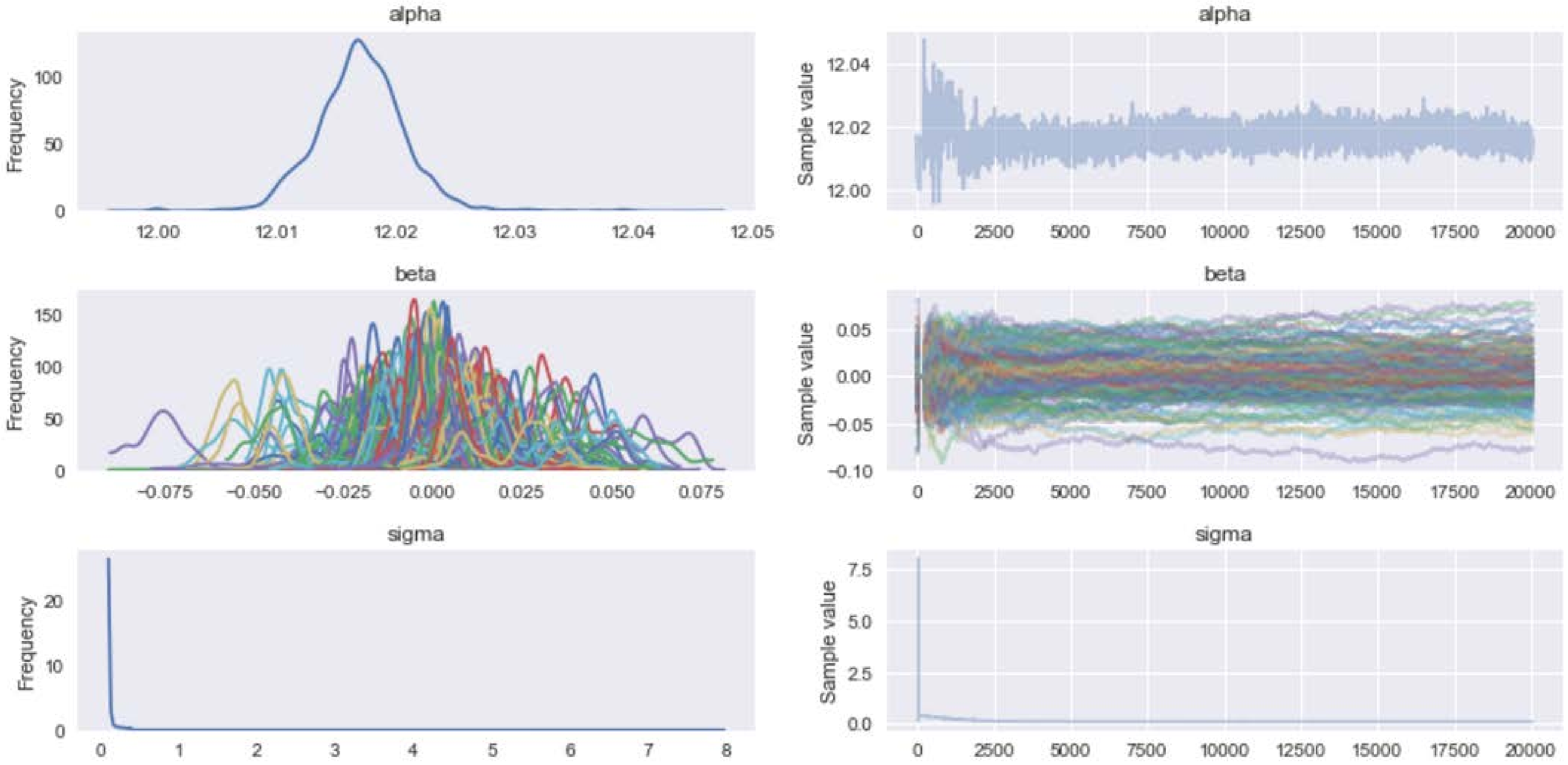
3. Simulate

```
from pymc3 import Metropolis, sample, find_MAP
from scipy import optimize

with regression_model:
    #map_estimate = find_MAP()
    step = Metropolis()
    db = backends.Text('trace')
    print ("Sampling now")
    trace = sample(20000, step, trace=db)
```

4. You get the trace

Trace



alpha
Frequency
100
50
0
12.00
12.01
12.02
12.03
12.04
12.05
alpha
Sample value
12.04
12.02
12.00
0
2500
5000
7500
10000
12500
15000
17500
20000
beta
Frequency
150
100
50
0
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
beta
Sample value
0.05
0.00
−0.05
−0.10
sigma
Frequency
20
10
0
1
2
3
4
5
6
7
8
sigma
Sample value
7.5
5.0
2.5
0.0

Ridge

$$p(B|D) = \frac{p(D|B)p(B)}{p(D)}$$

$$\propto p(D|B) \times p(B)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - B^T x_i)^2}{2\sigma^2}} \times \prod_{i=1}^{p} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(B)^2}{2\tau^2}}$$

$$B = \arg\max_B \; e^{-\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2}$$

$$= \arg\max_B -\frac{1}{2\sigma^2}\sum_1^n (y_i - \hat{y}_i)^2 - \frac{1}{2\tau^2}\sum_1^p \theta^2$$

$$= \arg\max_B -1\left(\sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2\right)$$

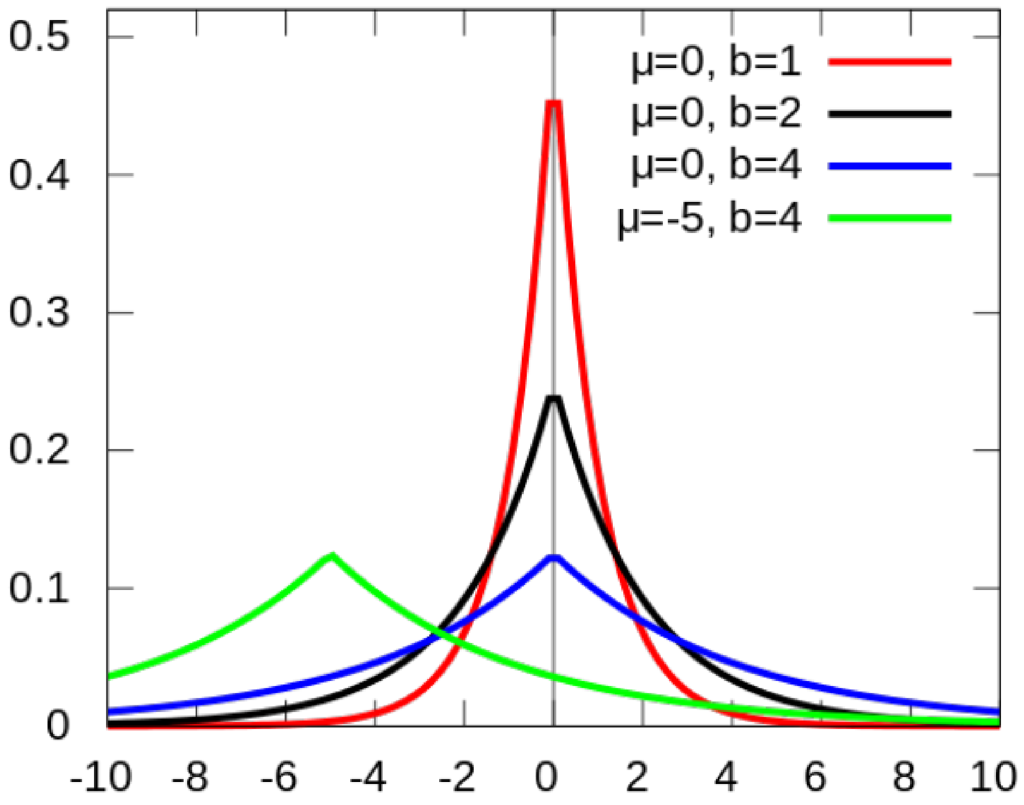
$$= \arg\min_B \sum_1^n (y_i - \hat{y}_i)^2 + \frac{\sigma^2}{\tau^2}\sum_1^p \theta^2$$

- Accuracy = 0.14976

Lasso

- Change Prior of parameter to strictly closer to zero distribution
- Laplace Distribution
- Accuracy : 0.11773

$$f(x \mid \alpha, \beta) = \frac{1}{2b} \exp\left\{-\frac{|x-\mu|}{b}\right\}$$



Custom Likelihood

```
class MyLogLogit(pm.Continuous):
    def __init__(self, a, b, *args, **kwargs):
        super(MyLogLogit, self).__init__(*args, **kwargs)
        self.a = tt.as_tensor_variable(a)
        self.b = tt.as_tensor_variable(b)
        self.median = tt.as_tensor_variable(a)
    def logp(self, x):
        a, b = self.a, self.b
        one = tt.log(b/a)
        two = (b - 1)*tt.log(x/a)
        three = 1. + (x/a)**b
        four = -2.*tt.log(three)
        ans = one + two + four
        return ans
```

$$p(x|a,b) = \frac{(b/a)(x/a)^{b-1}}{(1+(x/a)^b)^2}$$

$$logp = log(b/a) + (b-1)log(x/a) - 2log(1+(x/a)^b)$$

- We can extend base class and supply our own likelihood function
- We used Log Logistic Distribution

Thank you !

- BITS Wilp
  - Two years of amazing learning
- Adobe for work life balance
- My mentor Anirban Basu
  - Mathematics will carry along, subjective knowledge will not
- Teachers, professors and seniors
  - I think I am luck enough
- Parents and sisters
  - I think I am blessed

Questions and Suggestions !