Applied Data Science
fall 2017
Session 9: Bayesian inference. Linear regression revisited. Regularization - Ridge and Lasso

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

Probability: Frequentist vs Bayesian

- Frequentist: P(E) is a frequency of E

P(C=heads)=51/100=0.51

- Bayesian: probability P(E) is our degree of confidence that an event E will happen

P(C=heads)=0.5

www.3dartistonline.com

Limitations of frequentist thinking

- Lack of actual observations
- Lack of scalable framework incorporating beliefs and observations
- Non-intuitive interpretation of inference and hypothesis testing

Bayesian thinking

Being certain about uncertainty

"I can live with doubt and uncertainty... I have approximate answers and possible beliefs and different degrees of certainty about different things, and I'm not absolutely sure of anything..."

Richard Feynman

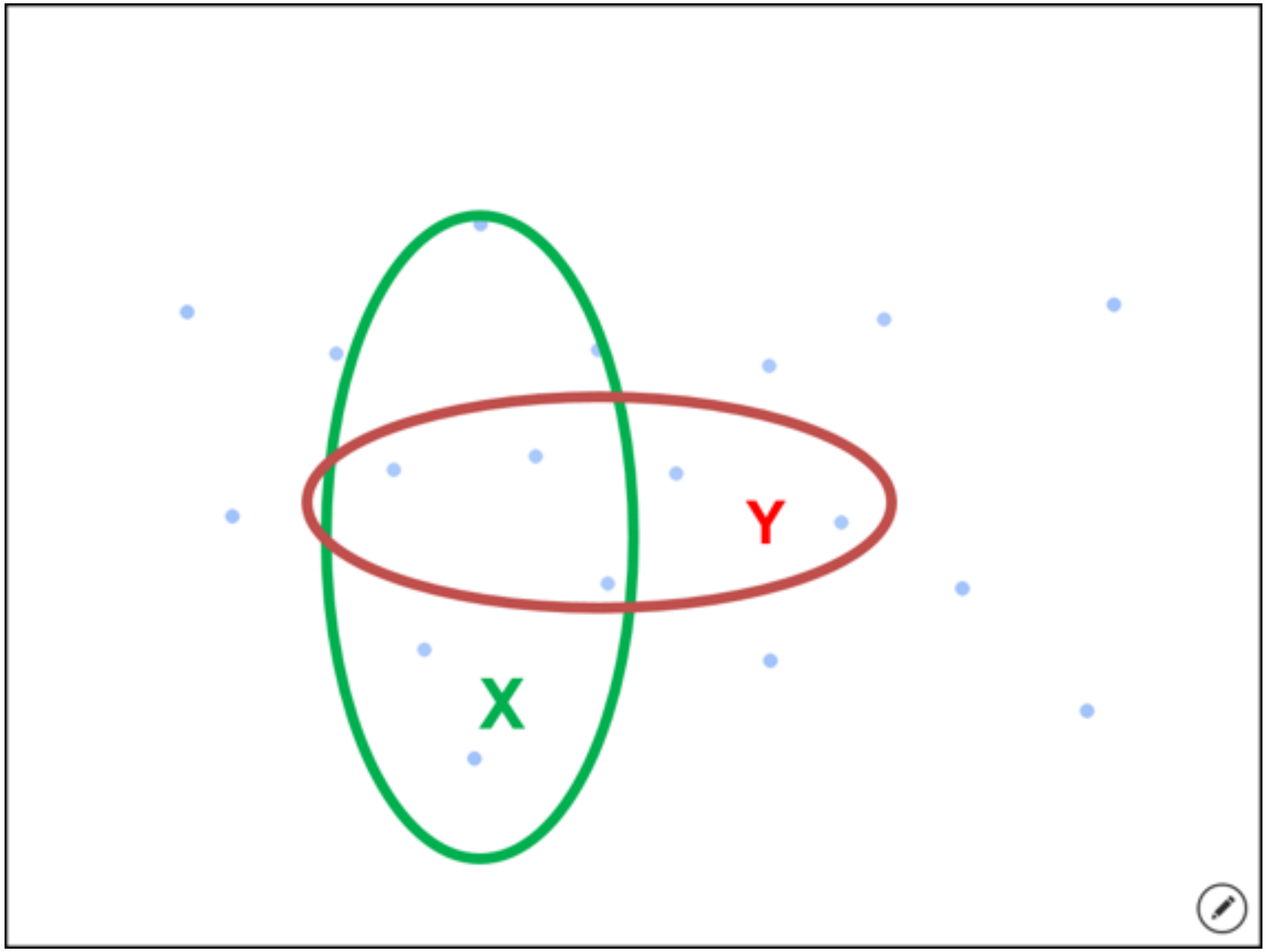
"Now Bayesian statistics are rippling through everything from physics to cancer research, ecology to psychology. Enthusiasts say they are allowing scientists to solve problems that would have been considered impossible just 20 years ago. And lately, they have been thrust into an intense debate over the reliability of research results."
"The essence of the frequentist technique is to apply probability to data. If you suspect your friend has a weighted [unfair] coin, for example, and you observe that it came up heads nine times out of 10, a frequentist would calculate the probability of getting such a result with an unweighted [fair] coin. The answer (about 1 percent) is not a direct measure of the probability that the coin is weighted [unfair]; it's a measure of how improbable the nine-in-10 result is — a piece of information that can be useful in investigating your suspicion."
"By contrast, Bayesian calculations go straight for the probability of the hypothesis, factoring in not just the data from the coin-toss experiment but any other relevant information — including whether you have previously seen your friend use a weighted [unfair] coin."

The New York Times, Sept. 29, 2014.
http://www.nytimes.com/2014/09/30/science/the-odds-continually-updated.html

Demand

- Wide-spread of data - analysis, statistics
- Variety of problems- intuitive unified framework
- Sufficient computation power

Conditional probability



$$P(X|Y)$$

$$P(X \cap Y) = P(X|Y)P(Y)$$

Conditional probability - Bayes theorem

$$P(X \cap Y) = P(X|Y)P(Y)$$

$$P(X \cap Y) = P(Y|X)P(X)$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

The core - Bayes theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Thomas Bayes 1740s Initial Belief + New Data -> Improved Belief

Pierre Simon Laplace 1774. the probability of a cause (given an event) is proportional to the probability of the event

1891, the Scottish mathematician George Chrystal urged: "[Laplace's principle] being dead” - declining subjectivity

Captain Dreyfus used it to demonstrate his innocence

Alan Turing used it to decode the German Enigma cipher

Example

Total of 60% of students were preparing for midterm

40% of those preparing got "A"

Only 15% of those not preparing got "A"

a) Consider random student. What is your confidence student was preparing? $P(prep)=60\%$

b) If now you learned the selected student got "A". How does it affect your confidence?

$$P(prep|A) = P(A|prep)P(prep)/P(A) = 0.4 \times 0.6/P(A)$$

$$P(!prep|A) = P(A|!prep)P(!prep)/P(A) = 0.15 \times 0.4/P(A)$$

$$P(prep|A) = 0.24/(0.24+0.06) = 4/5 = 80\%$$

Conceptual scheme

1. Express **prior** beliefs about the unknown parameters of interest

   P(prep)=60%

2. Collect data and evaluate its **likelihood** with respect to prior beliefs

   P(A|prep)=40%

3. Update your beliefs - **posterior**

$$P(prep|A) \sim P(A|prep)P(prep)$$

posterior~likelihood X prior

Prior beliefs

Important to consider once they exist

But where to get them?

- Theoretical understanding of the subject
- Previous observations
- Guess
- Possibility to model lack of knowledge

Bayesian inference - formalizm

$y \sim f(\alpha)$ $\quad$ $y \sim f(x, \alpha)$

$$D = \{(y_i, x_i), i = 1..N\}$$

discrete

$$P(\alpha|D) = \frac{P(D|\alpha)P(\alpha)}{P(D)}$$

$$P(D) = \sum_{\beta} P(D|\beta)P(\beta)$$

continous

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

$$p(D) = \int p(D|\beta)p(\beta)d\beta$$

$$p(\alpha|D) \sim p(D|\alpha)p(\alpha)$$

$$posterior \sim likelihood \ * \ prior$$

Tossing a coin

$$P(c=1)=\alpha \qquad P(c=0)=1-\alpha$$

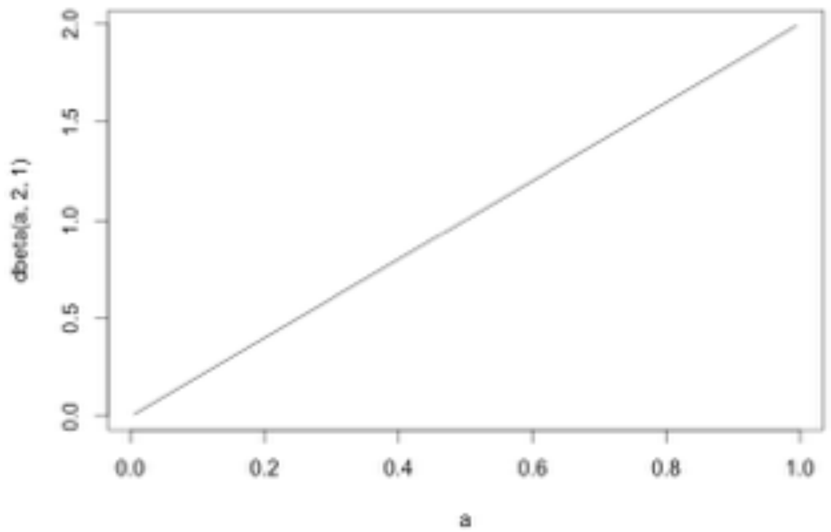
$$p(\alpha) - prior\ belief$$

$$uninformative\ prior:$$

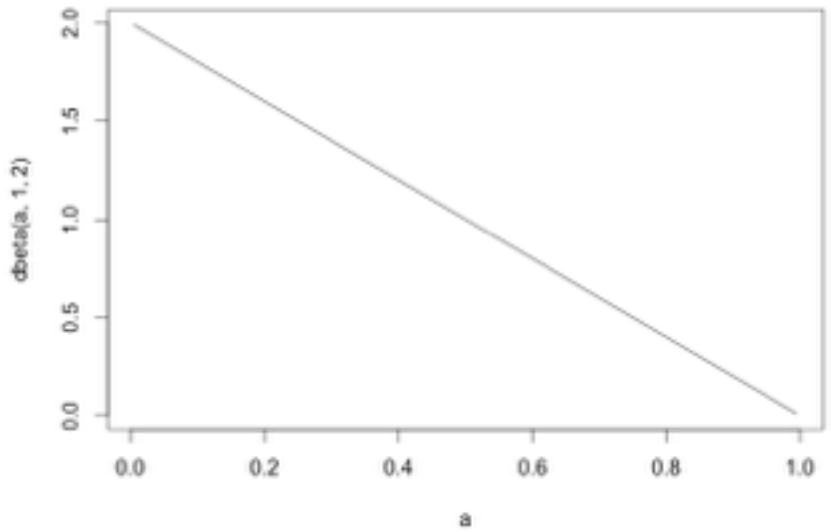
$$\alpha \sim Unif(0,1) \qquad p(\alpha)=1$$

Tossing coin - once

$$p(\alpha|c=1) \sim P(c=1|\alpha)p(\alpha) = \alpha p(\alpha)$$



$$p(\alpha|c=0) \sim P(c=0|\alpha)p(\alpha) = (1-\alpha)p(\alpha)$$

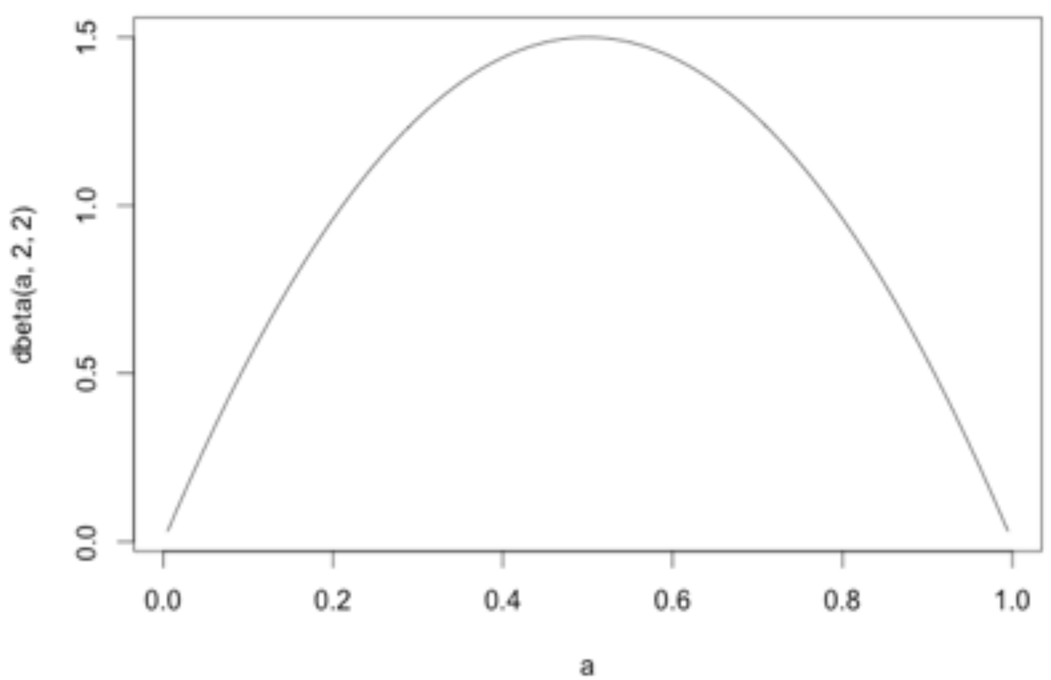


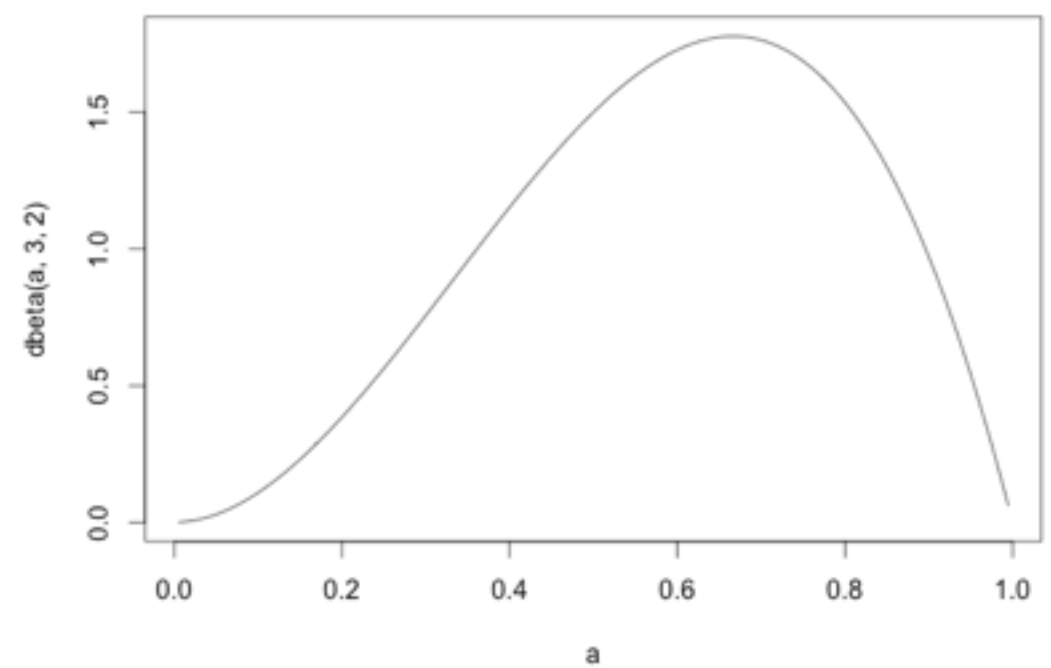
Tossing a coin two times

$p(\alpha) = 1$

I. c=0 $p(\alpha|c=0) = (1-\alpha)$

II. c=1 $p(\alpha|c=1) \sim P(c=1|\alpha)p(\alpha) = \alpha p(\alpha) = \alpha(1-\alpha)$





Tossing coin - multiple times

$$c = y_i, i = 1..N$$

$$p(\alpha|D) \sim \alpha^{\sum y_i}(1-\alpha)^{N-\sum y_i}p(\alpha)$$

$$p(\alpha) \equiv 1$$

$$p(\alpha|D) \sim \alpha^{\sum y_i}(1-\alpha)^{N-\sum y_i}$$

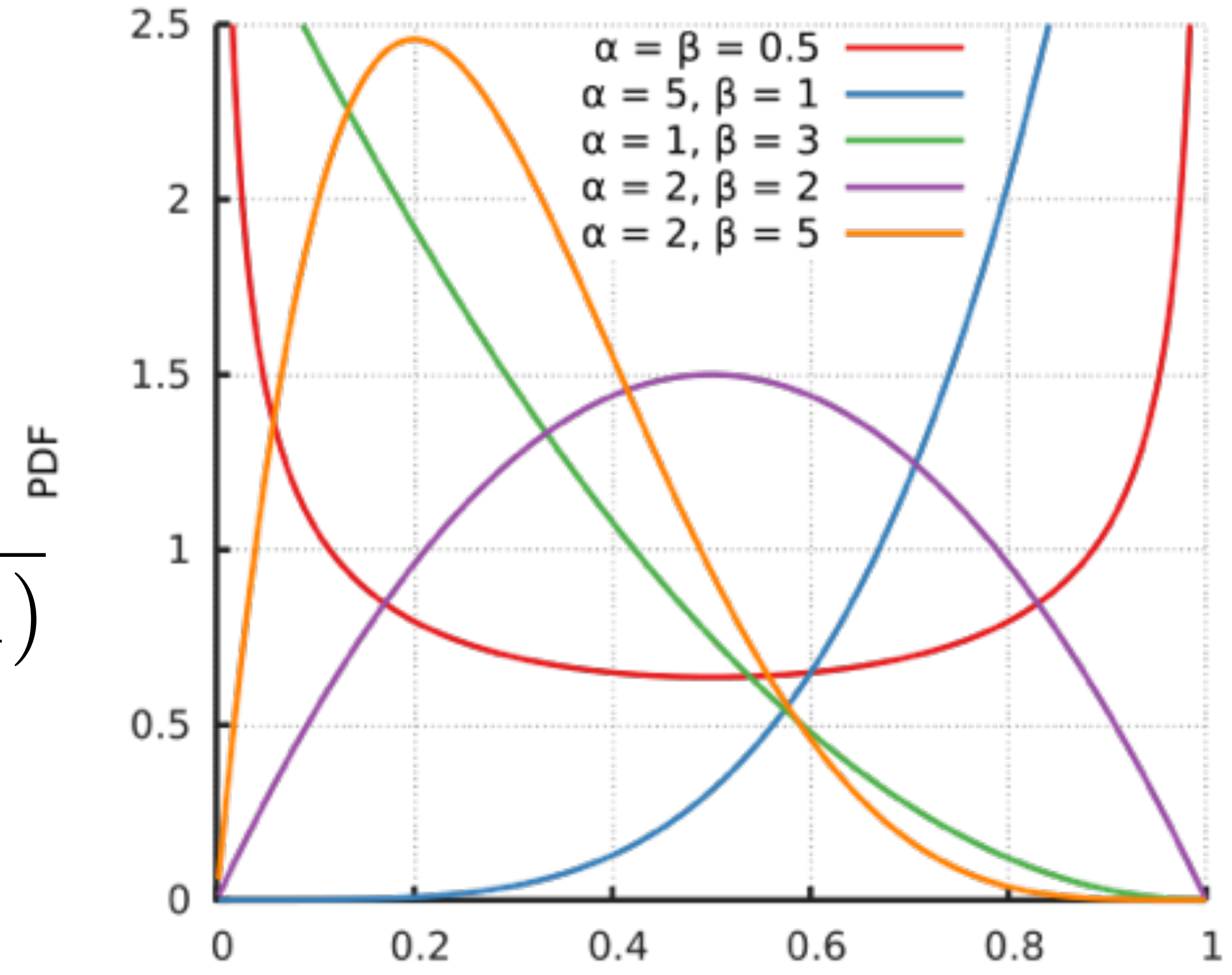
Beta distribution

$$Bpdf(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

$$\mu = \frac{a}{a+b}$$

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

wikipedia.org



Tossing coin - multiple times

$$p(\alpha) \equiv 1 \qquad p(\alpha|D) \sim \alpha^{\sum y_i}(1-\alpha)^{N-\sum y_i}$$

$$\alpha \sim B\left(1+\sum y_i, N+1-\sum y_i\right)$$

$$p(\alpha) = B(h+1, t+1) - prior$$

$$p(\alpha|y) = B(h+h_2+1, t+t_2+1)$$

Beta-distribution is conjugate to Bernoulli experiments

Univariate linear regression

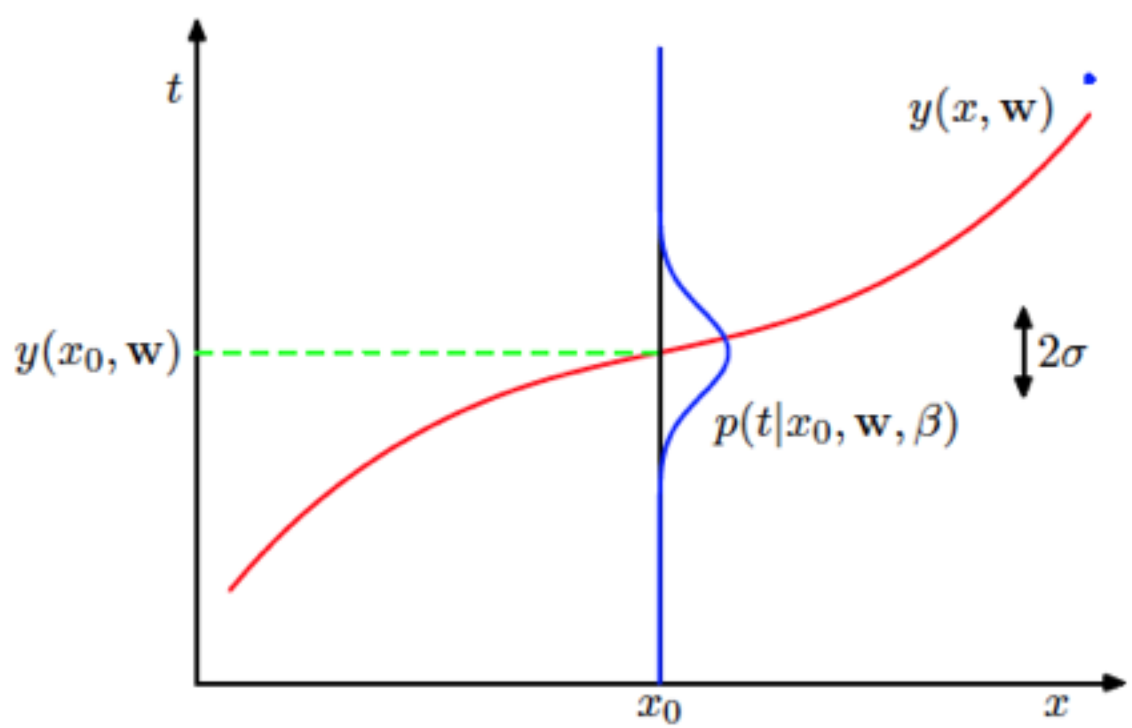
$$y = wx + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$y \sim \mathcal{N}(wx, \sigma^2) \qquad D = (Y, X) = \{(y_i, x_i), i = 1..N\}$$

ordinary least square (OLS) estimate $$\hat{w} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

$$\sigma = \sqrt{\frac{\sum_i (y_i - \hat{w}x_i)^2}{N}}$$

What if we know smth about w?



Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

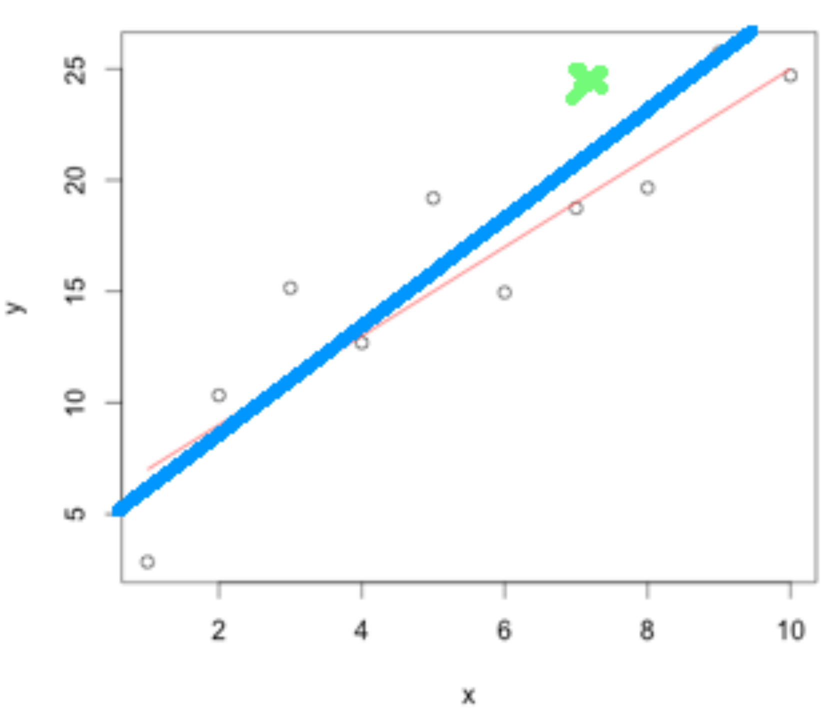
Bayesian approach

$$y \sim \mathcal{N}(wx, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-wx)^2}{2\sigma^2}} \quad D = (Y, X) = \{(y_i, x_i), i = 1..N\}$$

prior knowledge about w $$w \sim \mathcal{N}(w^*, (\sigma^*)^2) = \frac{1}{\sqrt{2\pi\sigma^*}} e^{-\frac{(w-w^*)^2}{2(\sigma^*)^2}}$$

posterior

$$p(w|y = y_i, x = x_i) \sim p(y = y_i|w, x = x_i)p(w) \sim e^{-\frac{(y_i - wx_i)^2}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}} \sim$$



$$\sim e^{-w^2 \frac{x_i^2(\sigma^*)^2 + \sigma^2}{2\sigma^2(\sigma^*)^2} + w \frac{(\sigma^*)^2 y_i x_i + \sigma^2 w^*}{2\sigma^2(\sigma^*)^2}} \sim e^{\frac{\left(w - \frac{\sigma^{-2} y_i x_i + (\sigma^*)^{-2} w^*}{\left(x_i^2 \sigma^{-2} + (\sigma^*)^{-2}\right)}\right)^2}{\frac{2}{(\sigma^*)^{-2} + x_i^2 \sigma^{-2}}}}$$

Bayesian approach - adding all observations at once

$$y \sim \mathcal{N}(wx, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-wx)^2}{2\sigma^2}}$$

$$D = (Y, X) = \{(y_i, x_i), i = 1..N\}$$

prior knowledge about w $\quad w \sim \mathcal{N}(w^*, \sigma^*)$

posterior

$$p(w|Y,X) \sim p(Y|w,X)p(w) = p(w)\prod_i p(y = y_i|w, x = x_i) \sim$$

$$\sim e^{-\sum_i \frac{(y_i - wx_i)^2}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}} = e^{-\frac{SSE(w)}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}}$$

$$\hat{w} = argmax_w p(w|Y,X) = argmin_w \left[\frac{RSS(w)}{2\sigma^2} + \frac{(w-w^*)^2}{2(\sigma^*)^2}\right]$$

What is the posterior after all?

posterior

$$p(w|Y,X) \sim p(Y|w,X)p(w) = p(w)\prod_i p(y=y_i|w,x=x_i) \sim$$

$$\sim e^{-\sum_i \frac{(y_i - wx_i)^2}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}} = e^{-\frac{RSS(w)}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}}$$

$$p(w|Y,X) \sim e^{-w^2\frac{\sum_i x_i^2(\sigma^*)^2+\sigma^2}{2\sigma^2(\sigma^*)^2} + w\frac{(\sigma^*)^2\sum_i y_i x_i + \sigma^2 w^*}{2\sigma^2(\sigma^*)^2}} \sim e^{\frac{\left(w - \frac{\sigma^{-2}\sum_i y_i x_i + (\sigma^*)^{-2} w^*}{\left(\sum_i x_i^2\sigma^{-2} + (\sigma^*)^{-2}\right)}\right)^2}{\frac{2}{(\sigma^*)^{-2}+\sum_i x_i^2\sigma^{-2}}}} \sim$$

$$\sim \mathcal{N}\left(\frac{\sigma^{-2}\sum_i y_i x_i + (\sigma^*)^{-2} w^*}{\left(\sum_i x_i^2\sigma^{-2} + (\sigma^*)^{-2}\right)}, \frac{1}{\sqrt{(\sigma^*)^{-2} + \sum_i x_i^2\sigma^{-2}}}\right)$$

Bayesian approach - uninformed prior

unfinformed prior with $\sigma^* \to \infty$

$$p(w|Y,X) \sim e^{-\sum_i \frac{(y_i - wx_i)^2}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}} = e^{-\frac{RSS(w)}{2\sigma^2} - \frac{(w-w^*)^2}{2(\sigma^*)^2}}$$

$$RSS(w) \to \min$$

$$w \sim \mathcal{N}\left(\frac{\sigma^{-2}\sum_i y_i x_i + (\sigma^*)^{-2} w^*}{\left(\sum_i x_i^2 \sigma^{-2} + (\sigma^*)^{-2}\right)}, \frac{1}{\sqrt{(\sigma^*)^{-2} + \sum_i x_i^2 \sigma^{-2}}}\right)$$

$$w \sim \mathcal{N}\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}, \frac{\sigma}{\sqrt{\sum_i x_i^2}}\right) \qquad \hat{\sigma}^2 = \frac{RSS(\hat{w})}{N}$$

Multivariate case

$$y \sim \mathcal{N}(w^T x, \sigma^2) \qquad w = (w_1, w_2, \ldots w_n)$$

$$D = (Y, X) = \{(y_i, x_i), i = 1..N\}$$

prior knowledge about w $\qquad w_j \sim \mathcal{N}(w_j^*, \sigma_j^*)$

posterior

$$p(w|Y,X) \sim p(Y|w,X)p(w) = \prod_i p(y = y_i|w, x = x_i)p(w) = \prod_i p(y = y_i|w, x = x_i) \prod_j p(w_j) \sim$$

$$\sim \prod_i e^{\frac{-(y_i - w^T x_i)^2}{2\sigma^2}} \prod_j e^{-\frac{(w_j - w_j^*)^2}{2(\sigma_i^*)^2}} = e^{-\frac{RSS(w)}{2\sigma^2} - \sum_j \frac{(w_j - w_j^*)^2}{2(\sigma_i^*)^2}}$$

$$\hat{w} = argmax_w p(w|Y,X) = argmin_w \left[\frac{RSS(w)}{\sigma^2} + \sum_j \frac{(w_j - w_j^*)^2}{(\sigma_j^*)^2}\right]$$

Overfitting

Area	Month	Price
1010	1	90000
2000	2	200000
2990	3	310000

`Price=-1000*Area+1.100.000*Month`

Area	Month	Price
1500	12	150000

`Price=-1000*1500+`
`1.1M*12=11.7M?`

Ridge and Lasso regression

Require coefficients to be not too big

$$\hat{w} = argmax_w p(w|Y, X) = argmin_w \left[\frac{RSS(w)}{\sigma^2} + \sum_j \frac{(w_j - w_j^*)^2}{(\sigma_j^*)^2} \right]$$

$$w_j \sim \mathcal{N}(0, \sigma/\sqrt{\lambda})$$

Ridge $\hat{w} = argmin_w \left[RSS(w) + \lambda ||w||_2^2\right]$ $||w||_2 = \sqrt{\sum_j w_j^2}$

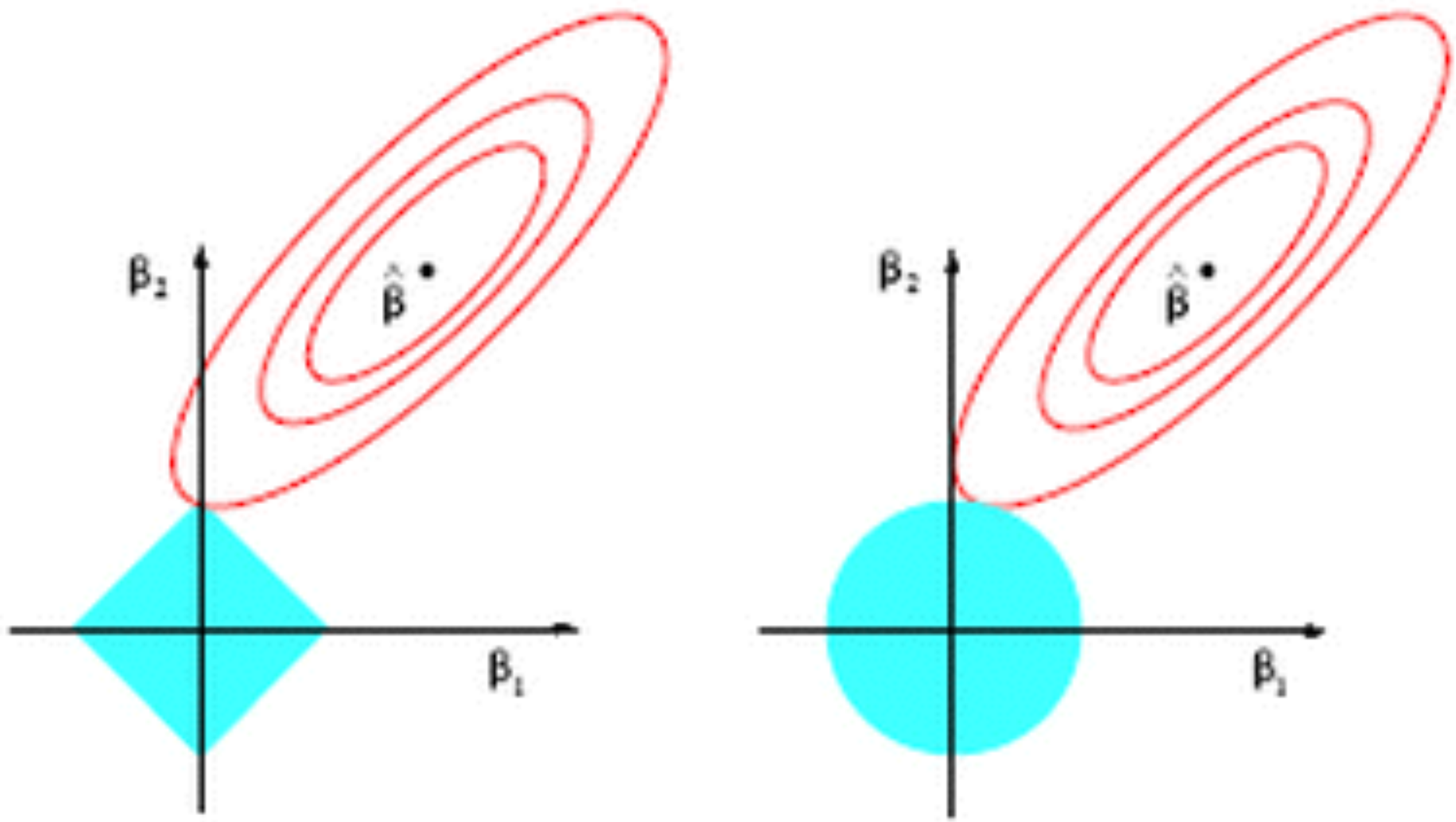
Laplacian prior distribution $p(w_j) \sim e^{-\lambda |w_j| / \sigma}$

Lasso $\hat{w} = argmin_w \left[RSS(w) + \lambda ||w||_1\right]$ $||w||_1 = \sum_j |w_j|$

$$RSS(w) \to min, \ ||w||_p \leq \alpha$$

Ridge regression admits solution in the closed form $\hat{w} = (X^T X + \lambda I)^{-1} X^T Y$

Ridge and Lasso regression



http://jamesmccammon.com/

Fighting overfitting

Area	Month	Price
1010	1	90000
2000	2	200000
2990	3	310000

```
Lasso (lambda=1): Price=88.34*Area+13212.5*Month
Lasso (lambda=50): Price=98.77*Area+2787.1*Month
Lasso (lambda=100): Price=101.56*Area+0*Month
```

Area	Month	Price
1500	12	150000

Criticism

- Dependence on the subjective prior
- Where do we get the prior?
- Computational complexity
- Universal? No free lunch!