Overview

- Probabilistic Interpretation of Linear Regression
- Maximum Likelihood Estimation
- Bayesian Estimation
- MAP Estimation

Probabilistic Interpretation: Linear Regression

- Assume output y is generated by:

$$Y = \theta^T x + M = h_\theta(x) + M$$

  where $h_\theta(x) = \theta^T x$ and M is Gaussian noise with mean 0 and variance 1. As usual, we use capitals for random variables.

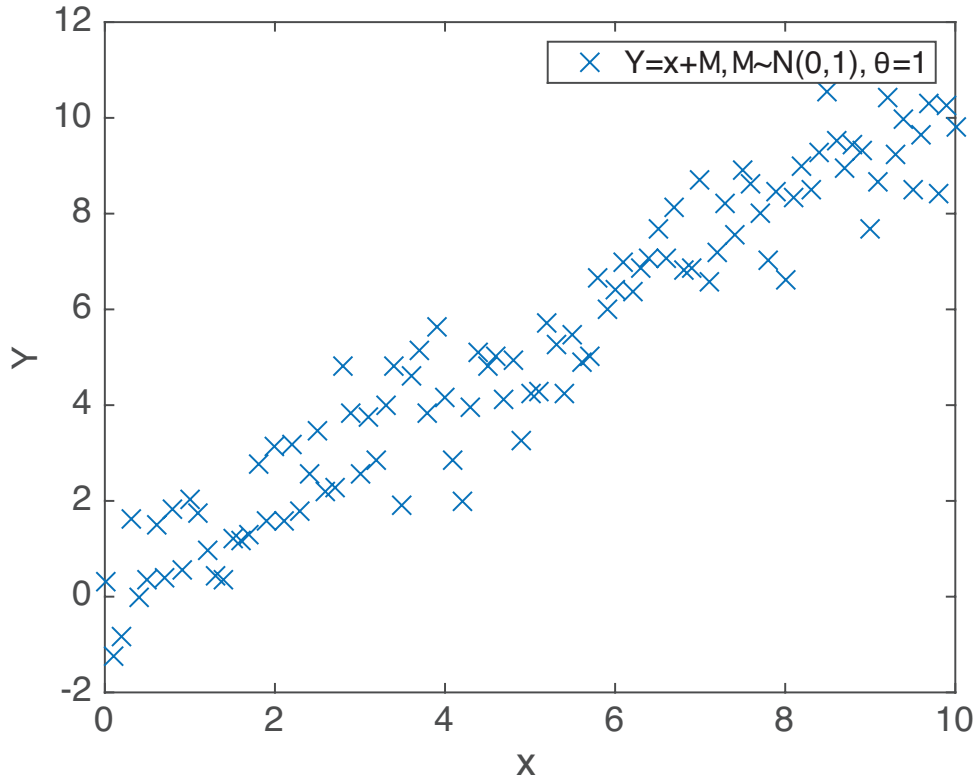
- So training data d is:

$$\{(x^{(1)}, h_\theta(x^{(1)}) + M^{(1)}), (x^{(2)}, h_\theta(x^{(2)}) + M^{(2)}), \cdots, (x^{(m)}, h_\theta(x^{(m)}) + M^{(m)})\}$$

  where $M^{(1)}, M^{(2)}, \ldots, M^{(m)}$ are **independent** random variables each of which is Gaussian with mean 0 and variance 1.

Example

- Training data generated using Matlab code: *x=[0:0.1:10]; y=x+randn(1,length(x)); plot(x,y,'+')*
- In this example θ has just one element ($m = 1$) with value $\theta_1 = 1$. In practice we don't know θ_1 in advance, and given the training data we want to estimate its value.



Probabilistic Interpretation: Linear Regression

- A Gaussian RV Z with mean μ and variance σ^2 has pdf $f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z-\mu)^2}{2\sigma^2}}$.
- So we are assuming: $f_M(m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{m^2}{2}}$, $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-h_\theta(x))^2}{2}}$.
- The **likelihood** $f_{D|\Theta}(d|\theta)$ of the training data d is therefore:

$$f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}}$$

$$= \frac{1}{\sqrt{2\pi}^m} e^{-\sum_{i=1}^{m} \frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}}$$

- Taking logs: $\log f_{D|\Theta}(d|\theta) = \log \frac{1}{\sqrt{2\pi}^m} - \sum_{i=1}^{m} \frac{(y^{(i)}-h_\theta(x^{(i)}))^2}{2}$

Parameter Estimation

- Recall Bayes Rule for PDFs

$$f_{\Theta|D}(\vec{\theta}|d) = \frac{f_{D|\Theta}(d|\vec{\theta}) f_{\Theta}(\vec{\theta})}{f_D(d)}$$

posterior *likelihood* *prior*

- The **maximum likelihood estimate** of θ maximises the likelihood. Equivalently, it maximises the **log-likelihood** (why ?). We can also drop the fixed scaling factor $\frac{1}{\sqrt{2\pi}}$ (why ?).
- In linear regression case we select θ to maximise:

$$-\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))^2$$

i.e. minimise

$$\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))^2$$

Probabilistic Interpretation: Who Cares ?

- Since probability is about reasoning under uncertainty it would be v odd indeed if our machine learning algorithms did not make good sense from a probability/statistics point of view.
- Casting an ML approach within a statistical framework clarifies the assumptions that have been made (perhaps implicitly). E.g. in linear regression:
  - Noise is additive $Y = \theta^T x + M$
  - Noise on each observation is independent and identically distributed
  - Noise is Gaussian – it is this which leads directly to the use of a square loss $(y - h_\theta(x))^2$. Changing the noise model would lead to a different loss function.

Closed-form solution

It turns out we can find the least square solution in closed form (not just by using gradient descent). Let's work through example with linear model and only one input ($m = 1$).

- Select θ to maximise $\log L(\theta) = -\frac{1}{2}\sum_{j=1}^{n}(y_j - \theta x_j)^2$
- Compute derivative with respect to θ:

$$\frac{dL}{d\theta} = \sum_{j=1}^{n}(y_j - \theta x_j)x_j = \sum_{j=1}^{n} y_j x_j - \theta \sum_{j=1}^{n} x_j^2$$

- Set derivative equal to 0 and solve for θ:

$$\sum_{j=1}^{n} y_j x_j - \theta \sum_{j=1}^{n} x_j^2 = 0$$

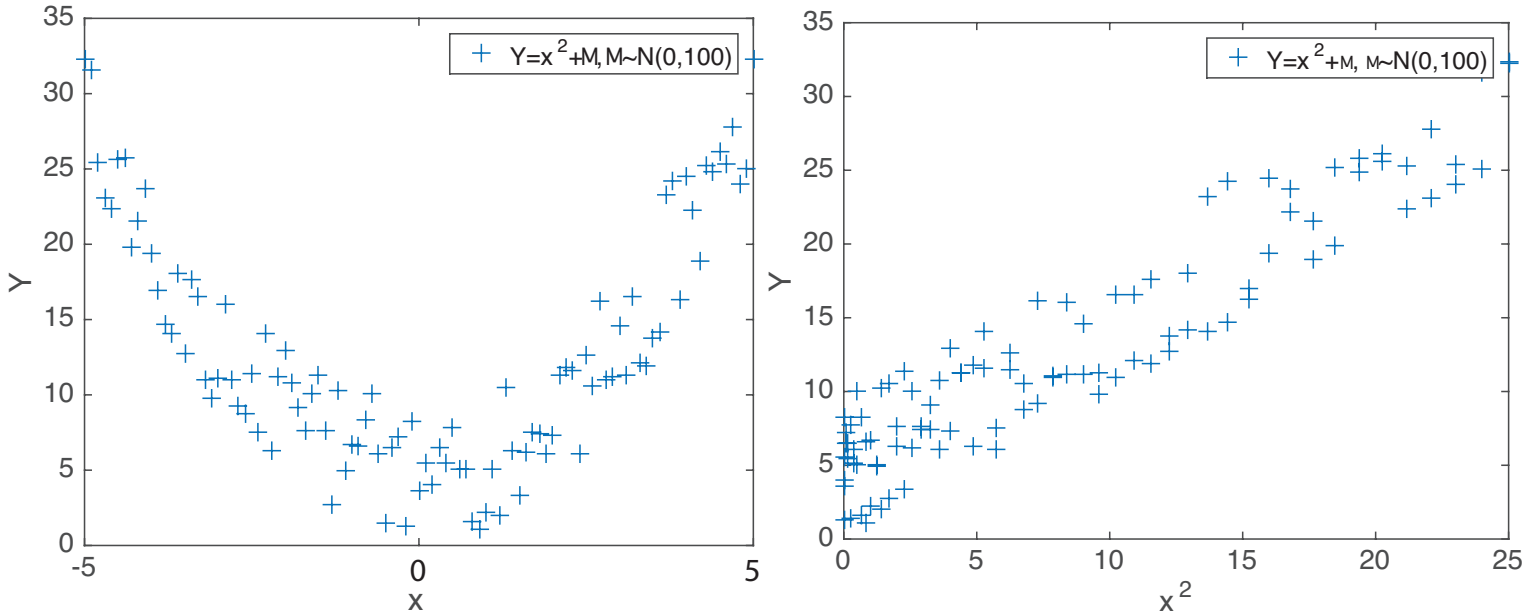
$$\Rightarrow \qquad \theta = \frac{\sum_{j=1}^{n} y_j x_j}{\sum_{j=1}^{n} x_j^2}$$

- Can extend to multiple inputs, but we won't do it in this course.

Fitting nonlinear curves: choosing features

Example:

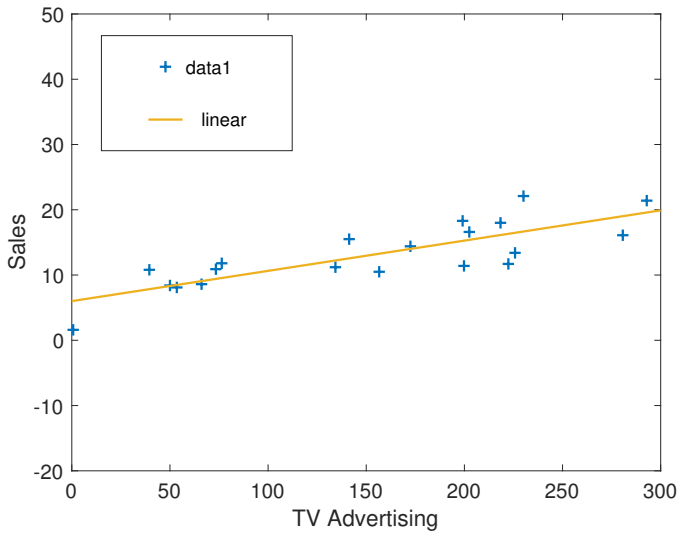
- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x^2$
- Define feature vector $\vec{z}$ with $z = x^2$. This new z can be computed given input x, so its known.
- Using this new feature vector our hypothesis can be rewritten as $h_\theta(z) = \theta_0 + \theta_1 z$, so can directly apply all the ideas we've just discussed.

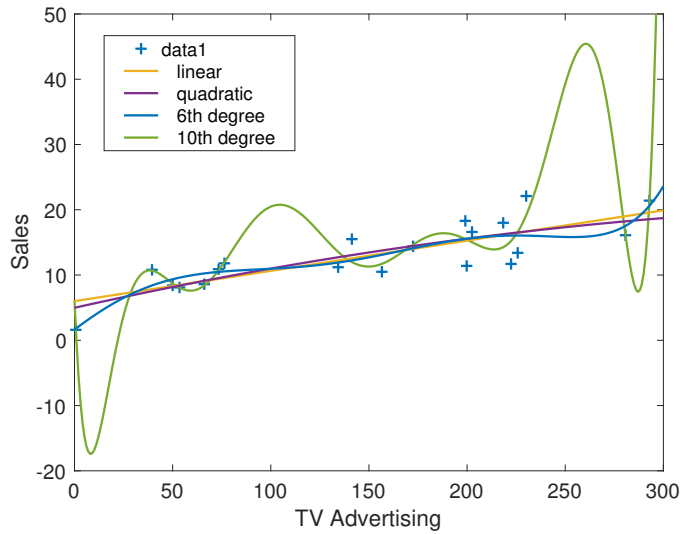


Regularisation & Model Selection

Advertising example again. Thin out data by taking every 10th point. Try a few different hypothesis:

- $h_\theta(x) = \theta_0 + \theta_1 x$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_6 x^6$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_{10} x^{10}$

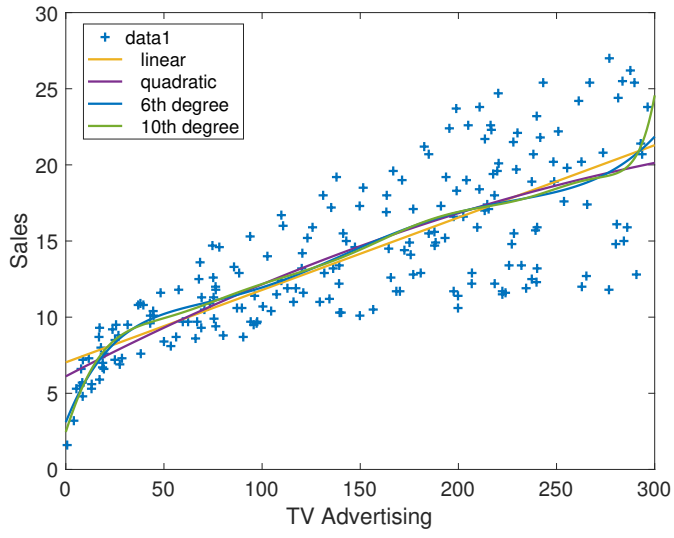


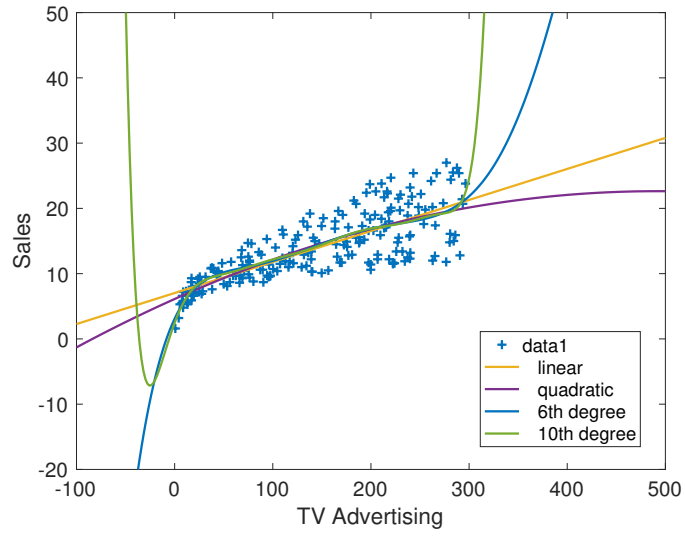


- As we add more parameters, we start to fit the "noise" in the training data, called **overfitting**.
- But of use too few parameters then will get a poor fit, **underfitting**.
- How to strike the right balance between these ? This is an example of the **bias-variance trade-off**.

Regularisation & Model Selection

More data can help, e.g. when don't thin out data:





- But even with more data, still our hypothesis doesn't generalise well i.e. doesn't predict well for data outside the training set.

Bayesian Estimation

- Estimate the posterior $f_{\Theta|D}(\theta|d)$, rather than likelihood $f_{D|\theta}(d|\theta)$
- A distribution rather than just a single value.

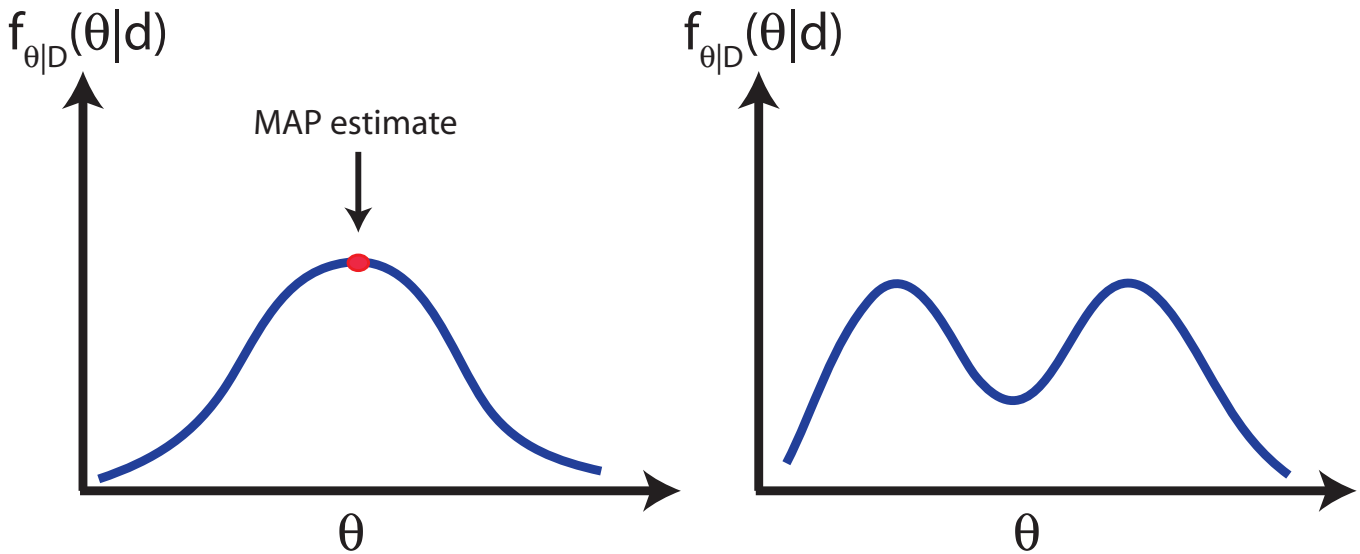
Example:

- $Y = \sum_{i=1}^{m} \Theta_i x_i + M$, $M \sim N(0,1)$, $\Theta_i \sim N(0,\lambda)$.
- Bayes theorem: $f_{\Theta|D}(\theta|d) = \frac{f_{D|\theta}(d|\theta) f_\theta(\theta)}{f_D(d)}$
- We already know that likelihood $f_{D|\Theta}(d|\theta) \propto \exp(-\sum_{j=1}^{n}(y^{(j)} - \sum_{i=1}^{m} \theta_i x_i^{(j)})^2/2)$.
- From our model we have prior $f_{\Theta_i}(\theta) \propto \exp(-\theta^2/2\lambda)$.
- $f_D(d)$ is a normalising constant (so area under PDF $f_{\Theta|D}(\theta|d)$ is 1).
- Combining these using Bayes Rule gives:

$$f_{\Theta|D}(\theta|d) \propto \exp(-\sum_{j=1}^{n}(y^{(j)} - \sum_{i=1}^{m} \theta_i x_i^{(j)})^2/2) \times \exp(-\theta^2/2\lambda)$$

MAP Estimation

- Maximum a posteriori (MAP) estimation[1]
- Select θ that maximises posterior $f_{\Theta|D}(\theta|d)$ (back to a single value rather than a distribution).



- Runs into trouble if distribution has > 1 peak.

[1] Challenge: what does this Latin mean ?

MAP Estimation For Linear Regression: Ridge Regression

Taking logs (the same value of θ maximises $f_{\Theta|D}(\theta|d)$ and $\log f_{\Theta|D}(\theta|d)$, why ?), select θ to maximise

$$-\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 - \frac{1}{\lambda}\sum_{j=1}^{n}\theta_j^2$$

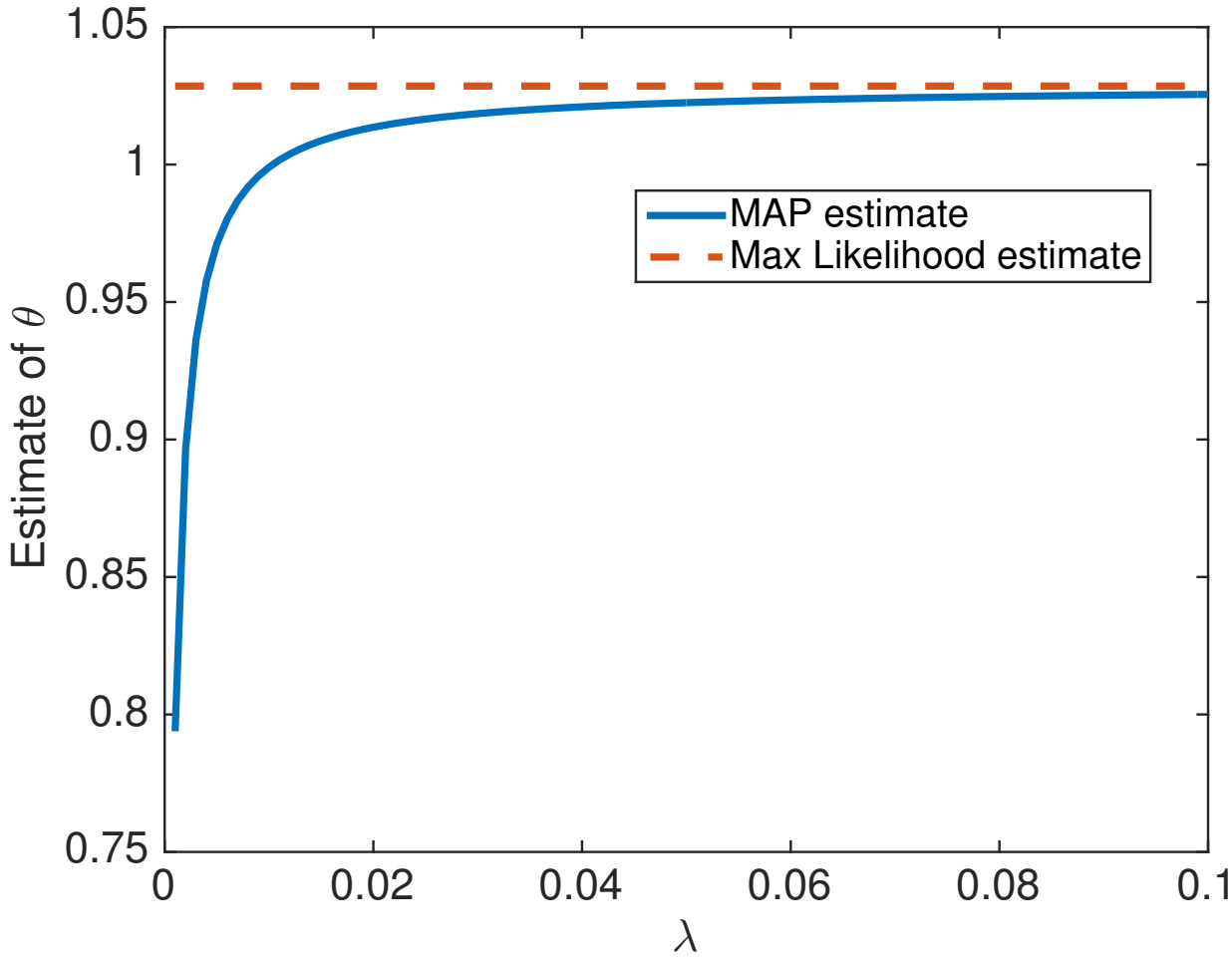
i.e. to minimise

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{1}{\lambda}\sum_{j=1}^{n}\theta_j^2$$

This is called **ridge regression**.

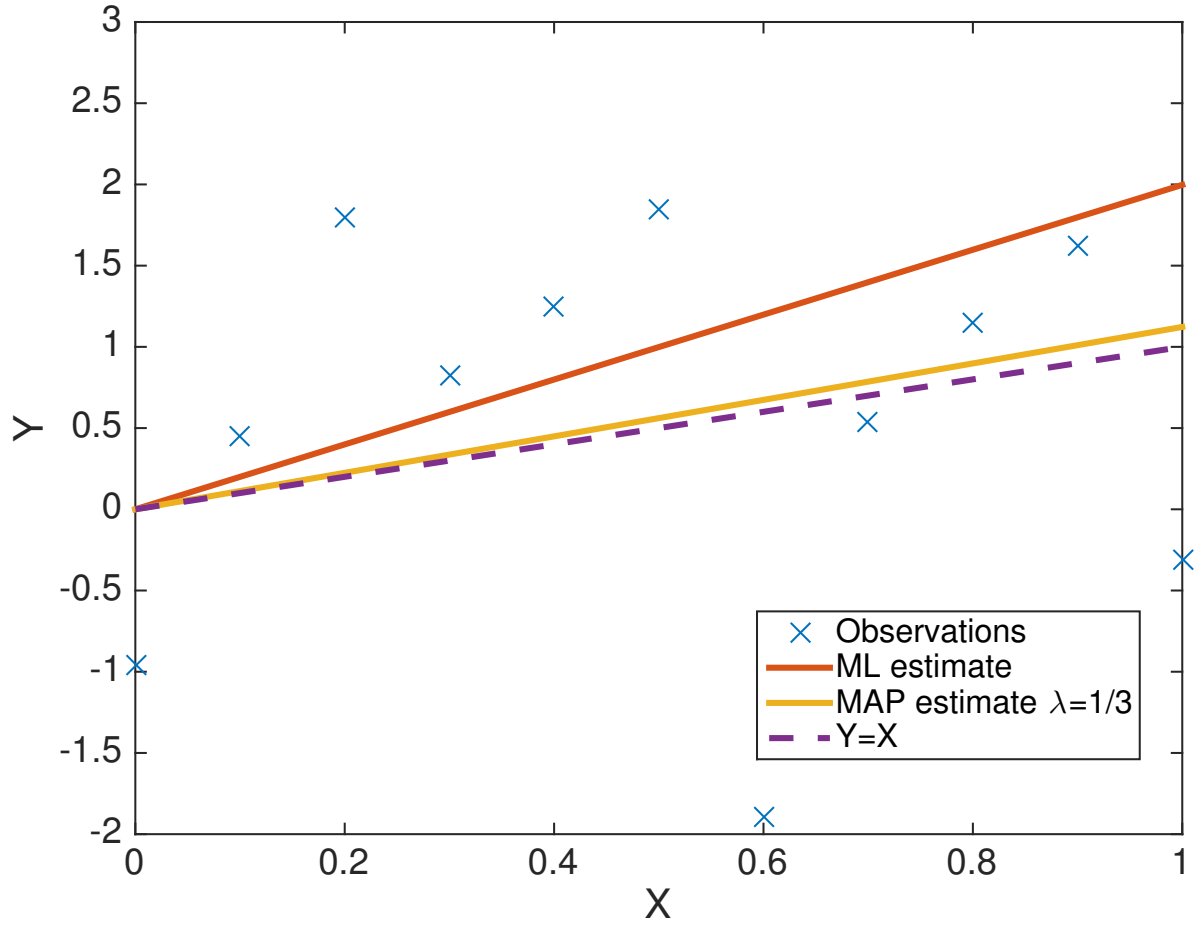
- When $\lambda \to 0$, then we are saying that we are certain $\theta = 0$.
- When λ is large we are saying that we know little about the value of θ prior to making the observations. The penalised estimate is then close to the non-penalised estimate.

MAP Estimation



MAP vs Maximum Likelihood Estimation

Difference between MAP and ML really kicks in when we only have a small number of observations, yet still need to make a prediction. Our prior beliefs are then especially important.



MAP vs Maximum Likelihood Estimation

- But as number N of observations grows, impact of prior on posterior tends to decline.