Machine Learning - Waseda University
Lecture 4: Bayesian approaches to parameter estimation and model selection

AD

June 2011

Bayesian linear regression

- We have the model

$$\widehat{t}\,(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

  and in a matrix-vector form

$$\mathbf{t} = \boldsymbol{\phi}\mathbf{w} + \boldsymbol{\epsilon}$$

- We want to present a full Bayesian analysis in this context.

Likelihood and Prior

- Likelihood

$$p(D|\mathbf{w}, \sigma^2) = \mathcal{N}(D; \boldsymbol{\phi}\mathbf{w}, \sigma^2 I);$$

  i.e. we assume $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ where $\lambda = 1/\sigma^2$ is the associated so-called precision

- Prior

$$p(\mathbf{w}, \sigma^2) = p(\mathbf{w}, \sigma^2)p(\sigma^2)$$

  where

$$p(\mathbf{w}|\,\sigma^2) = \mathcal{N}(\mathbf{w}; m, \sigma^2 V)$$

  and

$$p(\sigma^2) = \mathcal{IG}\left(\sigma^2; a, b\right)$$

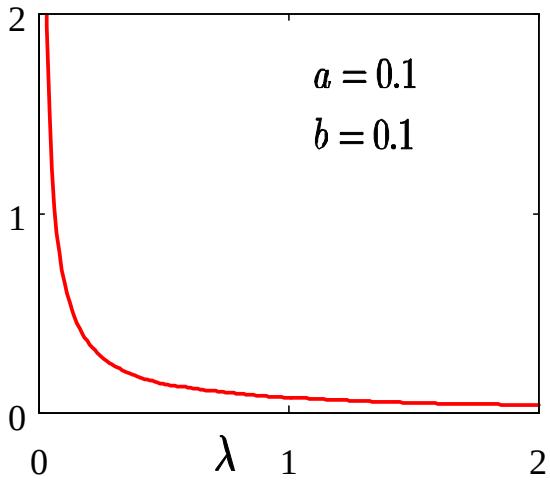
  which is written as

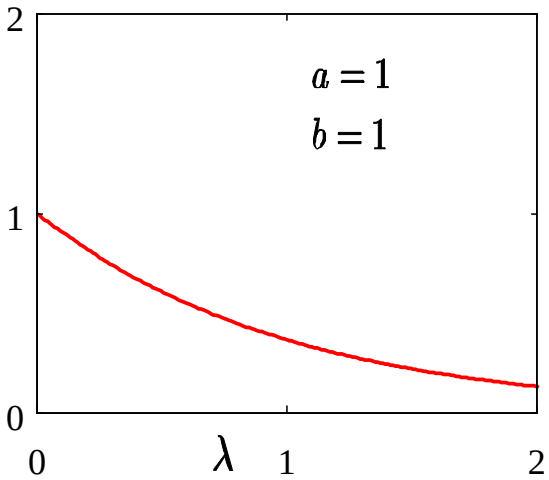
$$p(\mathbf{w}, \sigma^2) = \mathcal{NIG}(\mathbf{w}, \sigma^2; m, V, a, b)$$

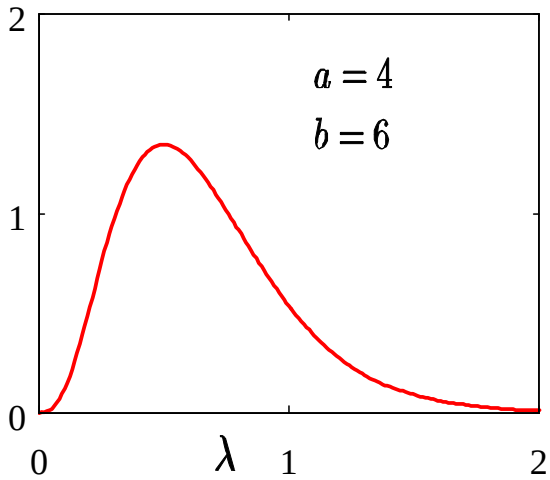
Gamma distribution

- Gamma with shape $a > 0$ and rate (inverse scale) $b > 0$

$$p(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \ \ \lambda^{a-1} \exp(-b\lambda)$$







Gamma or inverse gamma?

- We can either put a prior on the variance σ^2 or on the precision $\lambda = 1/\sigma^2$.
- An easy to handle prior for λ is $\lambda \sim Ga(a, b)$, $a > 0$ is shape, $b > 0$ is inverse scale

$$\begin{aligned} \mathcal{G}(\lambda; a, b) &= \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \\ \mathbb{E}[\lambda] &= a/b \end{aligned}$$

- The conjugate prior for σ^2 is $\sigma^2 \sim \mathcal{IG}(a, b)$, $a > 0$ is shape, $b > 0$ is scale

$$\begin{aligned} \mathcal{IG}(\sigma^2; a, b) &= \frac{1}{\Gamma(a)} b^a (\sigma^2)^{-(a+1)} \exp(-b/(\sigma^2)) \\ \mathbb{E}[\sigma^2] &= b/(a-1) \end{aligned}$$

Posterior Distribution for Bayesian Linear Regression

- We have

$$p(\mathbf{w}, \sigma^2 | D) = \frac{p(D|\mathbf{w}, \sigma^2) p(\mathbf{w}, \sigma^2)}{p(D)}.$$

- After tiedous calculations, we obtain

$$p(\mathbf{w}, \sigma^2 | D) = \mathcal{NIG}(\mathbf{w}, \sigma^2; m^*, V^*, a^*, b^*)$$

with

$$\begin{aligned} m^* &= V^*(V^{-1}m + \boldsymbol{\phi}^T \mathbf{t}) \\ V^* &= \left(V^{-1} + \boldsymbol{\phi}^T \boldsymbol{\phi}\right)^{-1} \\ a^* &= a + N/2 \\ b^* &= b + \frac{1}{2}(m^T V^{-1} m + \mathbf{t}^T \mathbf{t} - (m^*)^T (V^*)^{-1} m^*) \end{aligned}$$

Evidence

- Marginal likelihood or evidence is given by

$$\begin{aligned} p(D) &= \int p(D|\mathbf{w},\sigma^2)p(\mathbf{w},\sigma^2)d\mathbf{w} \\ &= \frac{|V^*|^{1/2}b^a\Gamma(a^*)}{|V|^{1/2}(b^*)^{a^*}\Gamma(a)\pi^{N/2}} \end{aligned}$$

- As evidence is available in closed form, additional hyper-parameters (a, b, m, V) can be estimated if necessary by maximizing $p(D)$.
- Take a simple form if $V = \delta^2 \left(\boldsymbol{\phi}^T\boldsymbol{\phi}\right)^{-1}$; this is the so-called g-prior. We have

$$\begin{aligned} p(D) &= \frac{|\left(\delta^{-2}+1\right)\left(\boldsymbol{\phi}^T\boldsymbol{\phi}\right)^{-1}|^{1/2}b^a\Gamma(a^*)}{|\delta^2\left(\boldsymbol{\phi}^T\boldsymbol{\phi}\right)^{-1}|^{1/2}(b^*)^{a^*}\Gamma(a)\pi^{N/2}} \\ &= \left(\frac{1}{1+\delta^2}\right)^{M/2}\frac{b^a\Gamma(a^*)}{(b^*)^{a^*}\Gamma(a)\pi^{N/2}} \end{aligned}$$

Posterior predictive distribution

- The posterior predictive density is a Student or t-distribution

$$\begin{aligned} p(t|\mathbf{x}, D) &= \int p(t|\mathbf{x}, \mathbf{w}, \sigma^2) p(\mathbf{w}, \sigma^2|D) d\mathbf{w} d\sigma^2 \\ &= St(t|\boldsymbol{\phi}^T(\mathbf{x}) m^*, b^*(1 + \boldsymbol{\phi}^T(\mathbf{x}) V^* \boldsymbol{\phi}(\mathbf{x})), a^*) \end{aligned}$$

  and

$$St(t|\mu, v, c) = \frac{\Gamma(c/2 + 1/2)}{\Gamma(c/2)\sqrt{\pi v}} \left[1 + \frac{(t-\mu)^2}{v}\right]^{-(c+1)/2}$$

  where $\mathbb{E}(T) = \mu$ and $\mathbb{V}(T) = v/(c-2)$.
- I follow the parameterization of Denison p29. This is different from Bishop!

Student distribution is a mixture of Gaussians

- The Student distribution is an infinite mixture of Gaussians with different variances

$$St(t;\mu,\lambda,\nu) = \int \mathcal{N}(t;\mu,\tau)\mathcal{G}(\tau;a,b)d\tau$$

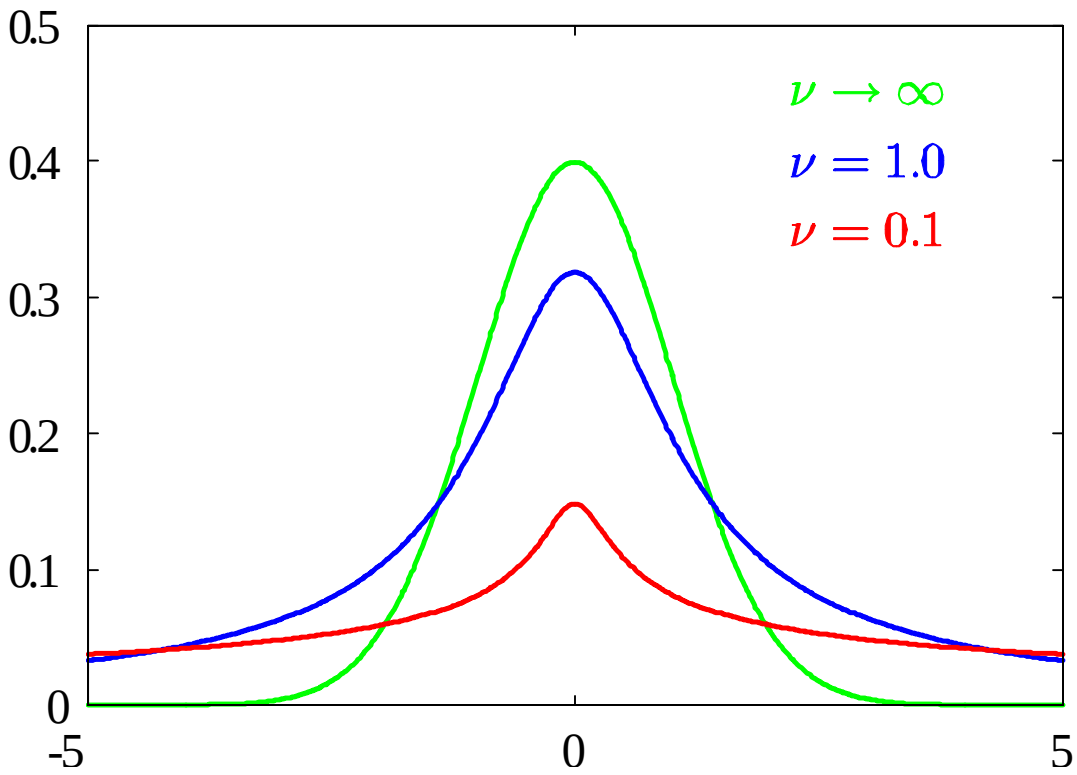
where $\nu = 2a$ and $\lambda = a/b$ and St_B is Bishop's parameterization

$$St_B(t;\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(t-\mu)^2}{\nu}\right]^{-(\nu+1)/2}$$

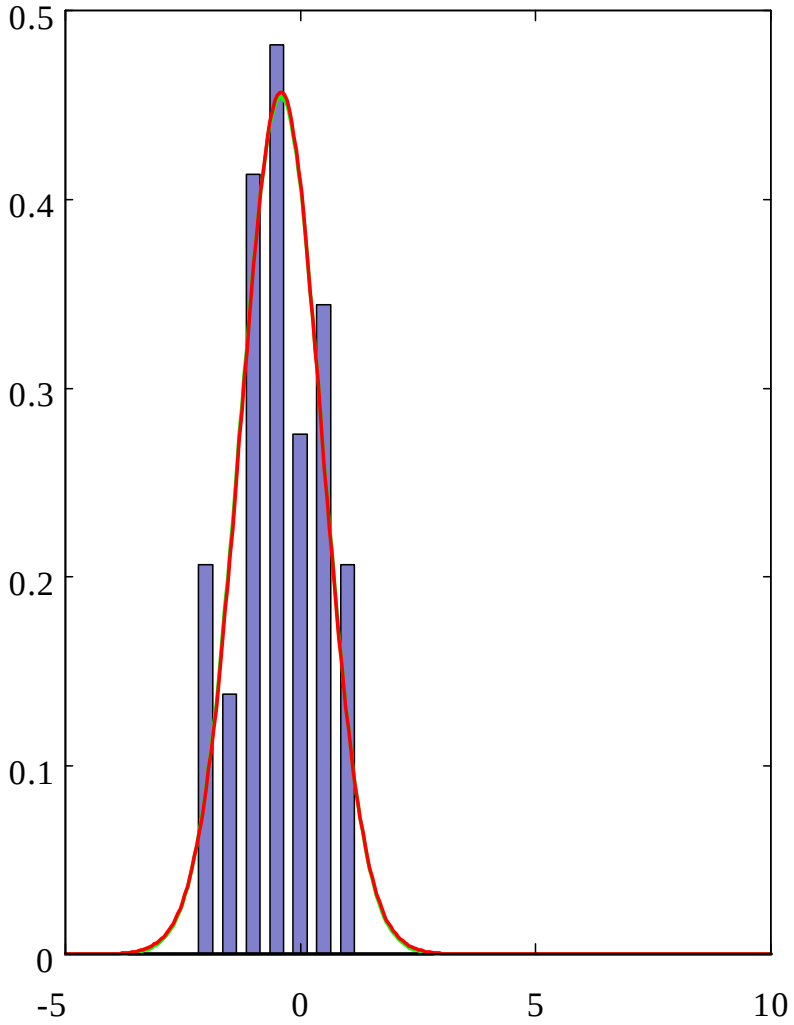
where $\mathbb{E}(T) = \mu$ and $\mathbb{V}(T) = \frac{1}{\lambda}\frac{\nu}{\nu-2}$.

- Hence a student distribution has wider tails than a Gaussian.
- As $\nu \to \infty$, $St(y;\mu,\lambda,\nu) \to \mathcal{N}(y;\mu,\text{precision} = \lambda)$.

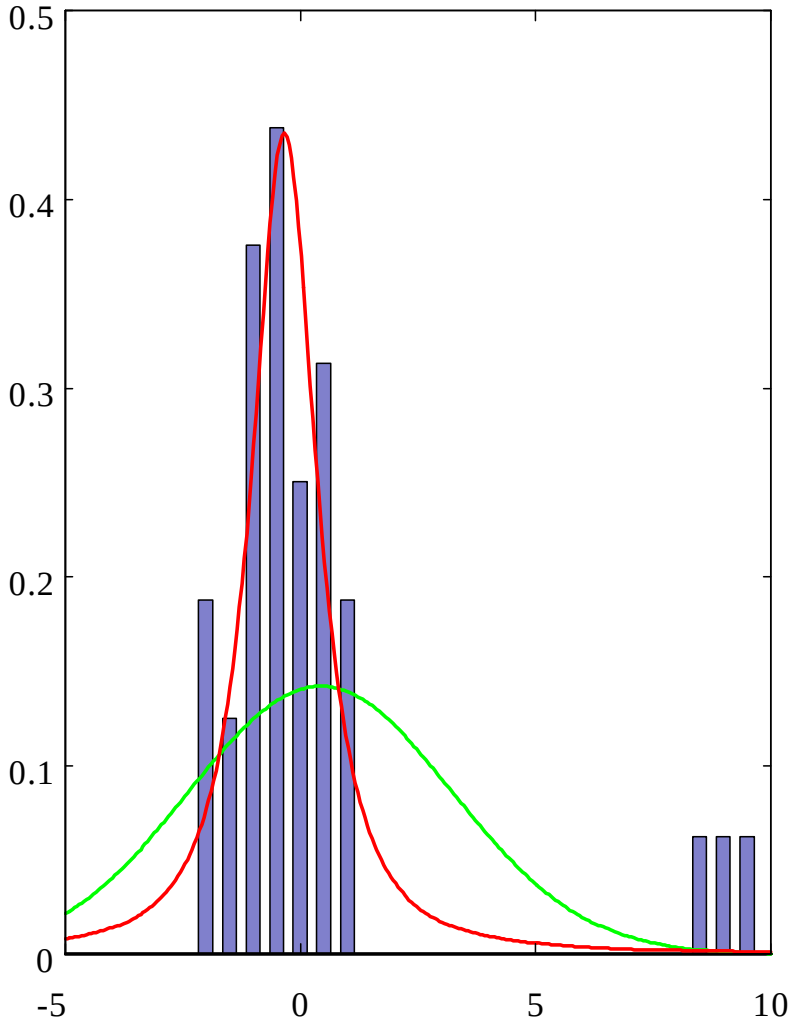
Student has wider tails than Gaussian



Robustness of student distribution to outliers



(a)



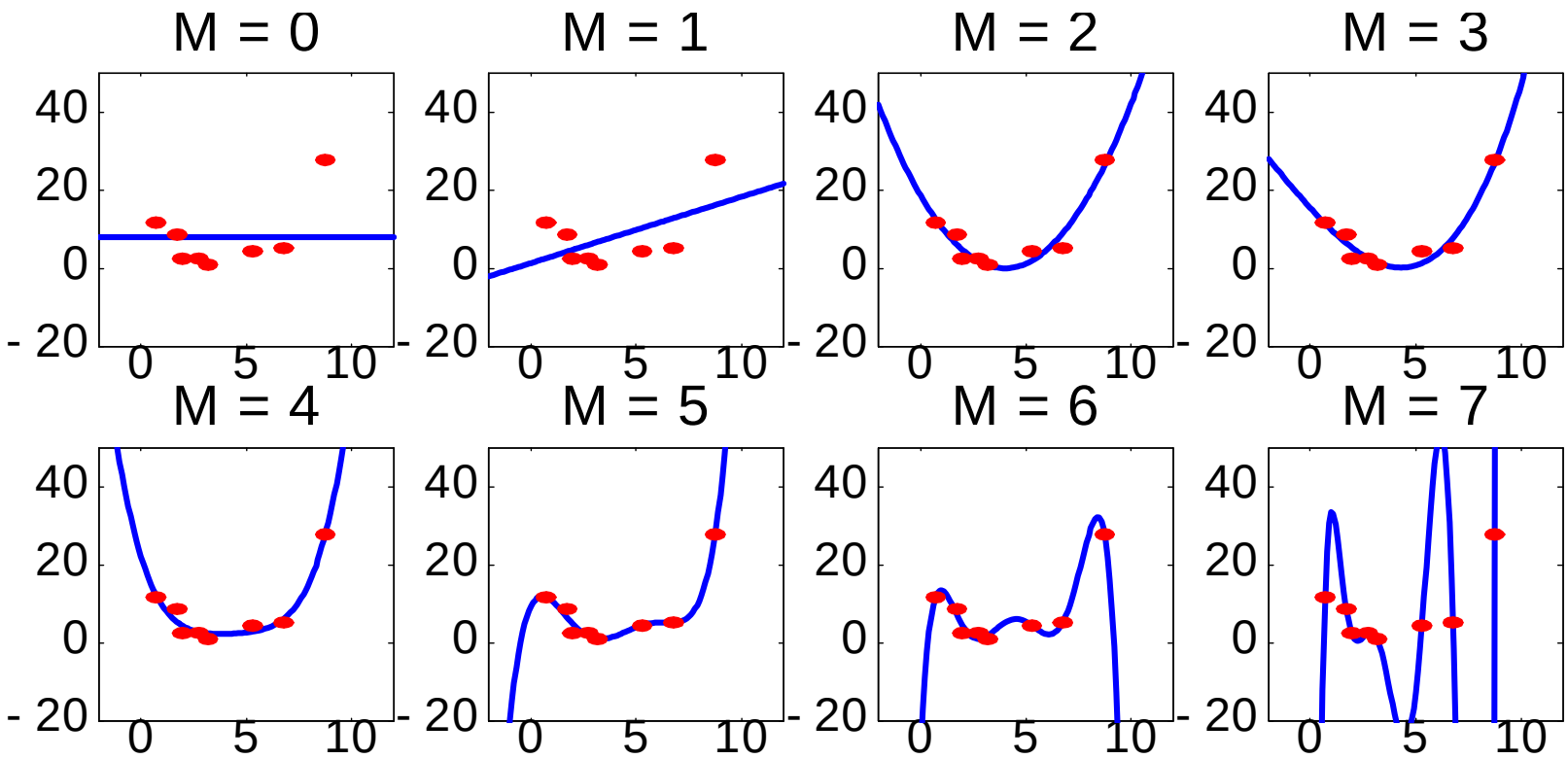
(b)

Model Selection

- Let model M_k $(\Leftrightarrow M = k)$ be polynomial regression of order k:

$$\widehat{t}\,(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}) = \sum_{j=0}^{M} w_j x^j$$

- Which model should we choose?



Bayesian Model Selection

- We select $V_M = \delta^2 I_{M+1}$ where $\delta^2 = 10$, $a = b = 1$.
- Now we have

$$p\left(M_k \middle| D\right) = \frac{p\left(D \middle| M_k\right) p\left(M_k\right)}{\sum_{l=0}^{M_{\max}} p\left(D \middle| M_l\right) p\left(M_l\right)}$$

- If we defined $p\left(M_k\right) = \frac{1}{k_{\max}+1}$ then picking the model having the highest posterior proba is picking the model have highest evidence.

Bayesian Model Comparison

- If we wish to compare two models, M_i and M_j, we can compute their posterior odds

$$\frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M=i)}{p(D|M=j)} \times \frac{p(M=i)}{p(M=j)}$$

- We can cancel out any prior preference of model i to j by computing the Bayes factor

$$BF(M_i, M_j) = \frac{p(M_i|D)}{p(M_j|D)} / \frac{p(M_i)}{p(M_j)} = \frac{p(D|M_i)}{p(D|M_j)}$$

- If the prior on models is uniform, so $p(M_i) = p(M_j)$, and if each model has prior $p(\mathbf{w}, \sigma^2|M_i) = \mathcal{NIG}(\mathbf{w}, \sigma^2; m_i, V_i, a, b)$, then

$$BF(M_i, M_j) = \frac{|V_j|^{1/2}|V_i^*|^{1/2}(b_j^*)^{a*}}{|V_i|^{1/2}|V_j^*|^{1/2}(b_i^*)^{a*}}$$

where $a^* = a_i^* = a_j^* = a + N/2$.

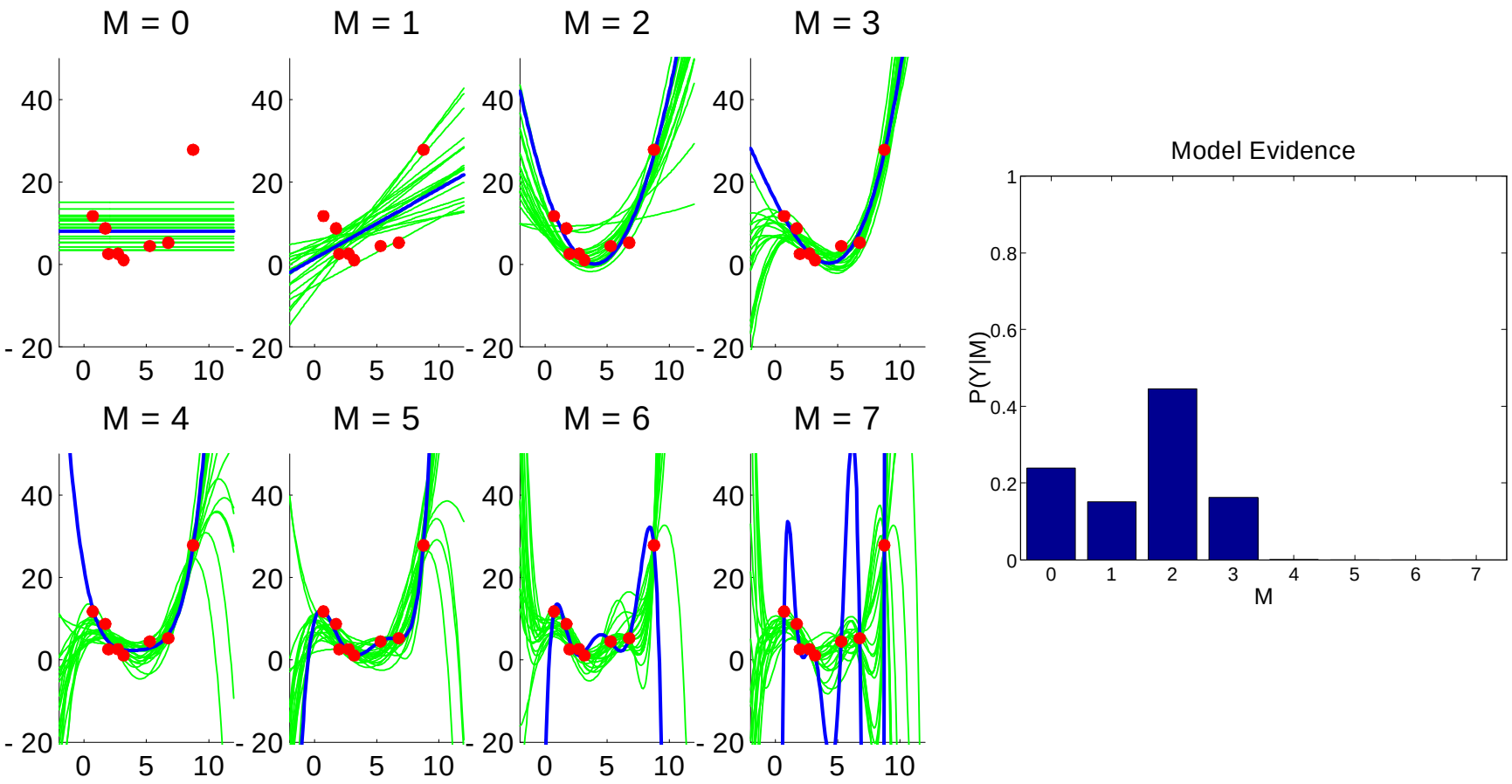
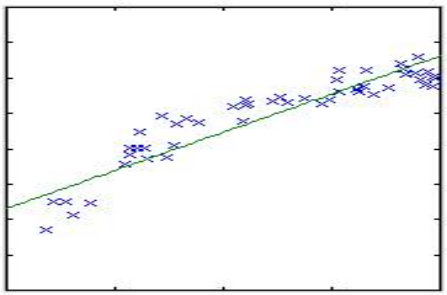


Figure: Evidence $p(D|M=k)$ and regression functions for random draws from $p(\mathbf{w}|D, M=k)$.

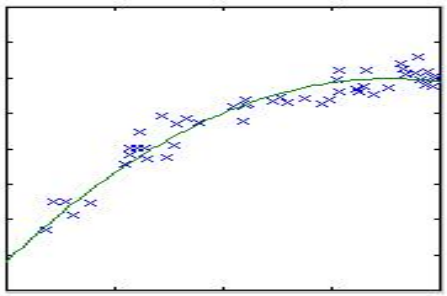
Bayesian Occam's razor

- Amazingly, even if we have no explicit penalty on complex models (so $P(M = k)$ is uniform), merely by integrating over all possible parameter values (i.e., by using $P(D|M = k) = \int P(D, \mathbf{w}, \sigma^2|M = k)d\mathbf{w}d\sigma^2$), we automatically prefer models that are not too complex (provided they fit the data well).
- This is called the Bayesian Occam's razor. (Occam's razor says: "if two models are equally good at predicting, pick the simpler one".)

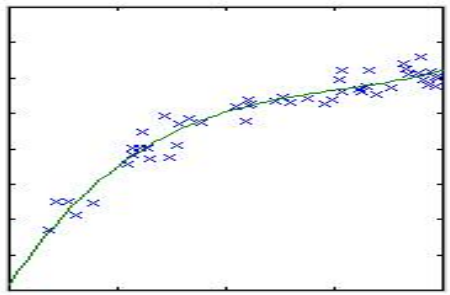
Another Example



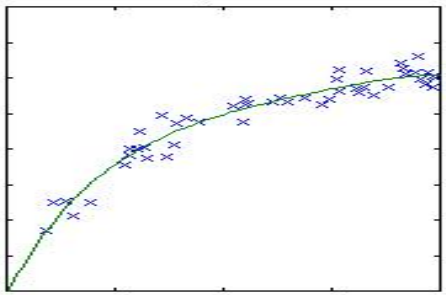
degree 1



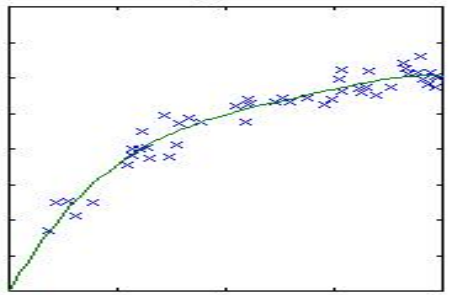
degree 2



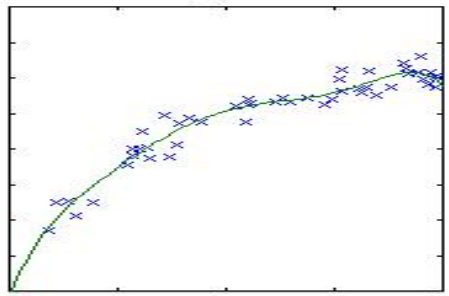
degree 3



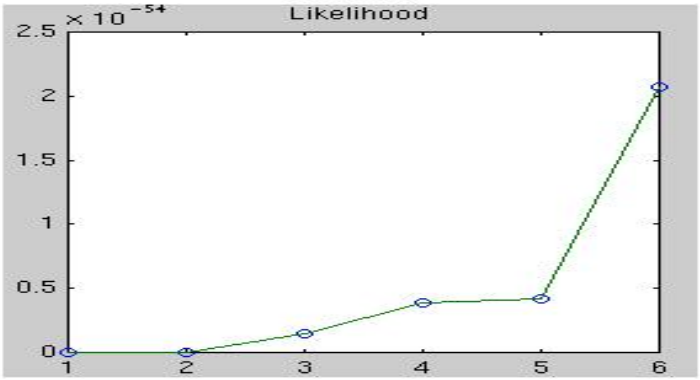
degree 4

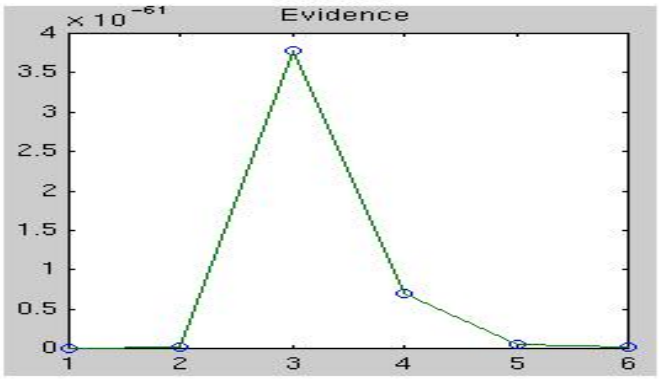


degree 5



degree 6





Beyond the Bayesian Folkore

- Testing hypothesis in a Bayesian way is attractive.... but be careful to vague priors!!!
- Assume you have $X|\left(\mu,\sigma^2\right)\sim\mathcal{N}\left(\mu,\sigma^2\right)$ where σ^2 is assumed known but μ (the parameter θ) is unknown. We want to test $H_0:\mu=0$ vs $H_1:\mu\sim\mathcal{N}\left(0,\tau^2\right)$ then

$$\begin{aligned} B_{10}(x) &= \frac{p(x|H_1)}{p(x|H_0)}=\frac{\int\mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{N}\left(\mu;0,\tau^2\right)d\mu}{\mathcal{N}\left(x;0,\sigma^2\right)} \\ &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right) \end{aligned}$$

So what??

Lindley's paradox

- You might be tempted to use vague priors; that is take $\tau^2 \longrightarrow \infty$.
- However, we have *for any* x

$$\lim_{\tau^2 \to \infty} B_{10}(x) = 0$$

- So we will always select hypothesis H_0 even if $x = 10^9$.
- Vague priors should be banned for Bayesian hypothesis testing/model selection.

Lindley's paradox for Bayesian linear regression

- Consider the polynomial regression case where model M_i corresponds to polynom of degree i.
- Consider the case where we take $V = \delta^2 \left(\boldsymbol{\phi}^T \boldsymbol{\phi}\right)^{-1}$ then

$$\begin{aligned} BF(M_i, M_j) &= \frac{|V_j|^{1/2}|V_i^*|^{1/2}(b_j^*)^{a*}}{|V_i|^{1/2}|V_j^*|^{1/2}(b_i^*)^{a*}} \\ &= \left(\frac{1}{1+\delta^2}\right)^{\frac{(i-j)}{2}} \frac{(b_j^*)^{a*}}{(b_i^*)^{a*}} \end{aligned}$$

- It follows that if $i > j$ then

$$\lim_{\delta^2 \to \infty} BF(M_i, M_j) = 0$$

- With a vague prior $\left(\delta^2 \to \infty\right)$, we would always select M_0! But you can pick δ^2 maximizing the evidence!