CS540: Machine Learning I

Lecture 4: Bayesian parameter estimation and model selection for linear Gaussian models

Kevin Murphy

Wednesday September 21, 2005[1]

[1] Slides last updated on September 21, 2005

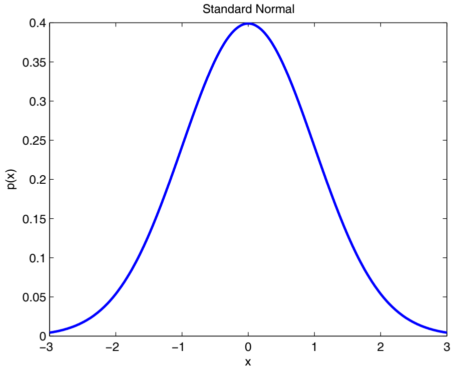
Daphne Koller's talk

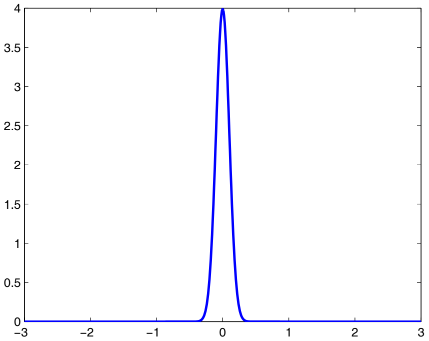
- Probablistic Models for Complex Domains: Cells, Bodies and Webpages
- Thursday 22nd, 4pm, Dempster 310

Gaussian (Normal) density in 1D

- If $X \sim N(\mu, \sigma^2)$, the probability density function (pdf) of X is defined as
  $$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$
  We will often use the precision $\lambda = 1/\sigma^2$ instead of the variance σ^2.
- Note that a density evaluated at a point can be bigger than 1!
- Here is how we plot the pdf in matlab

```
xs=-3:0.01:3;  plot(xs,normpdf(xs,mu,sigma))
```





The Gaussian maximizes the differential entropy

- Recall that the entropy of a discrete random variable is defined as

$$H[X] = -\sum_x p(x) \log p(x)$$

- The maximum entropy distribution is uniform (for discrete RVs).
- The differential entropy of a continuous random variable is defined as

$$E[X] = -\int p(x) \log p(x) dx$$

- If we maximize this subject to the following constraints (using Lagrange multipliers) $\int_{\infty}^{\infty} p(x)dx = 1$, $\int_{\infty}^{\infty} xp(x)dx = \mu$ and $\int_{\infty}^{\infty}(x-\mu)^2 p(x)dx = \sigma^2$, we get (Bishop p69)

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

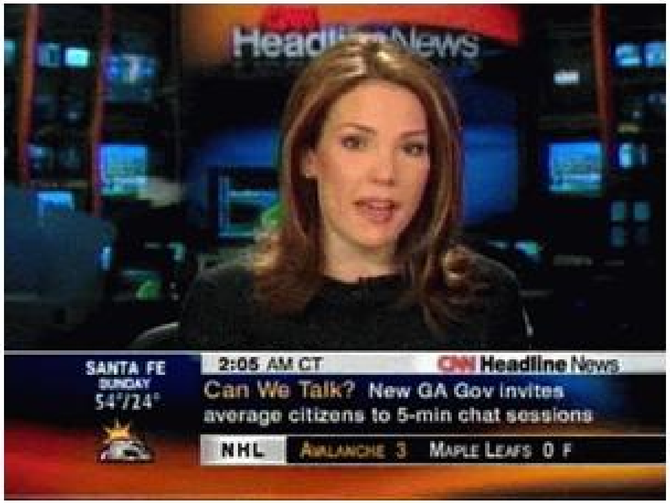
Entropy of a 1D Gaussian

- The differential entropy of a 1D Gaussian is

$$H[X] = \frac{1}{2}\left\{1 + \log(2\pi\sigma^2)\right\}$$

- Hence differential entropy can be negative (if $\sigma^2 < 1/(2\pi e)$).

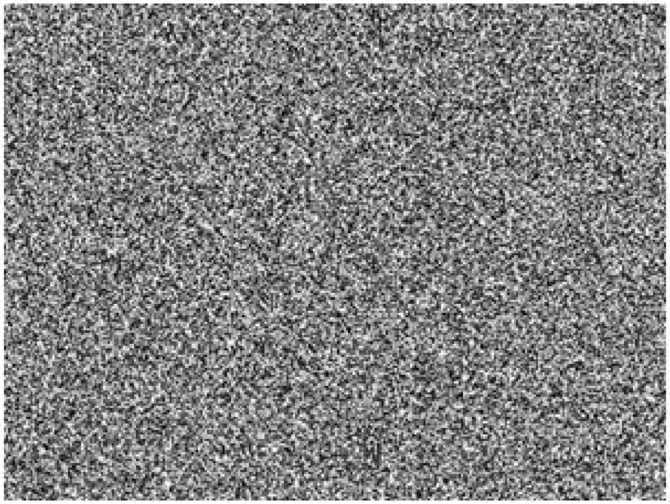
White snow paradox



TV news, sports, music, action movies, etc

≈ 0.3 MByte/s

(640x480, MPEG4, avg. 46,000 frames)



Greyscale snow

≈ 5.0 MByte/s

A surprise theory of attention

- Laurent Itti (USC) and Pierre Baldi (UCI), CVPR 2005.
- Surprising events are ones that changes your beliefs the most

$$S(D|M) \overset{\text{def}}{=} KL(\ P(M|D)||P(M)\ ) = \sum_m P(m|D) \log \frac{P(m|D)}{P(m)}$$

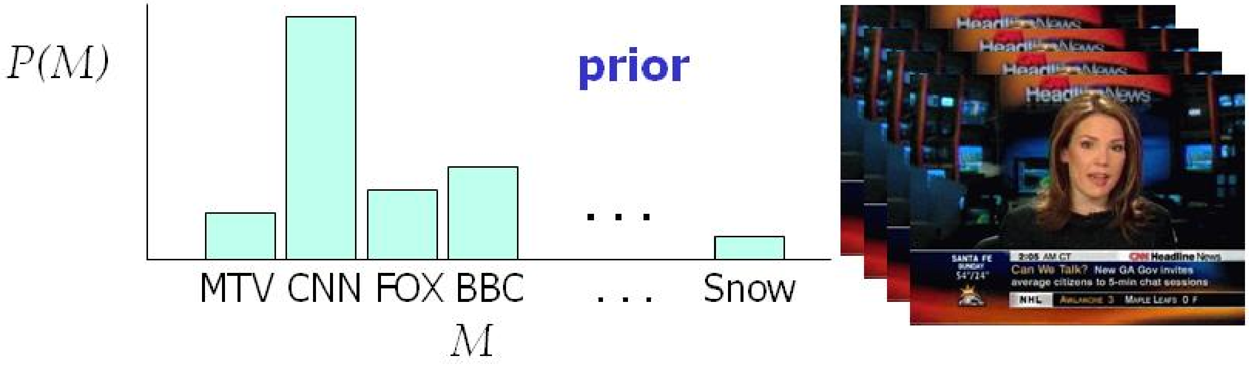
  where $P(M)$ are your prior beliefs in model M.

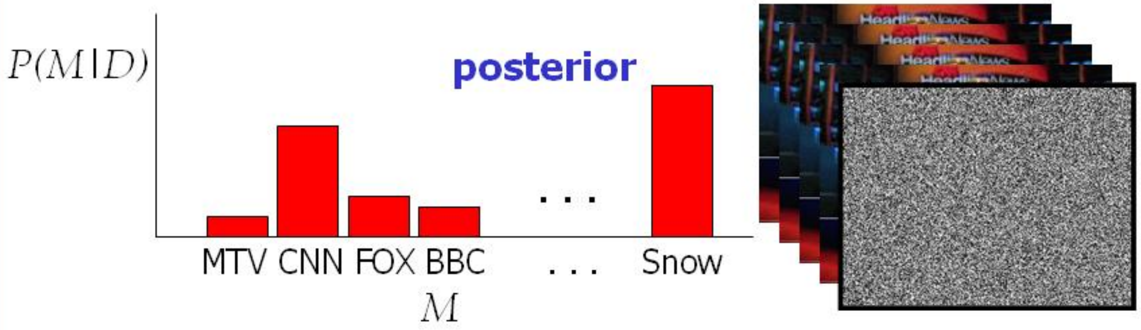
- Entropy just refers to data, not models

$$I(D) \overset{\text{def}}{=} -\sum_d P(d) \log P(d)$$

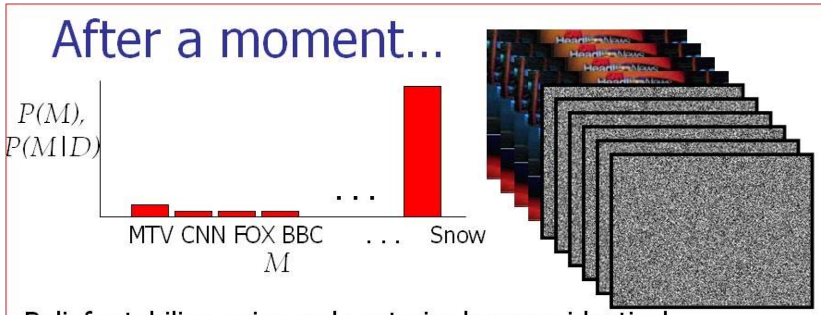
- Itti and Baldi show that the KL model is able to predict what visual events humans pay attention to better than looking for events with high “information” content or which are “salient” (local outliers) wrt low-level visual cues.

A surprise theory of attention





After a moment...



Beliefs stabilize, prior and posterior become identical, and additional snow frames carry no surprise

Multivariate Gaussian

- If $X \in \mathbb{R}^d$ is a jointly gaussian random vector, then its pdf is

$$p(x) = N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}|} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

- The quantity $\Delta^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is called the Mahalanobis distance between x and μ.
- The first and second moments are

$$E[X] = \mu, \quad \mathsf{Cov}[X] = \Sigma$$

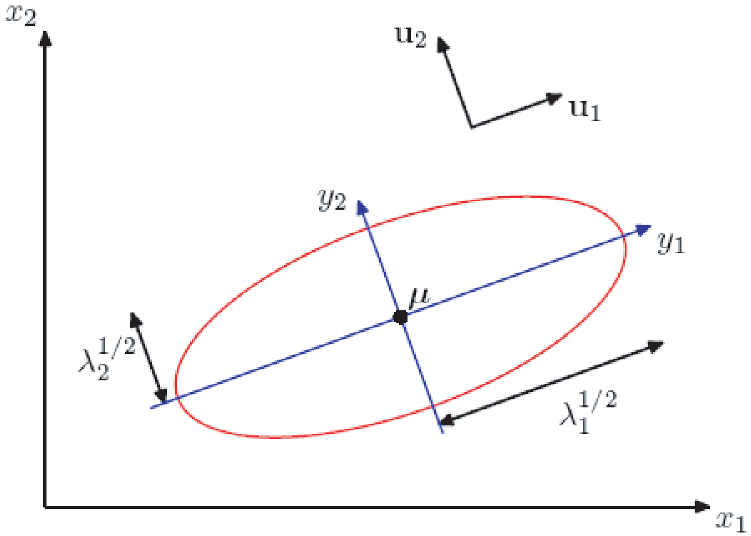
- Sometimes we will use the precision matrix Σ^{-1} instead of the covariance matrix Σ.

Quick linear algebra review (Bishop sec B.4)

- We can compute the eigenvectors u_i and eigenvalues λ_i of any square matrix A:
$$Au_i = \lambda_i u_i$$
- We can write this in matrix form as
$$A = U \Lambda U^T$$
where the columns of U are the u_i and $\Lambda = \text{diag}(\lambda_i)$. This is called diagonalizing A.
- If A is real and symmetric, then the eigenvalues are real and the eigenvectors are orthonormal, so that
$$u_i^T u_j = I_{ij}$$
or
$$U^T U = I$$
- The rank of A is the number of non-zero eigenvalues. If all $\lambda_i \geq 0$, then A is positive semi definite (psd), i.e., $x^T A x \geq 0$ for all x.

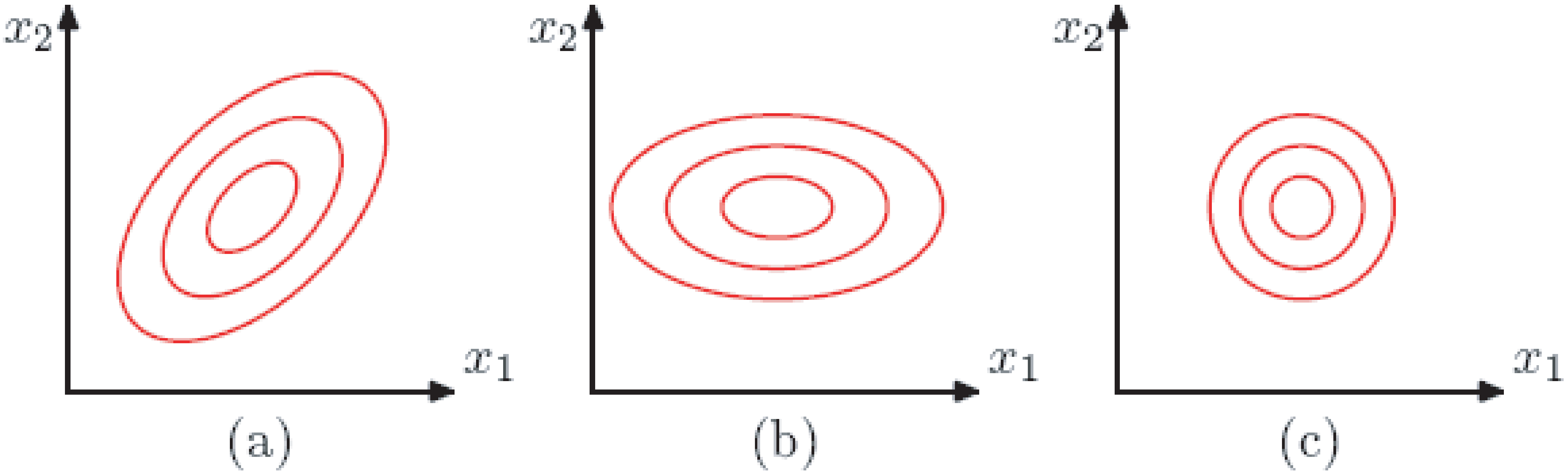
Visualizing the covariance matrix

- By diagonalizing $\Sigma = U\Lambda U^T$, we get $\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T$ so the Mahalanobis distance can be rewritten as $\Delta^2 = \sum_i \frac{y_i^2}{\lambda_i}$, where $y_j = u_j^T(x-\mu)$.
- The surfaces of constant probability satisfy $y_j = u_j^T(x-\mu) =$ const, which are ellipsoids.



Restricting the covariance matrix

- A full covariance matrix has $d(d+1)/2$ parameters.
- We can restrict Σ to be diagonal; this has d parameters.
- Or we can use a spherical covariance, $\Sigma = \sigma^2 I$.
- Later we will see how to use graphical models to represent other kinds of sparse parameterizations.



Manipulating Gaussians (Bishop sec 2.3.1-2.3.2)

- Suppose $x = (x_a, x_b)$ is jointly Gaussian with parameters

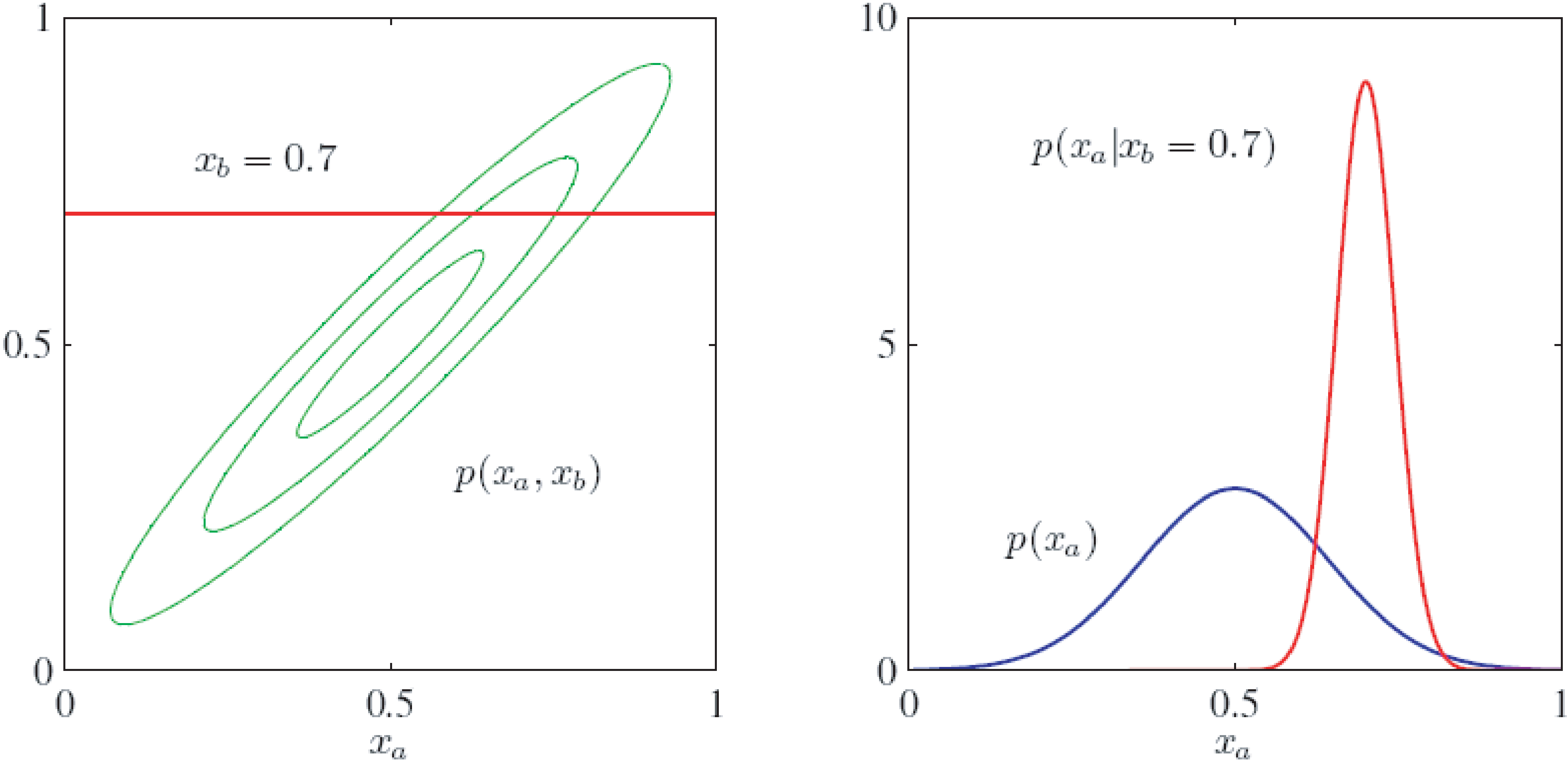
$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix},$$

- It can be shown that $P(X_a|x_b) = N(X_a; \mu_{a|b}, \Sigma_{a|b})$ where

$$\begin{aligned} \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \end{aligned}$$

- Note that the new mean is a linear function of x_a, and the new covariance is independent of x_a.
- Similarly, the marginal $P(X_a) = N(X_a; \mu_a, \Sigma_{aa})$.
- You should memorize these equations!

Marginals and conditionals of a 2D Gaussian



MLE for a Gaussian mean

- Given N iid datapoints x_n stored in rows of X, the log-likelihood is

$$\begin{aligned}\log p(X|\mu,\Sigma) &= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| \\ & -\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\end{aligned}$$

- Using the following two results (Sam Roweis 5a, 5b)

$$\frac{\partial(a^Tx)}{\partial x} = a, \quad \frac{\partial(x^TAx)}{\partial x} = (A+A^T)x$$

we can show (homework!)

$$\frac{\partial}{\partial\mu}\log p(X|\mu,\Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(x_n-\mu)$$

so

$$\mu_{ML} = \frac{1}{N}\sum_n x_n$$

MLE for a Gaussian covariance

- It can be shown that the MLE for Σ is

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^T = \frac{1}{N}S$$

  where the scatter matrix is

$$S = \sum_n (x_n - \bar{x})(x_n - \bar{x})^T = (\sum_n x_n x_n^T) - N\bar{x}\bar{x}^T$$

- The sufficient statistics are $\sum_n x_n$ and $\sum_n x_n x_n^T$.
- Note that $X^T X$ may not be full rank (eg. if $N < D$), in which case Σ_{ML} is not invertible.

Bayesian parameter estimation for a Gaussian

- There are various reasons to pursue a Bayesian approach
  - The MLE for Σ may not be full rank if we don't have enough data.
  - We would like to update our estimates sequentially over time.
  - We may have prior knowledge about the expected magnitude of the parameters.
- We will restrict our attention to conjugate priors.
- We will consider various cases, in order of increasing complexity:
  - Known σ, unknown μ
  - Known μ, unknown σ
  - Unknown μ and σ

Unknown μ, known σ

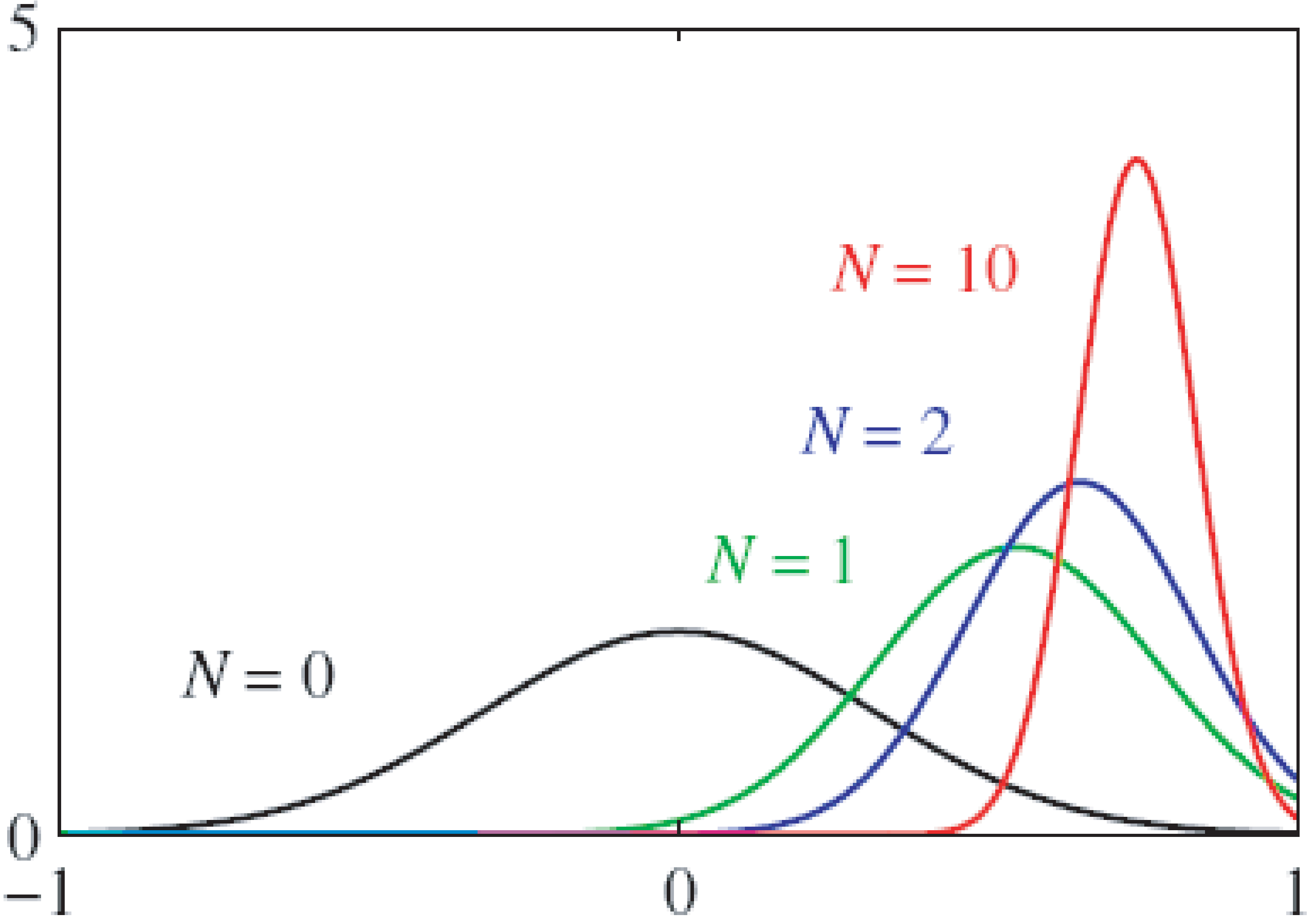
- The likelihood is $\prod_n N(x_n|\mu,\sigma)$.
- The conjugate prior is $p(\mu|\mu_0,\sigma_0^2)$.
- By completing the square, it can be shown that the posterior is

$$
\begin{aligned}
p(\mu|D) &= N(\mu|\mu_N,\sigma_N^2) \\
\mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML} \\
\frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}
\end{aligned}
$$

- The posterior mean is a convex combination of the prior and the MLE, with weights proportional to the relative noise levels.
- The precision of the posterior $1/\sigma_N^2$ is the precision of the prior $1/\sigma_0^2$ plus one contribution of data precision $1/\sigma^2$ for each observed data point.

Sequentially updating the mean

$\mu^* = 0.8$ (unknown), $(\sigma^2)^* = 0.1$ (known)



Effect of single data point

- The posterior mean is a convex combination of the prior and the observation x, with weights proportional to the relative noise levels.

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + \sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}x$$

- The posterior mean is the prior mean adjusted towards x:

$$\mu_1 = \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$$

- The posterior mean is the data 'shrunk' towards the prior mean:

$$\mu_1 = x - (x - \mu_0)\frac{\sigma^2}{\sigma^2 + \sigma_0^2}$$

Uninformative (reference) prior for μ

- The posterior is

$$\begin{aligned} p(\mu|D) &= N(\mu|\mu_N, \sigma_N^2) \\ \mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \\ \frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \end{aligned}$$

- Hence when $\sigma_0^2 \to \infty$ (vague/ flat prior), then $E[\mu|D] \to \mu_{ML}$.

KNOWN μ, UNKNOWN $\lambda = 1/\sigma^2$

- The likelihood is

$$p(D|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) \propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

- The conjugate prior is a Gamma with shape a_0 and rate (inverse scale) b_0

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

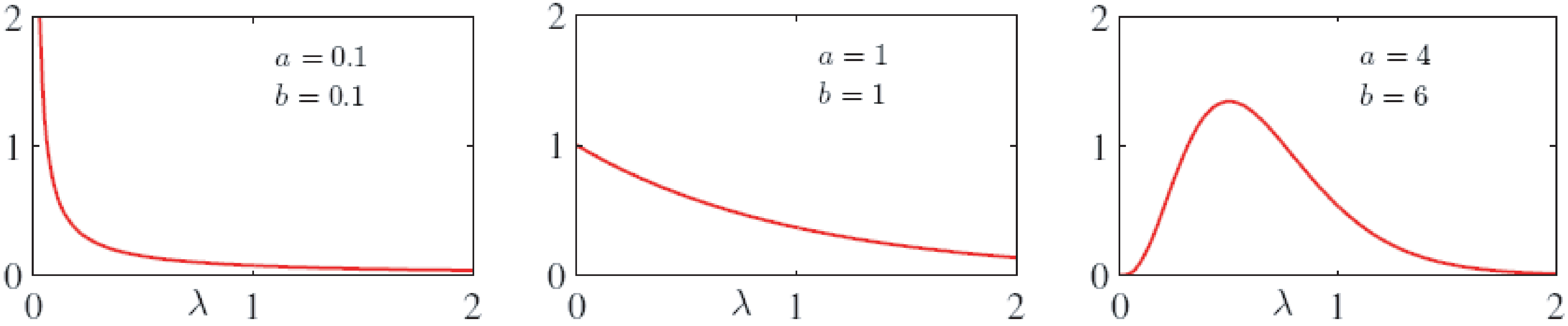
- The posterior is

$$\begin{aligned} p(\lambda|D) &= Ga(\lambda|a_N, b_N) \\ a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma^2_{ML} \end{aligned}$$

Gamma distribution

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

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



Uninformative (reference) prior for λ

- If $\lambda \sim Ga(a, b)$, then $E[\lambda] = a/b$.
- The posterior is

$$\begin{aligned} p(\lambda|D) &= Ga(\lambda|a_N, b_N) \\ a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma^2_{ML} \end{aligned}$$

- Hence the posterior mean is

$$E[\lambda|D] = \frac{a_0 + N/2}{b_0 + \frac{N}{2}\sigma^2_{ML}}$$

- Hence an uninformative prior is $a_0, b_0 \to 0$. Then $E[\lambda|D] \to 1/\sigma^2_{ML}$.

Gamma or inverse gamma?

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

$$\begin{aligned} Ga(\lambda|a, b) &= \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \\ E\lambda &= a/b \end{aligned}$$

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

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

UNKNOWN μ AND UNKNOWN σ^2

- The conjugate prior is Normal-Inverse-Gamma

$$\begin{aligned} P(\mu, \sigma^2) &= P(\mu|\sigma^2) P(\sigma^2) \\ &= \mathcal{N}(\mu|m, \sigma^2 V) \;\; IG(\sigma^2|a, b) \\ &\stackrel{\text{def}}{=} NIG(\mu, \sigma^2|m, V, a, b) \\ &= \frac{1}{Z(m, V, a, b)} (\sigma^2)^{-(a+(k/2)+1)} \\ &\quad \times \exp[-\left\{(\mu - m)^T V^{-1} (\mu - m) + 2b\right\} / (2\sigma^2)] \end{aligned}$$

where

$$1/Z(m, V, a, b) = \frac{b^a}{(2\pi)^{k/2} |V|^{1/2} \Gamma(a)}$$

Semi conjugate prior

- If we use a factorized prior,

$$\begin{aligned} P(\mu, \sigma^2) &= P(\mu)P(\sigma^2) \\ &= \mathcal{N}(\mu|\mu_0, V) IG(\sigma^2|a, b) \end{aligned}$$

  then the posterior $P(\mu, \sigma^2|D)$ is still coupled because of explaining-away ($\mu \rightarrow X \leftarrow \sigma^2$). Such a factored prior is called semi-conjugate.

Multivariate case[2]: likelihood

- Likelihood

$$p(x_{1:N}|\mu,\Sigma) = |\Sigma|^{-N/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)$$
$$= |\Sigma|^{-N/2}\exp\left(-\frac{1}{2}Tr(\Sigma^{-1}S_0)\right)$$

where S_0 is the "sum of squares" relative to μ:

$$S_0 = \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

[2] Here I follow Gelman et al p85–87

Multivariate case: prior

- The conjugate prior is Normal-Inverse-Wishart

$$
\begin{aligned}
P(\mu, \Sigma) &= P(\mu|\Sigma)P(\Sigma) \\
&= \mathcal{N}(\mu|\mu_0, \frac{1}{\kappa_0}\Sigma)\ \mathcal{IW}(\Sigma|\Lambda_0^{-1}, \nu_0)
\end{aligned}
$$

where

$$
\mathcal{IW}(\Sigma|\Lambda_0^{-1}, \nu) = \frac{1}{Z}|\Sigma|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}Tr(\Lambda_0\Sigma^{-1})\right)
$$

and

$$
1/Z(\Lambda_0, \nu) = \left(2^{\nu d/2}\pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma(\frac{\nu+1-i}{2})\right)^{-1} |\Lambda_0|^{\nu/2}
$$

- $E(\Sigma) = \frac{\Lambda_0}{\nu-d-1}$

Multivariate case: posterior

- The posterior is Normal-Inverse-Wishart with parameters

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{x} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T
\end{aligned}$$

where S is the "sum of squares" relative to the sample mean

$$S = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

Multivariate case: uninformative prior

- The Jeffrey's prior is the limit of the conjugate case as $\kappa_0 \to 0$, $\nu_0 \to -1$, $|\Lambda_0| \to 0$:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

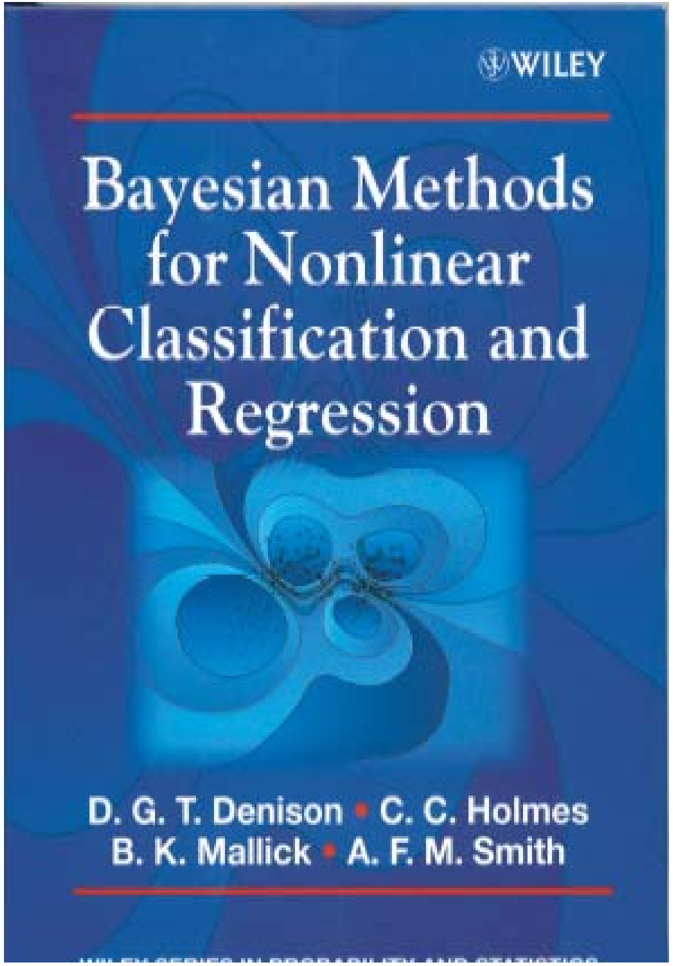
Bayesian linear regression

- So far, we have been considering *unconditional* density estimation.
- In many cases, we want to condition on known inputs $X \in \mathbb{R}^p$. In linear regression, we assume $E[Y|x]$ is a linear function

$$\mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- The linear assumption is fairly limiting, but is easy to overcome by defining a set of *fixed* basis functions $B_1(x), \ldots, B_k(x)$.
- The basis functions can be polynomials, splines, etc.

Denison book



WILEY

Bayesian Methods for Nonlinear Classification and Regression

D. G. T. Denison • C. C. Holmes
B. K. Mallick • A. F. M. Smith

WILEY SERIES IN PROBABILITY AND STATISTICS

Bayesian linear regression

- The model is

$$y_i = \sum_{j=1}^{k} \beta_j B_j(\vec{x}_i) + \epsilon_i$$

or, in matrix notation

$$Y = B\beta + \epsilon$$

where $Y = (y_1, \ldots, y_n)$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, and

$$B = \begin{pmatrix} B_1(x_1) & \cdots & B_k(x_1) \\ \vdots & \ddots & \vdots \\ B_1(x_n) & \cdots & B_k(x_n) \end{pmatrix}$$

- Standard linear regression can be modelled using $B_p(\vec{x}) = x_p$ and $B_{p+1} = 1$.
- An unconditional 1D Gaussian can be modelled using $B_1 = 1$, $\vec{\beta} = \mu$ and $\sigma^2 = V$.

Prior, likelihood, posterior, evidence

- Prior

$$p(\beta, \sigma^2) = NIG(\beta, \sigma^2 | m, V, a, b)$$

- Likelihood

$$p(D|\beta, \sigma^2) = \mathcal{N}(B\beta, \sigma^2 I)$$

- Posterior

$$\begin{aligned} p(\beta, \sigma^2|D) &= NIG(\beta, \sigma^2 | m^*, V^*, a^*, b^*) \\ m^* &= (V^{-1} + B'B)^{-1}(V^{-1}m + B'Y) \\ V^* &= (V^{-1} + B'B)^{-1} \\ a^* &= a + N/2 \\ b^* &= b + \frac{1}{2}(m^T V^{-1} m + Y^T Y - (m^*)^T (V^*)^{-1} m^*) \end{aligned}$$

- Marginal likelihood

$$p(D) = \frac{|V^*|^{1/2} b^a \Gamma(a^*)}{|V|^{1/2} (b^*)^{a^*} \Gamma(a) \pi^{n/2}}$$

Posterior predictive distribution

- If $P(\beta) = NIG(\beta|m, V, a, b)$, then the posterior predictive density is a Student or t-distribution

$$\begin{aligned} p(y|x, D) &= \int p(y|x, \beta, \sigma^2) p(\beta, \sigma^2|D) d\beta d\sigma^2 \\ &= St_D(y|u^T m^*, b^*(I + u^T V^* u), a^*) \end{aligned}$$

  where $u = (B_1(x), \ldots, B_k(x))$ and

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

  where $EY = \mu$ and Var $Y = v/(c-2)$.

- I follow the parameterization of Denison p29. This is different from Bishop p115!

Student distribution is a mixture of Gaussians

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

$$St_B(y|\mu, \lambda, \nu) = \int \mathcal{N}(y|\mu, \tau) Ga(\tau|a, b) d\tau$$

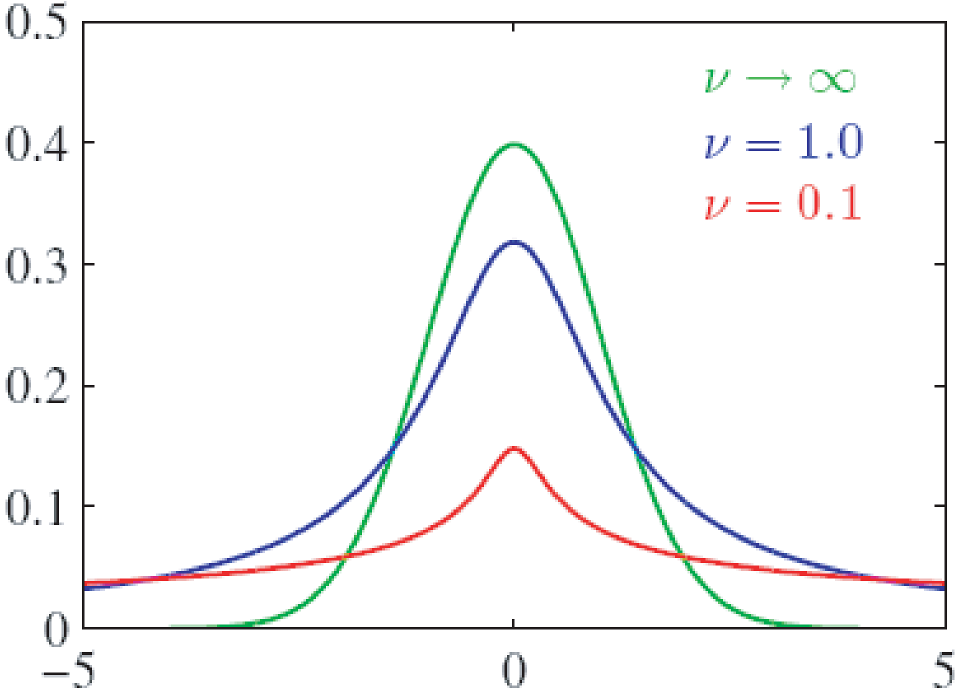
  where $\nu = 2a$ and $\lambda = a/b$ and St_B is $\mathbf{B}$ishop's parameterization

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

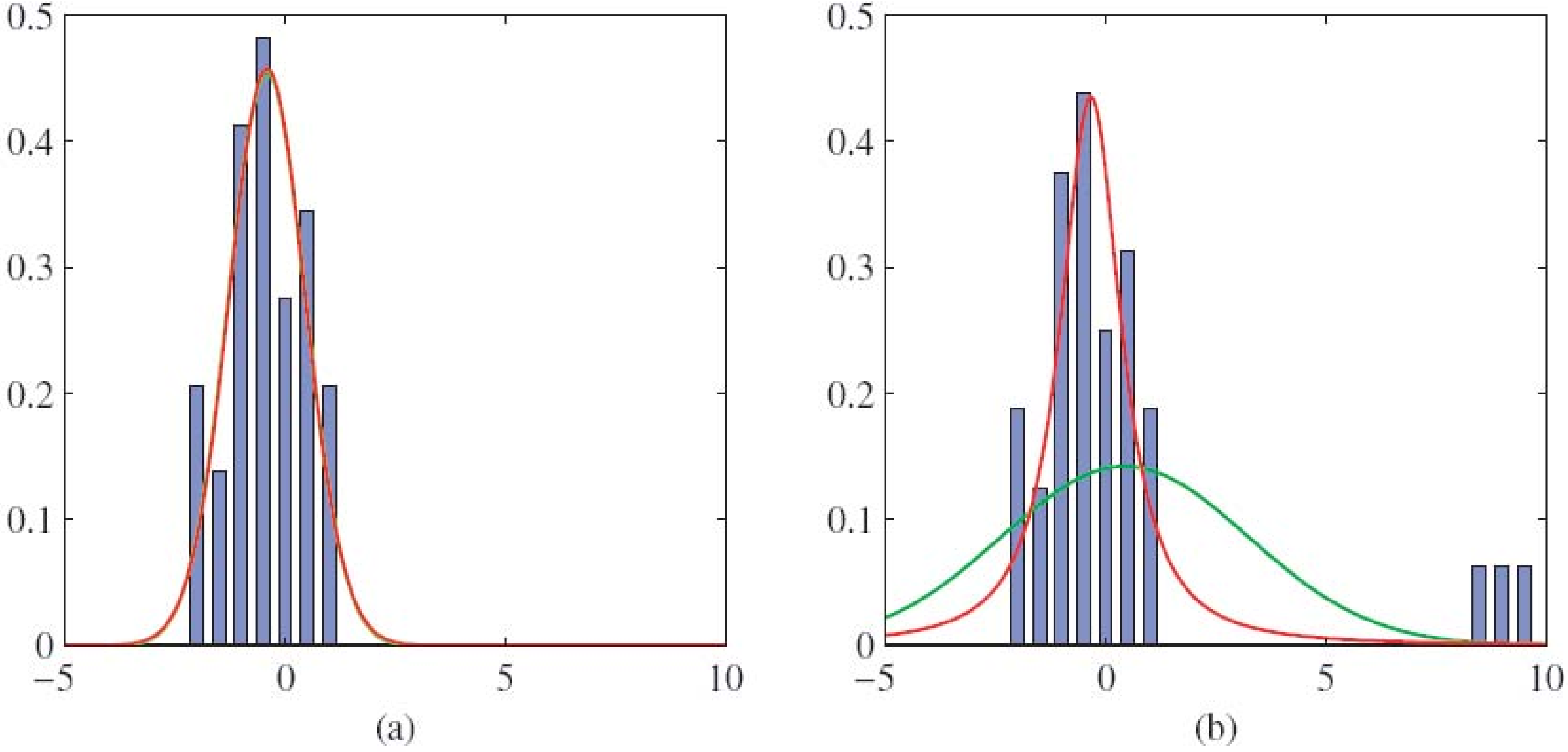
  where $EY = \mu$ and Var$Y = \frac{1}{\lambda}\frac{\nu}{\nu-2}$.

- 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

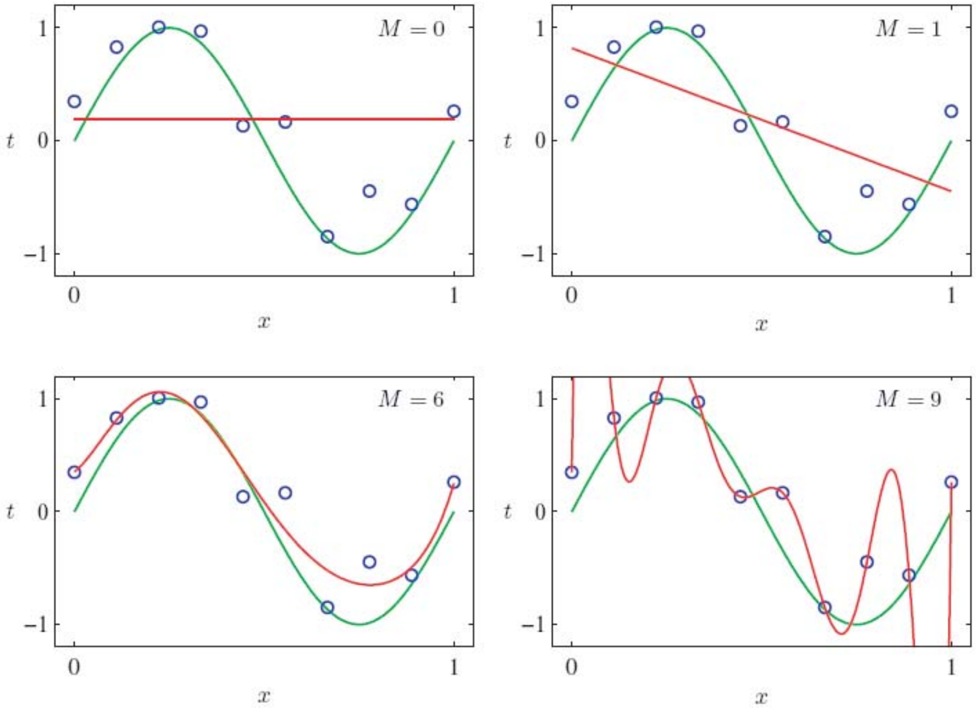


Model selection

- Let model M_k be polynomial regression of order k:

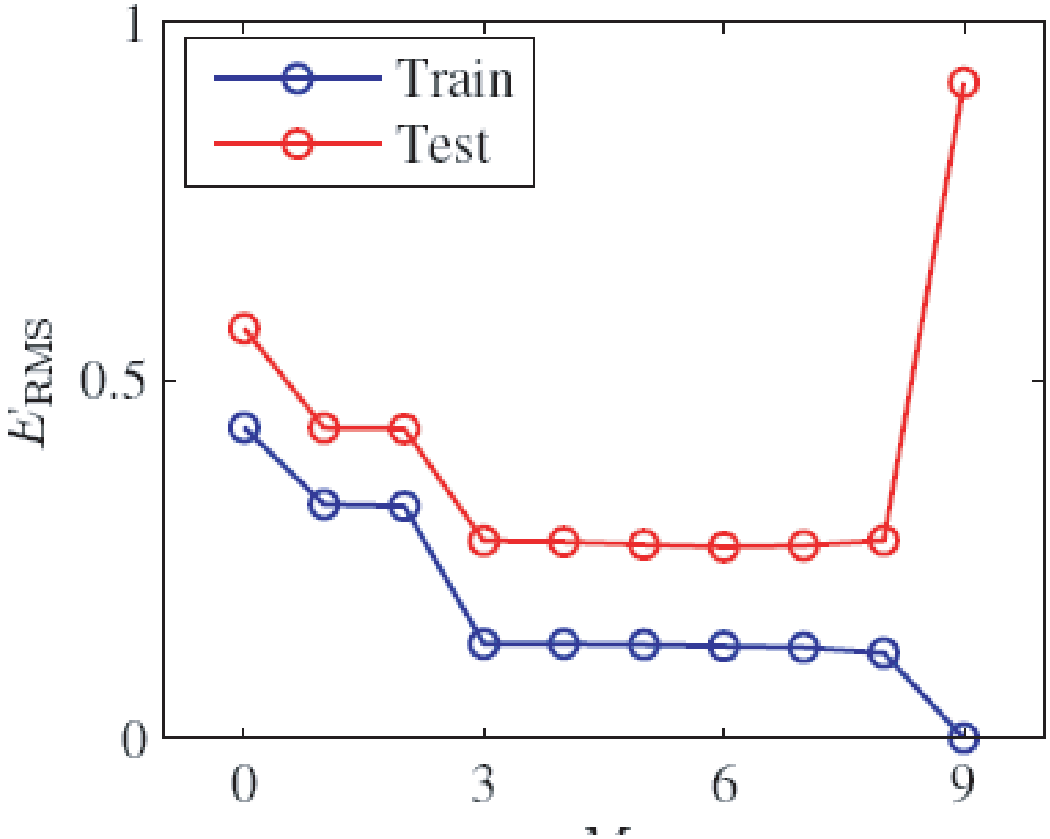
$$Ey = \beta_0 + \sum_{i=1}^{k} \beta_i x^i$$

- Which model should we choose?



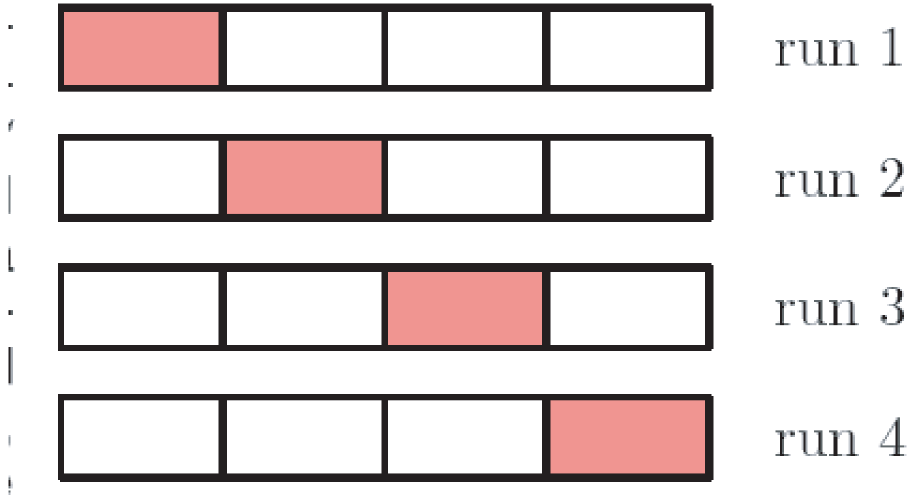
Generalization error

- A complex model will always fit the training data better, but may not generalize to test data. This is called overfitting.



Cross validation

- A simple approach to picking the right model is to compare performance of the different models on a holdout/ validation set.
- If data is scarce, we can use K-fold cross validation, which uses $K/(K-1)$ of the data for training and the rest for testing.
- If $K = N$, this is called leave-one-out cross validation.
- Unfortunately, this is slow, especially if there are many parameters.



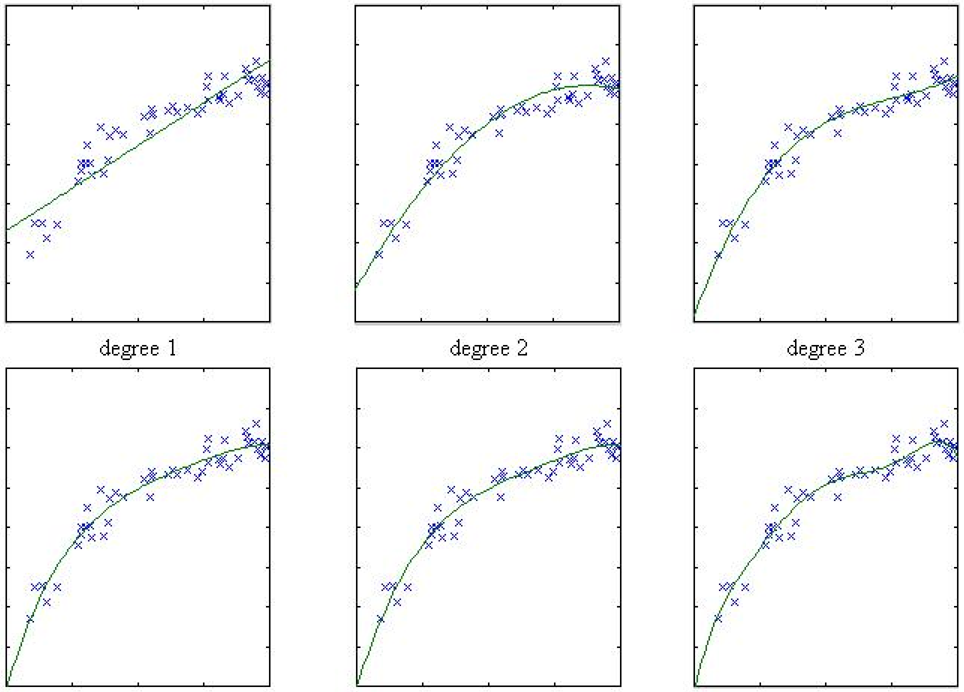
Bayesian model comparison

- 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 $p(\beta, \sigma^2|M_i) = NIG(m_i, V_i, a, b)$, 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 $a^* = a_i^* = a_j^* = a + n/2$.

Bayesian Occam's razor

- Amazingly, even if we have no explicit penalty on complex models (so $P(M_i)$ is uniform), merely by integrating over all possible parameter values (i.e., by using $P(D|M_i) = \int P(D, \theta|M_i) d\theta$), 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".)

Using the marginal likelihood to select K



degree 1

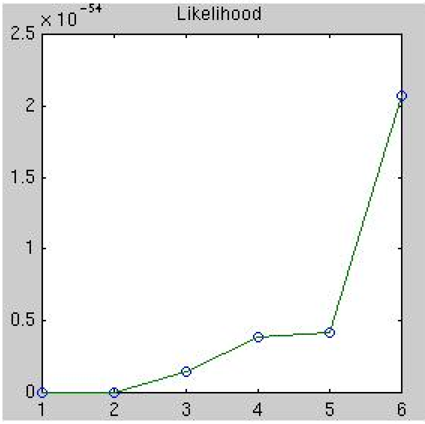
degree 2

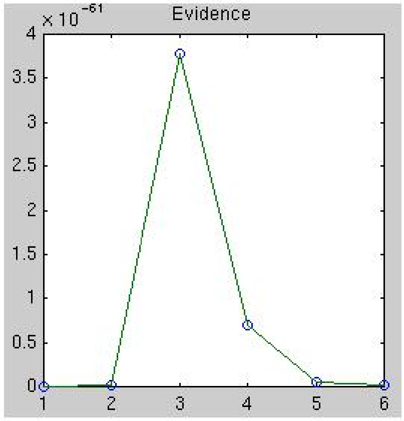
degree 3

degree 4

degree 5

degree 6

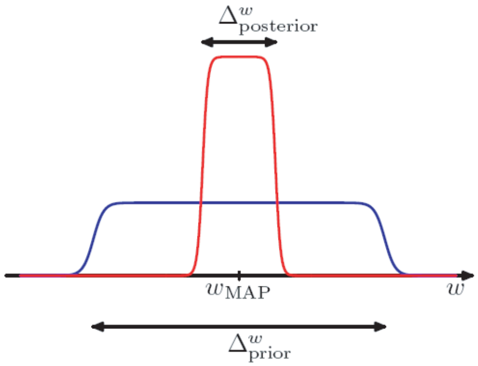




Bayesian Occam's razor

- Let us evaluate the quality of a model with one parameter w using the evidence (marginal likelihood) $p(D) = \int p(D|w)p(w)dw$.
- Suppose the posterior $P(w|D) \propto P(D|w)P(w)$ is sharply peaked around w_{MAP} and has width Δw_{post}. Then we may approximate the integral by the peak times the width.
- Also, suppose the prior is flat with width Δw_{prior}, so $p(w) = 1/\Delta w_{prior}$. Then

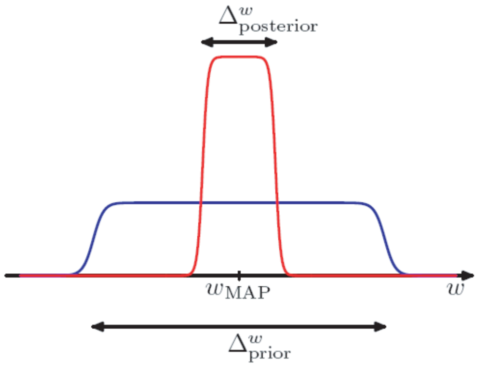
$$p(D) = \int p(D|w)p(w)dw \approx p(D|w_{MAP})\frac{\Delta w_{post}}{\Delta w_{prior}}$$



Occam factors

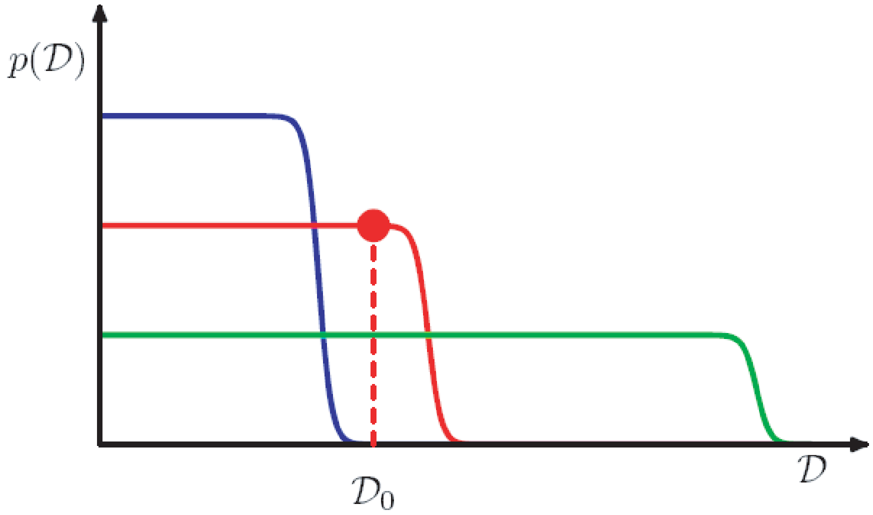
- The ratio $\frac{\Delta w_{post}}{\Delta w_{prior}}$ of posterior accessible volume of the parameter space compared to the prior is called the Occam factor.
- This measures the degree to which the hypothesis space shrinks on arrival of data.
- If in the posterior the parameters have to be finely tuned, then the penalty is large (since $\Delta w_{post}/\Delta w_{prior} \ll 1$).
- If there are M parameters, we may approximate

$$\log p(D) = \log p(D|w_{MAP}) + M \log \frac{\Delta w_{post}}{\Delta w_{prior}}$$



Too simple, too complex, just right

- An overly simple model M_1 has low $P(D|M_1)$ since it has poor fit to the data.
- An overly complex model M_3 has lower $P(D)$ than a medium model M_2, since a complex model spreads its probability mass over more possible datasets.
- We trust an expert who predicts a few *specific* (and correct!) things more than an expert who predicts many things.



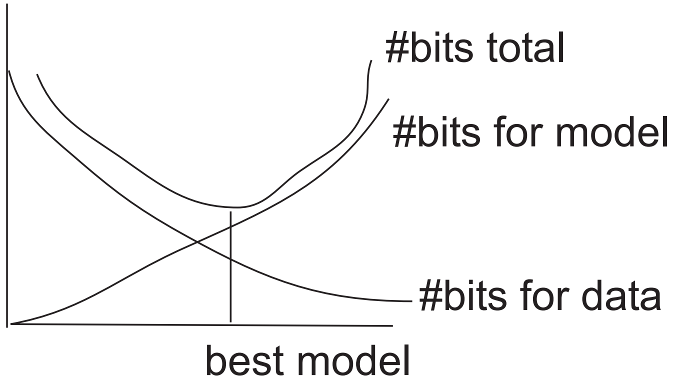
Minimum description length (MDL)

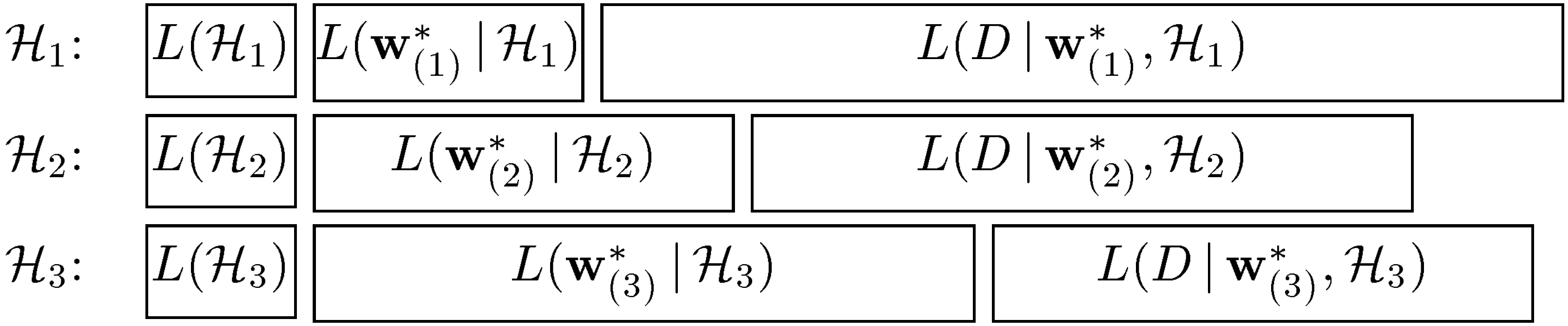
- Another way of thinking about Bayesian Occam's razor is in terms of information theory.
- To losslessly send a message about an event x with probability $P(x)$ takes $L(x) = -\log_2 P(x)$ bits.
- Suppose instead of sending the raw data, you send a model and then the residual errors (the parts of the data not predicted by the model).
- This takes $L(D, H)$ bits:

$$L(D, H) = -\log P(H) - \log(P(D|H))$$

- The best model is the one with the overall shortest message.

Minimum description length (MDL)





$\mathcal{H}_1$: $L(\mathcal{H}_1)$ $L(\mathbf{w}^*_{(1)}\,|\,\mathcal{H}_1)$ $L(D\,|\,\mathbf{w}^*_{(1)},\mathcal{H}_1)$

$\mathcal{H}_2$: $L(\mathcal{H}_2)$ $L(\mathbf{w}^*_{(2)}\,|\,\mathcal{H}_2)$ $L(D\,|\,\mathbf{w}^*_{(2)},\mathcal{H}_2)$

$\mathcal{H}_3$: $L(\mathcal{H}_3)$ $L(\mathbf{w}^*_{(3)}\,|\,\mathcal{H}_3)$ $L(D\,|\,\mathbf{w}^*_{(3)},\mathcal{H}_3)$

Bayesian image interpretation

- How many boxes behind the tree?
- The intrepretation that the tree is in front of one box is much more probable than there being 2 boxes which happen to have the same height and color (suspicious coincidence).
- This can be formalized by assuming (uniform) priors on the box parameters, and computing the Occam factors.