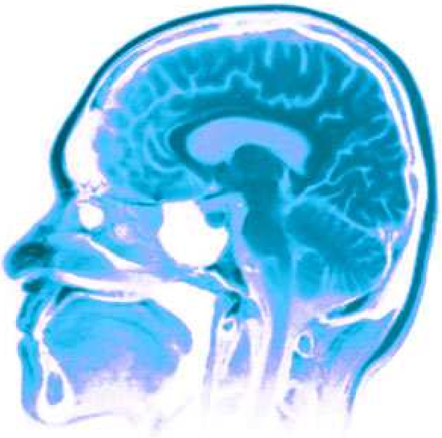
CPSC540



Bayesian Learning

Nando de Freitas

2011

KPM Book Sections: *4*

Learning from positive examples

- Consider the problem of learning to understand the meaning of a word, such as “dog”. Presumably, as a child, one’s parents point out **positive examples** of this **concept**, saying such things as, “look at the cute dog!”, or “mind the doggy”, etc.

- However, it is very unlikely that they provide **negative examples**, by saying “look at that non-dog”.

- Certainly, negative examples may be obtained during an active learning process — the child says “look at the dog” and the parent says “that’s a cat, dear, not a dog” — but psychological research has shown that people can learn concepts from positive examples alone.

The number game

- Consider a simple example of concept learning called the **number game** (Josh Tenenbaum).

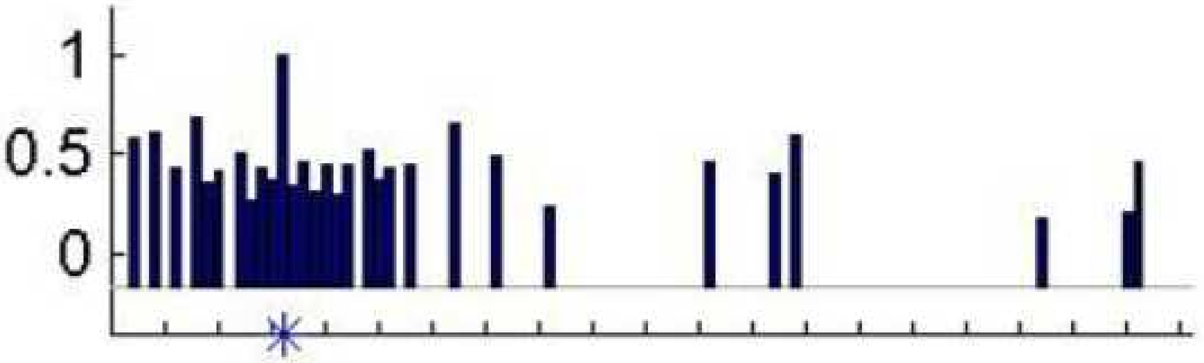
- Suppose I tell you I am thinking of some simple arithmetical concept C, such as "prime number" or "an even number".

- I give you a series of randomly chosen positive examples $\mathcal{D} = \{x_1, \ldots, x_n\}$ drawn from C.

- Then, ask you whether any other test cases $\tilde{x}$ belong to C (i.e., I ask you to classify test examples $\tilde{x}$).

The number game

- Suppose, for simplicity, that all numbers are integers between 1 and 100.
- Suppose I tell you "16" is a positive example of the concept.
- What other numbers do you think are positive? 17? 6? 32? 99?
- 17 is similar, because it is "close by", 6 is similar because it has a digit in common, 32 is similar because it is also even and a power of 2, but 99 does not seem similar. Thus some numbers are more likely than others.



Likelihood

- We must explain why we chose h_{two} :="powers of two", and not, say, h_{even} := "even numbers" after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, given that both hypotheses are consistent with the evidence.

- The key intuition is that we want to avoid **suspicious coincidences**. If the true concept was even numbers, how come we only saw numbers that happened to be powers of two?

- To formalize this, let us assume that examples are sampled uniformly at random from the **extension** of a concept. (The extension of a concept is just the set of numbers that belong to it, e.g., the extension of h_{even} is $\{2, 4, 6, \ldots, 98, 100\}$; the extension of "numbers ending in 9" is $\{9, 19, \ldots, 99\}$.) Tenenbaum calls this the **strong sampling assumption**.

Likelihood

- Given the previous assumption, the probability of independently sampling n items (with replacement) from h is given by

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^n = \left[\frac{1}{|h|}\right]^n$$

- This crucial equation embodies what Tenenbaum calls the **size principle**, which means the model favors the simplest (smallest) hypothesis consistent with the data; this is of course the same as **Occam's razor**.

- To see how it works, let $\mathcal{D} = \{16\}$. Then $p(\mathcal{D}|h_{two}) = 1/6$, since there are only 6 powers of two less than 100, but $p(\mathcal{D}|h_{even}) = 1/50$, since there are 50 even numbers.

- So the likelihood that $h = h_{two}$ is higher than if $h = h_{even}$. After 4 examples, the likelihood of h_{two} is $1/6^4 = 7.7 \times 10^{-4}$, whereas the likelihood of h_{even} is $1/50^4 = 1.6 \times 10^{-7}$.

- This is a **likelihood ratio** of almost 5000:1 in favor of h_{even}.

Priors

- Suppose $D = \{16, 8, 2, 64\}$. The concept $h' =$"powers of two *except* 32" is more likely than $h =$"powers of two".

- h' is the maximum likelihood estimate, since this is the smallest hypothesis consistent with the data.

- However, h' might seem "conceptually unnatural". We can capture such intuition by assigning low prior probability to unnatural concepts.

- Of course, your prior might be different than mine. This **subjective** aspect of Bayesian reasoning is a source of much controversy, since it means, for example, that a child and a math professor will reach different answers.

Posterior

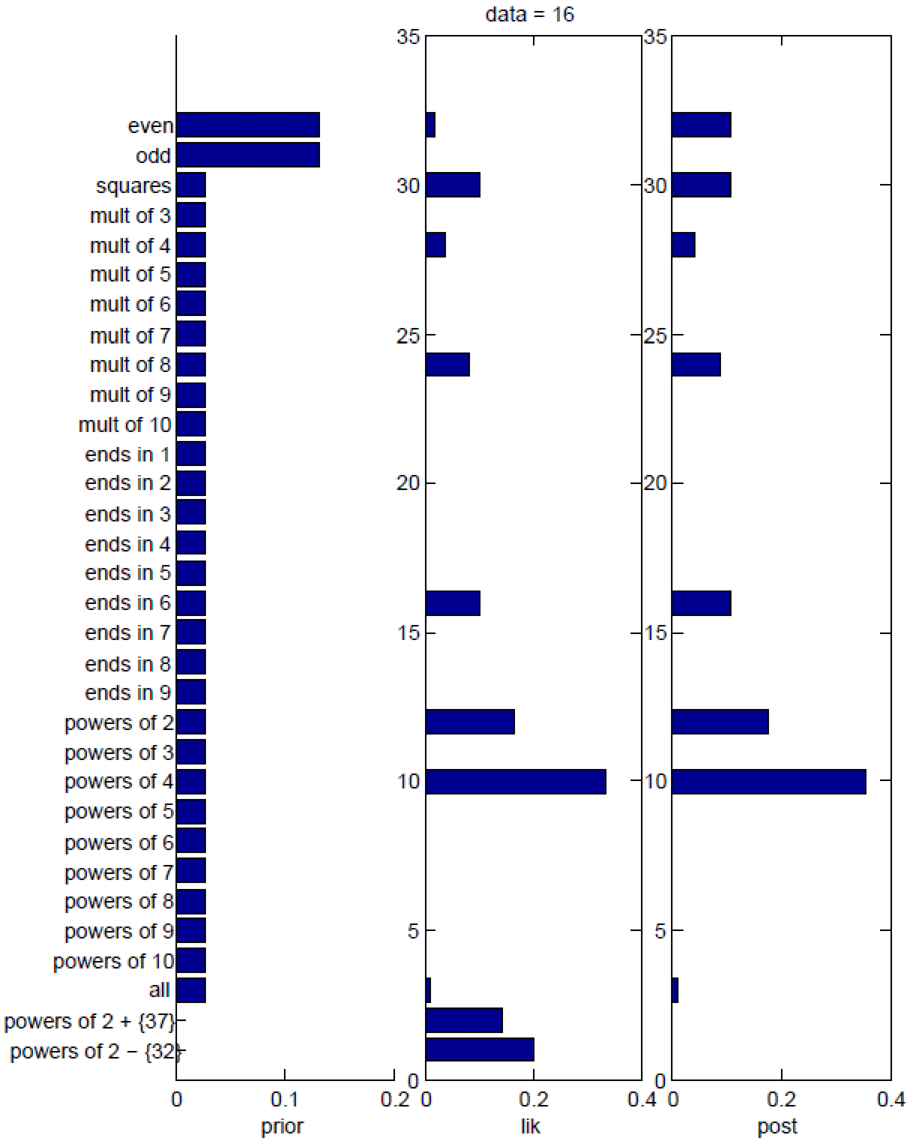
- The posterior is simply the likelihood times the prior, normalized:

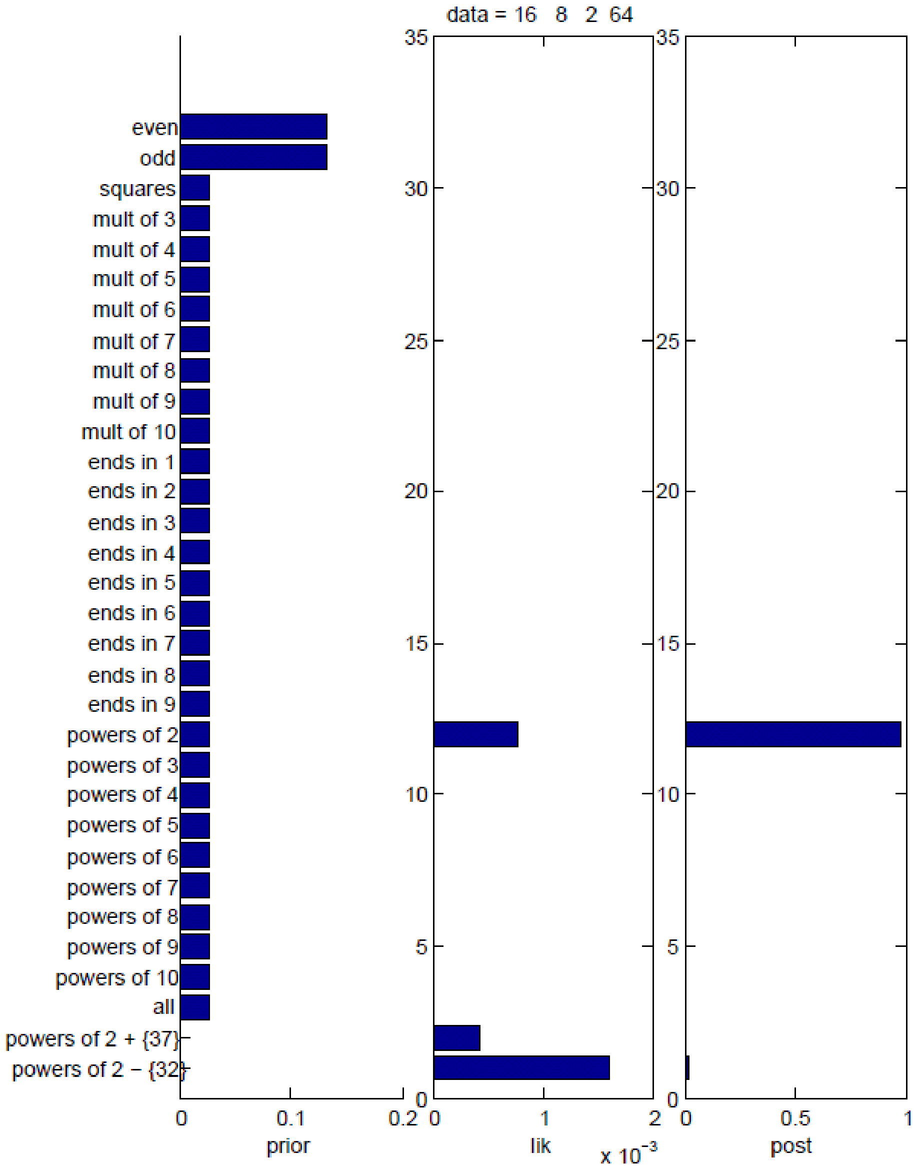
$$p(h|\mathcal{D}) \quad = \quad \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')} = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^n}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(D \in h')/|h'|^n}$$

- For illustration, we use a simple prior which puts support on 30 simple arithmetical concepts, such as “even numbers”, “numbers ending in 9”, etc.

- We include two “unnatural” concepts, namely “powers of 2, plus 37” and “powers of 2, except 32”, but give them low prior weight.

Bayesian updating





Posterior predictive distribution

- The posterior is our internal **belief state** about the world.
- The way to test if our beliefs are justified is to use this to predict observable quantities.
- Specifically, the **posterior predictive distribution** of, say 42, in this context is given by

$$p(y = 42|\mathcal{D}) = \sum_h p(y = 42|h)p(h|\mathcal{D})$$

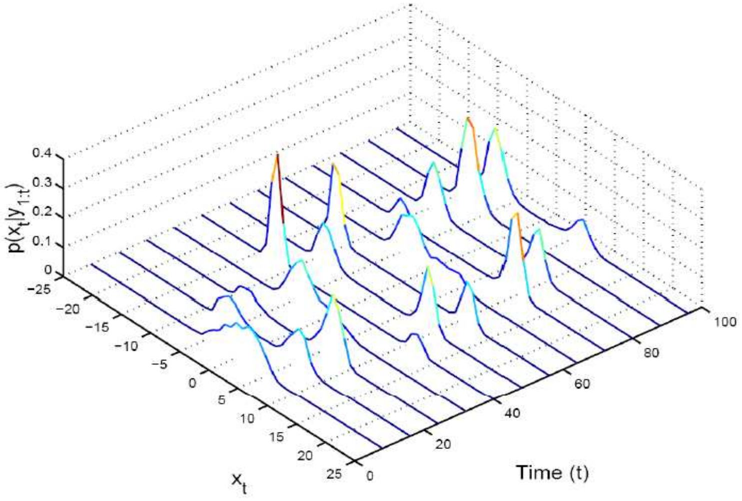
This is just a **weighted average** of the predictions of each individual hypothesis.

Posterior predictive distribution

- If
$$p(h|\mathcal{D}) \approx \delta_{\hat{h}}(h),$$
by the sifting property of delta functions the predictive distribution becomes
$$p(y = 42|\mathcal{D}) = \sum_h p(y = 42|h)\delta_{\hat{h}}(h) = p(y = 42|\hat{h})$$

- This is called a **plugin approximation** to the predictive density and is very widely used, due to its simplicity.

- But note that it can lead to disastrous predictions.



Posterior predictive distribution

- The posterior is our internal **belief state** about the world.
- The way to test if our beliefs are justified is to use this to predict observable quantities.
- Specifically, the **posterior predictive distribution** in this context is given by

  $$p(y=1|\tilde{x},\mathcal{D})=\sum_h p(y=1|\tilde{x},h)p(h|\mathcal{D})$$

  This is just a **weighted average** of the predictions of each individual hypothesis.

Conjugate Bayesian analysis

- We will focus on the use of a special kind of prior known as a **conjugate prior**. We say a prior $p(\boldsymbol{\theta}) \in \mathcal{F}$ is conjugate to a likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ if the resulting posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is also in $\mathcal{F}$.

Likelihood	Prior
Binomial/ Bernoulli	Beta
Multinomial/ multinoulli	Dirichlet
Poisson	Gamma
MVN (fixed $\boldsymbol{\Sigma}$)	MVN
MVN (fixed $\boldsymbol{\mu}$)	Wishart
MVN (general case)	MVN-Wishart
Exponential family	Conjugate

Beta-Bernoulli

- Suppose $X_i \sim \text{Ber}(\theta)$, so $X_i \in \{0, 1\}$. We know that the likelihood has the form

  $$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

  where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails.

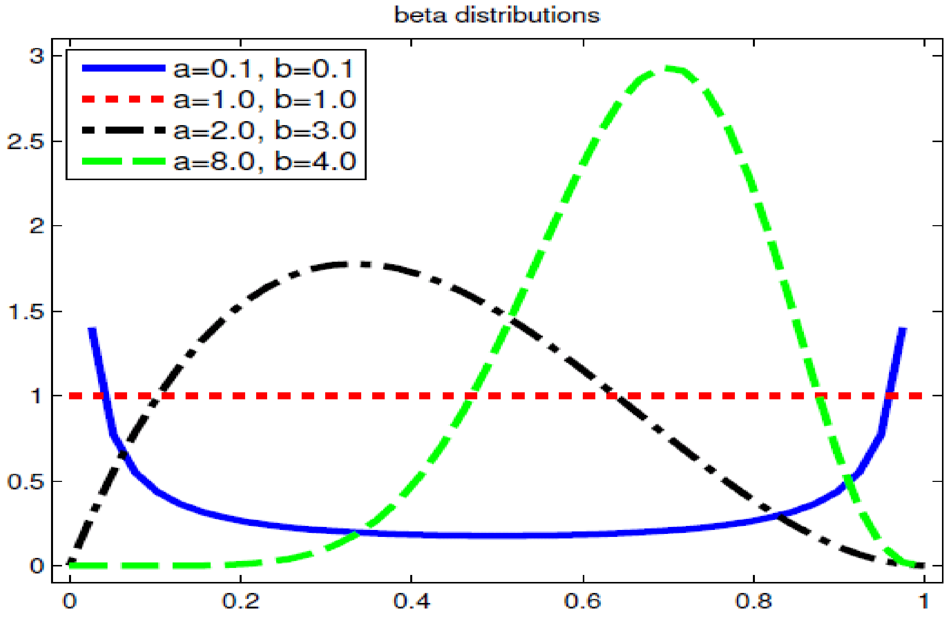
Beta-Bernoulli

- The conjugate prior must be defined over $[0,1]$ and have the form $p(\theta) \propto \theta^{\text{some power}}(1-\theta)^{\text{some power}}$.

- Fortunately, there is such a distribution, known as **beta distribution**. Its pdf is defined as follows:

$$\text{Beta}(\theta|\alpha_1,\alpha_2) = \frac{1}{B(\alpha_1,\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Here $B(p,q)$ is the **beta function**: $B(a,b) := \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.



Beta-Bernoulli

- We require $\alpha_2 > 0$ and $\alpha_1 > 0$ to ensure the distribution is integrable (i.e., to ensure $B(\alpha_1, \alpha_2)$ exists).
- The distribution has the following properties:

$$\text{mean} = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \ \text{mode} \ = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}, \ \text{Var} = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}$$

- The parameters of the prior are called **hyper-parameters**. We can set them in order to encode our prior beliefs.

What is the posterior?

Beta-Bernoulli

- If we multiply the Bernoulli likelihood by the beta prior we get

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)p(\theta) \\ &\propto [\theta^{N_1}(1-\theta)^{N_2}][\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}] \\ &= \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1} \\ &\propto \mathrm{Beta}(\theta|N_1+\alpha_1, N_2+\alpha_2) \end{aligned}$$

- We see that the posterior has the same functional form (beta) as the prior (beta), since it is conjugate.

- The posterior is obtained by adding the prior hyper-parameters α_k to the empirical counts N_k. For this reason, the α_k hyper-parameters are known as **pseudo counts**. The strength of the prior, also known as the **effective sample size** of the prior, is the sum of the pseudo counts, $\alpha_1 + \alpha_2$; this plays a role analogous to the data set size, $N_1 + N_2 = N$.

Beta-Bernoulli

- Updating the posterior sequentially is equivalent to updating in a single batch.

- Suppose we have two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ with sufficient statistics N_1^a, N_2^a and N_1^b, N_2^b. Let $N_1 = N_1^a + N_1^b$, $N_2 = N_2^a + N_2^b$ and $N = N_1 + N_2$.

- In batch mode we have

$$\begin{aligned} p(\theta|\mathcal{D}_1, \mathcal{D}_2) &\propto \mathrm{Bin}(\theta|N_1, N_1 + N_2)\mathrm{Beta}(\theta|\alpha_1, \alpha_2) \\ &\propto \mathrm{Beta}(\theta|N_1 + \alpha_1, N_2 + \alpha_2) \end{aligned}$$

- In sequential mode, we have

$$\begin{aligned} p(\theta|\mathcal{D}_1, \mathcal{D}_2) &\propto p(\mathcal{D}_2|\theta)p(\theta|\mathcal{D}_1) \\ &\propto \mathrm{Bin}(\theta|N_1^b, N_1^b + N_2^b)\mathrm{Beta}(\theta|N_1^a + \alpha_1, N_2^a + \alpha_2) \\ &\propto \mathrm{Beta}(\theta|\ N_1^a + N_1^b + \alpha_1, N_2^a + N_2^b + \alpha_2) \end{aligned}$$

- This makes Bayesian inference particularly well-suited to **online learning**, as we will see later.

Beta-Bernoulli

- The posterior mode, or **MAP estimate**, is given by

$$\hat{\theta}_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N - 2}$$

- By contrast, the posterior mean is given by,

$$\overline{\theta} = \frac{\alpha_1 + N_1}{\alpha_1 + \alpha_2 + N}$$

- If we use a uniform prior, $\alpha_k = 1$, then the MAP estimate reduces to the MLE, but the posterior mean estimate does not.

Beta-Bernoulli

- We will now show that the posterior mean is **convex combination** of the prior mean and the MLE.

- Let the prior mean be $m_1 = \alpha_1/\alpha_0$, where $\alpha_0 = \alpha_1 + \alpha_2$ controls the strength of the prior.

- Then the posterior mean is

$$\mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\alpha_0 m_1 + N_1}{N + \alpha_0} = \frac{\alpha_0}{N + \alpha_0} m_1 + \frac{N}{N + \alpha_0} \frac{N_1}{N} = \lambda m_1 + (1 - \lambda)\hat{\theta}_{ML}$$

where

$$\lambda = \frac{\alpha_0}{N + \alpha_0}$$

is the ratio of the prior to posterior precision (sample size).

Beta-Bernoulli

- Consider predicting the probability of heads in a single future trial under a $\mathrm{Beta}(\alpha_1, \alpha_2)$ posterior.

- We have

$$\begin{aligned} p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\theta) p(\theta|\mathcal{D}) d\theta \\ &= \int_0^1 \theta \, \mathrm{Beta}(\theta|\alpha_1, \alpha_2) d\theta = \mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\alpha_1}{\alpha_1 + \alpha_2} \end{aligned}$$

- With a uniform prior, we have $\alpha_1 = N_1 + 1$ and $\alpha_2 = N_2 + 1$, which gives **Laplace's rule of succession**

$$p(\tilde{x} = 1|\mathcal{D}) = \frac{N_1 + 1}{N_1 + N_2 + 2}$$

This justifies the common practice of adding 1 to the empirical counts.

Beta-Binomial

- Suppose now we were interested in predicting the number of heads, x, in M future trials. This is given by

$$\begin{aligned} p(x|\mathcal{D}, M) &= \int_0^1 \mathrm{Bin}(x|\theta, M)\mathrm{Beta}(\theta|\alpha_1, \alpha_2)d\theta \\ &= \binom{M}{x} \frac{1}{B(\alpha_1, \alpha_2)} \int_0^1 \theta^x (1-\theta)^{M-x} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1} d\theta \end{aligned}$$

  This simplifies to the **beta-binomial** distribution:

$$Bb(x|\alpha_1, \alpha_2, M) := \binom{M}{x} \frac{B(x+\alpha_1, M-x+\alpha_2)}{B(\alpha_1, \alpha_2)}$$

  This distribution has the following mean and variance

$$\begin{aligned} \mathbb{E}[x] &= M \frac{\alpha_1}{\alpha_1 + \alpha_2} \\ \mathrm{var}[x] &= \frac{M\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2} \frac{(\alpha_1 + \alpha_2 + M)}{\alpha_1 + \alpha_2 + 1} \end{aligned}$$

Dirichlet-Multinomial

- The likelihood has the form

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k}$$

where $N_k = \sum_{i=1}^{N} \mathbb{I}(y_i = k)$ is the number of times event k occured.

Dirichlet-Multinomial

- The conjugate prior is the **Dirichlet distribution** which is the natural generalization of the beta distribution to multiple dimensions.

- The pdf is defined as follows:

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \quad := \quad \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\mathbb{I}(\mathbf{x}\in S_K)$$

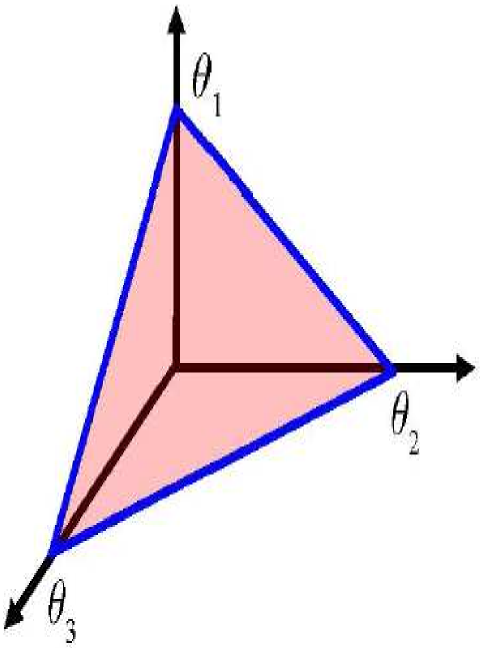
  where S_K is the K-dimensional **probability simplex**, which is the set of vectors such that $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K}\theta_k = 1$.

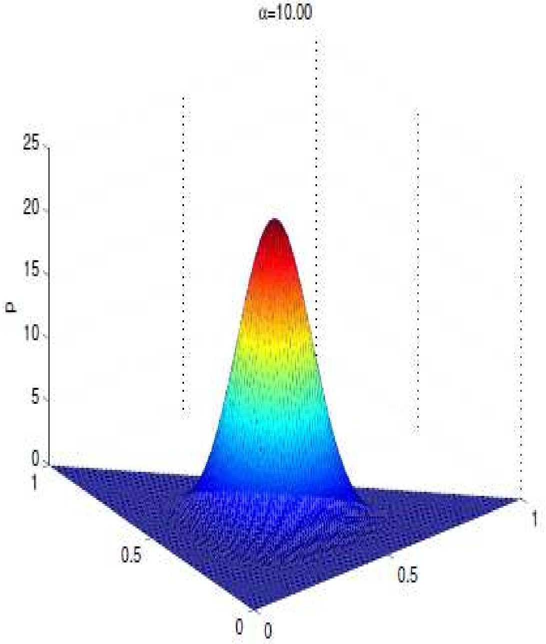
- In addition, $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables:

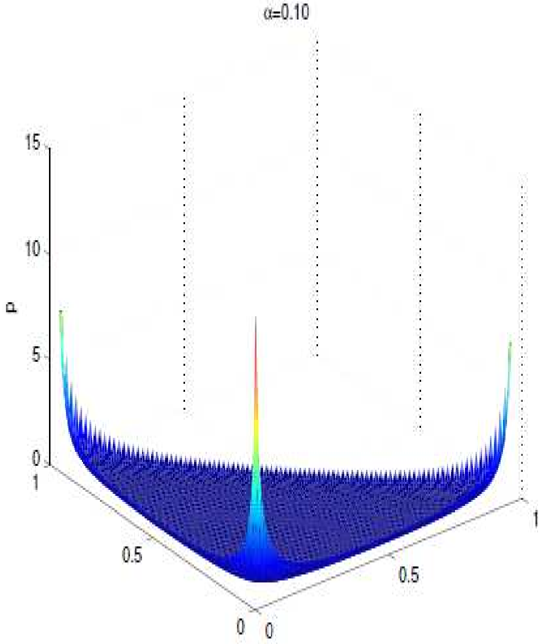
$$B(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

  where $\alpha_0 := \sum_{k=1}^{K}\alpha_k$.

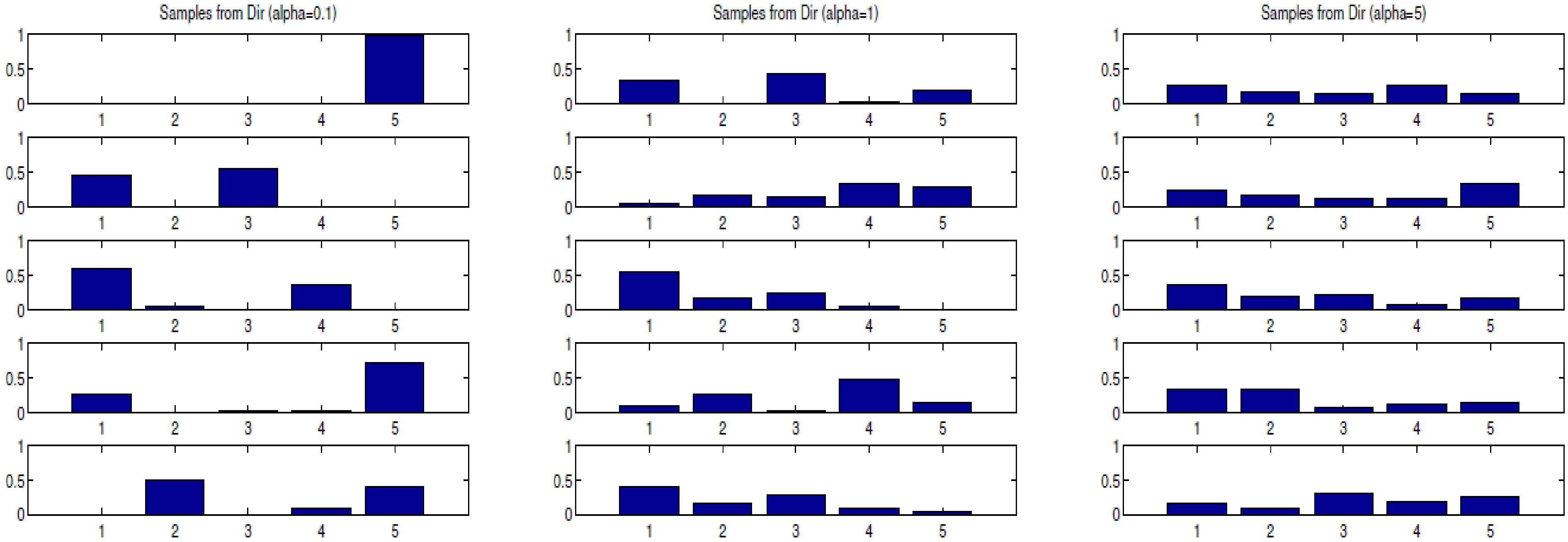
Dirichlet-Multinomial







Dirichlet-Multinomial



Dirichlet-Multinomial

- Multiplying the likelihood by the prior, we find that the posterior is also Dirichlet:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta}) \\ &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \theta_k^{N_k} = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} \\ &= \mathrm{Dir}(\boldsymbol{\theta}|\alpha_1 + N_1, \ldots, \alpha_K + N_K) \end{aligned}$$

- We see that the posterior is obtained by adding the prior hyper-parameters (pseudo-counts) α_k to the empirical counts N_k.

Dirichlet-Multinomial

- The posterior predictive distribution for a categorical variable is given as follows:

$$\begin{aligned} p(X=j|\mathcal{D}) &= \int p(X=j|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &= \int p(X=j|\theta_j)\left[\int p(\boldsymbol{\theta}_{-j},\theta_j|\mathcal{D})d\boldsymbol{\theta}_{-j}\right]d\theta_j \\ &= \int \theta_j p(\theta_j|\mathcal{D})d\theta_j = \mathbb{E}\left[\theta_j|\mathcal{D}\right] \\ &= \frac{\alpha_j+N_j}{N+\sum_k \alpha_k} \end{aligned}$$

  where $E[\theta_j|\mathcal{D}]$ is the j'th component of the posterior mean, and $\boldsymbol{\theta}_{-j}$ are all the components of $\boldsymbol{\theta}$ except θ_j.

- This expression avoids the zero-count problem.

Marginal likelihood (evidence)

- Let $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z_0$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior.

- Let $p(\mathcal{D}|\boldsymbol{\theta}) = q(\mathcal{D}|\boldsymbol{\theta})/Z_\ell$ be the likelihood, where Z_ℓ contains any constant factors in the likelihood.

- Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = q(\boldsymbol{\theta}|\mathcal{D})/Z_N$ be our posterior, where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\boldsymbol{\theta})q(\mathcal{D}|\boldsymbol{\theta})$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior.

- We have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ \frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} &= \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_\ell Z_0 p(\mathcal{D})} \\ p(\mathcal{D}) &= \frac{Z_N}{Z_0 Z_\ell} \end{aligned}$$

Marginal likelihood (evidence)

- Let us apply the above result to the Beta-Binomial model. Since we know $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|\alpha_1', \alpha_2')$, where $\alpha_1' = \alpha_1 + N_1$ and $\alpha_2' = \alpha_2 + N_2$, we know the normalization constant of the posterior is $B(\alpha_1', \alpha_2')$. Hence

$$
\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \\
&= \frac{1}{p(\mathcal{D})} \left[\frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1} \right] \left[\binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} \right] \\
&= \binom{N}{N_1} \frac{1}{p(\mathcal{D})} \frac{1}{B(\alpha_1, \alpha_2)} \left[\theta^{\alpha_1 + N_1 - 1} (1-\theta)^{\alpha_2 + N_2 - 1} \right]
\end{aligned}
$$

Hence

$$
\begin{aligned}
\frac{1}{B(\alpha_1 + N_1, \alpha_2 + N_2)} &= \binom{N}{N_1} \frac{1}{p(\mathcal{D})} \frac{1}{B(\alpha_1, \alpha_2)} \\
p(\mathcal{D}) &= \binom{N}{N_1} \frac{B(\alpha_1 + N_1, \alpha_2 + N_2)}{B(\alpha_1, \alpha_2)}
\end{aligned}
$$

Marginal likelihood (evidence)

- The marginal likelihood for the Dirichlet-Categorical model is given by

$$p(\mathcal{D}) = \frac{B(\mathbf{N} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

- From earlier, we have

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$$

- Hence we can rewrite the above result in the following form, which is what is usually presented in the literature:

$$p(\mathcal{D}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}$$

Bayesian model selection / hypothesis testing

- Suppose we have a set of models $\mathcal{M}$ and we want to know which one is the best. The most natural approach is to compute

  $$m^* = \arg\max_{m \in \mathcal{M}} p(m|\mathcal{D})$$

  This is called **Bayesian model selection**.

- Suppose our prior on models is uniform, $p(m) \propto 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood, $\arg\max p(\mathcal{D}|m)$.

Bayesian model selection / hypothesis testing

- Now suppose we just have two models we are considering, call them the **null hypothesis**, M_0, and the **alternative hypothesis**, M_1.

- Define the **Bayes factor** as the ratio of marginal likelihoods:

$$BF_{1,0} := \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} / \frac{p(M_1)}{p(M_0)}$$

- This is like a **likelihood ratio**, except we integrate out the parameters, which allows us to compare models of different complexity.

- If $BF_{1,0} > 1$ then we prefer model 1, otherwise we prefer model 0.

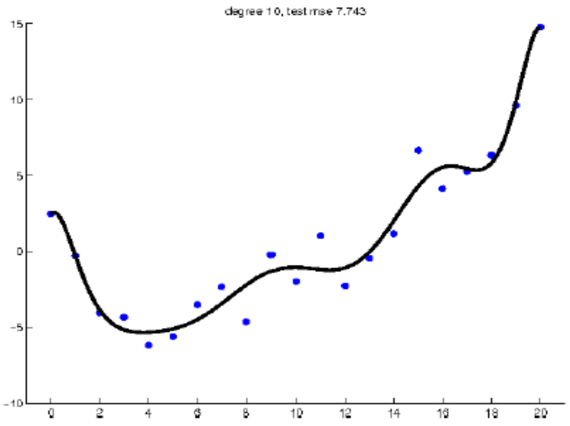
Model selection

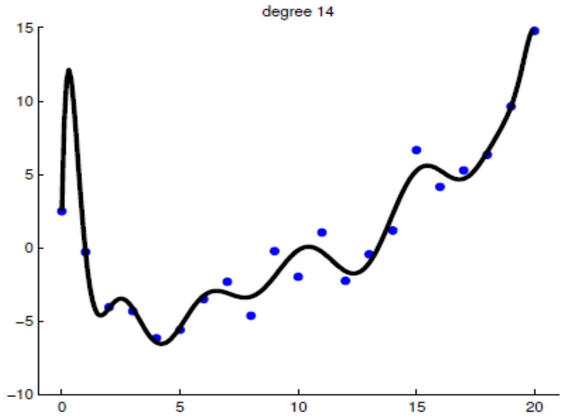
- The obvious approach to picking from a set $\mathcal{M}$ of possible models (also known as the **hypothesis space**) is to fit each one to data (i.e., compute $\hat{\boldsymbol{\theta}}_m$), and then pick the one that fits the best:

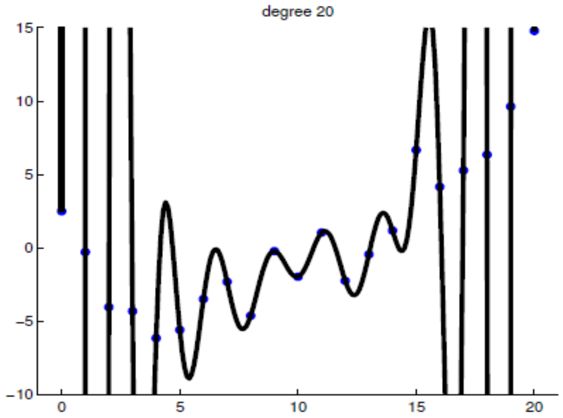
$$m^* = \arg\max_{m \in \mathcal{M}} p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m)$$

- Unfortunately this will not work, since it will always pick the most complex model.

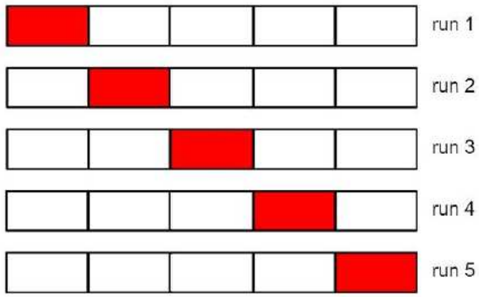
- To see why, note that a model with more parameters has more "capacity" to memorize the training data, and hence can achieve a higher likelihood. However, complex models overfit.







Cross-validation



- A simple fix to the problem is to fit the model on the training set, but to evaluate it on the test set, since we have already seen that models that are too simple or too complex predict poorly on the test set (the U-shaped curve). But how to get a test set?

- A simple but popular solution to this is to use **cross validation** (**CV**).

- The idea is simple: we split the training data into K **folds**; then, for each fold $k \in \{1, \ldots, K\}$, we train on all the folds but the k'th, and test on the k'th, in a round-robin fashion.

- It is common to use $K = 5$; this is called 5-fold CV.

- If we set $K = N$, then we get a method called **leave-one out cross validation**, or **LOOCV**, since in fold i, we train on all the data cases except for i, and then test on i.

Example: Ridge regression

- In polynomial regression, we can encourage the parameters to be small by optimizing the following **penalized least squares** objective function

  $$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2 + \lambda||\mathbf{w}||_2^2$$

  where $||\mathbf{w}||_2^2 = \sum_{j=1}^{D} w_j^2$ is the ℓ_2 norm of the weight vector.

- (Note that the offset term w_0 is not regularized, since this just affects the height of the function, not its complexity.)

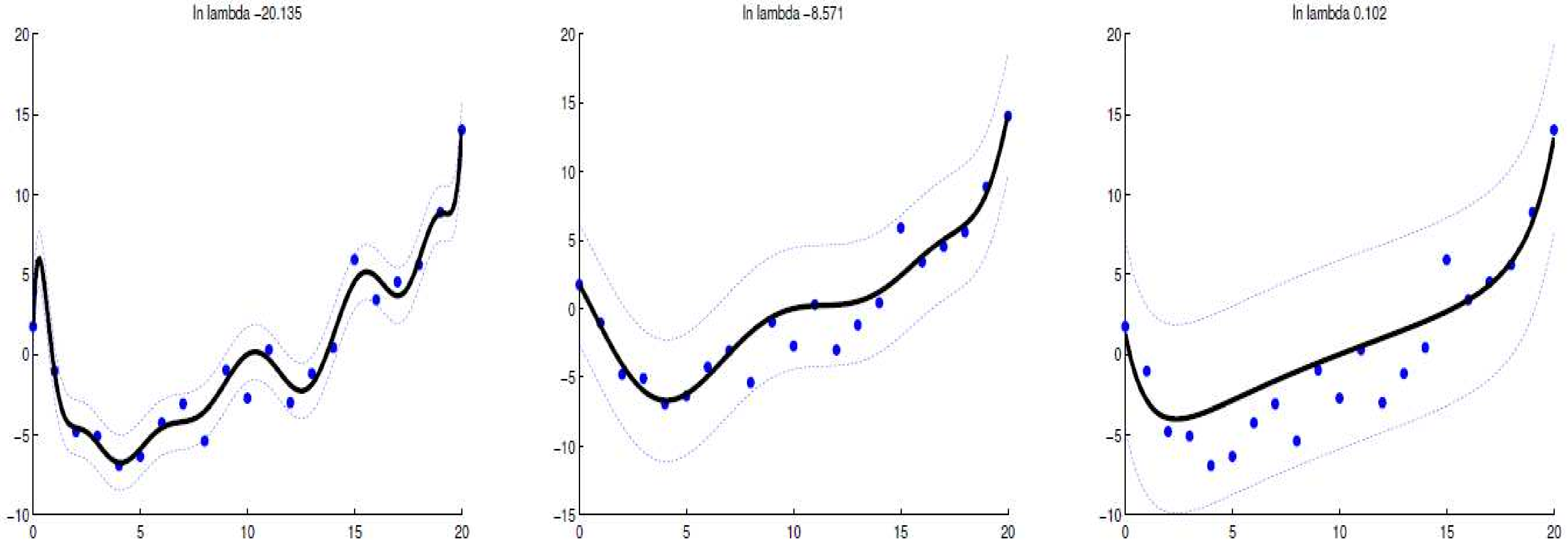
- Here the first term is the MSE as usual, and the second term is a complexity penalty.

- $\lambda \geq 0$ controls the strength of the penalty.

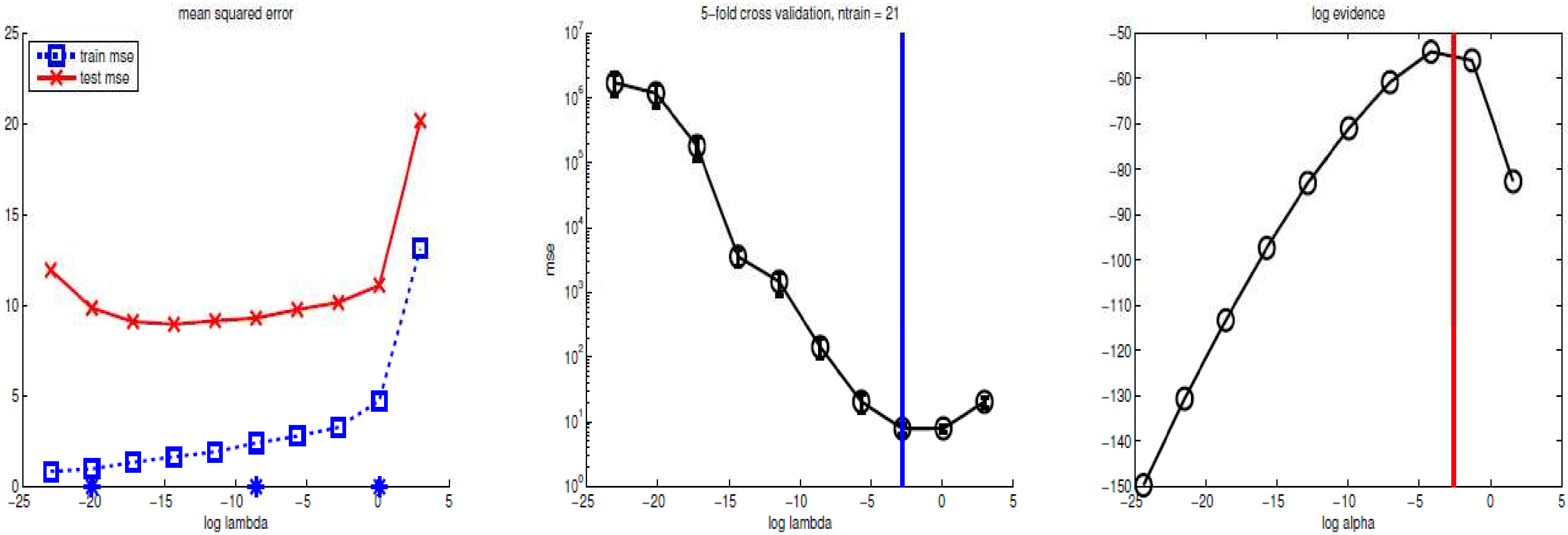
- This technique is known as **ridge regression**, and more generally as ℓ_2 **regularization** or **weight decay**.

Example: Ridge regression

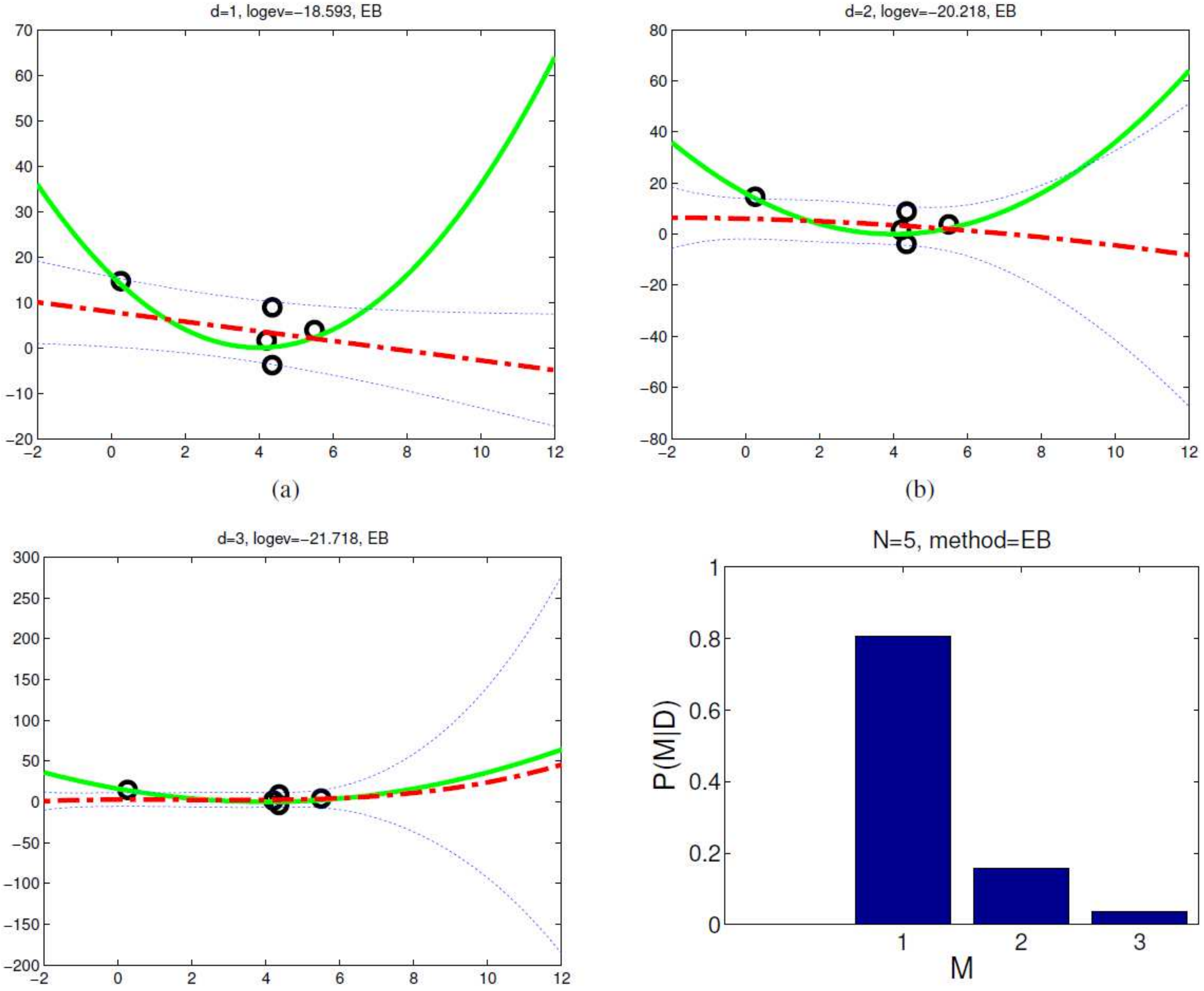
- Regression with polynomial of degree 14.



Example: Ridge regression



Example of Bayesian model selection



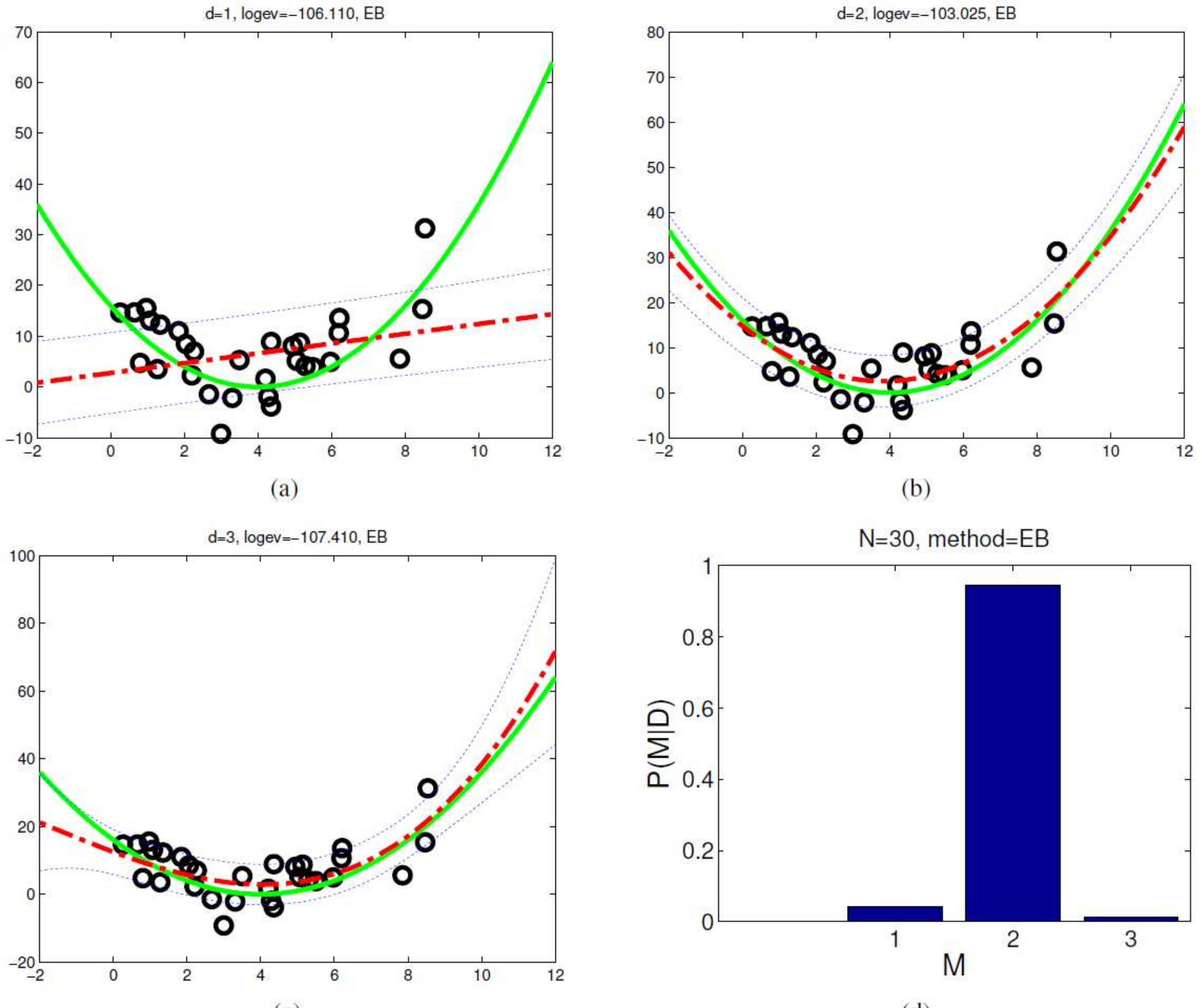
(a)

(b)

(c)

(d)

Example of Bayesian model selection



d=1, logev=−106.110, EB

d=2, logev=−103.025, EB

d=3, logev=−107.410, EB

N=30, method=EB

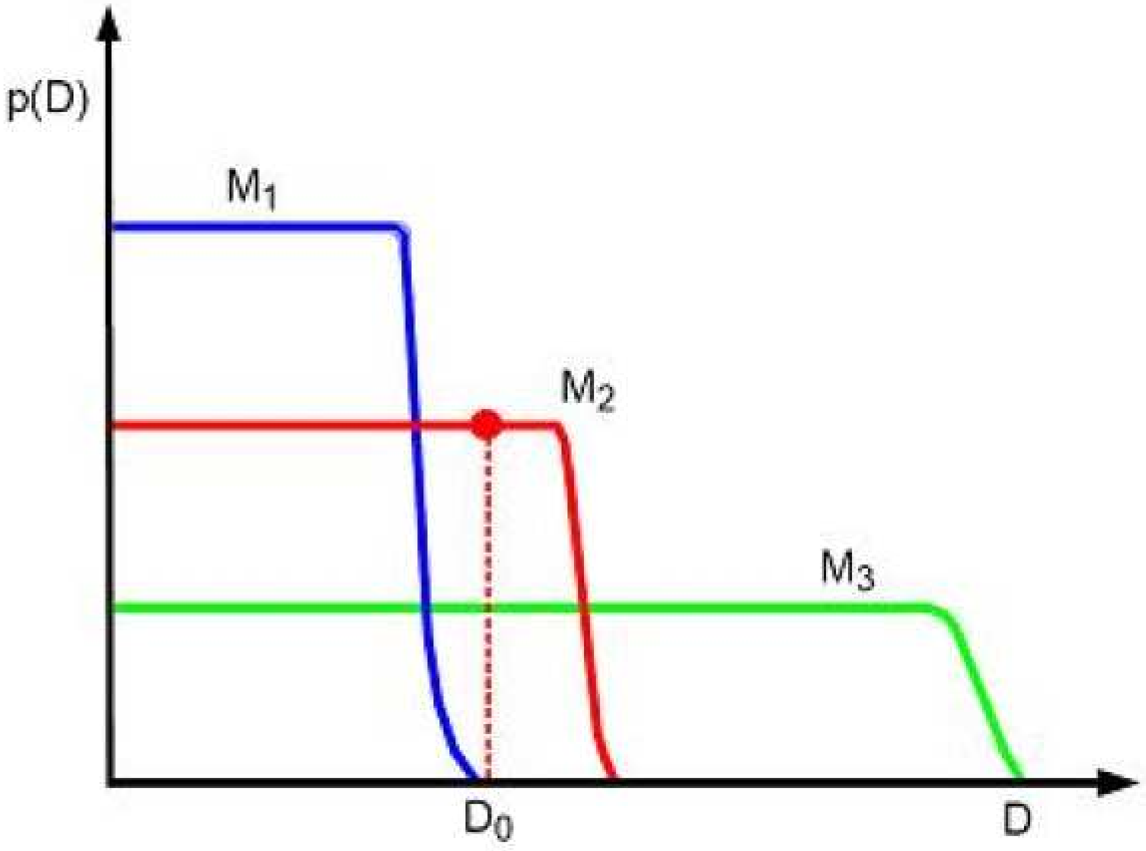
(a)

(b)

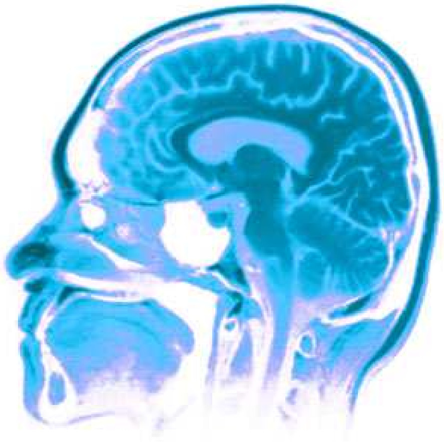
(c)

(d)

Occam’s razor



Next class



Bayesian Learning and the Multivariate Gaussian

Nando de Freitas

2011

KPM Book Sections: *5*