The probabilistic approach in data modeling

Maurizio Filippone

Department of Computer Science - The University of Sheffield

Machine Learning Group

October 21st, 2009

Outline

Random variables

- Random variables: results of non exactly reproducible experiments
- Either intrinsically random or the system is incompletely known, cannot be controlled precisely
- The probability of an experiment taking a certain value is the frequency with which that value is taken in the limit of infinite experimental trials
- Alternatively, we can take probability to be our belief that a certain value will be taken

Informal definition

- A probability distribution is a rule associating a number $0 \leq p(x) \leq 1$ to each state $x \in \Omega$, such that the sum of the probabilities of all possible experimental outcomes is 1:

$$\sum_{x \in \Omega} p(x) = 1$$

- Ω can be discrete or continuous

Rules

- Sum rule: the marginal probability $p(x)$ is given by summing the joint $p(x, y)$ over all possible values of y:

$$p(x) = \sum_{y \in \Omega} p(x, y)$$

- Product rule: the joint is the product of the conditional and the marginal:

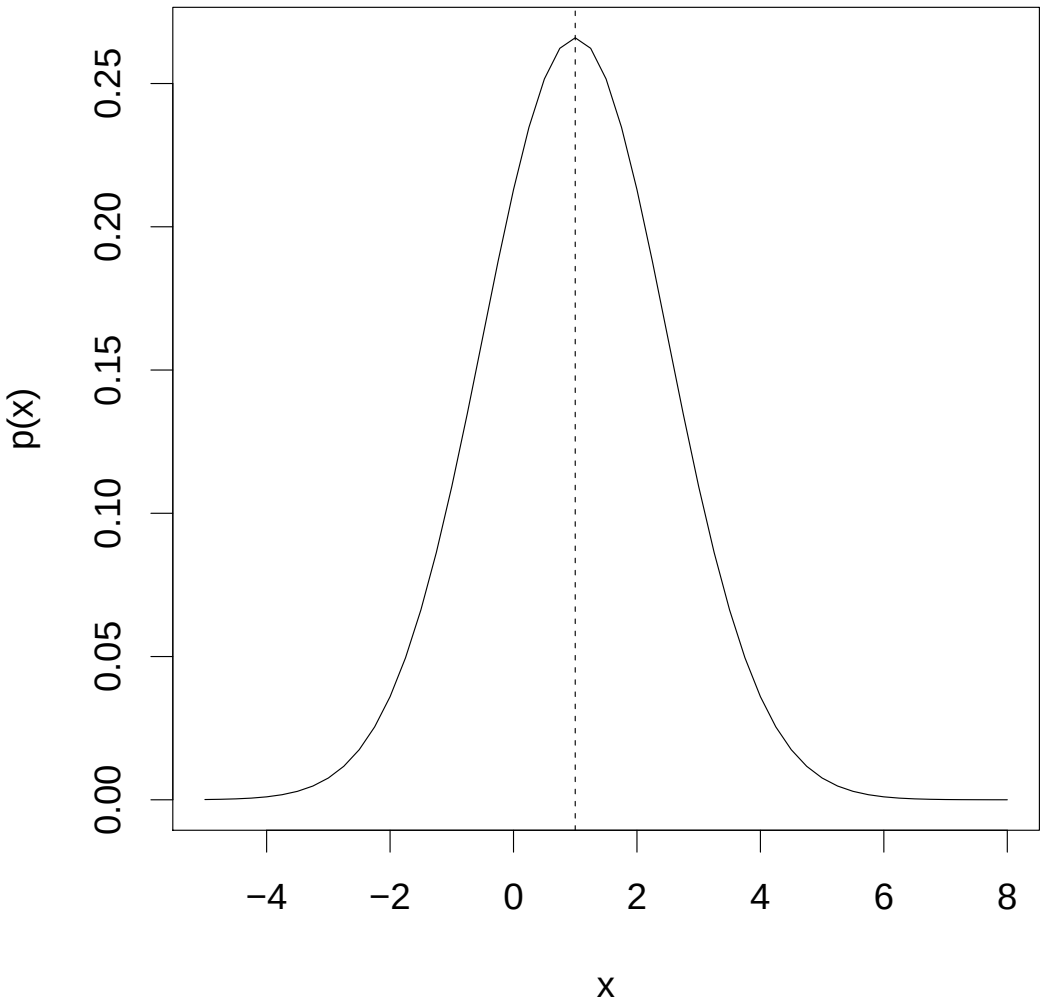
$$p(x, y) = p(x|y)p(y)$$

- Bayes' rule: the posterior is the ratio of the joint and the marginal:

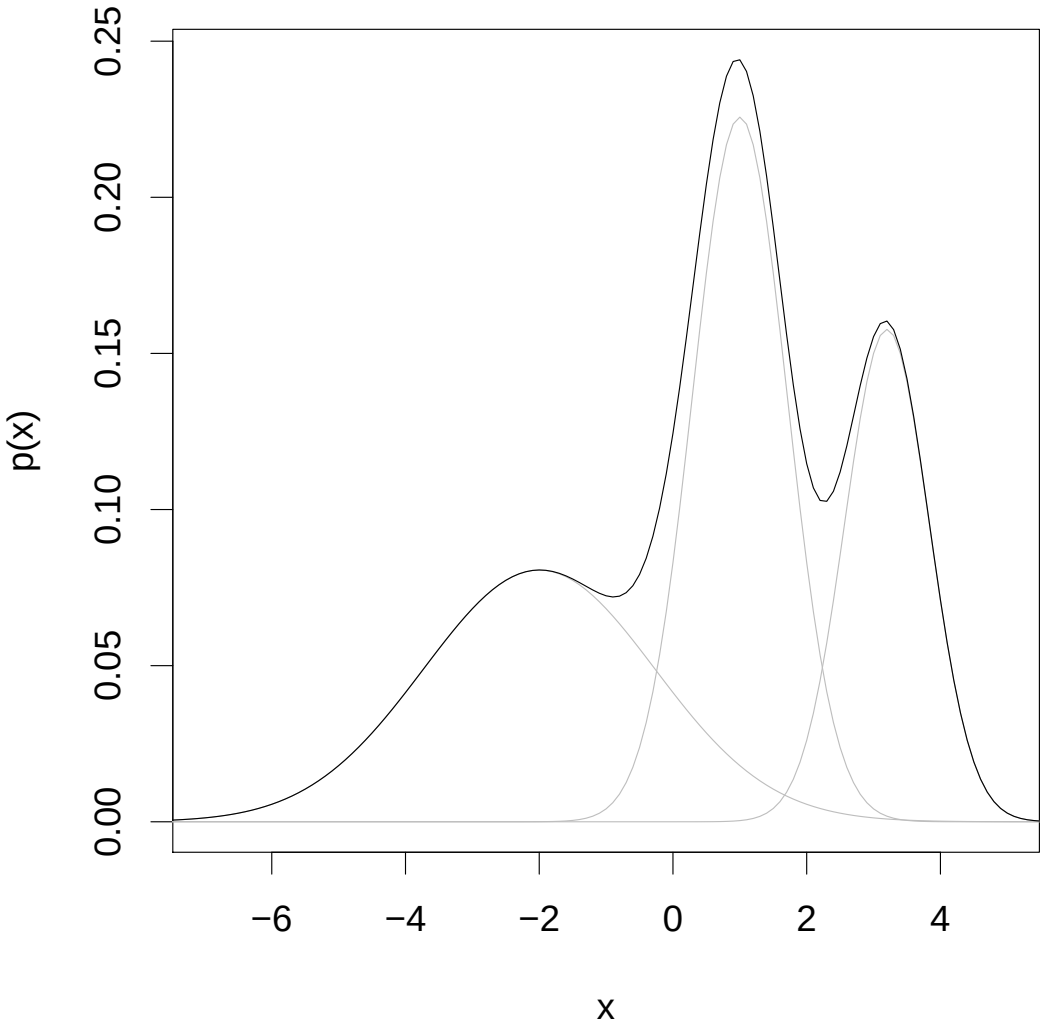
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- NOTE: Bayes' rule is a direct consequence of the product rule

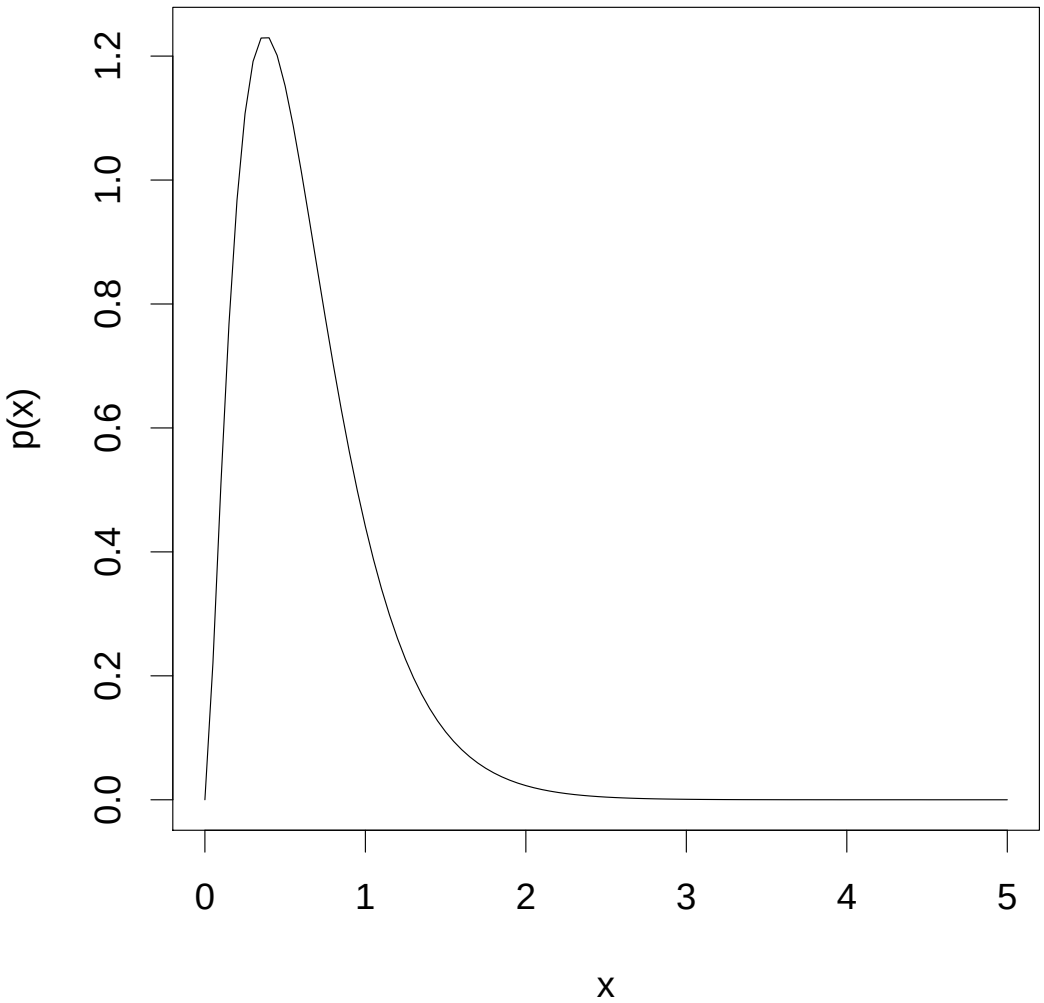
Examples of distributions - Gaussian



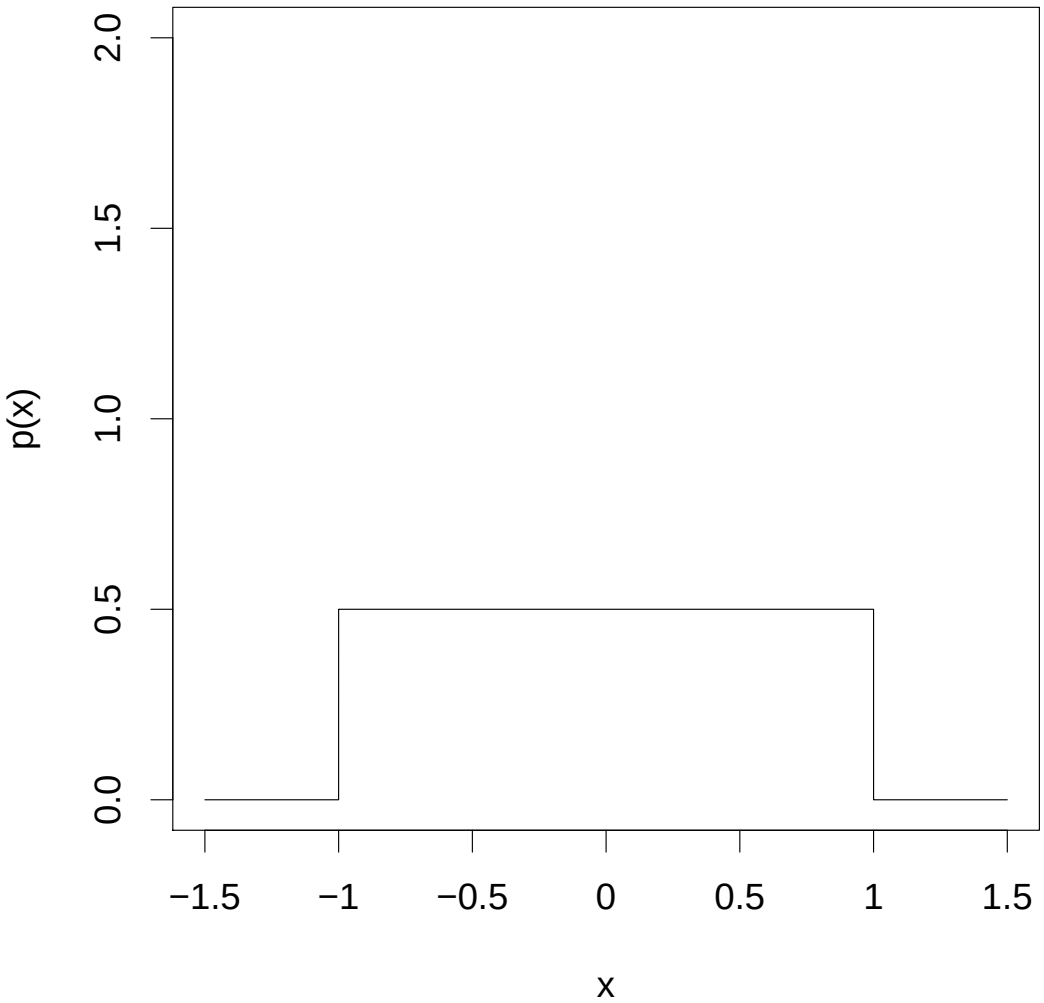
Examples of distributions - Mixture of Gaussians



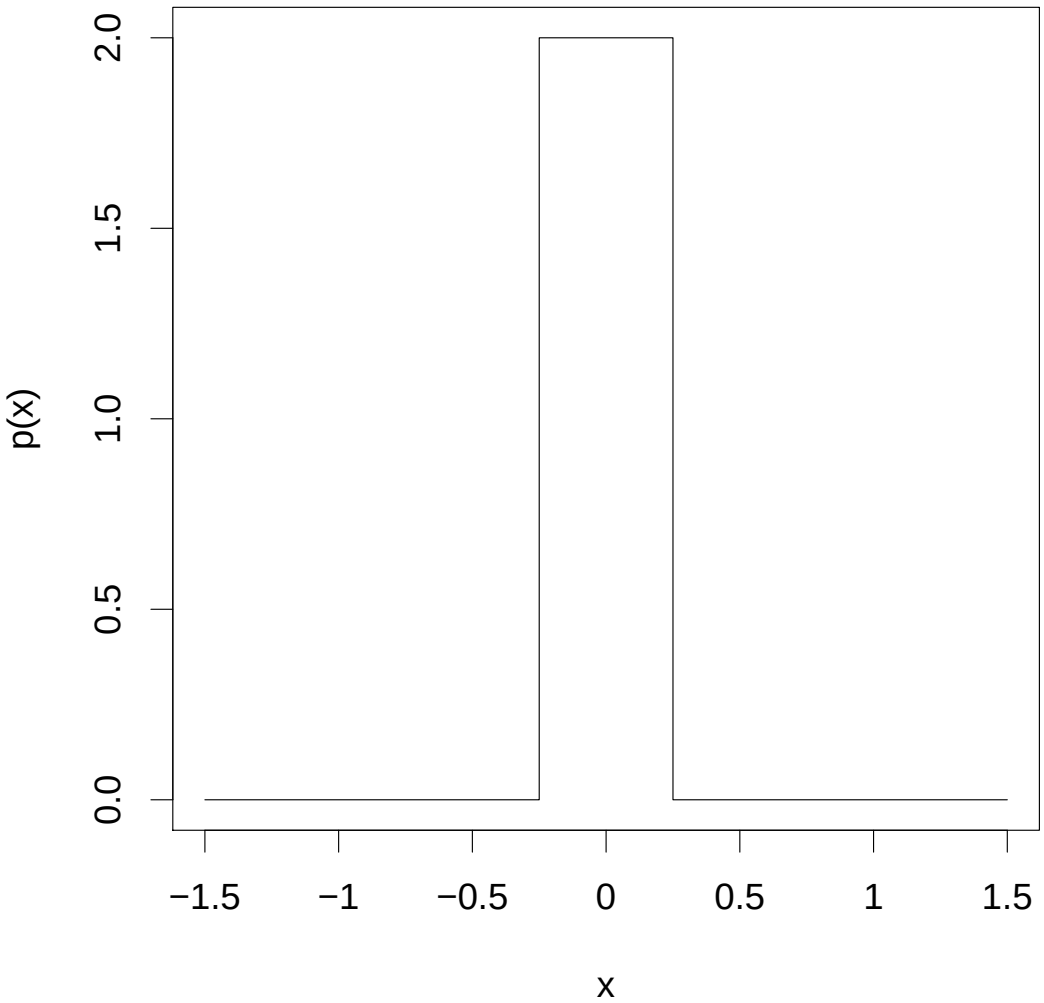
Examples of distributions - Gamma



Examples of distributions - Uniform



Examples of distributions - Uniform



Independence

- Two random variables x and y are independent if their joint probability factorizes in terms of marginals

$$p(x, y) = p(x)p(y)$$

- Using the product rule, this is equivalent to the conditional being equal to the marginal

$$p(x|y) = p(x)$$

What do we do in data modeling?

- fit probability distributions to a dataset
- regression
- classification
- clustering/dimensionality reduction

Likelihood

- Set of observations (in $\mathbb{R}$):

$$X = \{x_1, \ldots, x_n\}$$

- Assume a parametric form for $p(x)$, i.e. $p(x) = p(x|\theta)$
- The likelihood is defined as:

$$L = p(x_1, \ldots, x_n|\theta) = p(X|\theta)$$

Likelihood - i.i.d. case

- In the case of independent and identically distributed data:
  - independent:
$$p(x_i|x_j) = p(x_i) \qquad \forall i, j$$
  - identical:
$$p(x_i) = p(x_j) \qquad \forall i, j$$
- The likelihood becomes:
$$L = \prod_{i=1}^{n} p(x_i|\theta)$$

Maximum Likelihood Approach

- Given the likelihood
$$L = p(X|\theta)$$
- ML yields an estimate $\hat{\theta}$ of θ by maximizing L
- ML estimators converge to the true parameters as $n \to +\infty$

Problems with the ML approach

- does not take into account the size of the data set
- in many cases the direct maximization of L in intractable. In such cases usually we introduce latent variables and optimize the likelihood using an iterative algorithm: Expectation Maximization (EM)
- the likelihood can have several local maxima

Problems with the ML approach

- Let's toss a fair coin three times
- Let "TTT" be the sequence of coin tosses
- ML would estimate $p(H) = 0$ and $p(T) = 1$
- ML does not take into account the size of the data set

The frequentist view

- If we repeat an experiment many times, we will obtain different sets of observations and therefore we will estimate different values of the model parameters
- The frequentist view aims at describing the variability of parameters given to the finite sample size
- This leads naturally to statistical testing

Example

- Let's toss a fair coin three times obtaining "TTTT"
- We can ask the following question:
  How likely is to observe the sequence "TTTT" given that I assume the coin to be fair?

Example

- In the hypothesis H_0 that the coin is fair, the probability of "TTTT" is:

$$p_{value} = 1/16 = 0.0625$$

- the p_{value} is the probability to observe a particular dataset, given H_0

$$p_{value} = p(\text{“}TTTT\text{”}|H_0)$$

- if we set a significance level $\alpha = 0.05$ we can't reject H_0

Statistical testing

- In general, statistical testing requires a null hypothesis H_0 (and an alternative hypothesis H_1)
- Compute the statistics we want to test
- Set a significance level α and compute the p_{value}
- if the p_{value} is smaller than α then reject H_0

The Bayesian view

- The likelihood is:

$$L = p(x_1, \ldots, x_n|\theta) = p(X|\theta)$$

- We can use Bayes' theorem in the following way:

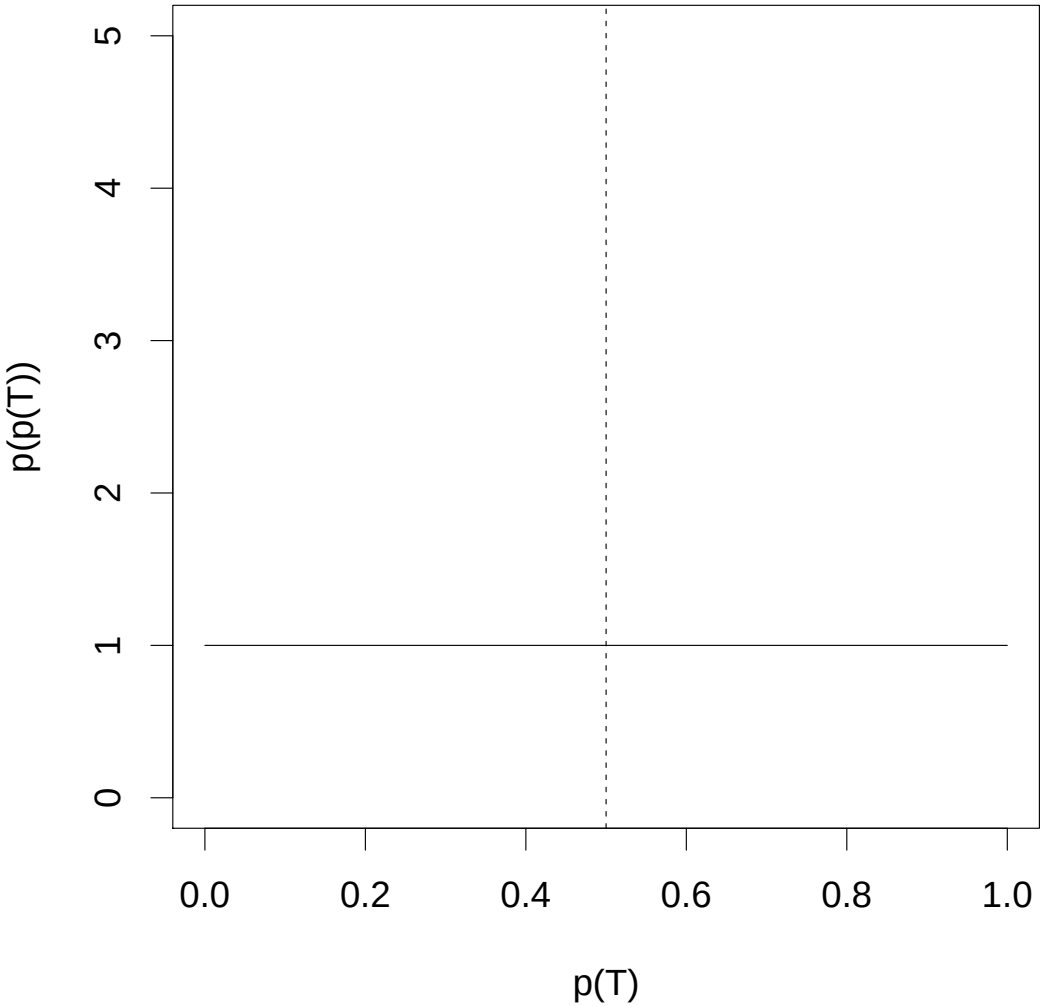
$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

- prior: $p(\theta)$
- posterior: $p(\theta|X)$
- normalization: $p(X)$

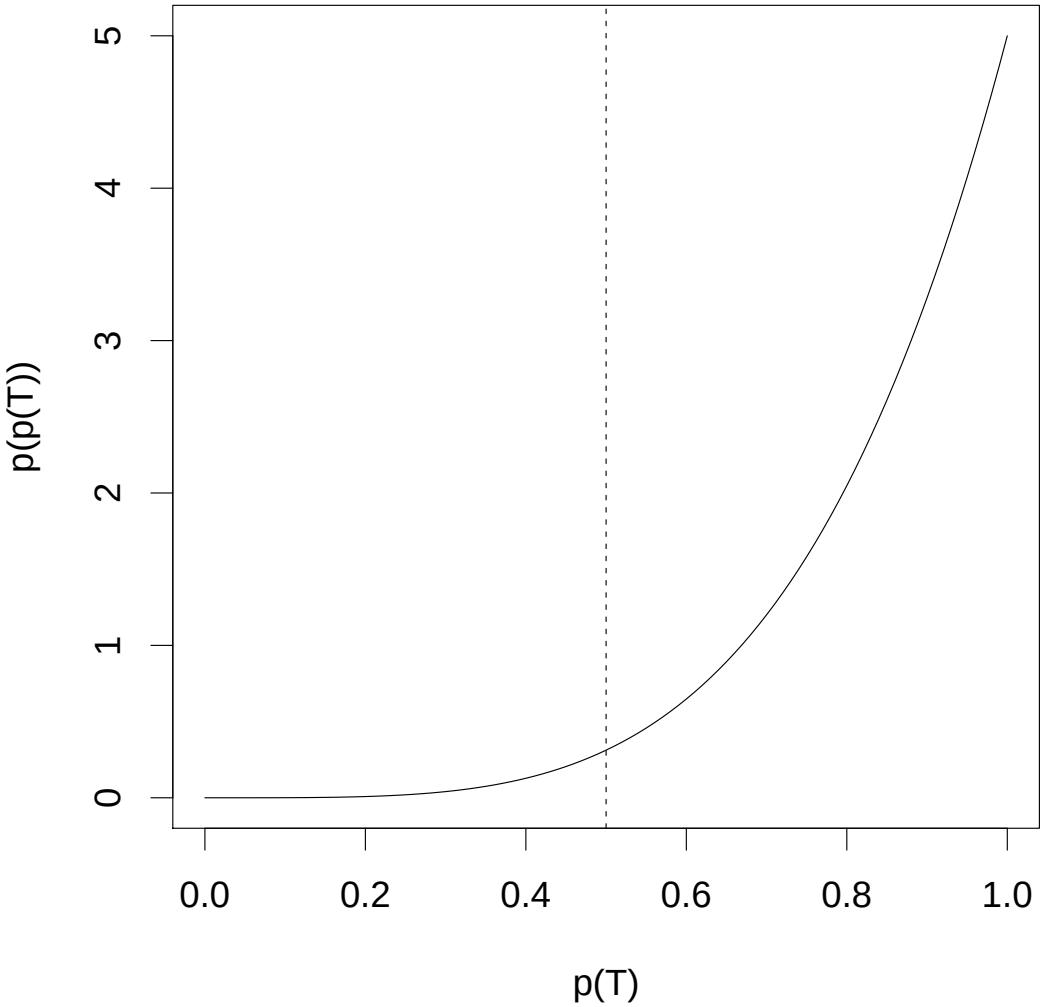
Advantages

- distribution over the parameters
- easy to incorporate prior knowledge
- incremental learning is natural (posterior becomes the prior for new observed data)
- no hand-tuning of extra parameters
- avoids overfitting

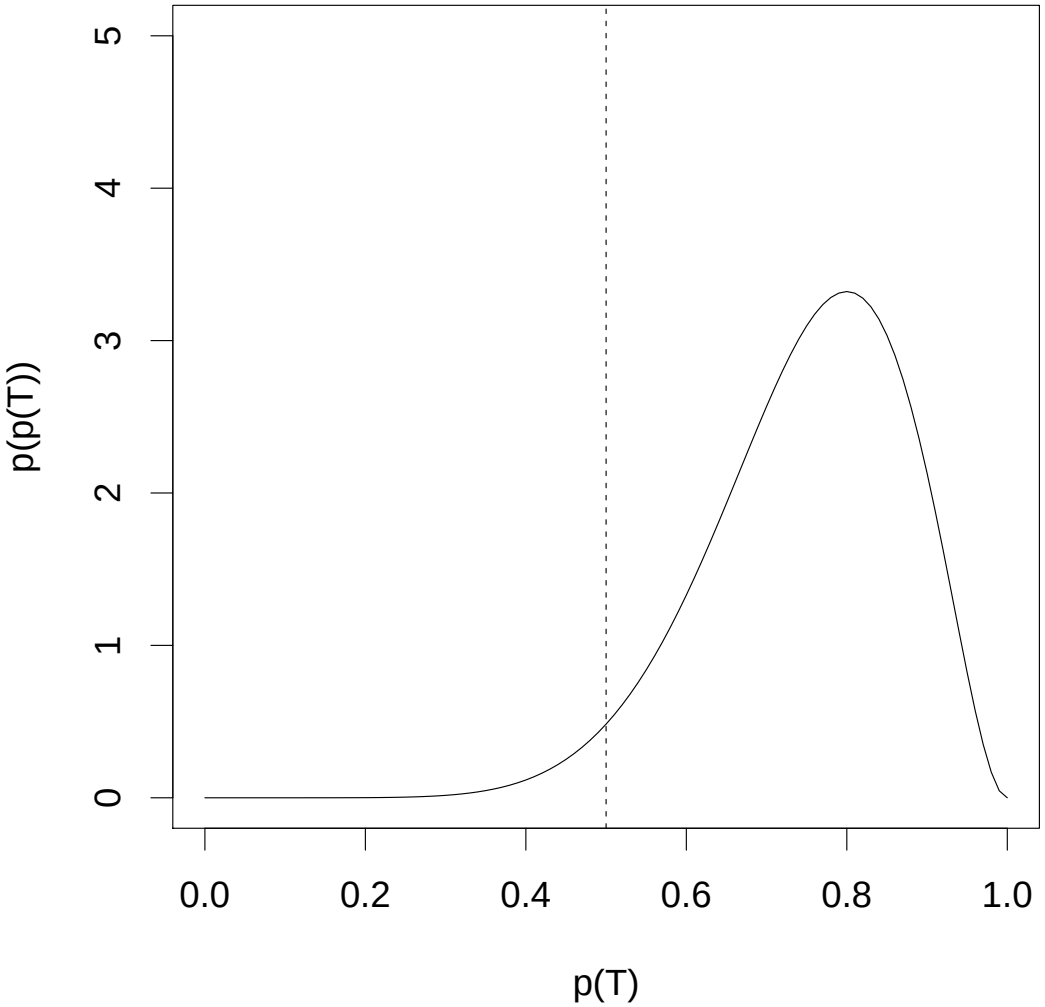
Example - Fair coin - Prior



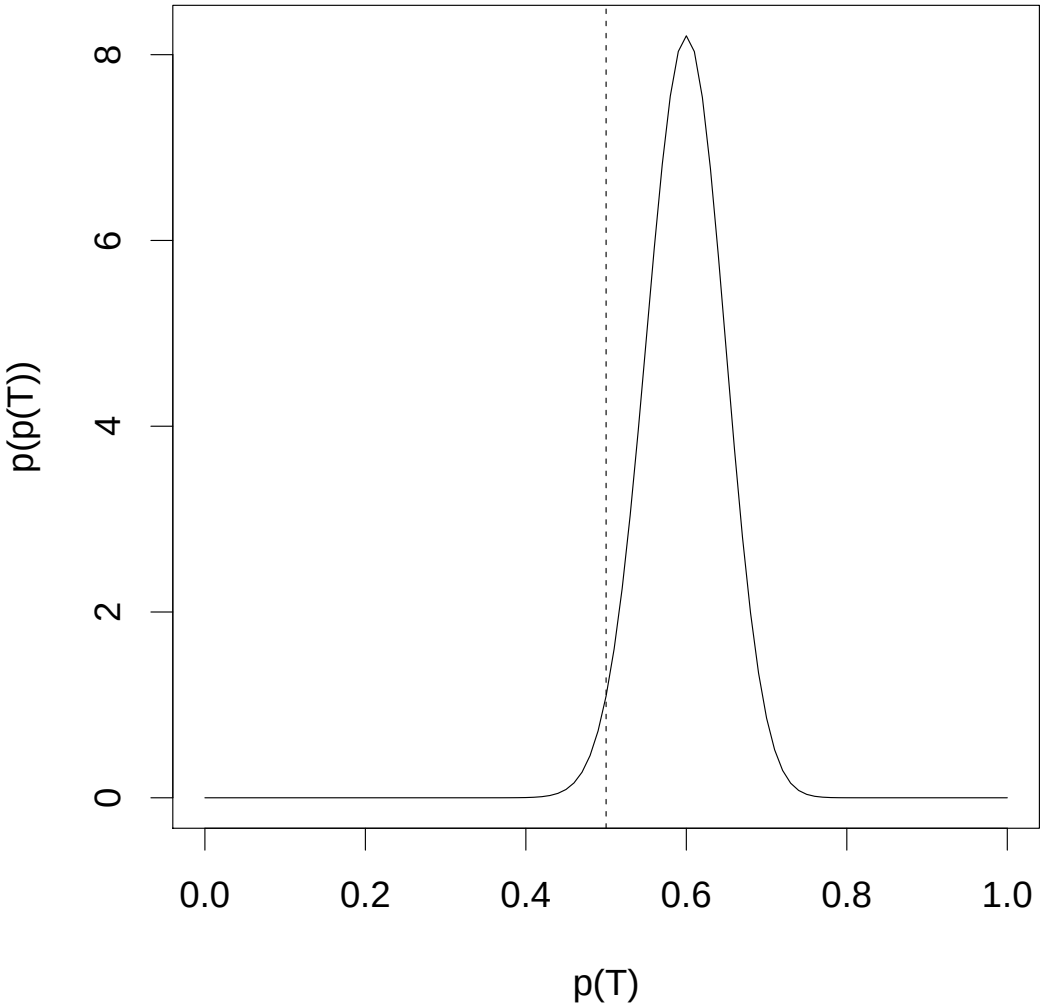
Example - Fair coin - Posterior after "TTTT"



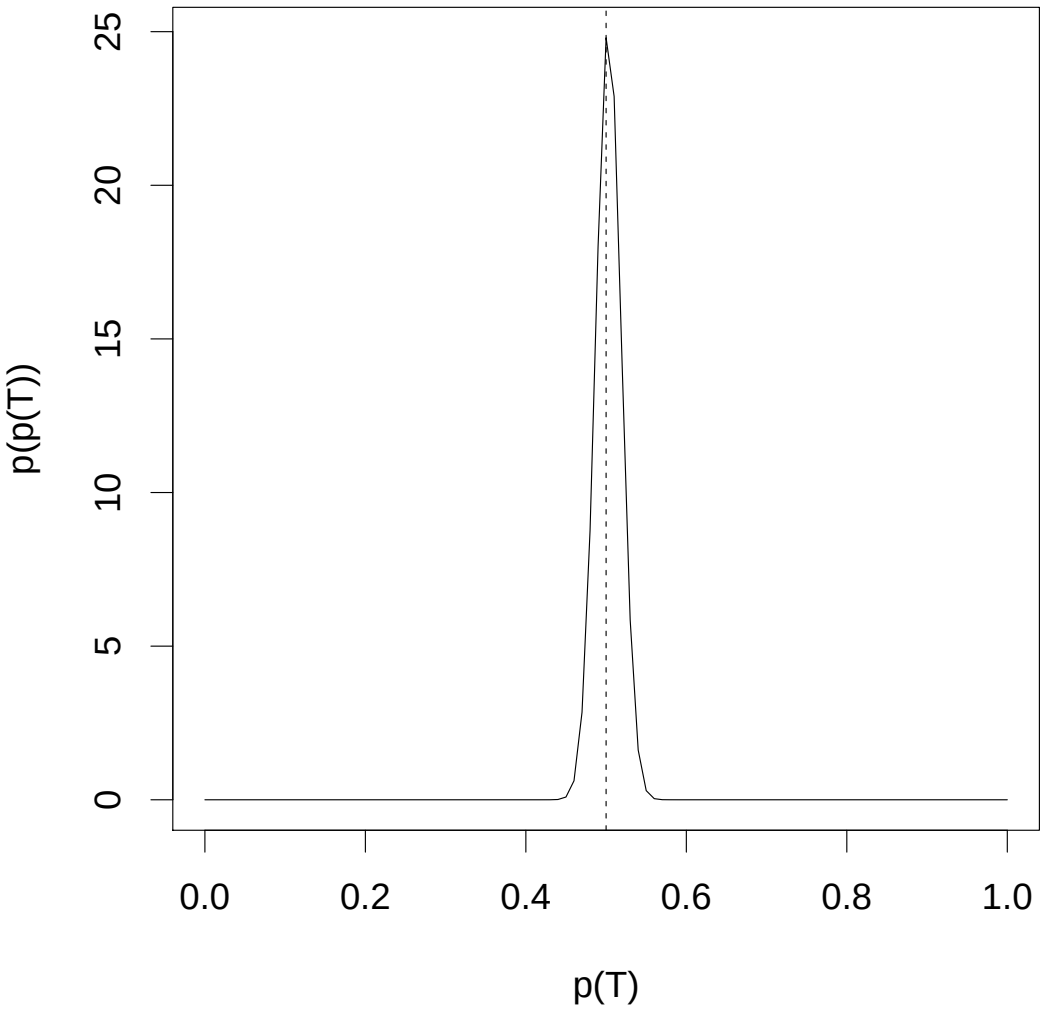
Example - Fair coin - Posterior after "TTTTTTTHTHT"



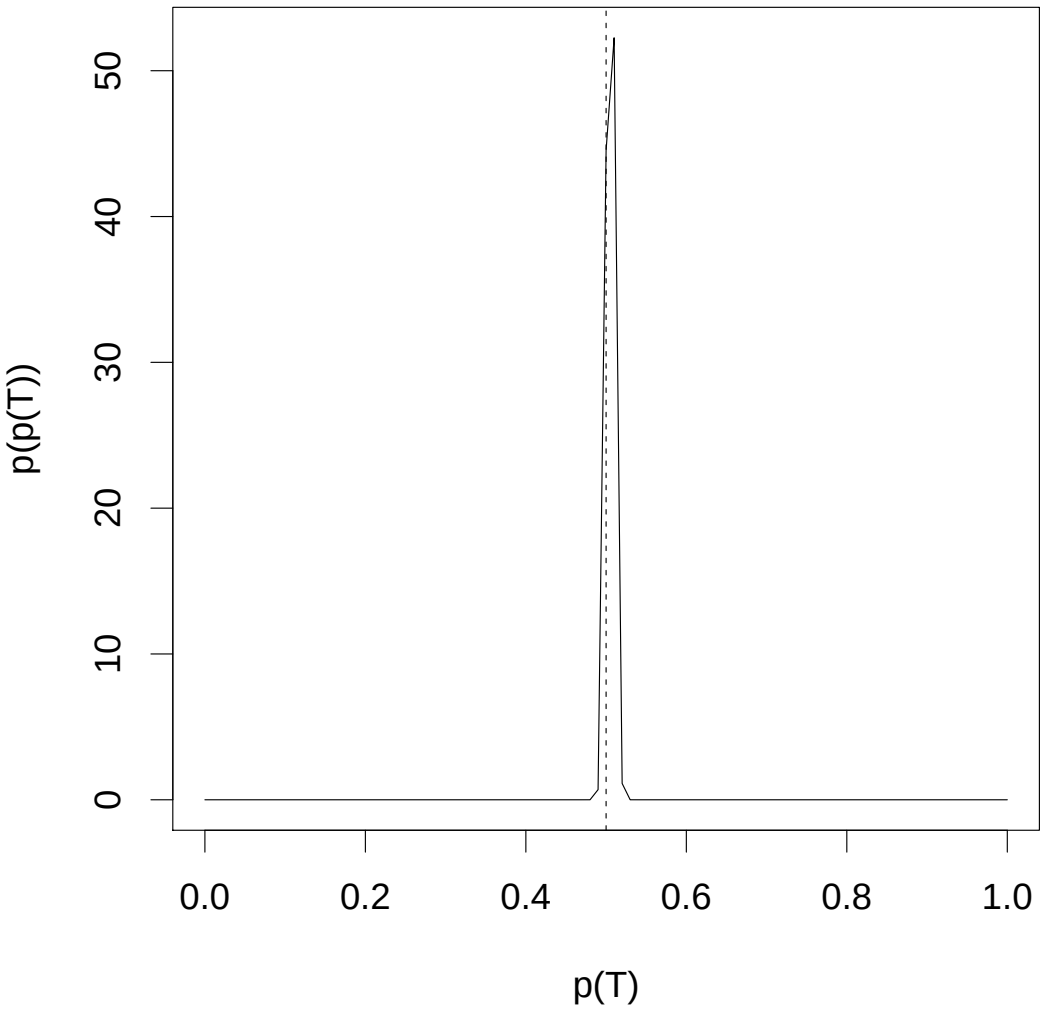
Example - Fair coin - Posterior after 100 tosses



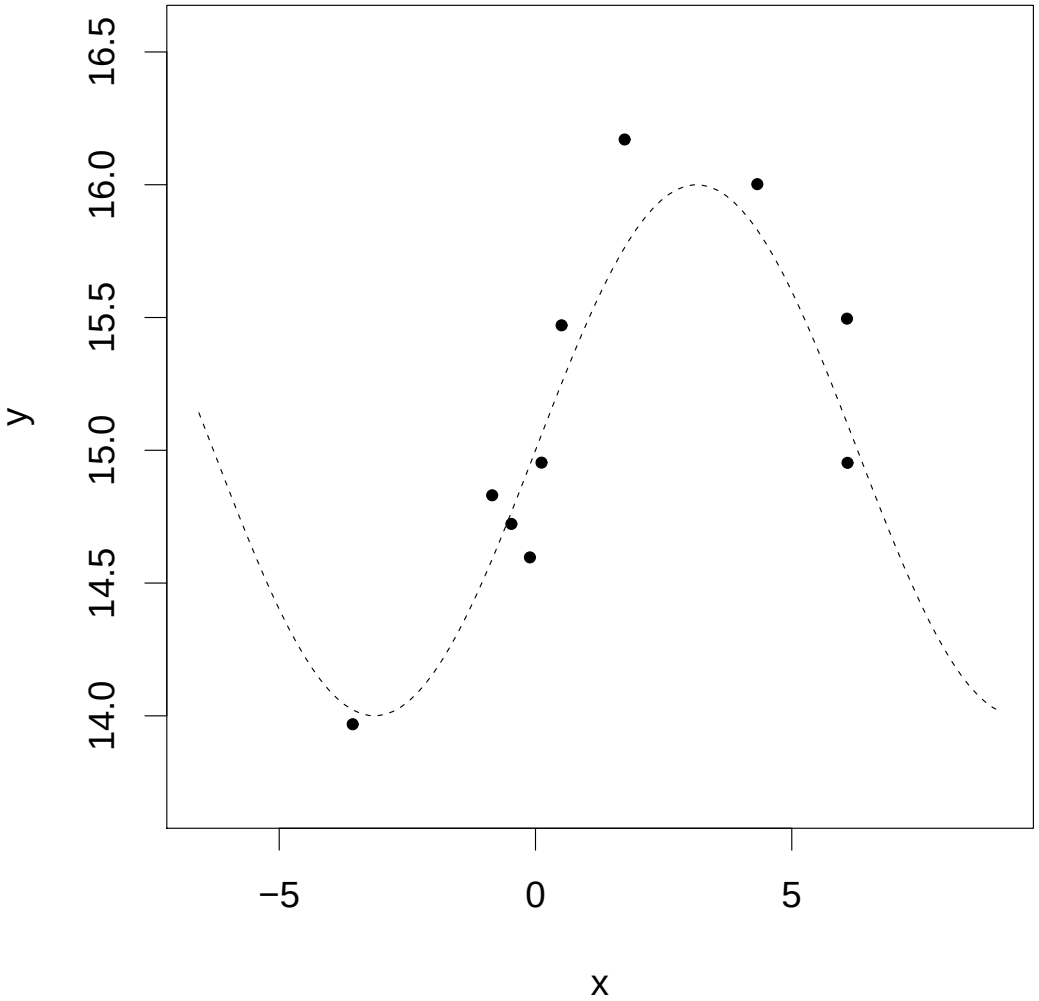
Example - Fair coin - Posterior after 1000 tosses



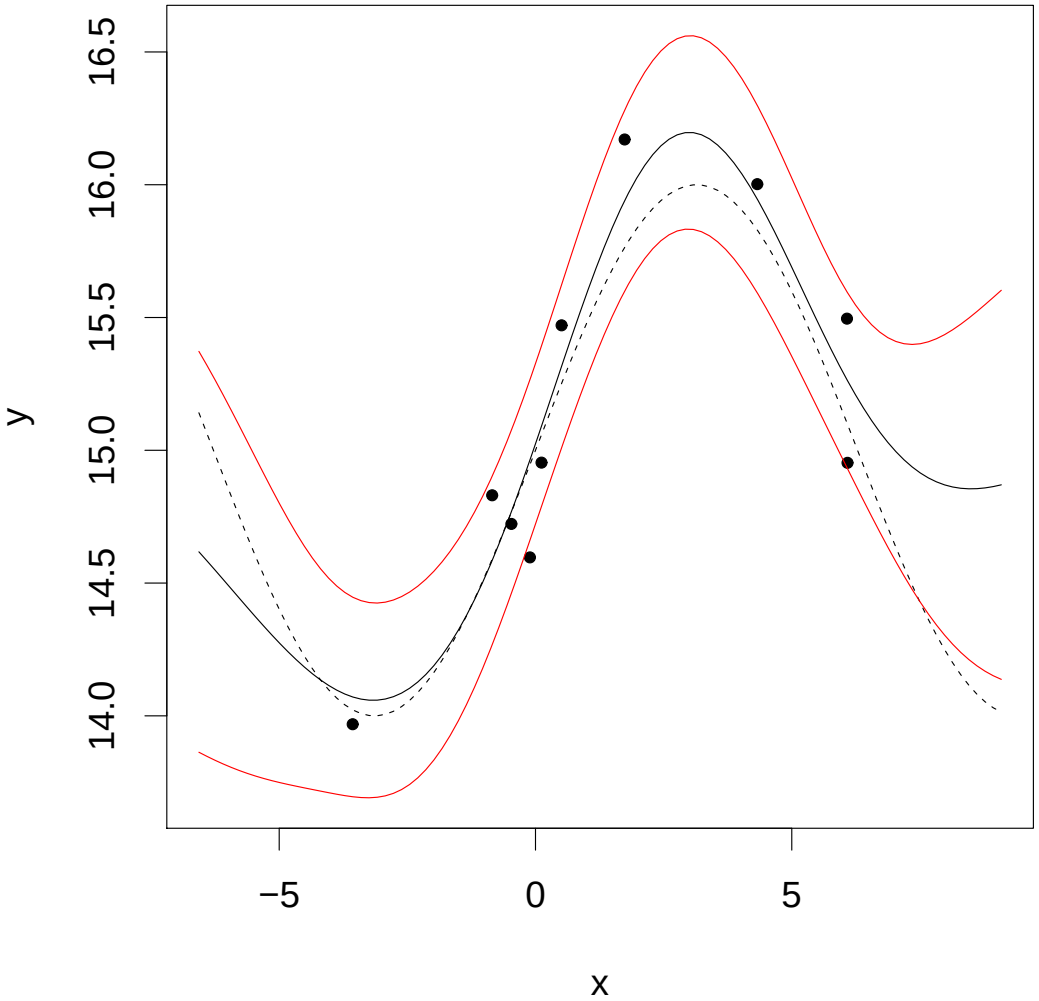
Example - Fair coin - Posterior after 10000 tosses



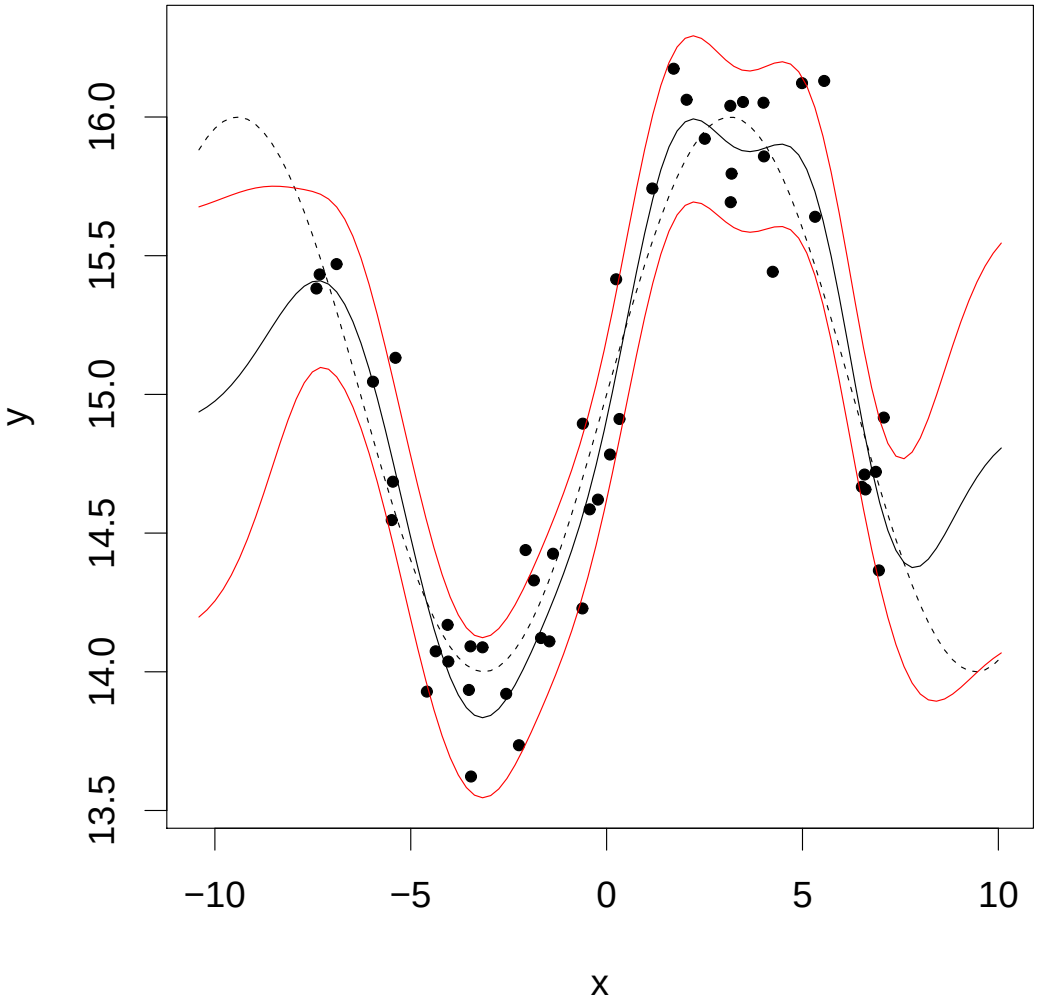
Example - Regression with Gaussian processes



Example - Regression with Gaussian processes



Example - Regression with Gaussian processes



Marginalization and model evidence

- Once we estimate the parameters of our model we can use the sum rule to marginalize out the parameters:

$$p(X) = \int_{\Theta} p(X|\theta)p(\theta)d\theta$$

- in fact $p(X)$ should read:

$$p(X) \equiv p(X|M)$$

  in words: the likelihood of X **given** the model assumption M
- $p(X|M)$ is the *model evidence*

Model selection

- Let's assume that we have a set of candidates models $\{M_i\}$ for our dataset X
- We would like to pick the model having the highest probability given X:

$$p(M_i|X) \propto p(M_i)p(X|M_i)$$

- If we don't have any prior knowledge on the best model $p(M_i) = p(M_j)$
- In this case, it means find the model with the highest *model evidence*:

$$p(X|M_i) = \int_{\Theta} p(X|M_i, \theta)p(\theta)d\theta$$

One of the main problems

- The model evidence:

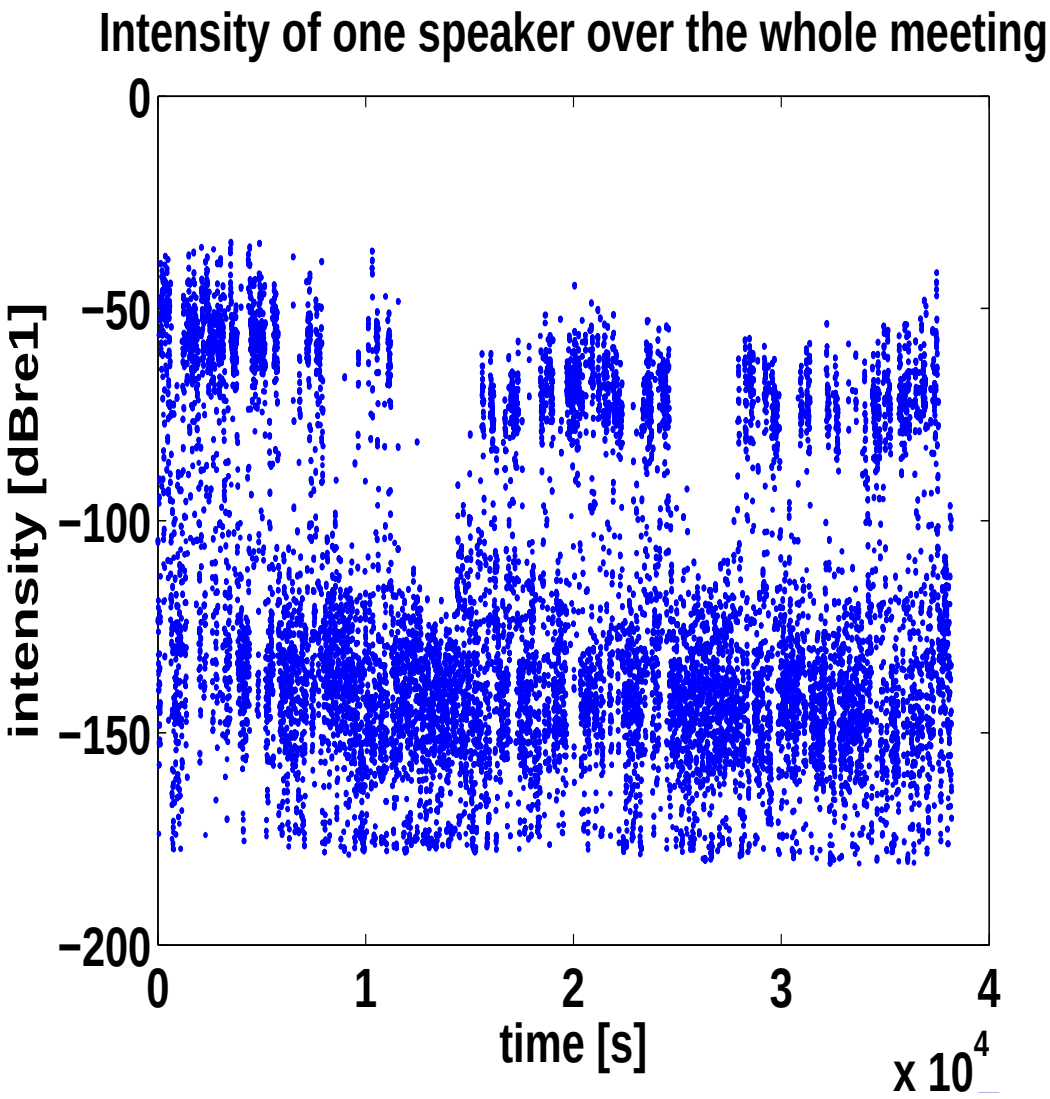
$$p(X|M) = \int_{\theta} p(X|M,\theta)p(\theta)d\theta$$

- In many cases this integral is analytically intractable
- Approximation schemes are needed (sampling, variational methods, ...)

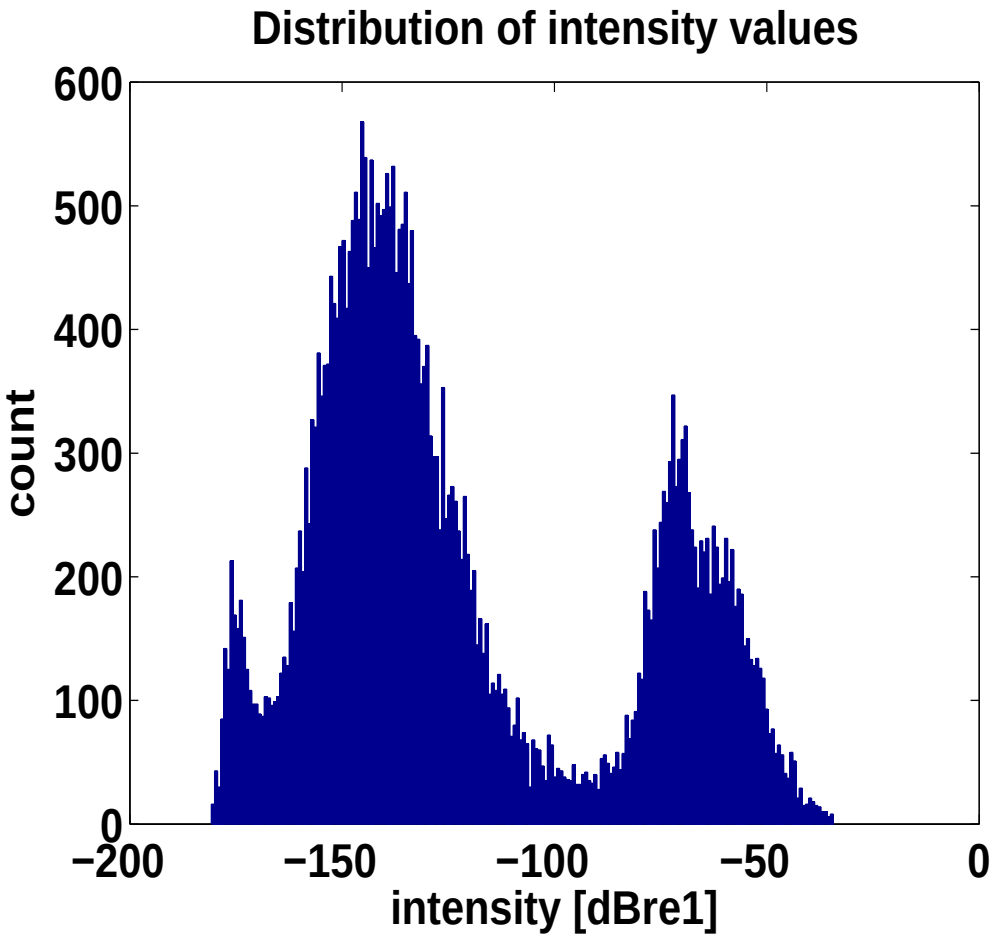
The problem

- Identify the intensity range of a speaker
- *Why?* Interactionally relevant property (besides f_0 movements, range, and others)
- *What do we want?* Compare a speaker's intensity to the speaker's range (e.g. lower or higher than the mean)
- *How?*
  - conversational data (sequences of no speech, cross talk, overlapping talk)
  - intensity contour
  - $\Rightarrow$ extraction of peaks

Intensity Peaks over time - Speaker 1

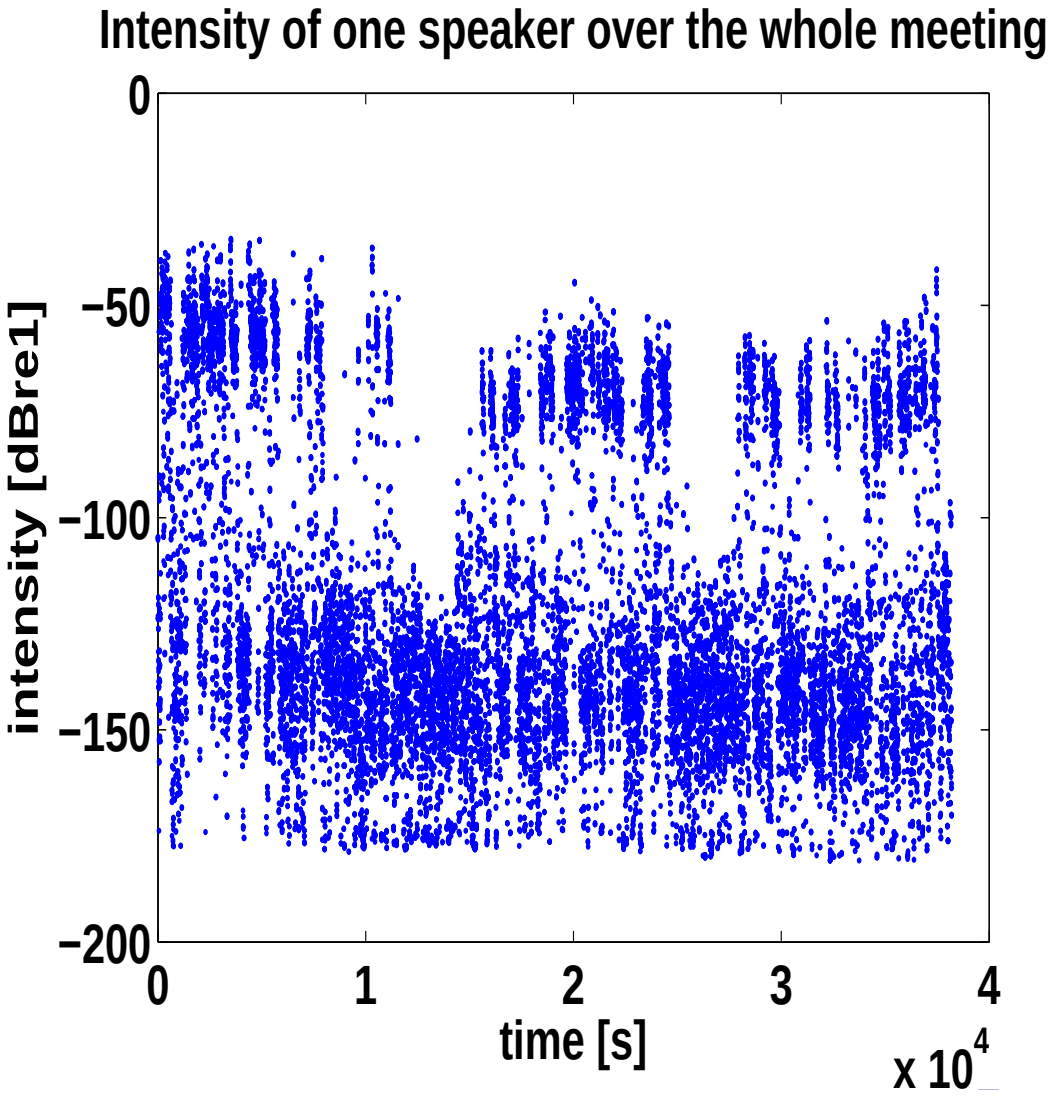


Intensity Peaks - Histogram - Speaker 1

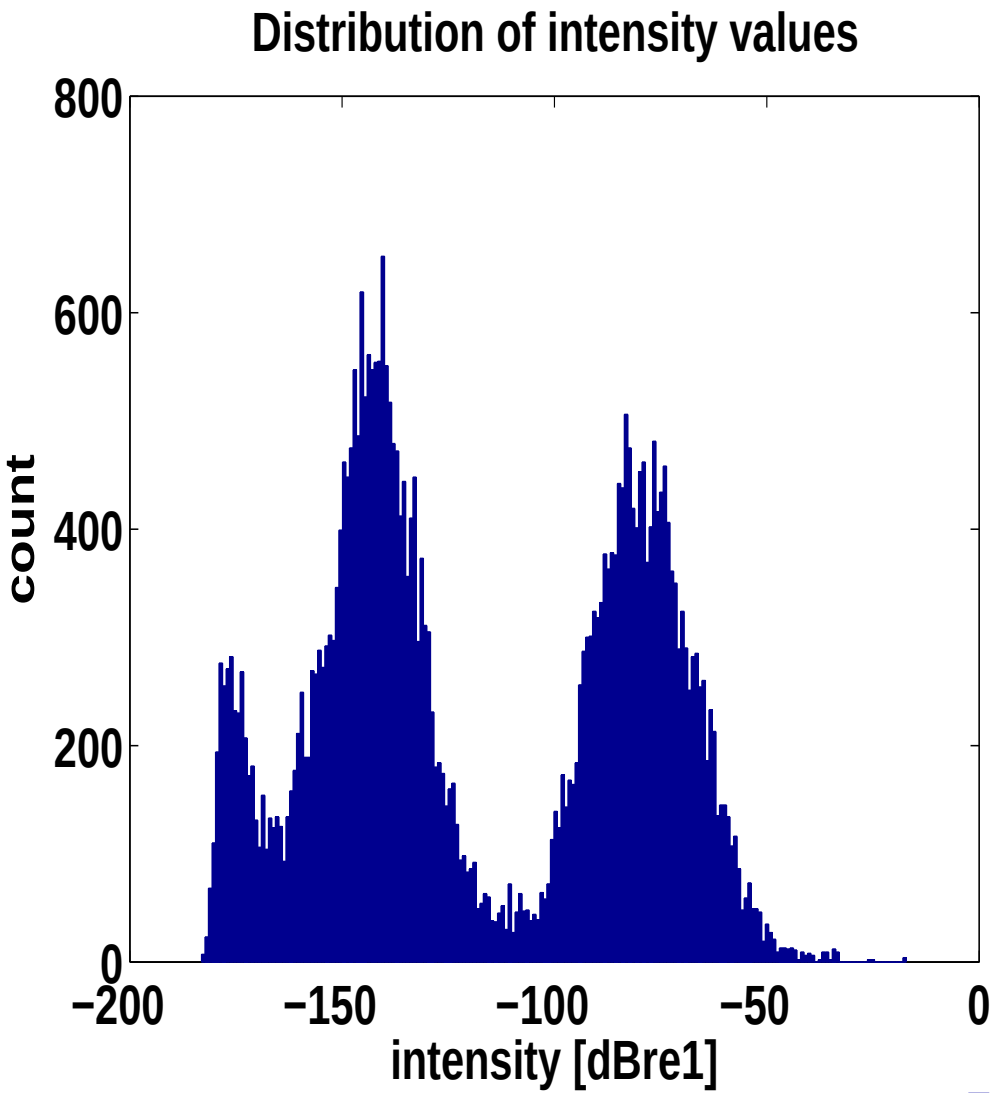


Where do we put the boundaries in the intensity histogram?

Intensity Peaks over time - Speaker 2



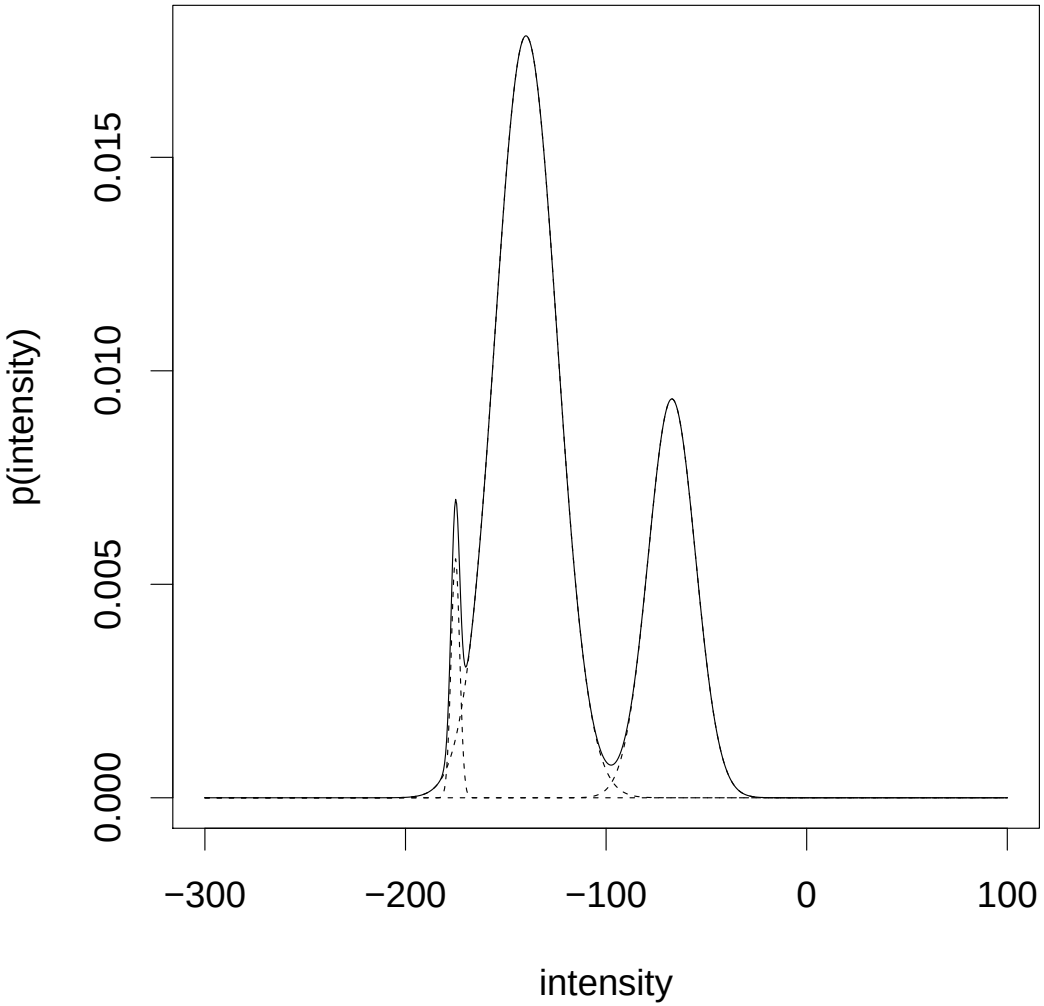
Intensity Peaks - Histogram - Speaker 2



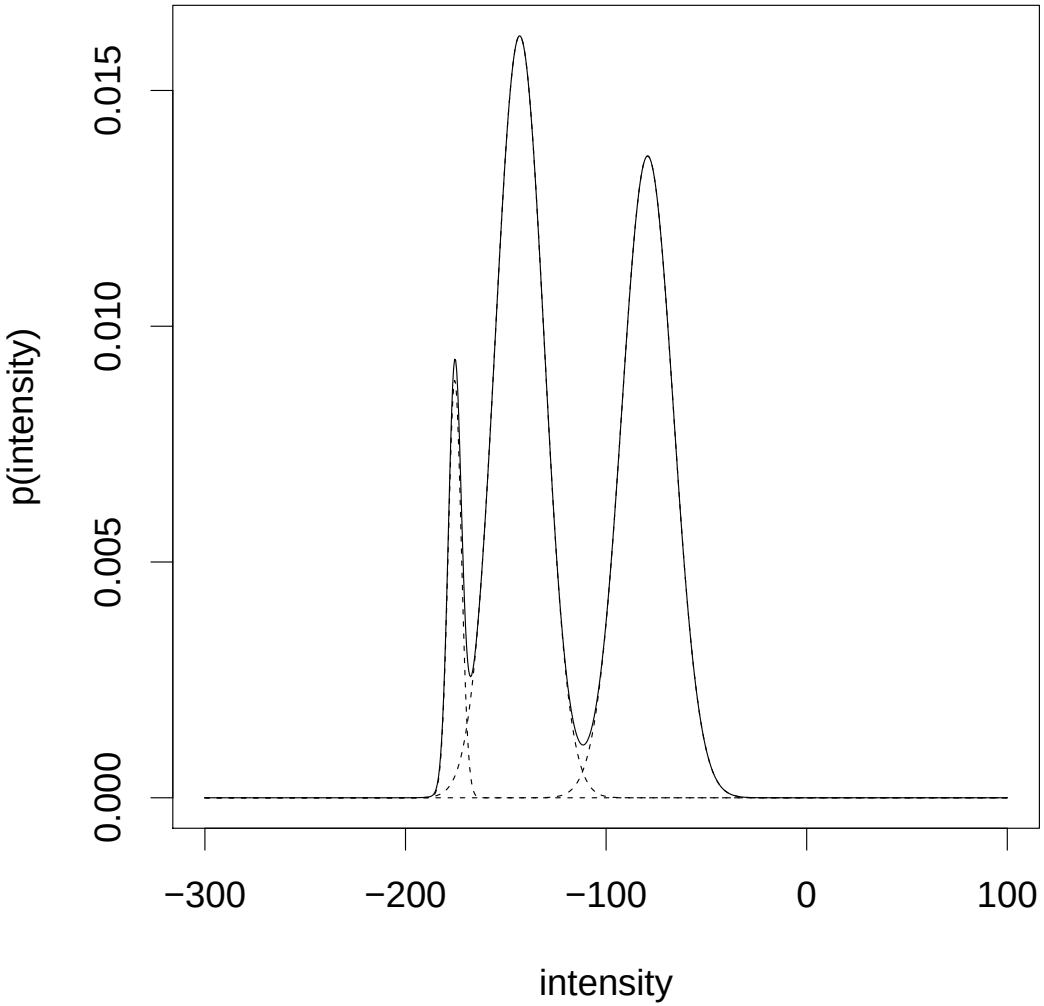
How can we deal with this problem?

- From the histogram it is reasonable to fit a mixture of Gaussians
- We can do it with ML, but the solution is not in closed form
- Nevertheless, we can use the Expectation Maximization algorithm that is iterative and maximizes the likelihood with respect to the parameters

Mixture of Gaussians - Speaker 1



Mixture of Gaussians - Speaker 2



Conclusions

- We have briefly reviewed some basics concepts underpinning probabilistic methods for data modeling
- Probabilistic methods provide a powerful tool to analyze data and solve a broad range of problems
- Depending on the problem, it is important to understand what we require from the learning method and the computational cost of the approach

References

- C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, August 2007.
- my web page:
  http://www.dcs.shef.ac.uk/∼filippone
- these slides:
  http://www.dcs.shef.ac.uk/∼filippone/Talks/tutorial_spandh09.pdf