Learning theory: Maximum Likelihood, Bayesian Learning, Model Selection

Shikui Tu
Shanghai Jiao Tong University

2021-04-06

Outline

- **Maximum Likelihood (ML) learning**
- Bayesian learning, Maximum A Posterior (MAP)
- Model Selection
  - Two-phase procedure: AIC, BIC
  - Automatic: Variational Bayes (VB)

An example

- If flipping a coin a few times, and get

- What is the probability it will fall with the head up?

You may say: 3/5

Because ….

Bernoulli distribution

The dataset $D = \{ x_t \}, t=1,\dots,N,\ x_t \in \{H, T\}$

$D =$ 

$$P(x = Head) = \theta$$

$$P(x = Tail) = 1 - \theta$$

Flipping coins are **i.i.d.**, i.e., independent identically distributed according to Bernoulli distribution

Question: What is the parameter θ that maximizes the probability of observed data?

Maximum Likelihood Estimation

- Choose parameter θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$ Identically distributed

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5** "Frequency of heads"

Number of heads Number of tails

Outline

- Maximum Likelihood (ML) learning
- **Bayesian learning, Maximum A Posterior (MAP)**
- Model Selection
  - Two-phase procedure: AIC, BIC
  - Automatic: Variational Bayes (VB)

Bayesian Learning

- Bayes rule

$$P(\Theta|X) = \frac{P(X|\Theta)P(\Theta)}{P(X)}$$

$P(X|\Theta)$: likelihood of data X given parameter Θ

$P(\Theta)$: prior distribution over the parameter Θ

$P(X)$: marginal distribution of data X

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

- Prior distribution
  - Represents expert knowledge
  - Uninformative priors: Uniform distribution
  - Conjugate priors: Closed-form representation of posterior, P(θ) and P(θ|D) have the same form

Bayesian learning

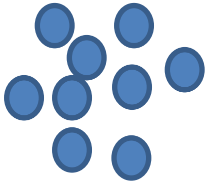
- Maximum A Posteriori (MAP)

$$\max_{\Theta} p(\Theta|X)$$

Equivalent to:

$$\log p(X,\Theta) = \log p(X|\Theta) + \log p(\Theta)$$

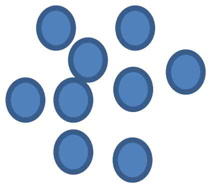
Consider a simple example:



$$p(x|\Theta) = G(x|\mu,\Sigma)$$

$$p(\mu) = G(\mu|\mu_0,\sigma_0^2)$$

Derivation



$$p(x|\Theta) = G(x|\mu, \Sigma)$$

$$p(\mu) = G(\mu|\mu_0, \sigma_0^2)$$

When is MAP the same as MLE?

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

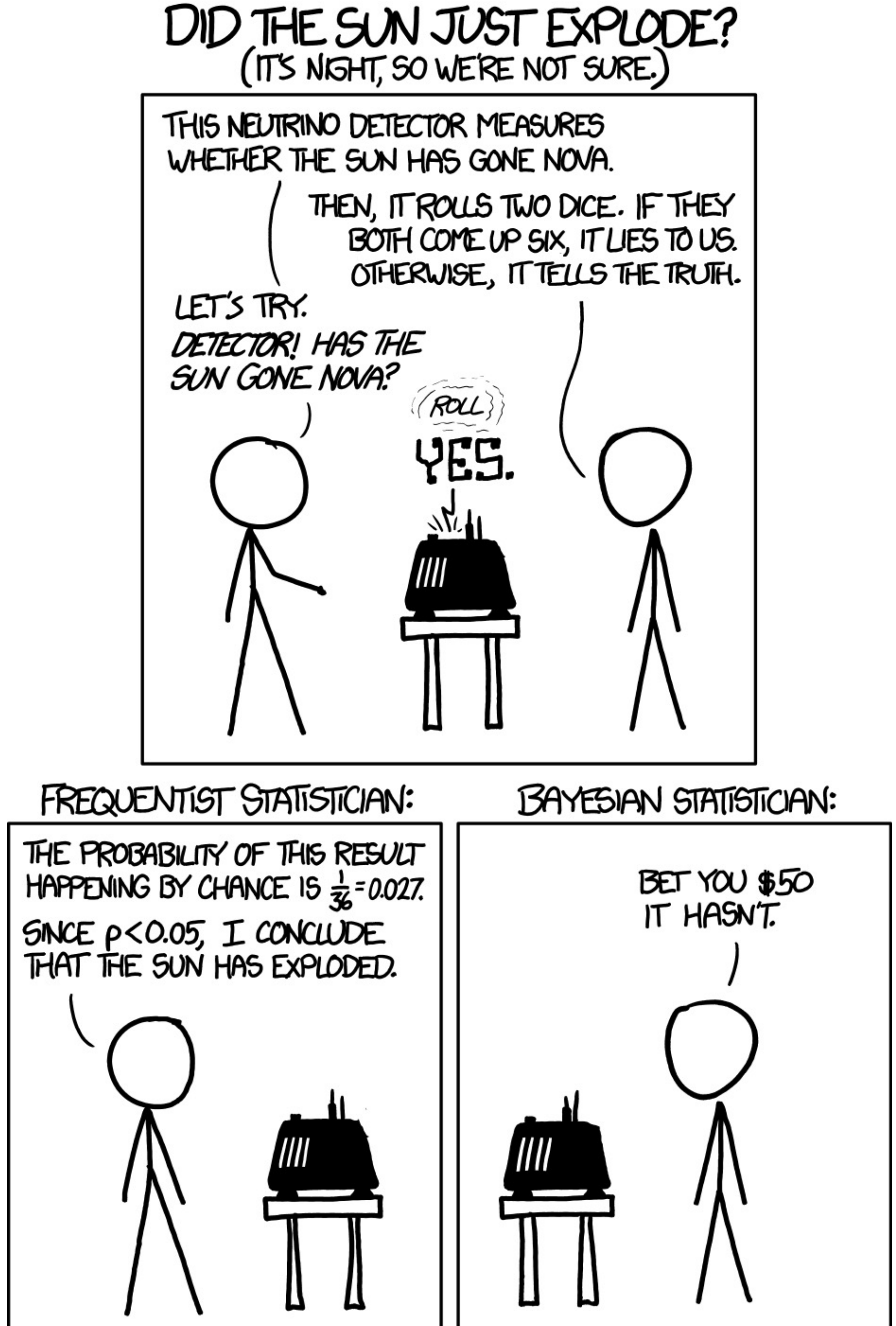
  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D)\\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Bayesians vs Frequentists

You are no good when sample is small

You give a different answer for different priors


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

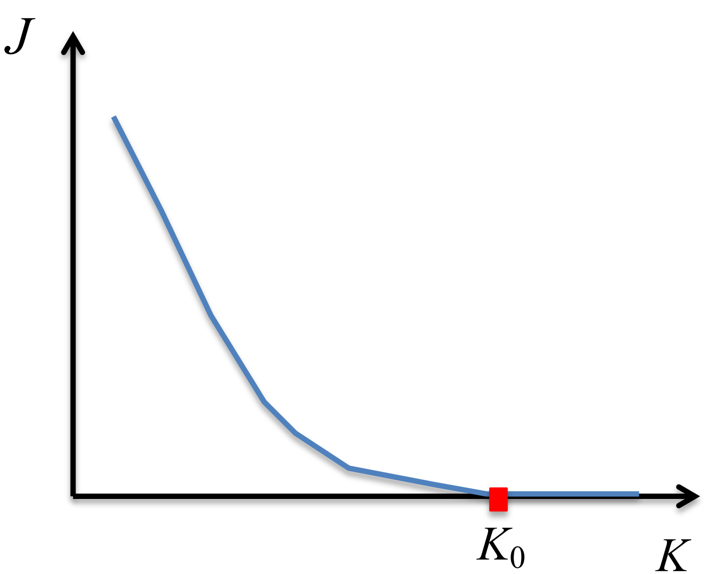
Outline

- Maximum Likelihood (ML) learning
- Bayesian learning, Maximum A Posterior (MAP)
- **Model Selection**
  - **Two-phase procedure: AIC, BIC**
  - **Automatic: Variational Bayes (VB)**

How to determine the cluster number K?

K-mean

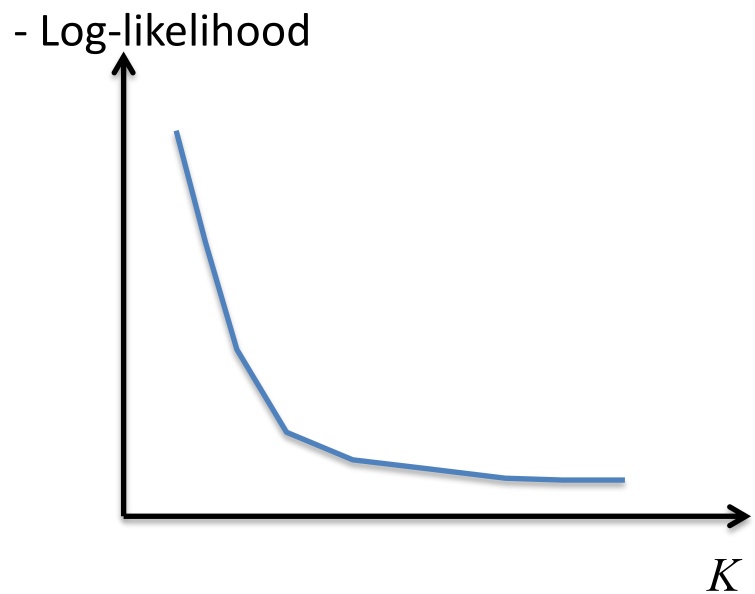
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$



J does not tell which K is better.

GMM

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$



Negative log-likelihood also decreases as K increases.

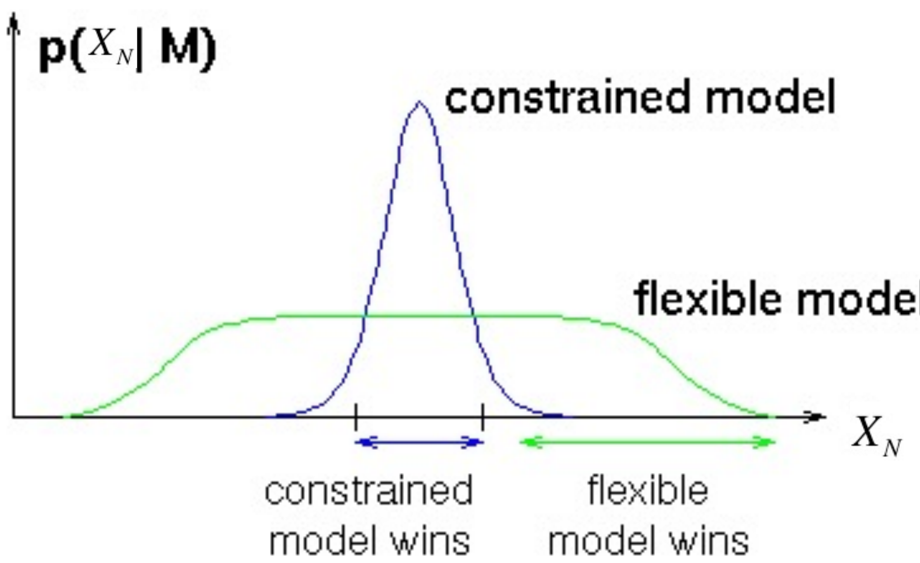
Model selection

Probabilistic model

$$p(X_N \mid \Theta_K)$$

Candidate models:

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K \subseteq \cdots$$



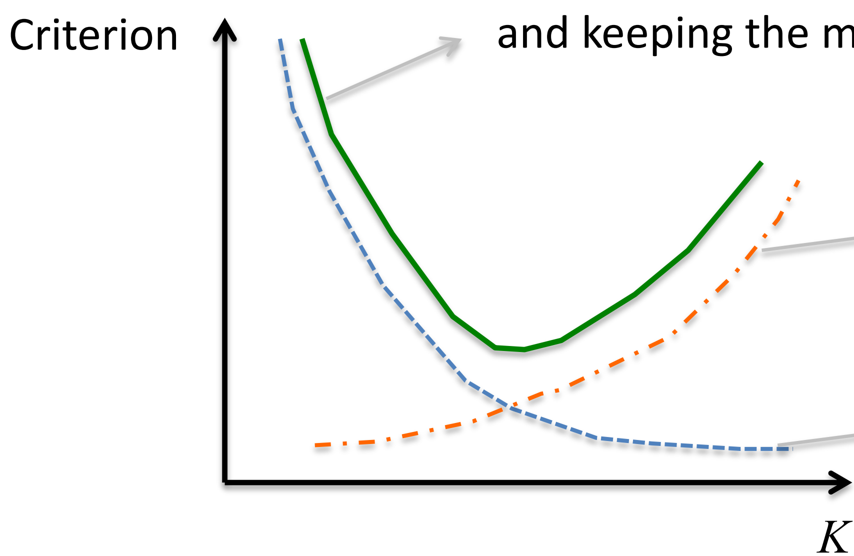
Model selection for generalization

Probabilistic model

$$p(X_N \mid \Theta_K)$$

Candidate models:

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K \subseteq \cdots$$



A trade-off between fitting the data well and keeping the model simple

Models become more and more complex as K increases.

Negative log-likelihood (or fitting error) to make sure the model fit the data well.

Akaike's Information Criterion (AIC) $\ln p(X_N \mid \hat{\Theta}_K) - d_k$

Bayesian Information Criterion (BIC) $\ln p(X_N \mid \hat{\Theta}_K) - \frac{1}{2} d_k \ln N$

d_k: number of free parameters
N: sample size

Two-phase method for model selection

- Assume the optimal K* is within the range [1,K_{max}].

- Phase (1): For each $k = 1, \ldots, K_{\max}$, compute the maximum likelihood estimator:

$$\widehat{\Theta}_{ML}(k) = \underset{\Theta}{\operatorname{argmax}} \log[P(X|\Theta, k)]$$

- Phase (2): Select the optimal K* by optimizing the values of the model selection criterion J, e.g., AIC, BIC:

$$K^* = \underset{k}{\operatorname{argmax}}\, J(\widehat{\Theta}_{ML}(k))$$

Akaike's Information Criterion (AIC)

$$\ln p(X_N \mid \hat{\Theta}_K) - d_k$$

Bayesian Information Criterion (BIC)

$$\ln p(X_N \mid \hat{\Theta}_K) - \frac{1}{2} d_k \ln N$$

Using Occam's Razor to Learn Model Structure

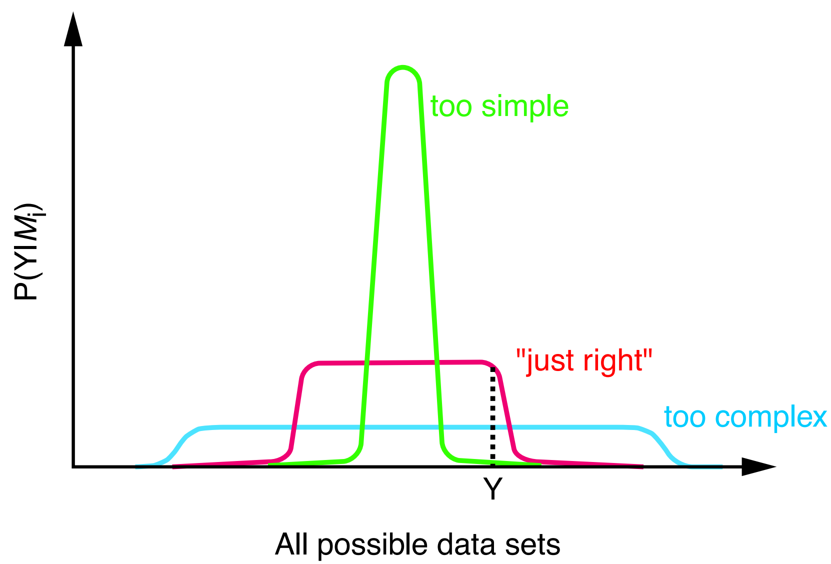
Compare model classes m using their posterior probability given the data:

$$P(m|\mathbf{y}) = \frac{P(\mathbf{y}|m)P(m)}{P(\mathbf{y})}, \qquad P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\boldsymbol{\theta}_m, m)P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

Interpretation of $P(\mathbf{y}|m)$: The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Bayesian model selection

- A model class m is a set of models parameterised by $\boldsymbol{\theta}_m$, e.g. the set of all possible mixtures of m Gaussians.

- The marginal likelihood of model class m:

$$P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\boldsymbol{\theta}_m, m) P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

  is also known as the Bayesian evidence for model m.

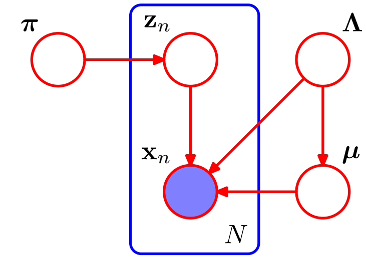
- The ratio of two marginal likelihoods is known as the Bayes factor:

$$\frac{P(\mathbf{y}|m)}{P(\mathbf{y}|m')}$$

- The Occam's Razor principle is, roughly speaking, that one should prefer simpler explanations than more complex explanations.

- Bayesian inference formalises and *automatically* implements the Occam's Razor principle.

VBEM for GMM



- Model descriptions:

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \quad p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

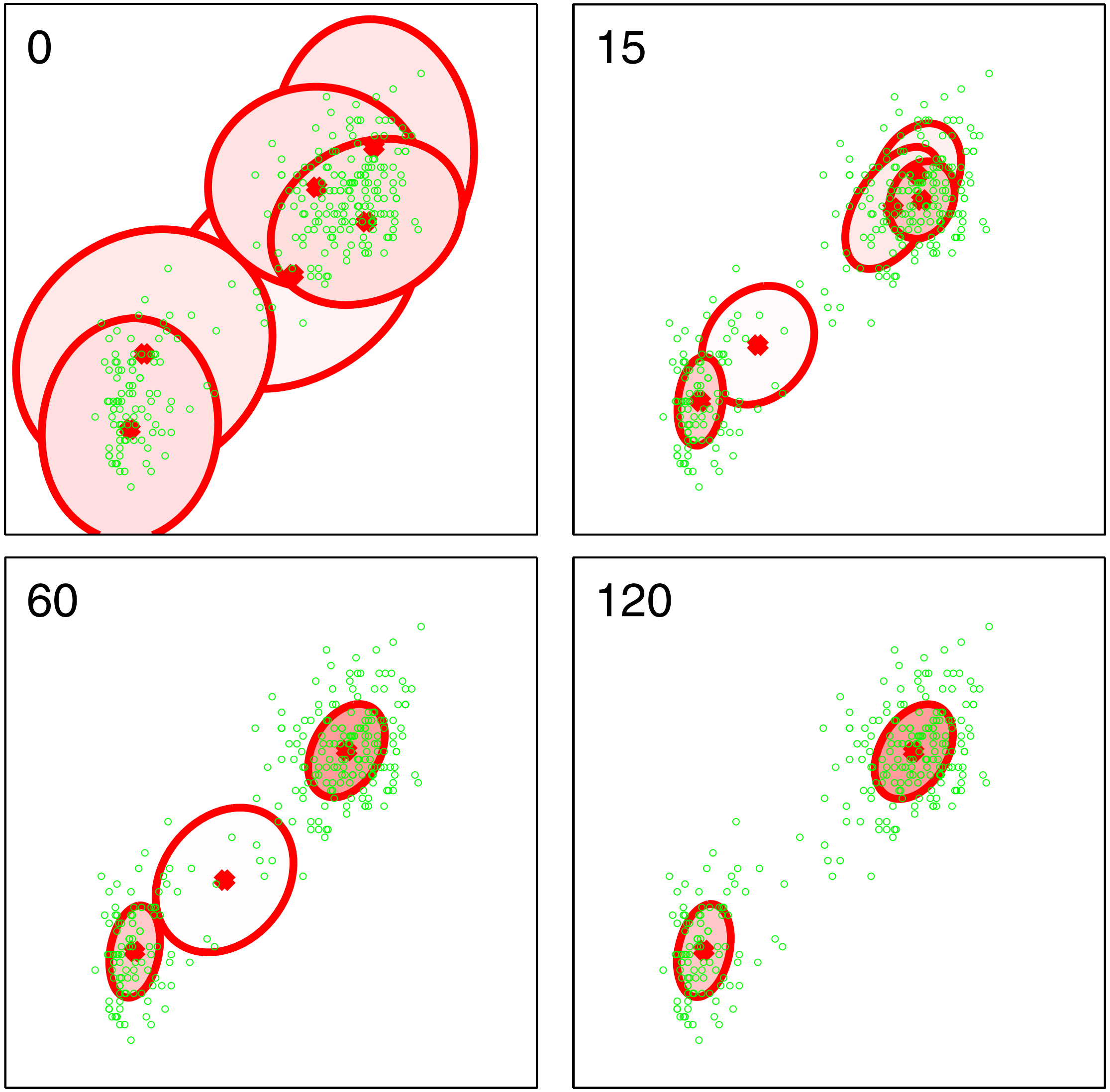
- Prior distributions over parameters:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$$

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0) \end{aligned}$$

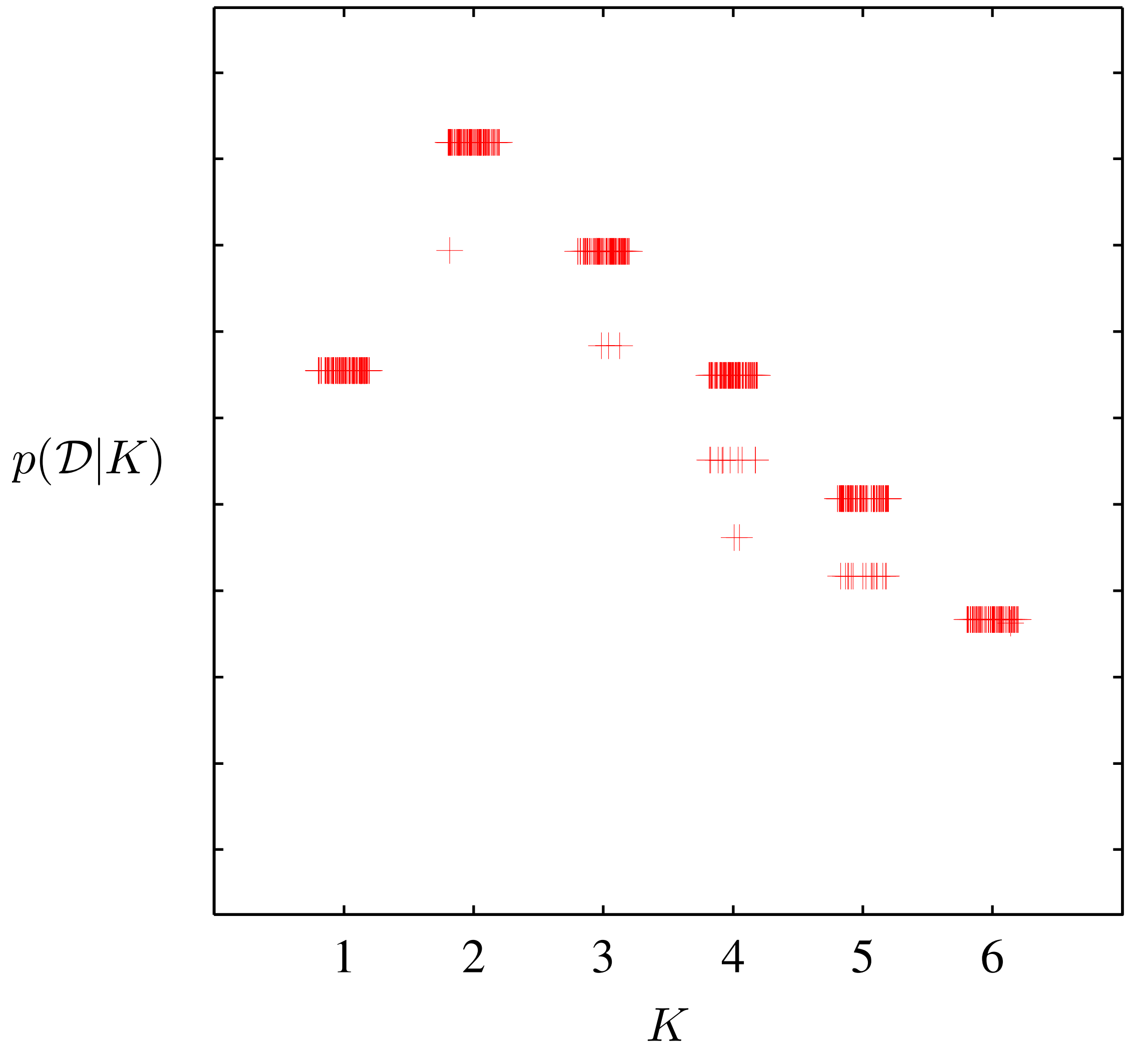
$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

How VBEM for GMM works



http://www.cs.ubc.ca/~murphyk/Software/VBEMGMM/index.html
http://scikit-learn.org/stable/modules/mixture.html

Determine K by the variational lower bound (free energy)



Thank you!