Infinite Mixture Models with Dirichlet Process

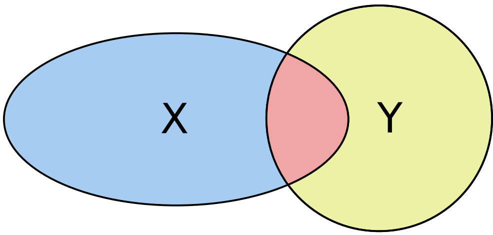
Rémi Emonet - 2015-04-30

Université Jean Monnet - Laboratoire Hubert Curien

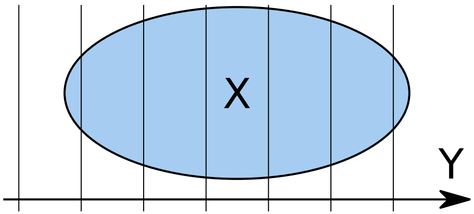
Just Some Reminders (Probably Almost Surely)

Probability Rules, Measure Theory

- Product rule
  - $p(X, Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$



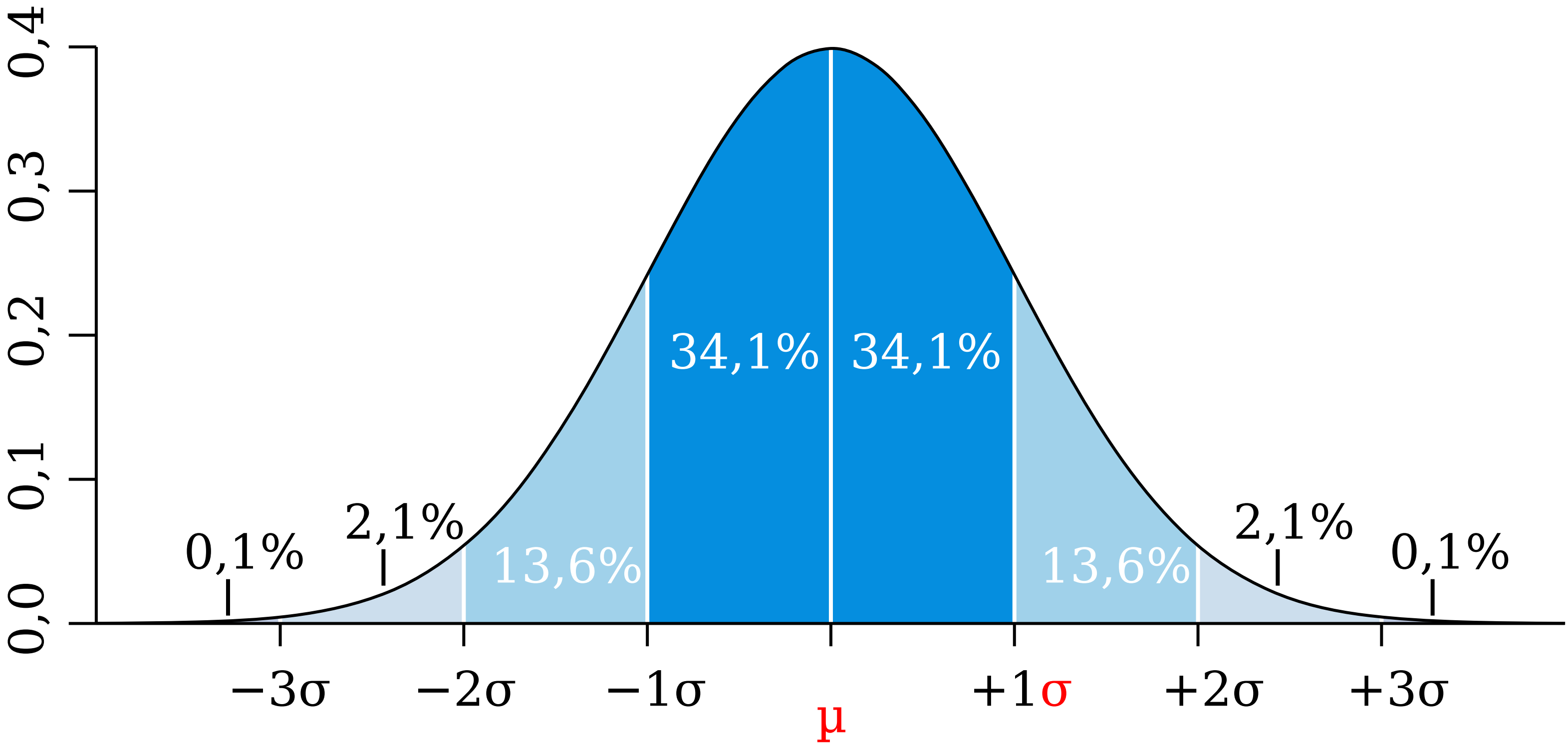
- Marginalization, Sum rule
  - $p(X) = \sum_{Y \in \mathcal{Y}} p(X, Y)$
  - $p(X) = \int_{Y \in \mathcal{Y}} p(X, Y)$



- Bayes rule
  - $p(Y|X) = \dfrac{p(X|Y)\ p(Y)}{p(X)}$
  - $p(X) = \sum_{Y \in \mathcal{Y}} p(X|Y)\ p(Y)$

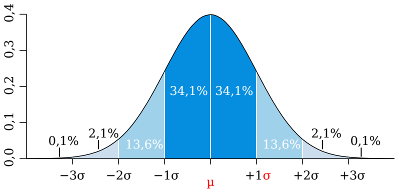
Gaussian/Normal Distribution

Gaussian/Normal Distribution (1D)



Gaussian/Normal Distribution (1D)

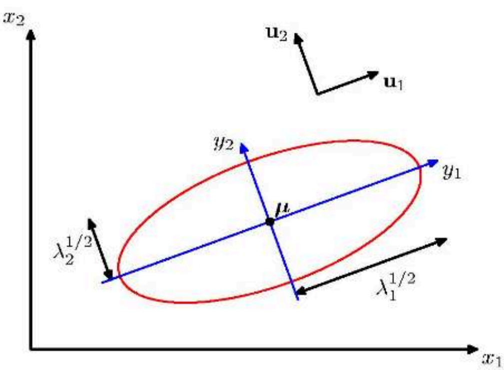
- Normal Distribution or Gaussian Distribution
  - $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - Is-a probability density
  - $\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2)\, dx = 1$
  - $\mathcal{N}(x|\mu, \sigma^2) > 0$



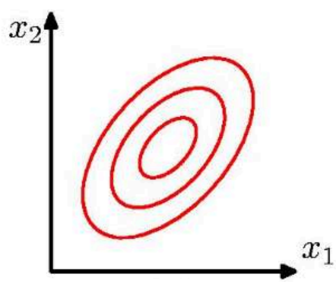
- Parameters
  - μ: mean, $\mathrm{E}[X] = \mu$
  - σ^2: variance, $\mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \sigma^2$

Multivariate Normal Distribution

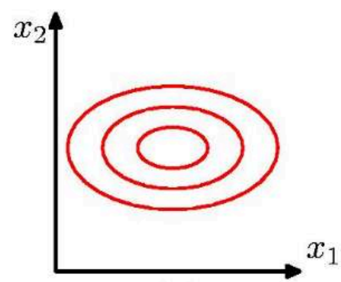
- D-dimensional space: $\mathrm{x} = \mathrm{x}_1, \ldots, \mathrm{x}_D$
- Probability distribution
  - $\mathcal{N}(\mathrm{x}|\mu, \Sigma) = \dfrac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp\left(-\dfrac{(\mathrm{x}-\mu)^T \Sigma^{-1} (\mathrm{x}-\mu)}{2}\right)$
  - Σ: covariance matrix



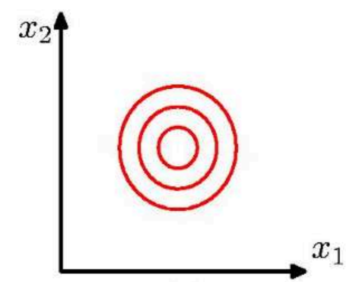
Elliptical equiprobability surfaces.
Major axes = eigenvectors of Σ



full cov. matrix
($\frac{D(D+1)}{2}$ parameters)



diagonal cov. matrix
(D parameters)

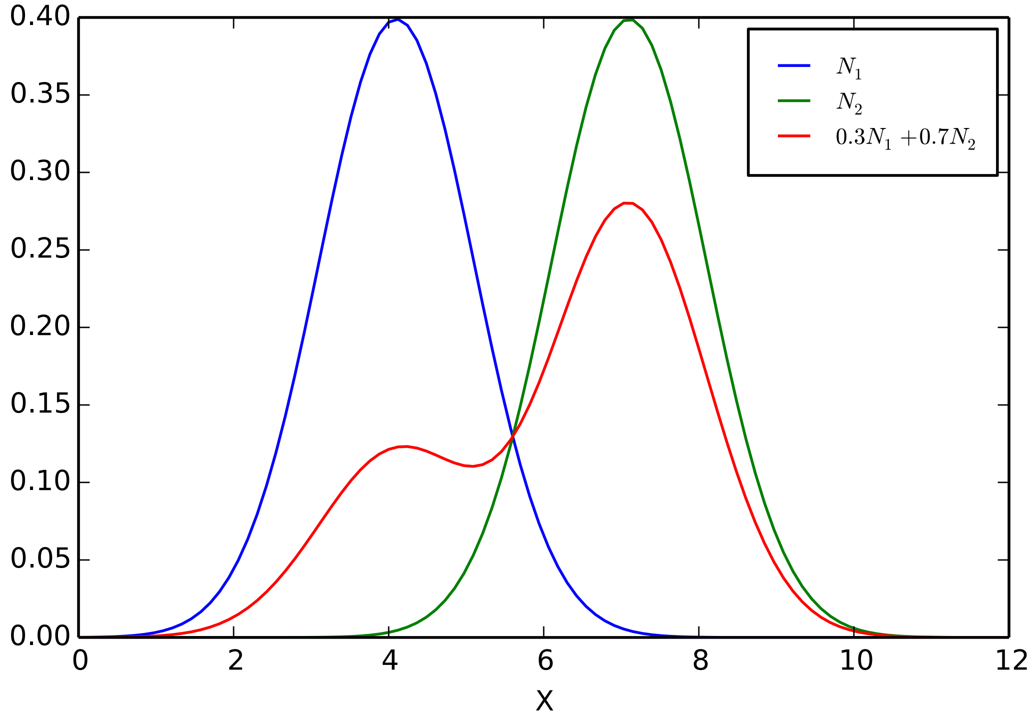


identity cov. matrix
(1 parameter)

Gaussian Mixture Models

Gaussian Mixture Model (1D)

- Density: $p\left(x\ \theta\right) = \sum_{k=1}^{K} w_k\ \mathcal{N}\left(x\ \mu_k, \sigma_k^2\right)$
- Parameters: $\theta = \left(w_k, \mu_k, {\sigma_k}^2\right)_{k=1..K}$



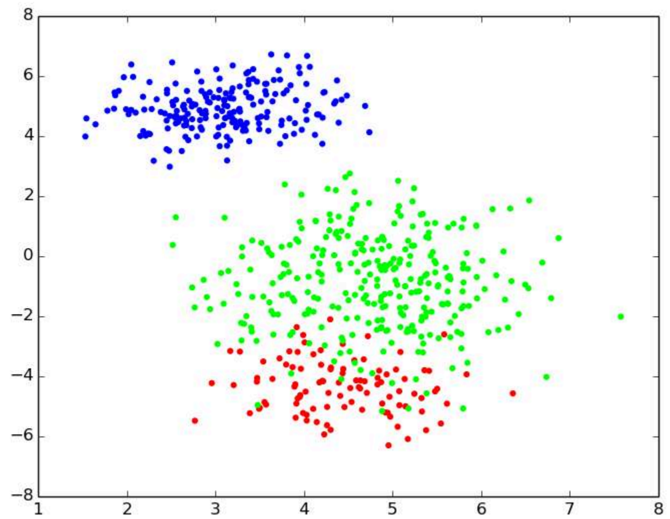
Sampling from a 2D GMM

- Density:

$$p(x|\theta) = \sum_{k=1}^{K} w_k \; \mathcal{N}(x|\mu_k, \Sigma_k)$$

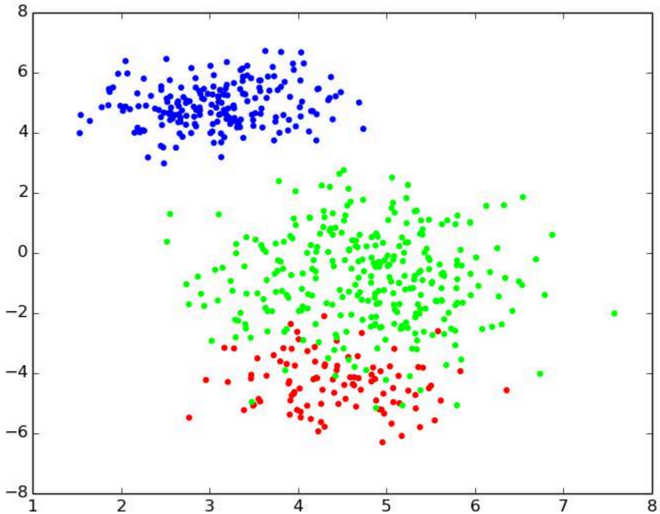
- Parameters:

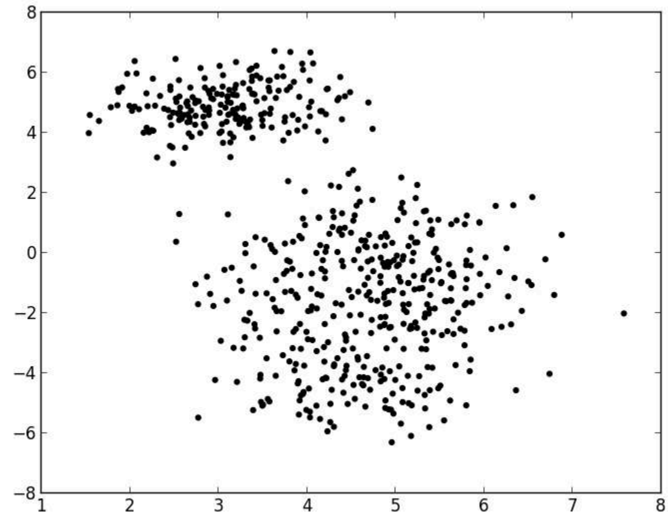
$$\theta = (w_k, \mu_k, \Sigma_k)_{k=1..K}$$



- For each point i to be generated
  - draw a component index (i.e., a color): $z_i \sim \text{Categorical}(w)$
  - draw its position from the component: $x_i \sim \mathcal{N}(x\, \mu_{z_i}, \Sigma_{z_i})$

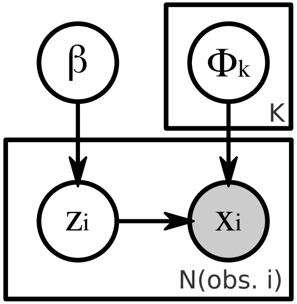
Only Positions are Observed





- For each point i
  - draw a component/color: $z_i \sim \text{Categorical}(w)$
  - draw a position: $x_i \sim \mathcal{N}(x\ \mu_{z_i}, \Sigma_{z_i})$
- Finding z seems impossible,
  finding the "best" θ might be feasible?

GMM: Probabilistic Graphical Model



- Renamed
  - $w \rightarrow \beta$ (vector of component weights)
  - $(\mu_k, \Sigma_k) \rightarrow \varphi_k$ (parameters of component k)
- Reminder: for each point i
  - draw a component/color: $z_i \sim \text{Categorical}(\beta)$
  - draw a position: $x_i \sim \mathcal{N}(x\ \varphi_{z_i})$

Learning/Inference
finding the best θ

Maximum Likehood in GMM

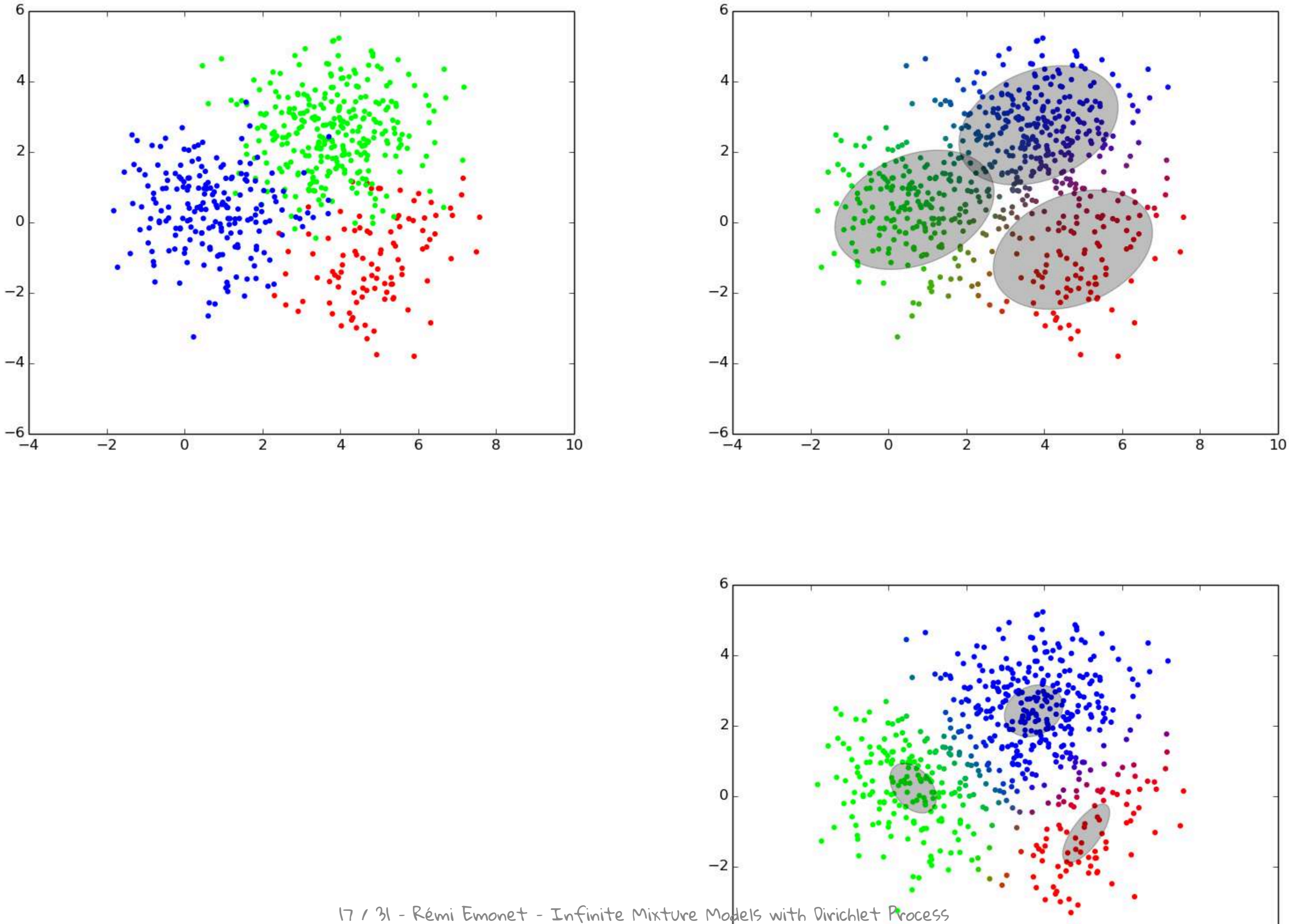
- From a set of observed points $x = \{x_i\}_{i=1..N}$
- Maximize the likelihood of the model
  - NB: $\mathscr{L}(\theta|x) = p(x|\theta)$
  - NB: $\arg\max_\theta \mathscr{L}(\theta|x) = \arg\max_\theta \log \mathscr{L}(\theta|x)$
  - indep: $\log p(x|\theta) = \log \prod_{i=1}^{N} p(x_i|\theta) = \sum_{i=1..N} \log p(x_i|\theta)$
  - mixture: $\log p(x|\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} w_k \mathscr{N}(x_i|\varphi_k)$
- No closed form expression → need to approximate

Expectation Maximization

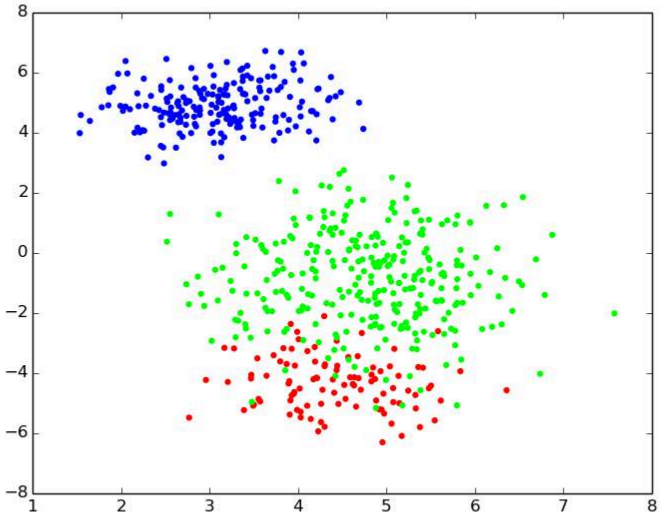
Expectation-Maximization

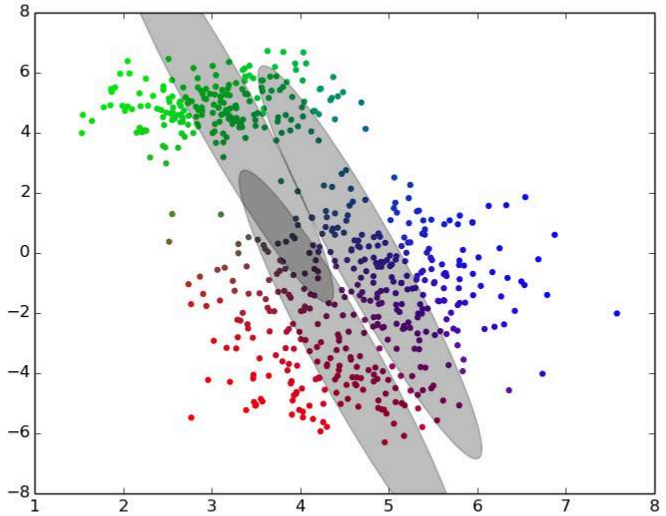
- Approximate inference by local optimization
  - converges to local optimum
  - needs a "good" initialization
- Handling parameters and latent variables differently
  - single (point) estimate of the parameters θ
  - distribution estimate of the latent variables z
- Two-step iterative algorithm
  - init: select a value for the parameters θ^0
  - E step:
    - compute the distribution over the latent variables
    - i.e., $\forall i, k \quad p\left(z_i = k \, \theta^t\right)$
    - these probabilities are called the "responsibilities"
  - M step:
    - find the best parameters θ given the responsibilities
    - i.e., $\theta^{t+1} = \ \arg\max_\theta$

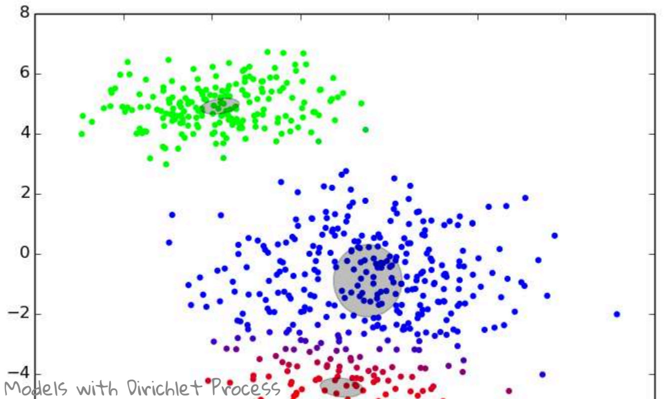
EM iterations



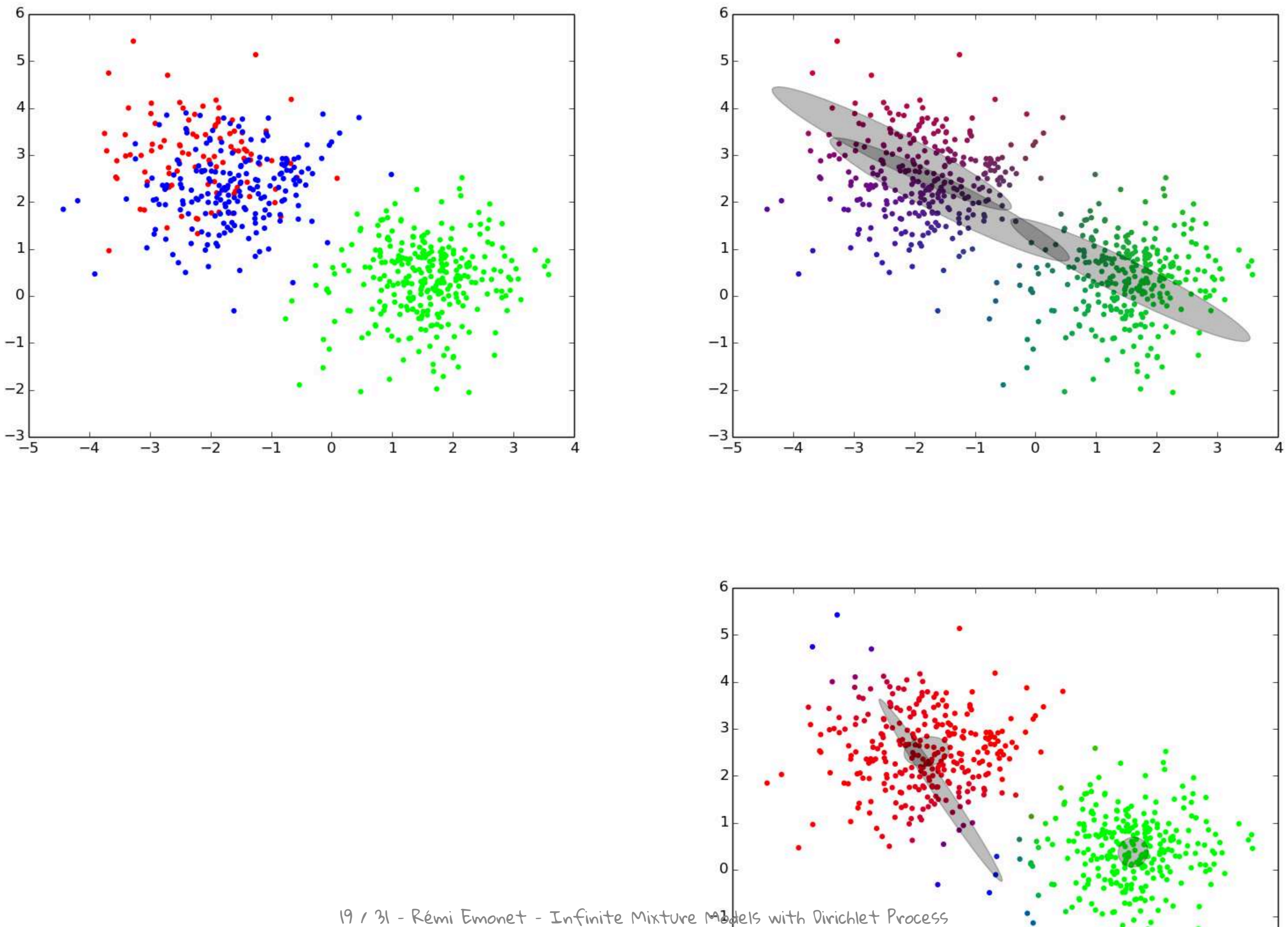
EM iterations



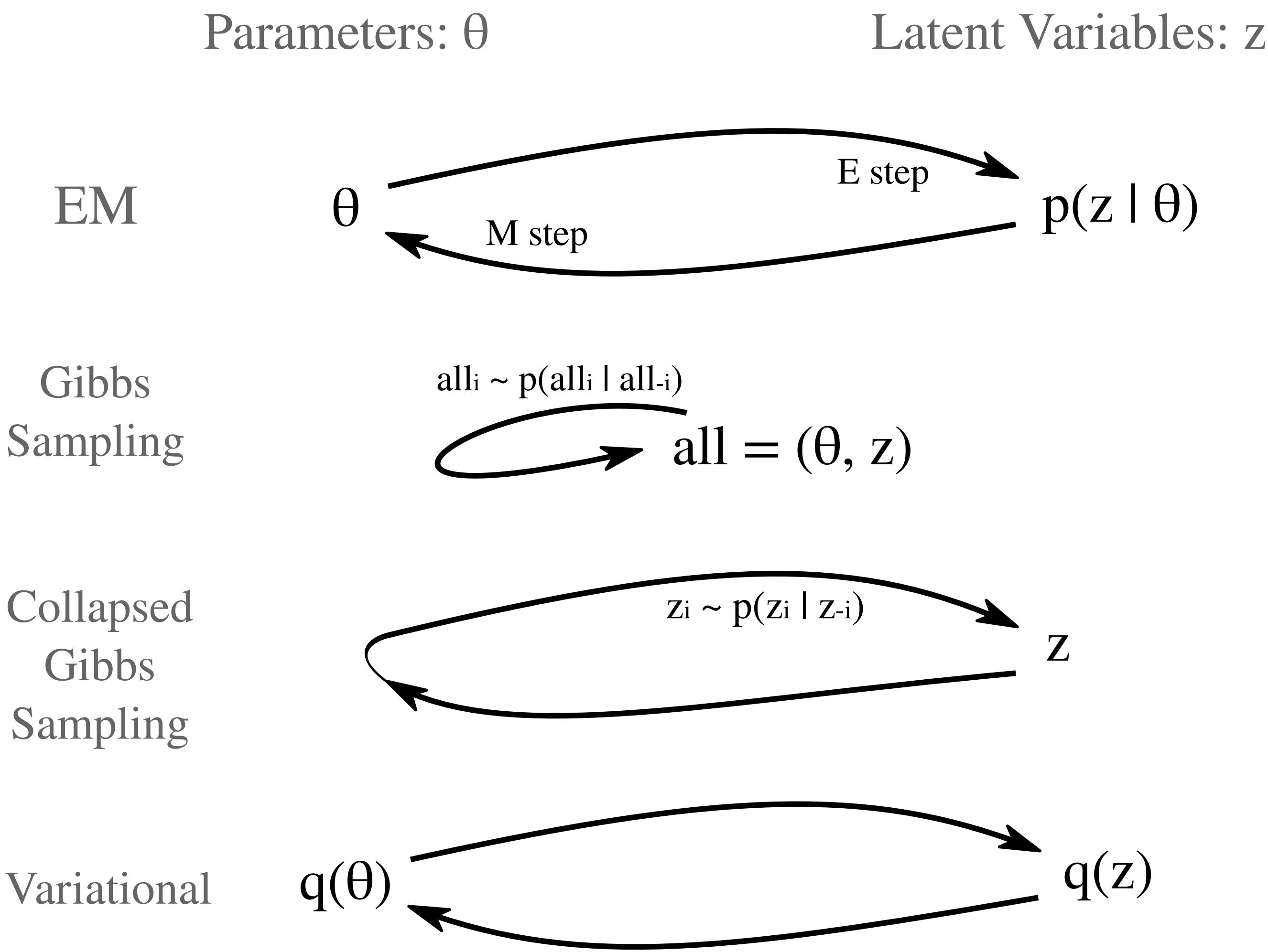




EM iterations



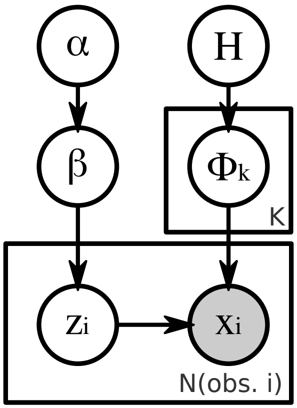
EM, Gibbs Sampling, Variational Inference

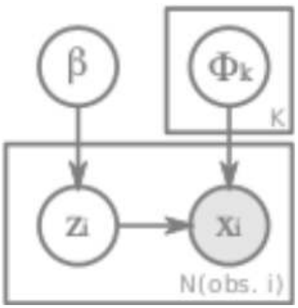


Beyond EM, using prior

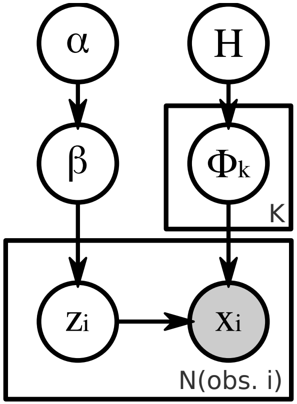
Prior on GMM

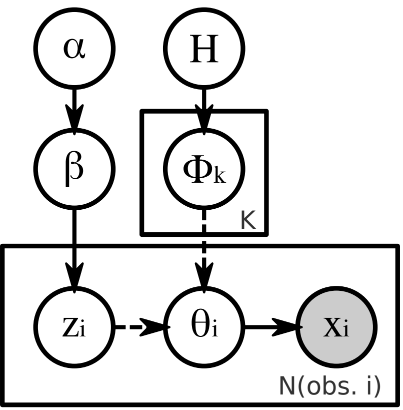
- It's all about prior
- Intution in EM
  - disappearing component
  - "minimal" weight
  - "minimal" variance





- We virtually add a few observations to each component
- α: causes the weights β to never be zero
  - Dirichlet distribution: $\beta \sim \text{Dirichlet}(\alpha)$
- H: adds some regularization on the variances Σ_k



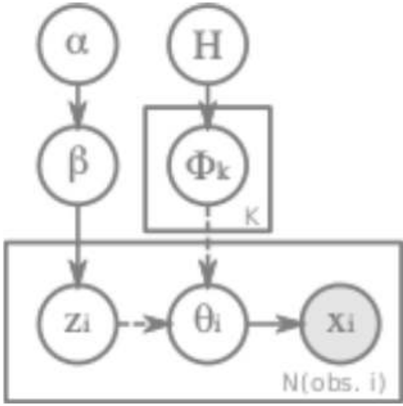


... Unfold

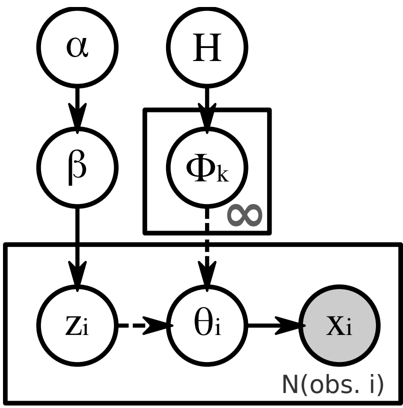
Dirichlet Processes

From GMM to DPGMM

- Plain GMM
  - the number of component K needs to be know
  - need to try multiple ones and do model selection
  - (prior are not handled by EM)



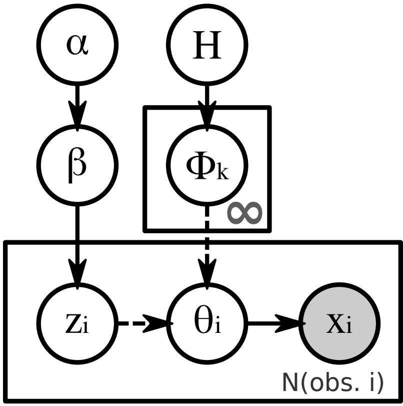
- Dirichlet Process GMM
  - a GMM with an infinity of components
  - with a proper prior
  - cannot use EM for inference (infinite vectors)



What is a Dirichlet Process, finally

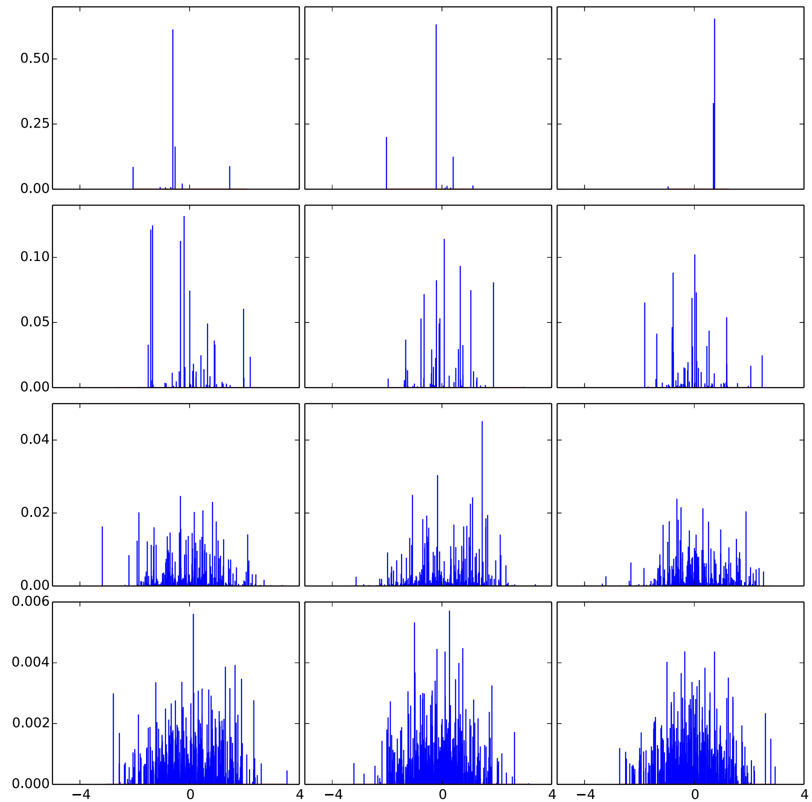
Dirichlet Process

- What?
  - a distribution over distributions
  - a prior over distributions
  - two parameters
    - a scalar α, the "concentration"
    - a "base" distribution H (any type)
  - a draw from a DP is a countably infinite sum of Diracs
- Related formulations
  - Definition: complicated :)
  - Stick breaking process (GEM)
    - as shown on the graphical model
    - e.g., a prior on the values of the weights
  - Chinese Restaurant Process (CRP)
    - how to generate the z_i, one by one, in sequence
  - Polya Urn

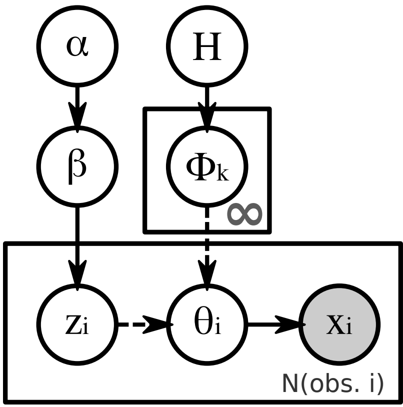


~~Definition~~ Example (Wikipedia)

Draws from the Dirichlet process DP(N(0,1), alpha). Each row uses a different alpha: 1, 10, 100 and 1000. A row contains 3 repetitions of the same experiment.



Stick breaking process (GEM)



- Each "atom" drawn from H
- Infinite vector drawn from a GEM process
  - $p_1 \sim Beta(1, \alpha)$; $\beta_1 = p_1$
  - $p_2 \sim Beta(1, \alpha)$; $\beta_2 = (1 - \beta_1) * p_2$
  - $p_3 \sim Beta(1, \alpha)$; $\beta_3 = (1 - \beta_1 - \beta_2) * p_3$
  - …
  - and $\forall k \quad \Phi_k \sim H$
- Denoted $\beta \sim GEM(\alpha)$
- then $G = \sum_{k=1}^{\infty} \beta_k \delta_{\Phi_k}$ is a draw from $DP(H, \alpha)$
- (GEM for Griffiths, Engen and McCloskey)

Polya Urn?

Chinese Restaurant Process (CRP)

- Gibbs sampling friendly
  - easy to get $p\left(z_i \, z^{-i}, \ldots\right)$
  - $p\left(z_i \, z^{-i}\right) = \ldots$