Lehrstuhl für Mathematische Statistik
Fakultät für Mathematik
Technische Universität München

Variational Bayesian Inference

Christoph Lienhard

München, 26. Januar 2018

Outline

Introduction
What is Variational Bayesian (VB) Inference
Applications

Basic Theory

Mixture Models
Introduction
Example with Normal Distribution
Application of VB Approach

Case Studies

Summary

What is Variational Bayesian Inference

Every Bayesian problem starts with:

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{\int p(\boldsymbol{y},\theta)d\theta}$$

Obstacle:

Complicated $p(\boldsymbol{y},\theta)$ and thereby intractable integral $\int p(\boldsymbol{y},\theta)d\theta$

Variational approach:

Approximate $p(\theta|\boldsymbol{y})$ with an easier distribution $q(\theta)$

Applications

Alternative to MCMC methods

In general: mixture models, e.g.

- model selection
- machine learning

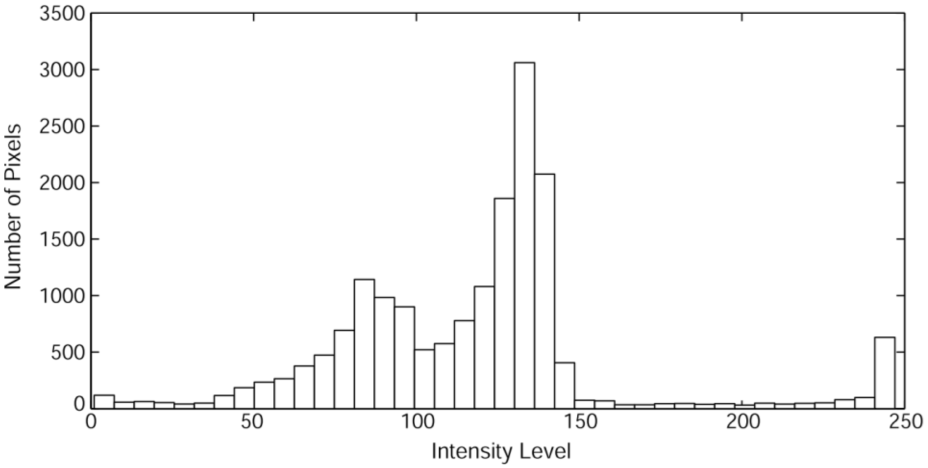
Basic Theory

Setup:

Data $\boldsymbol{y}$

A model with likelihood $p(\boldsymbol{y}|\theta_1, \theta_2)$
+ prior $p(\theta_1, \theta_2)$

Parameters θ_1, θ_2



Basic Theory

Goal:

Find posterior $p(\theta_1, \theta_2 | \boldsymbol{y})$

Find an approximation $q(\theta_1, \theta_2)$

⇒ minimize Kullback – Leibler divergence:

$$KL(q||p) = \int q(\theta_1, \theta_2) \log\left(\frac{q(\theta_1, \theta_2)}{p(\theta_1, \theta_2 | \boldsymbol{y})}\right) d\theta_1 d\theta_2$$

Basic Theory – Variational Calculus

Assume $q(\theta_1, \theta_2) = q_{\theta_1}(\theta_1) q_{\theta_2}(\theta_2)$

Applying variational calculus to $KL(q||p)$:

$$0 \stackrel{!}{=} \frac{\delta KL(q||p)}{\delta q_{\theta_1}}(\theta_1)$$

$$= \int q_{\theta_2}(\theta_2) \left(1 + \log\left(\frac{q_{\theta_1}(\theta_1) q_{\theta_2}(\theta_2)}{p(\theta_1, \theta_2|\boldsymbol{y})}\right)\right) d\theta_2$$

$$= 1 + \log\left(q_{\theta_1}(\theta_1)\right) + \int q_{\theta_2}(\theta_2) \left(\log\left(q_{\theta_2}(\theta_2)\right) - \log\left(p(\theta_1, \theta_2|\boldsymbol{y})\right)\right) d\theta_2$$

Basic Theory – Variational Calculus

Finally:

$$\log\left(q_{\theta_1}(\theta_1)\right) = \mathbf{E}_{q_{\theta_2}}\left(\log\left(p(\theta_1, \theta_2|\boldsymbol{y})\right)\right) + const.$$

Accordingly:

$$\log\left(q_{\theta_2}(\theta_2)\right) = \mathbf{E}_{q_{\theta_1}}\left(\log\left(p(\theta_1, \theta_2|\boldsymbol{y})\right)\right) + const.$$

Problem: q_{θ_1} dependent on q_{θ_2} and vice versa

=> Solution through iterative approach

Introduction

Basic Theory

Mixture Models

Introduction

Example with Normal Distribution

Application of VB Approach

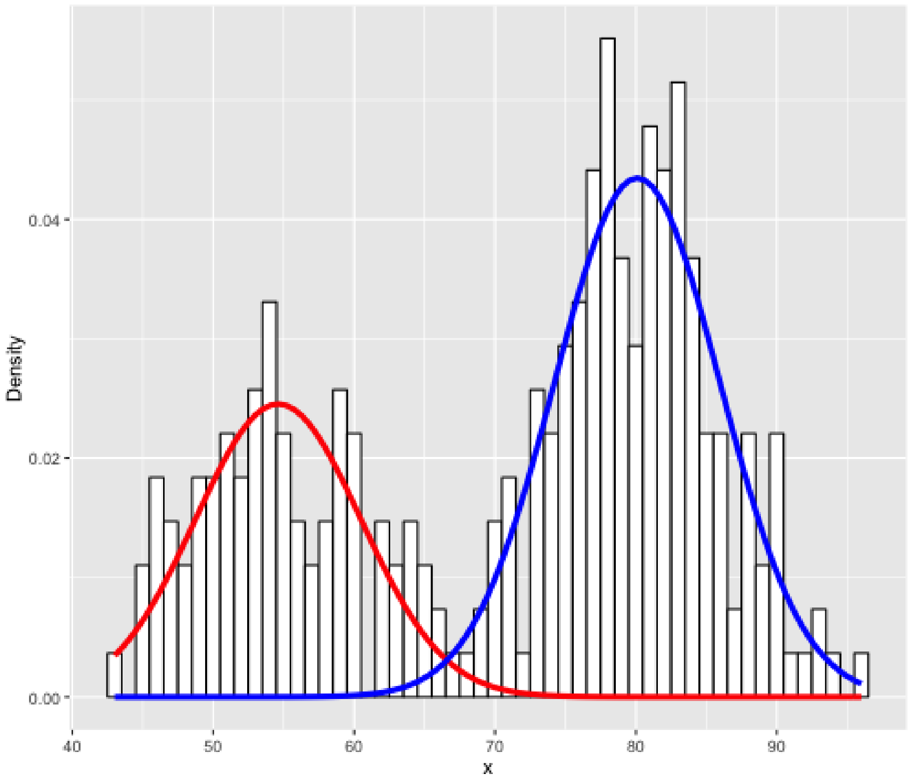
Case Studies

Summary

Mixture Models
Introduction

Subpopulations in overall population:

$$p(y_i|\lambda, \phi) = \sum_{j=1}^{K} \lambda_j \, p_j(y_i|\phi_j)$$



Mixture Models
Introduction

Introduce z_i:

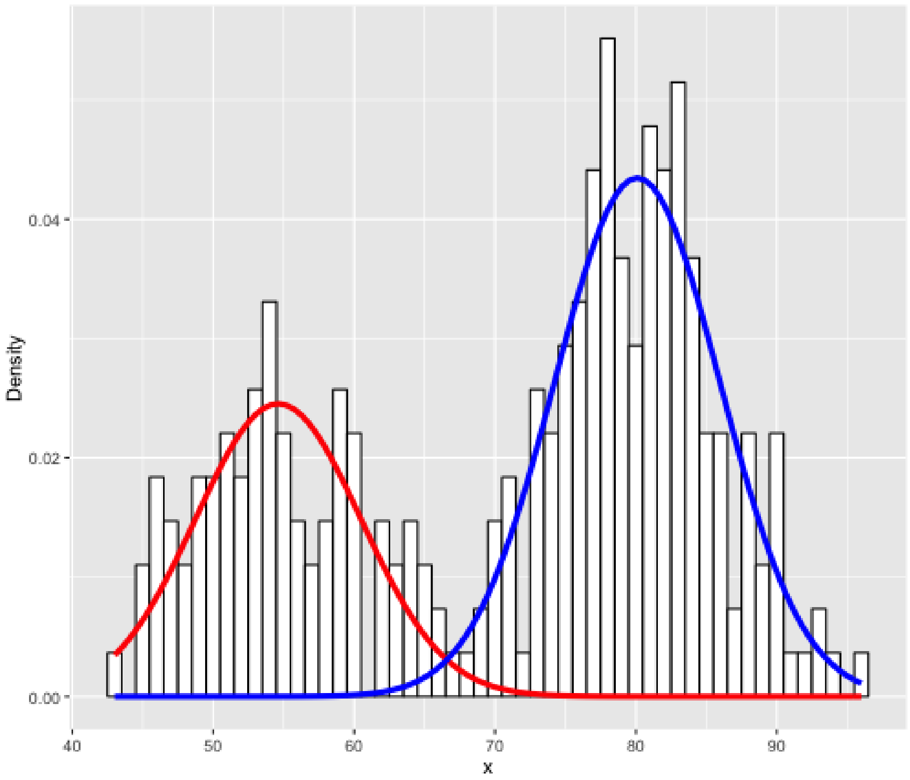
if observation y_i belongs to component m then

$$z_i = m \in \{1, 2, \dots, K\}$$

And:

$$p(y_i, z_i | \lambda, \phi) = \prod_{j=1}^{K} \left(\lambda_j \, p_j(y_i | \phi_j) \right)^{z_{ij}}$$

where $z_{ij} \coloneqq \delta_{j m_i}$



Mixture Models
Introduction

Introduce z_i:

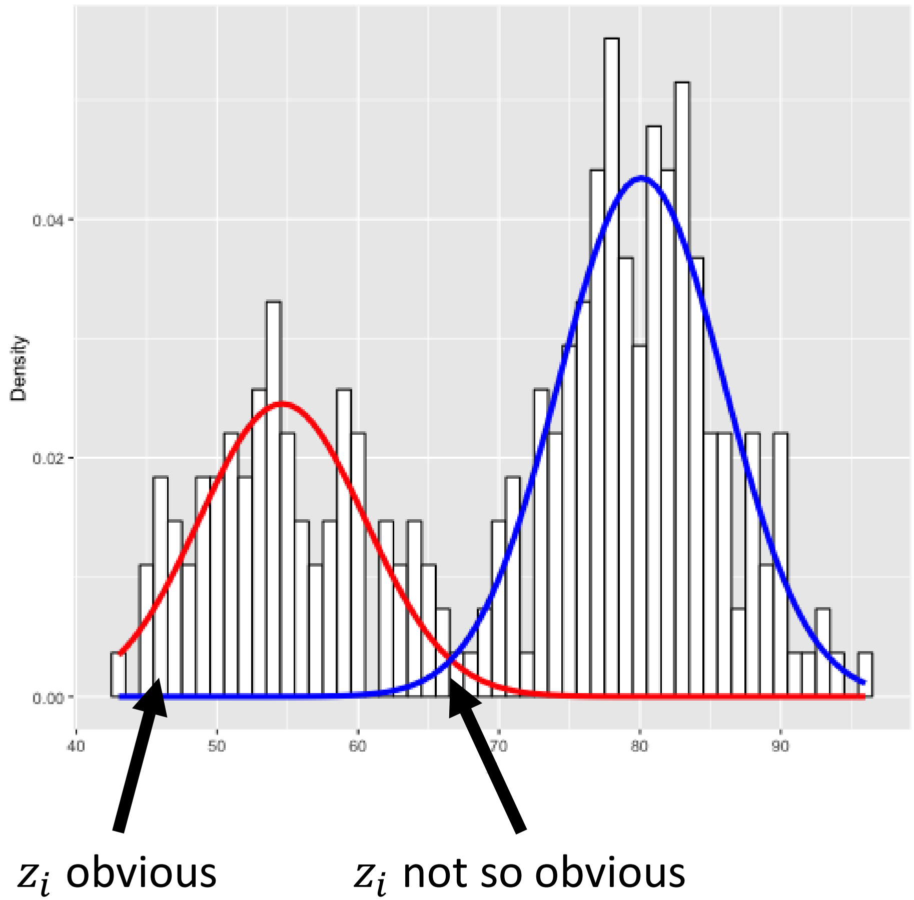
if observation y_i belongs to component m then

$$z_i = m \in \{1, 2, \dots, K\}$$

And:

$$p(y_i, z_i | \lambda, \phi) = \prod_{j=1}^{K} \left(\lambda_j \, p_j(y_i | \phi_j)\right)^{z_{ij}}$$

where $z_{ij} := \delta_{jm_i}$



Mixture Models
Example: Gaussians

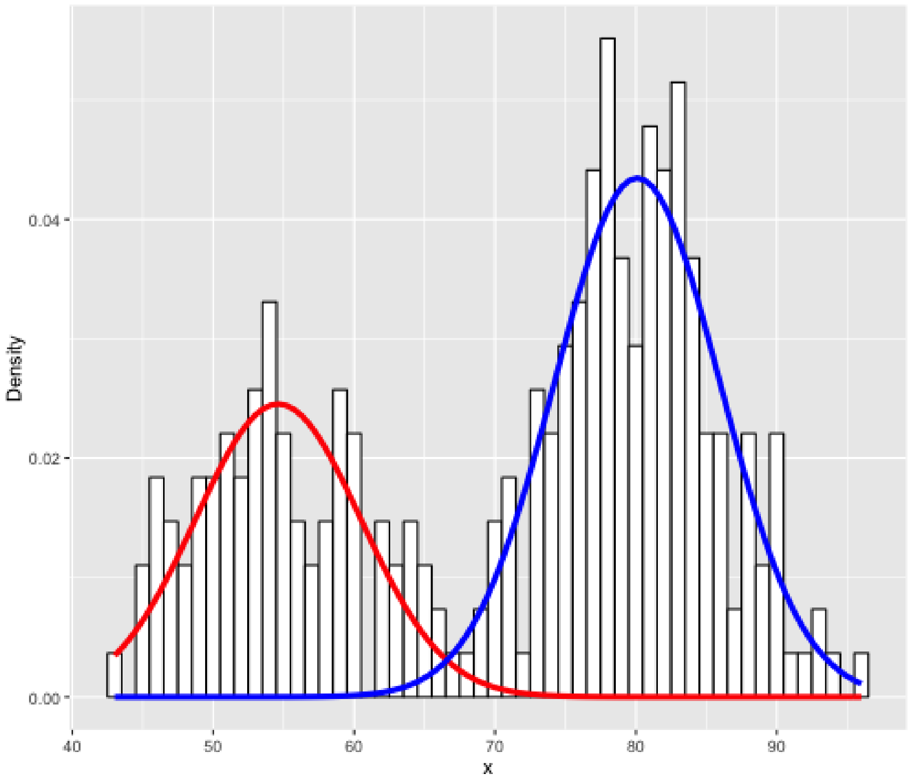
$$p(y_i, z_i | \lambda, \phi) = \prod_{j=1}^{K} \left(\lambda_j \, p_j\left(y_i | \phi_j\right)\right)^{z_{ij}}$$

Model Parameters:

$$\phi_j = \{\mu_j, \sigma_j^2\}$$

Set

$$p_j\left(y_i \middle| \phi_j\right) = N(y_i : \mu_j, \sigma_j^2)$$



Mixture Models
Example: Gaussians

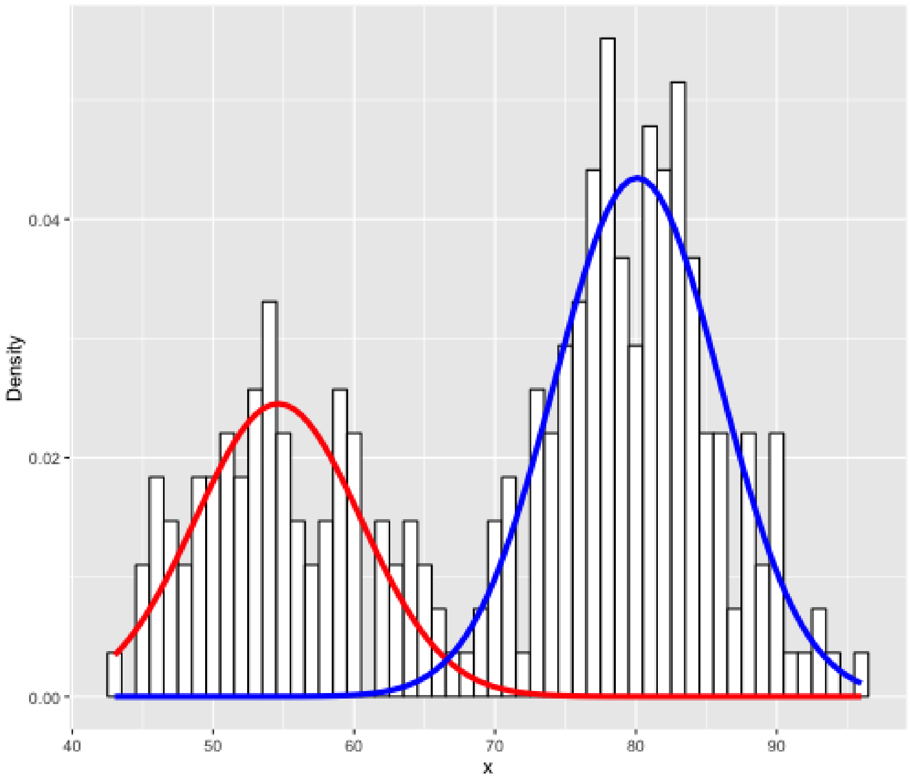
$$p(y_i, z_i | \lambda, \phi) = \prod_{j=1}^{K} \left(\lambda_j \, N\left(y_i : \phi_j\right)\right)^{z_{ij}}$$

Model Parameters:

$$\phi_j = \{\mu_j, \sigma_j^2\}$$

Prior factorization:

$$\begin{aligned} p(\lambda, \phi) &= p(\lambda) p(\mu, \sigma^2) \\ &= p(\lambda) p(\mu | \sigma^2) p(\sigma^2) \end{aligned}$$



Mixture Models
Example: Gaussians

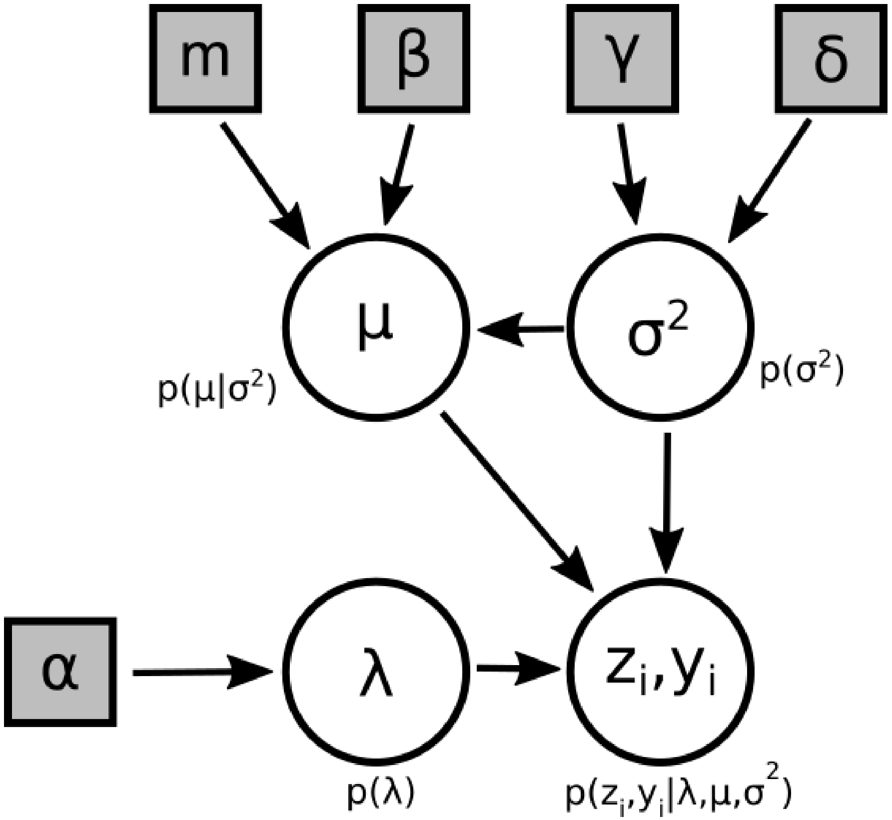
$$p(y_i, z_i | \lambda, \phi) = \prod_{j=1}^{K} \left(\lambda_j \; N\left(y_i : \mu_j, \sigma_j^2\right)\right)^{z_{ij}}$$

Prior distributions:

$$p(\lambda) \propto \prod_{j=1}^{K} \lambda_i^{\alpha_j - 1}$$

$$p(\sigma^2) \propto \prod_{j=1}^{K} \sigma_j^{-\gamma_j - 2} \exp\left(-\frac{\delta_j}{2\sigma_j^2}\right)$$

$$p(\mu | \sigma^2) \propto \prod_{j=1}^{K} N\left(\mu_j : m_j, \beta_j^{-1} \sigma_j^2\right)$$



Mixture Models
Example: Gaussians

Goal:

find $p(\lambda, \phi, \boldsymbol{z}|\boldsymbol{y})$

Bayesian Analysis:

$$p(\lambda, \phi, \boldsymbol{z}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{z}|\lambda, \mu, \sigma^2)p(\lambda)p(\mu|\sigma^2)p(\sigma^2)}{\int \sum_{\boldsymbol{z}} p(\boldsymbol{y}, \boldsymbol{z}, \lambda, \mu, \sigma^2)\, d\lambda d\mu d\sigma^2}$$

Easy, right?

Mixture Models
Example: Gaussians

$$p(\boldsymbol{y}, \boldsymbol{z}, \lambda, \mu, \sigma^2) \propto \prod_{j=1}^{K} \lambda_j^{\alpha_j - 1 + \sum_{i=1}^{N} z_{ij}} \prod_{j=1}^{K} \left[\sigma_j^{-\gamma_j - 1 - \sum_{i=1}^{N} z_{ij}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} z_{ij} \frac{\left(y_i - \mu_j\right)^2}{\sigma_j^2} \right) \right.$$

$$\left. \times \exp\left(-\frac{1}{2\sigma_j^2} \left(\beta_j \left(\mu_j - m_j\right)^2 + \delta_j \right) \right) \right]$$

Mixture Models
Application of VB approach

The Variational Approach:

$$q(\boldsymbol{z}, \lambda, \mu, \sigma^2) = q_\lambda(\lambda) q_\mu(\mu|\sigma^2) q_{\sigma^2}(\sigma^2) \prod_{i=1}^{N} q_{z_i}(z_i)$$

From basic theory:

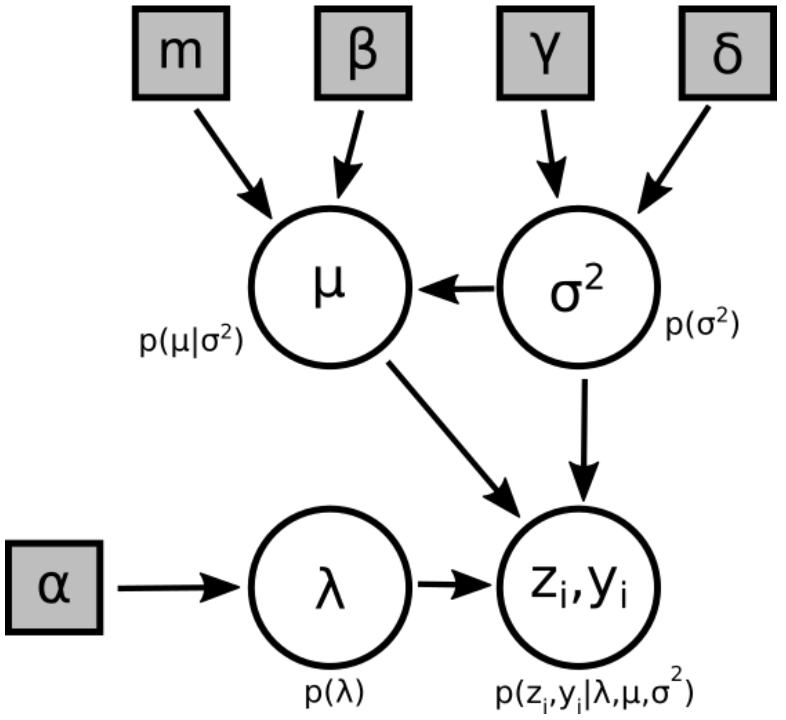
$$\log(q_i) = \mathbf{E}_{q_{j \neq i}}\left[\log\left(p(\boldsymbol{z}, \lambda, \mu, \sigma^2|\boldsymbol{y})\right)\right] + const.$$

Mixture Models
Application of VB approach

Turns out:

q_i's for model parameters same as priors, but updated parameters:

Parameter	Dependencies
α'	$\alpha, q_{\mathbf{z}}$
β'	$\beta, q_{\mathbf{z}}$
γ'	$\gamma, q_{\mathbf{z}}$
m'	$m, q_{\mathbf{z}}, \beta, \beta', y$
δ'	$\delta, q_{\mathbf{z}}, \beta, \beta', y, m, m'$



Mixture Models
Application of VB approach

Remaining question: What is $q_{\boldsymbol{z}}$?

$$\log(q_{\boldsymbol{z}}) = \mathbf{E}_{q_\phi}\left[\log\left(p(\boldsymbol{z}, \boldsymbol{y}, \lambda, \mu, \sigma^2)\right)\right] + const.$$

Define $q_{ij} \coloneqq q_{z_i}(z_i = j)$

$$q_{ij} \propto \exp\left\{\mathbf{E}_{q_\lambda}\left[\log\left(\lambda_j\right)\right] - \frac{1}{2}\mathbf{E}_{q_{\sigma^2}}\left[\log\left(\sigma_j^2\right)\right] - \mathbf{E}_{q_\mu, q_{\sigma^2}}\left[\frac{\left(y_i - \mu_j\right)^2}{2\sigma_j^2}\right]\right\}$$

Mixture Models
Application of VB approach

Use iterative algorithm:

Set initial number of components k

Set initial parameters α, β, y, m, δ

Set initial distribution q_{ij}

while not converged:

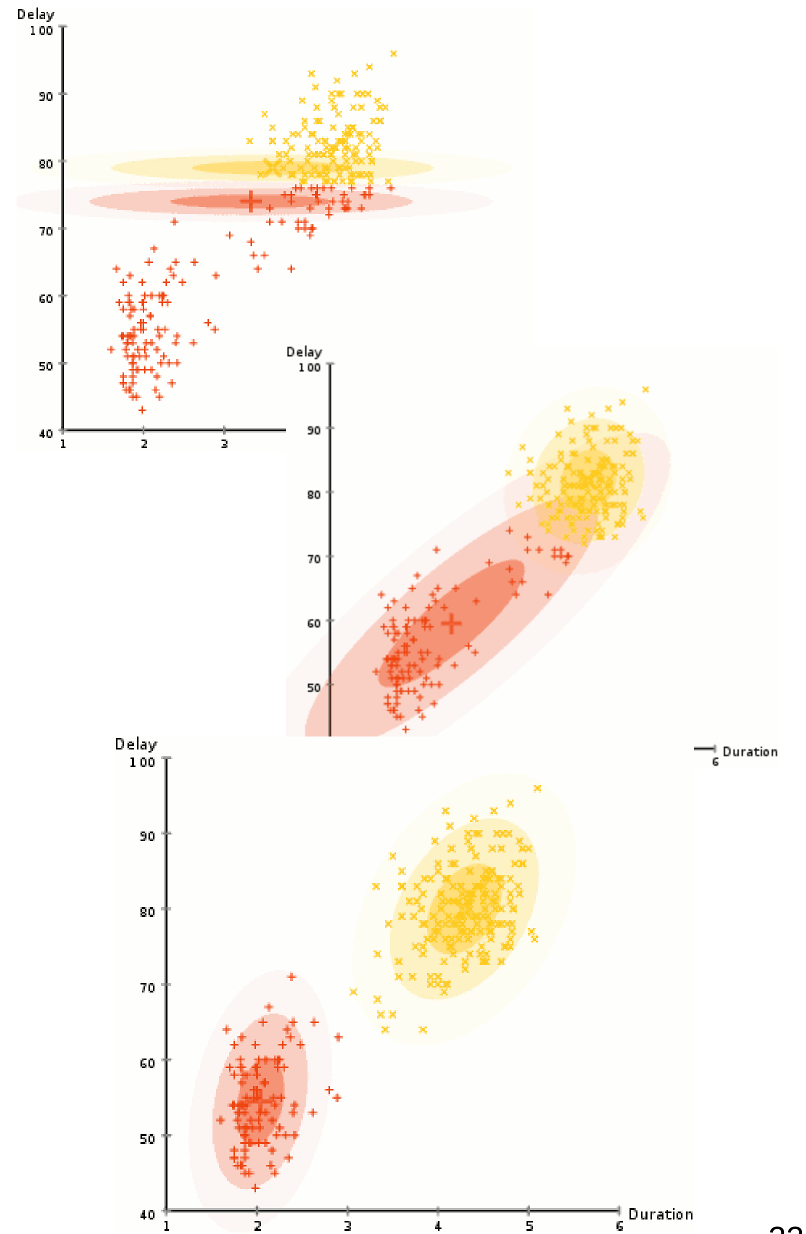
  update α', β', y', m', δ'

  update q_{ij}

  eliminate unimportant components

  check if converged

end while



Introduction

Basic Theory

Mixture Models

Case Studies

Pork Carcass

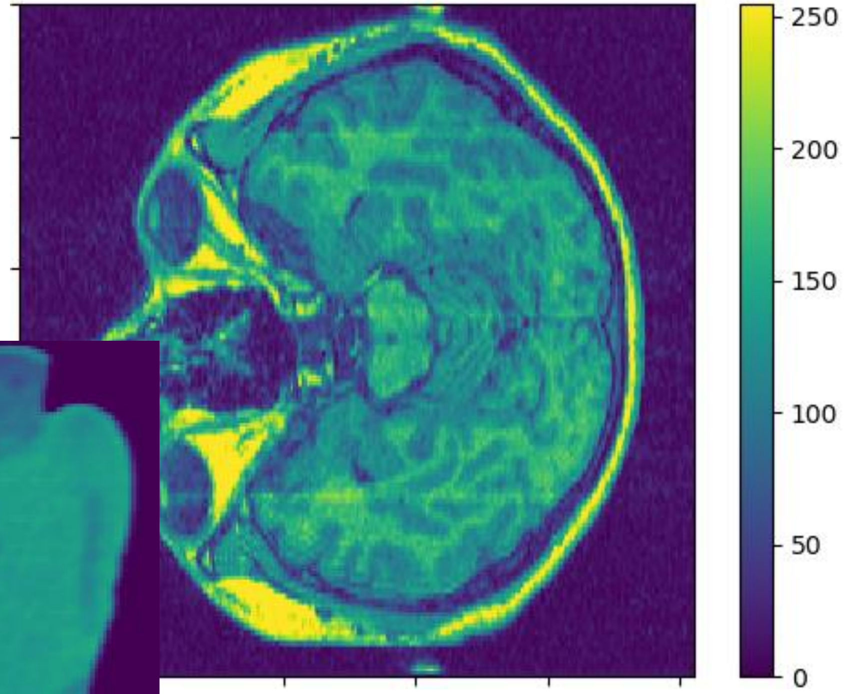
Human Head

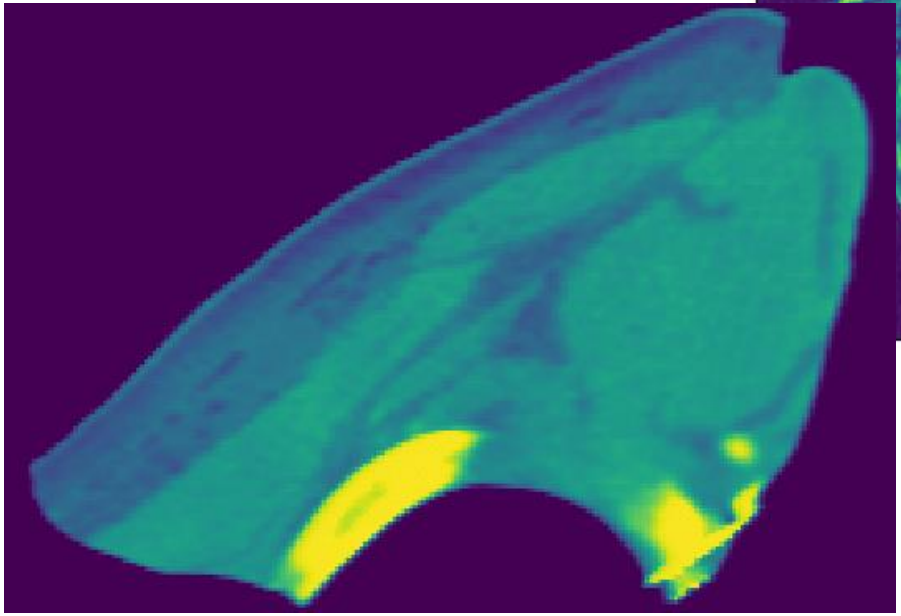
Summary

Case Studies

CT-Scans:

- Pork
- Human Head (3D)





Case Study - Pork
Setup

Dimensions:

148 x 218 pixels

i.e. $N = 32\,264$

Intensity Resolution:

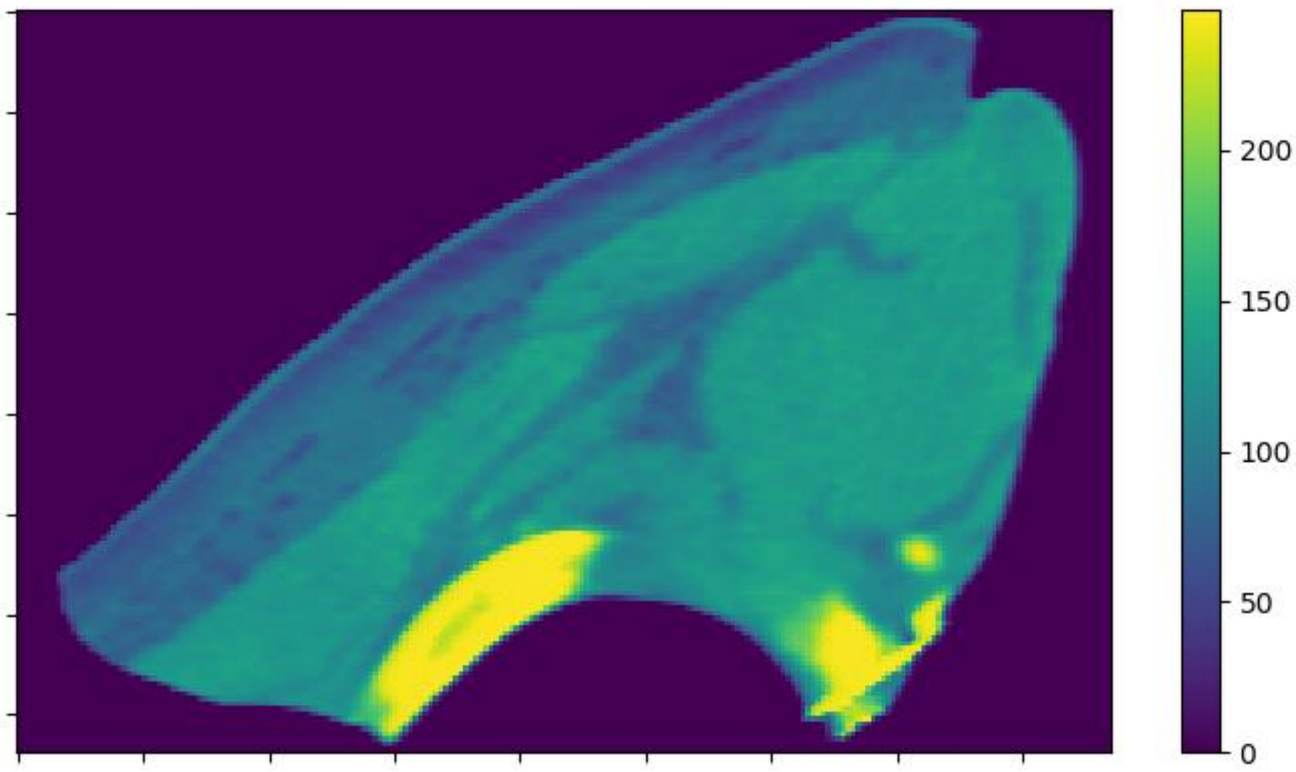
256

Expected Components:

bone

fat

muscle



Case Study - Pork Setup

Dimensions:

148 x 218 pixels

i.e. $N = 32\,264$

Intensity Resolution:

256

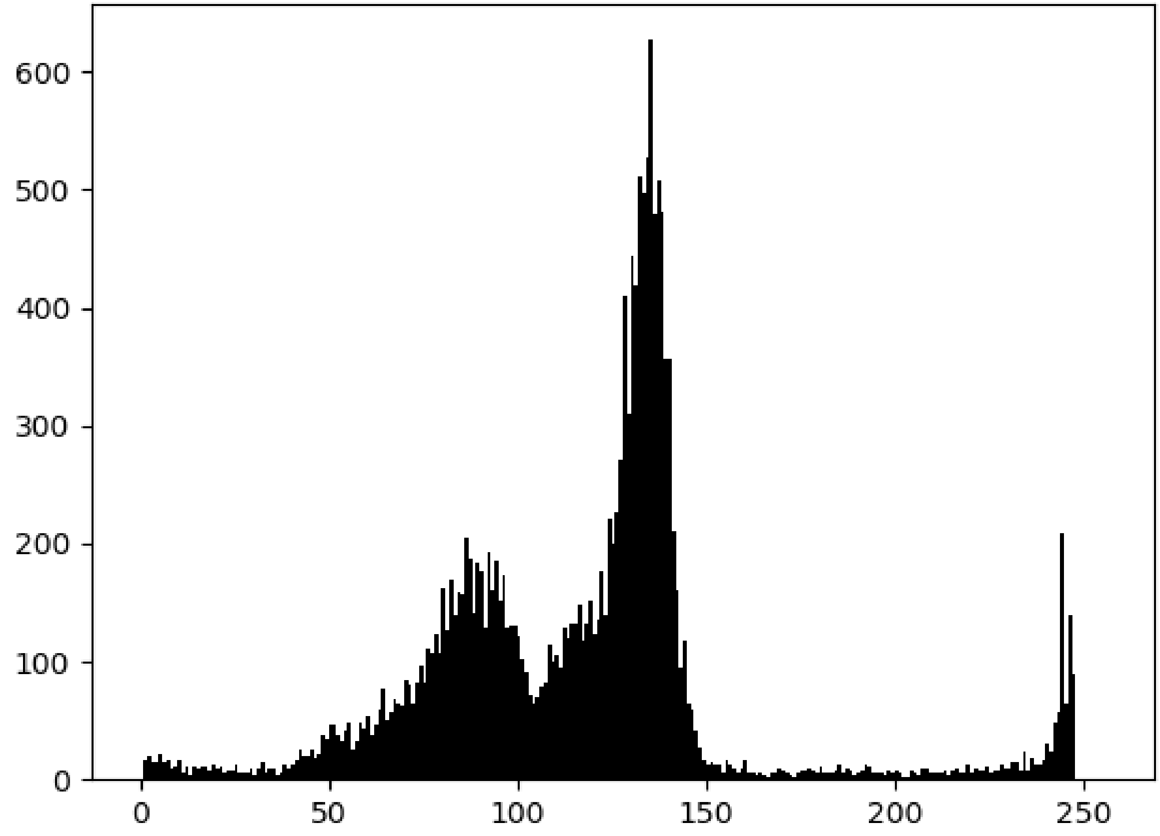
Expected Components:

bone

fat

muscle

Histogram of pixel intensities

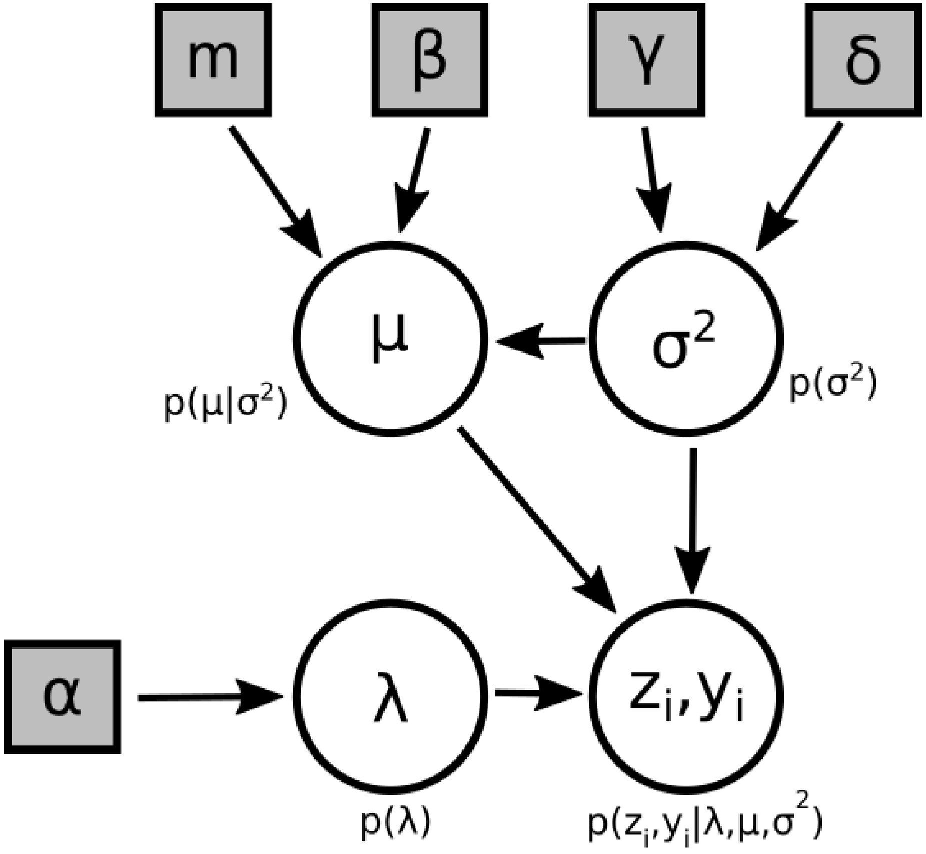


Case Study - Pork
Setup

Start with $k = 15$ components

Initial values for hyperparameters:

$\alpha_j^{(0)}$: all weights the same



Case Study - Pork
Setup

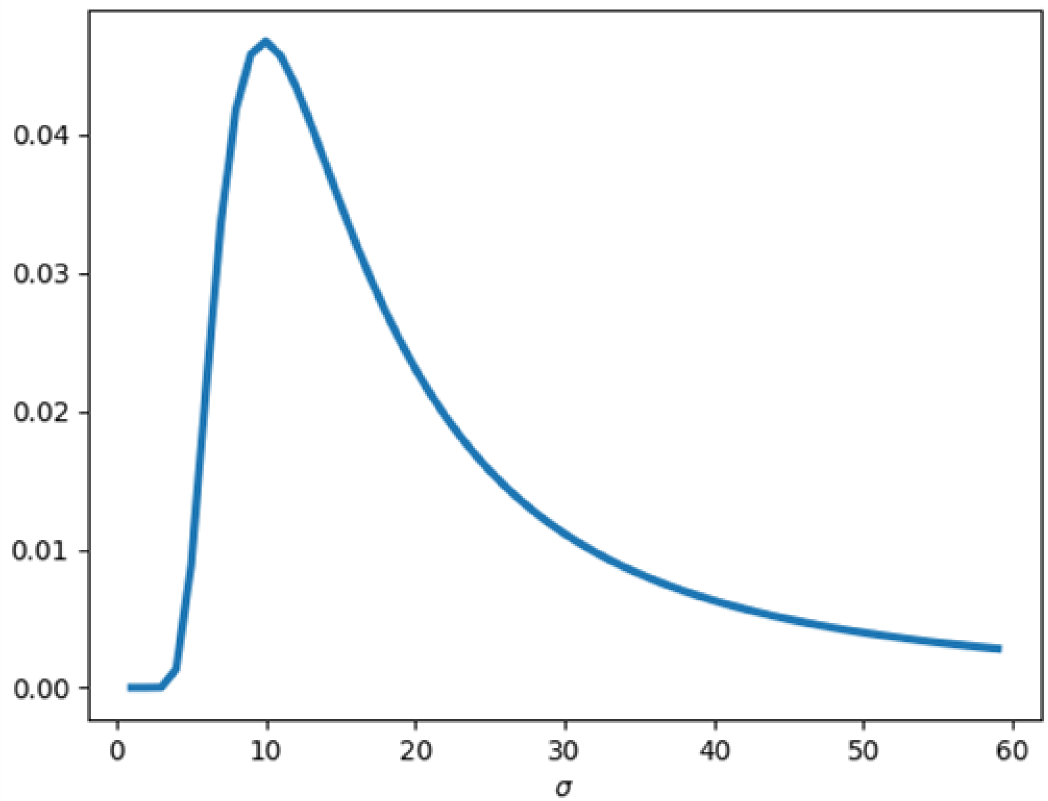
Start with $k = 15$ components

Initial values for hyperparameters:

$\alpha_j^{(0)}$: all weights the same

$\gamma_j^{(0)}$, $\delta_j^{(0)}$

Inverse Gamma for σ^2



Case Study - Pork Setup

Start with $k = 15$ components

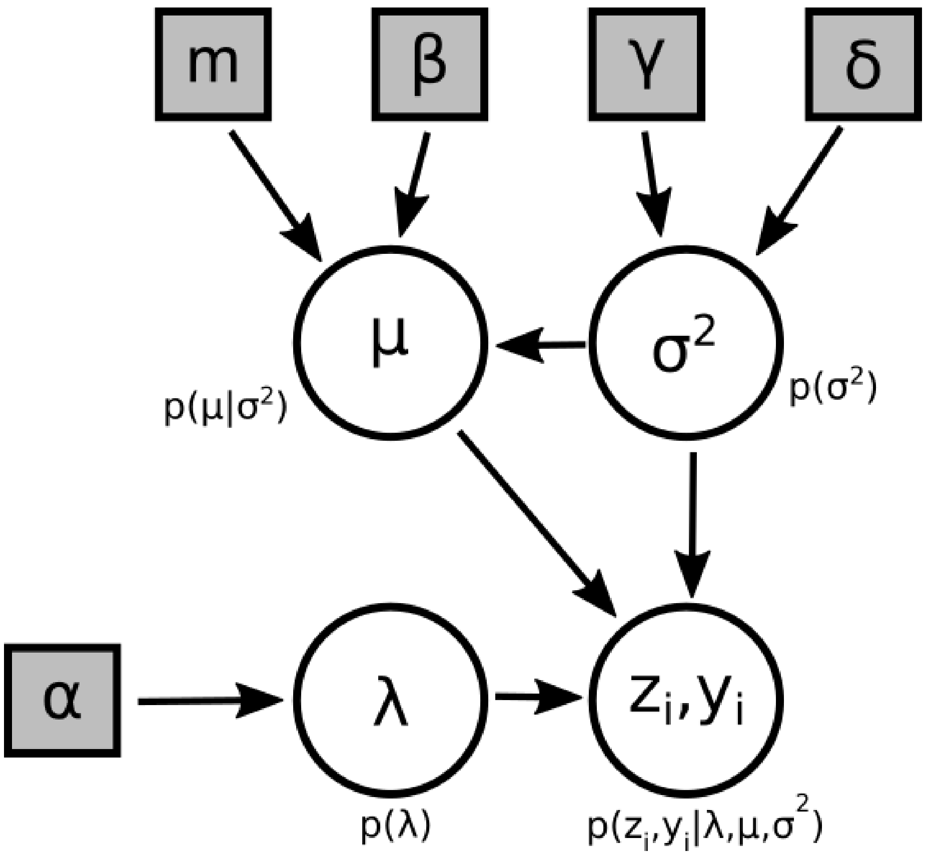
Initial values for hyperparameters:

$\alpha_j^{(0)}$: all weights the same

$\gamma_j^{(0)}$, $\delta_j^{(0)}$

$m_j^{(0)} = 125$

$\beta_j^{(0)} = 0.05$



Case Study - Pork
Setup

Start with $k = 15$ components

Initial values for hyperparameters:

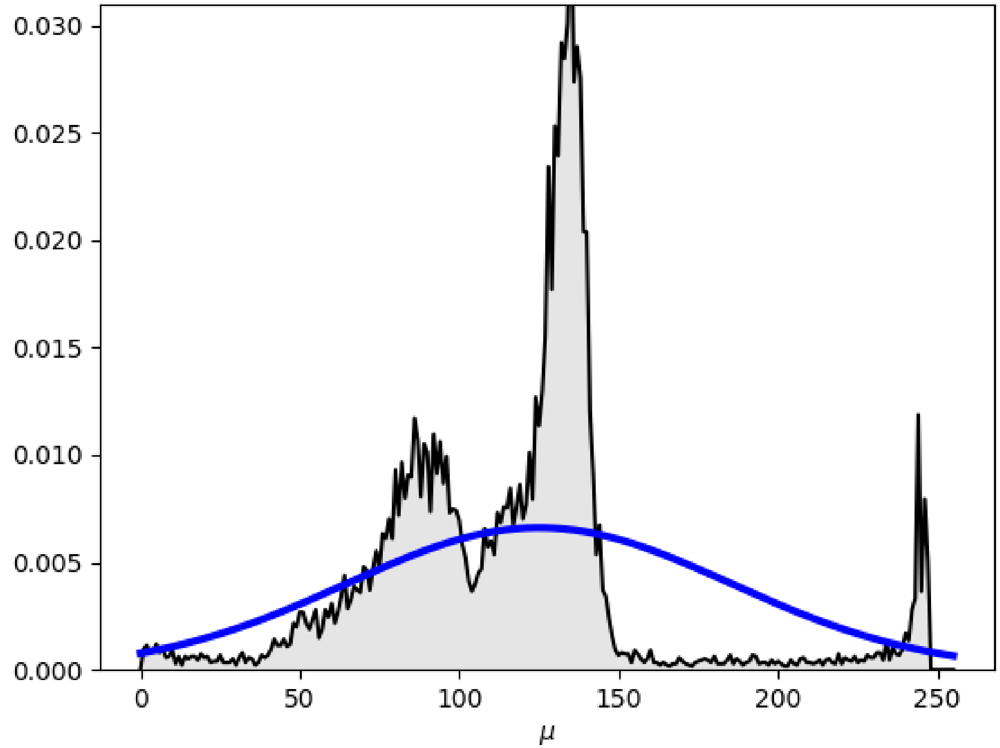
$\alpha_j^{(0)}$: all weights the same

$\gamma_j^{(0)}, \delta_j^{(0)}$

$m_j^{(0)} = 125$

$\beta_j^{(0)} = 0.05$

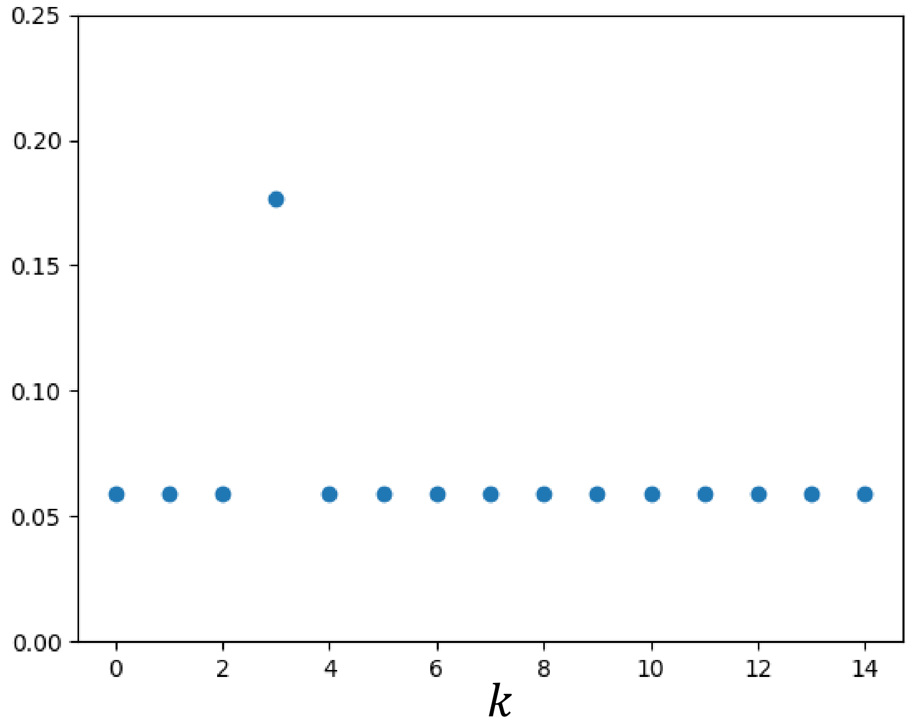
Normal Distribution for μ



Case Study - Pork
Setup

Initial allocation for data $\boldsymbol{y}$: Random with uniform distribution over k components

$q(z_i = j)$:

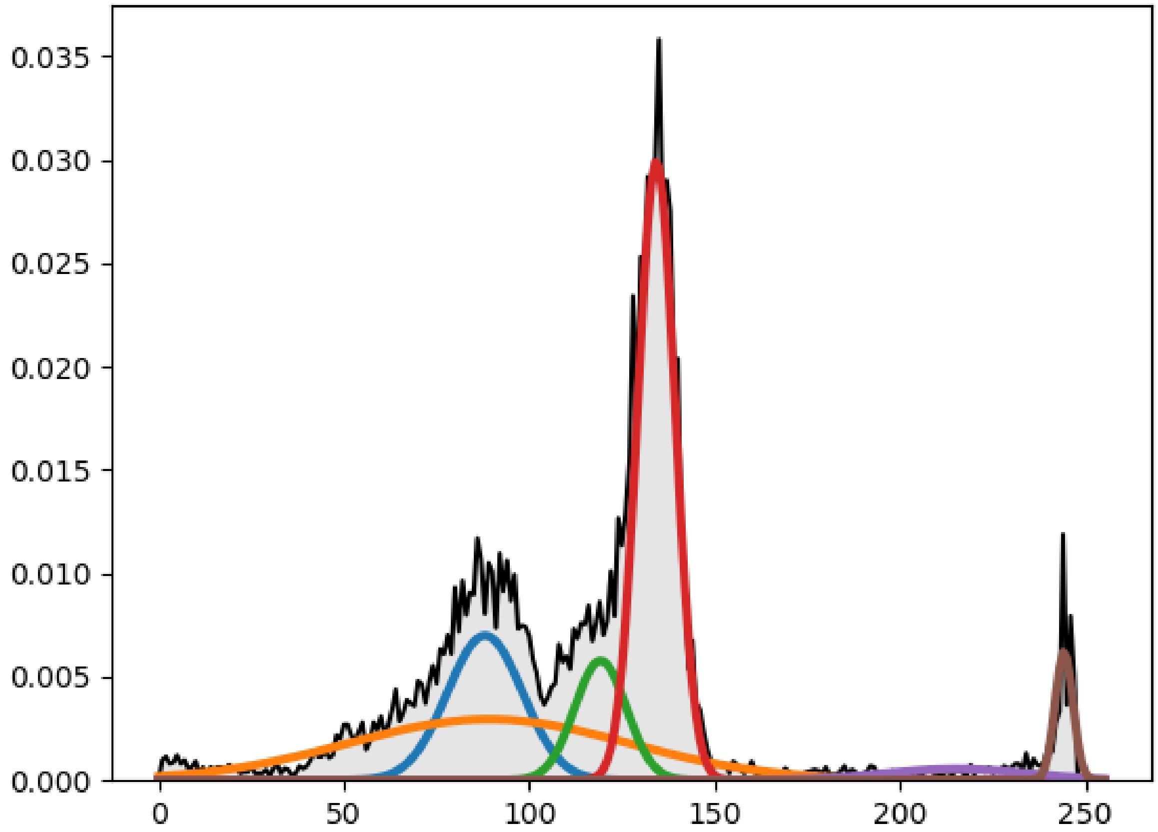


Case Study - Pork Solution

$$\lambda_j q_j(y_i|\mu_j, \sigma_j) = \lambda_j N(y_i : \mu_j, \sigma_j)$$

6 Components

μ_j	σ_j	λ_j
88.1	10.1	0.18
89.1	37.8	0.28
119.2	7.0	0.10
134.3	5.0	0.38
215.3	19.3	0.03
244.2	2.4	0.04

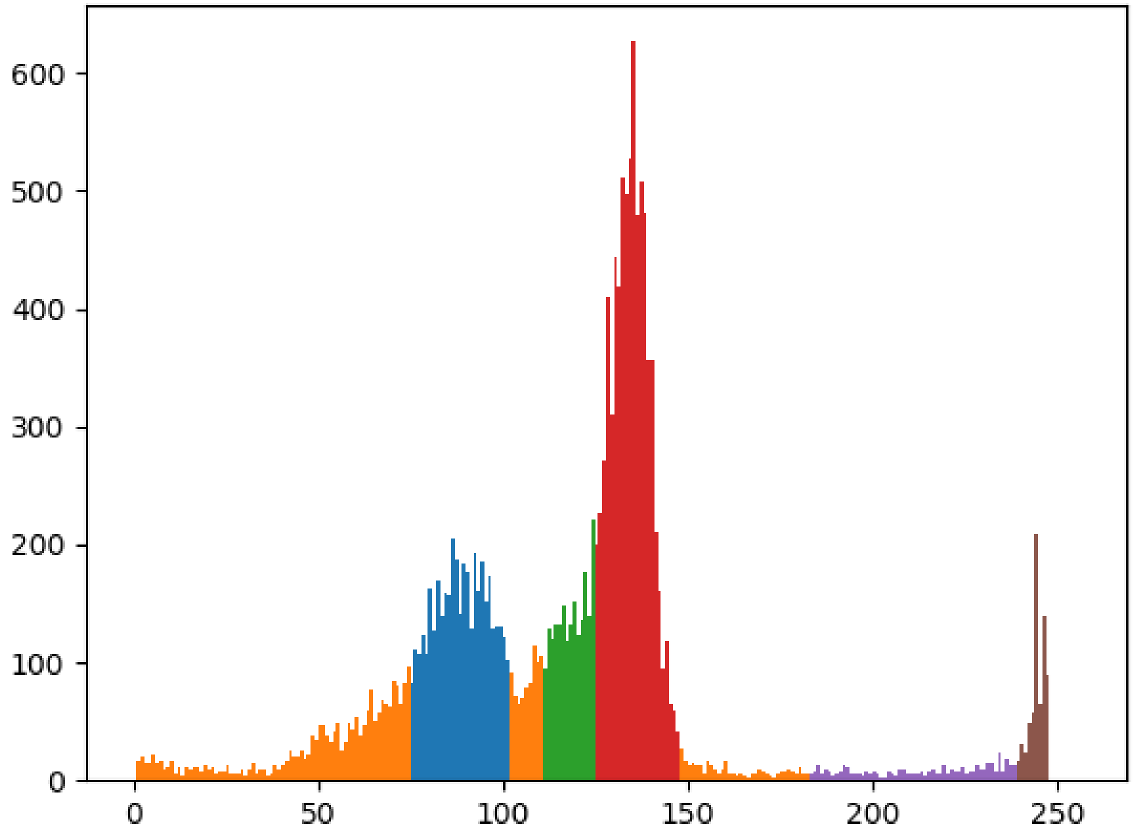


Case Study - Pork Solution

6 Components

μ_j	σ_j	λ_j
88.1	10.1	0.18
89.1	37.8	0.28
119.2	7.0	0.10
134.3	5.0	0.38
215.3	19.3	0.03
244.2	2.4	0.04

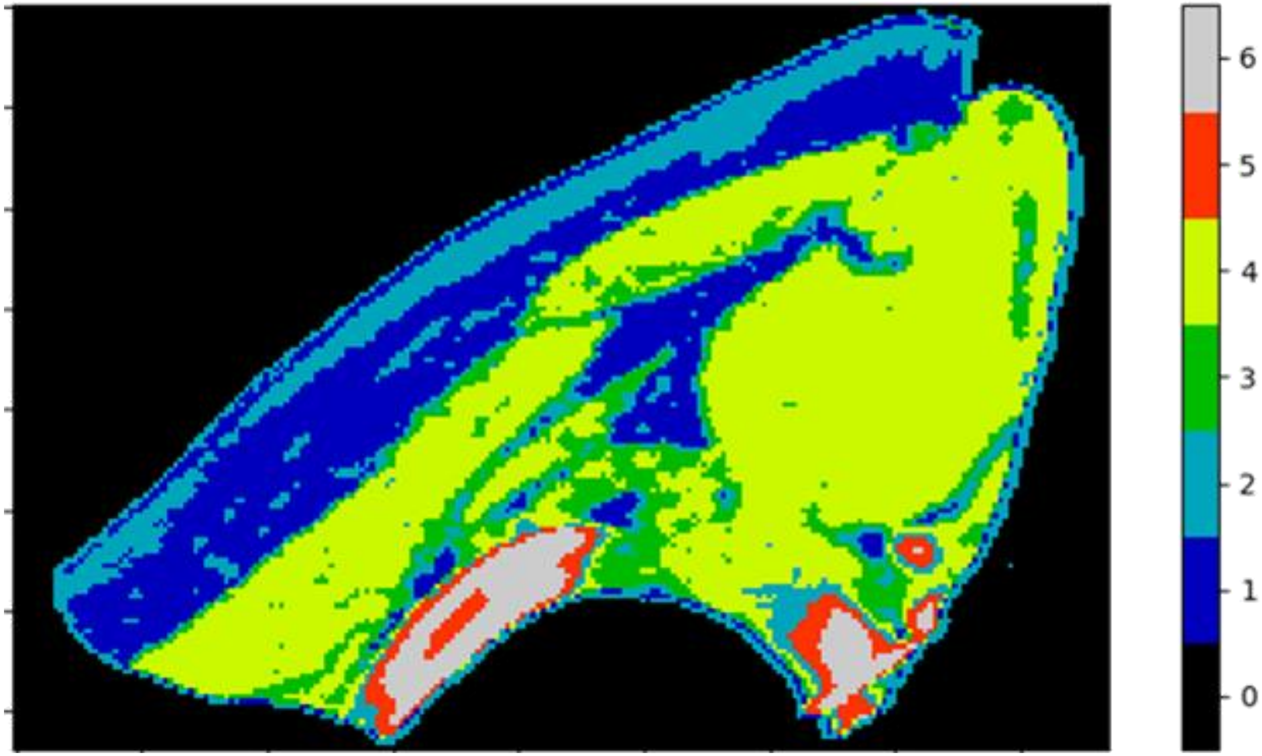
$$\underset{j}{\operatorname{argmax}}\, q_{ij}$$



Case Study - Pork
Solution

6 Components

μ_j	σ_j	λ_j
88.1	10.1	0.18
89.1	37.8	0.28
119.2	7.0	0.10
134.3	5.0	0.38
215.3	19.3	0.03
244.2	2.4	0.04

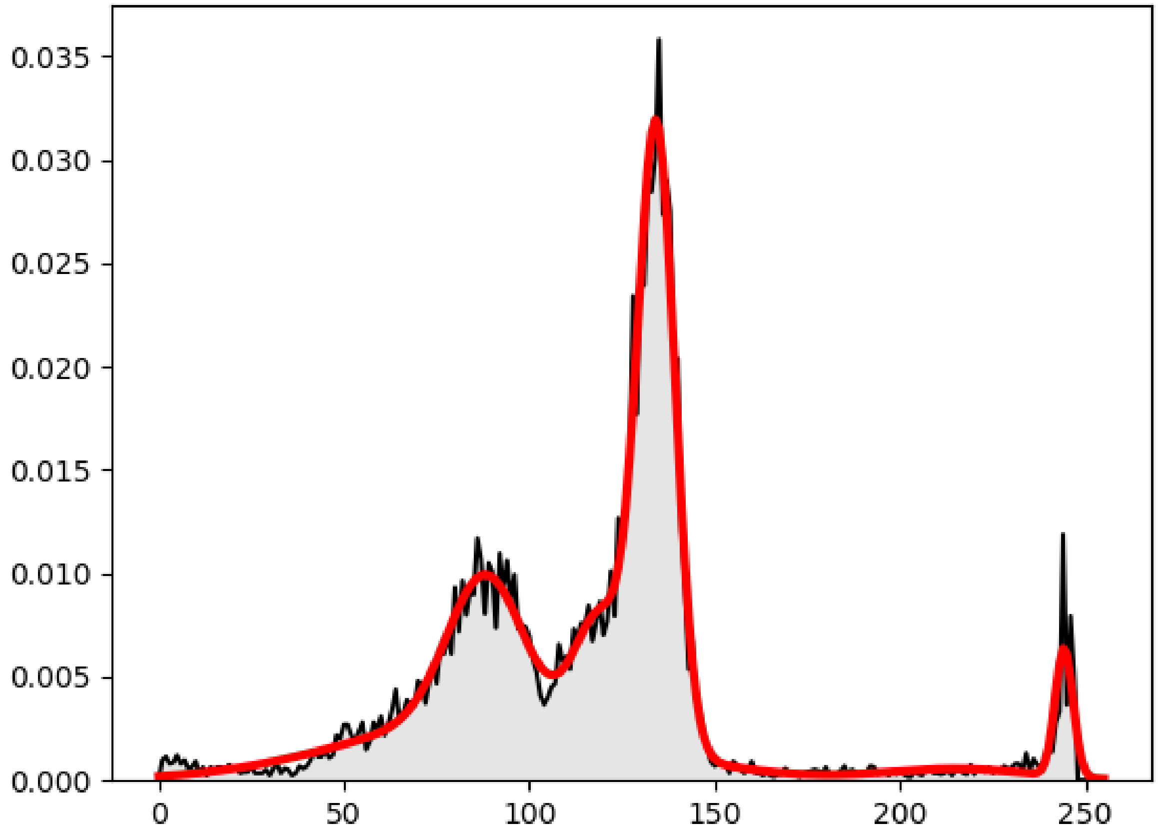


Case Study - Pork Solution

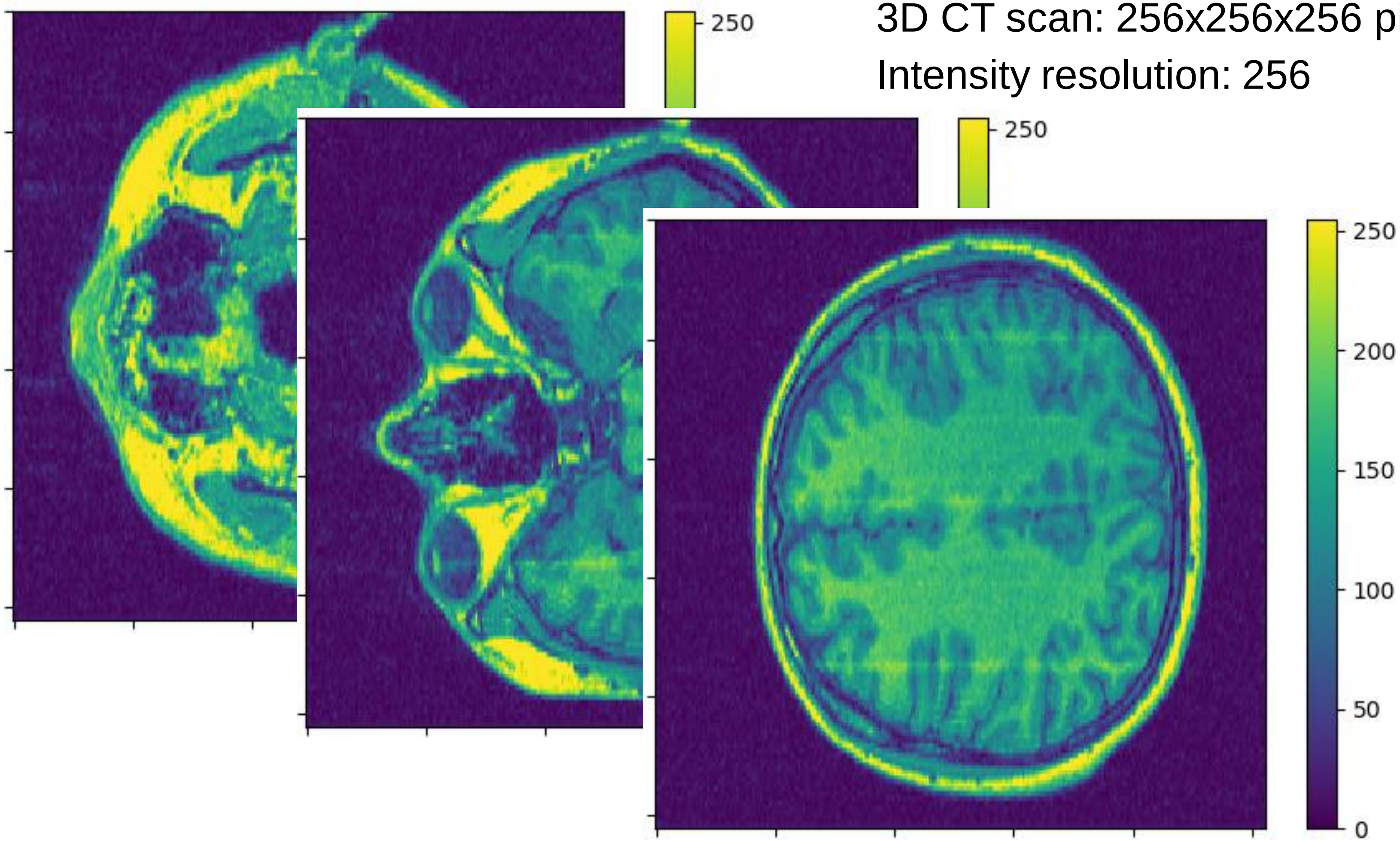
$$q(y_i|\mu,\sigma) = \sum_{j=1}^{k} \lambda_j N(y_i : \mu_j, \sigma_j)$$

6 Components

μ_j	σ_j	λ_j
88.1	10.1	0.18
89.1	37.8	0.28
119.2	7.0	0.10
134.3	5.0	0.38
215.3	19.3	0.03
244.2	2.4	0.04



Case Study – Human Head



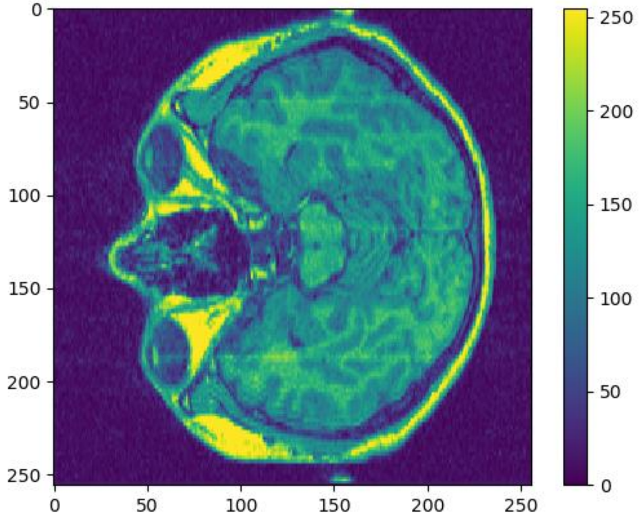
3D CT scan: 256x256x256 pixels

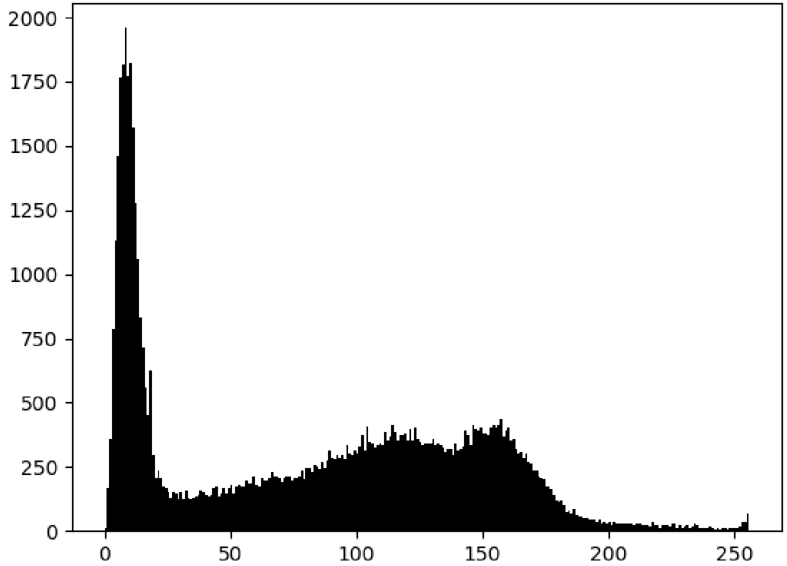
Intensity resolution: 256

Case Study – Human Head
Setup

Huge amount of data points => Use just one slice for approximation

Same initial parameters as pork data

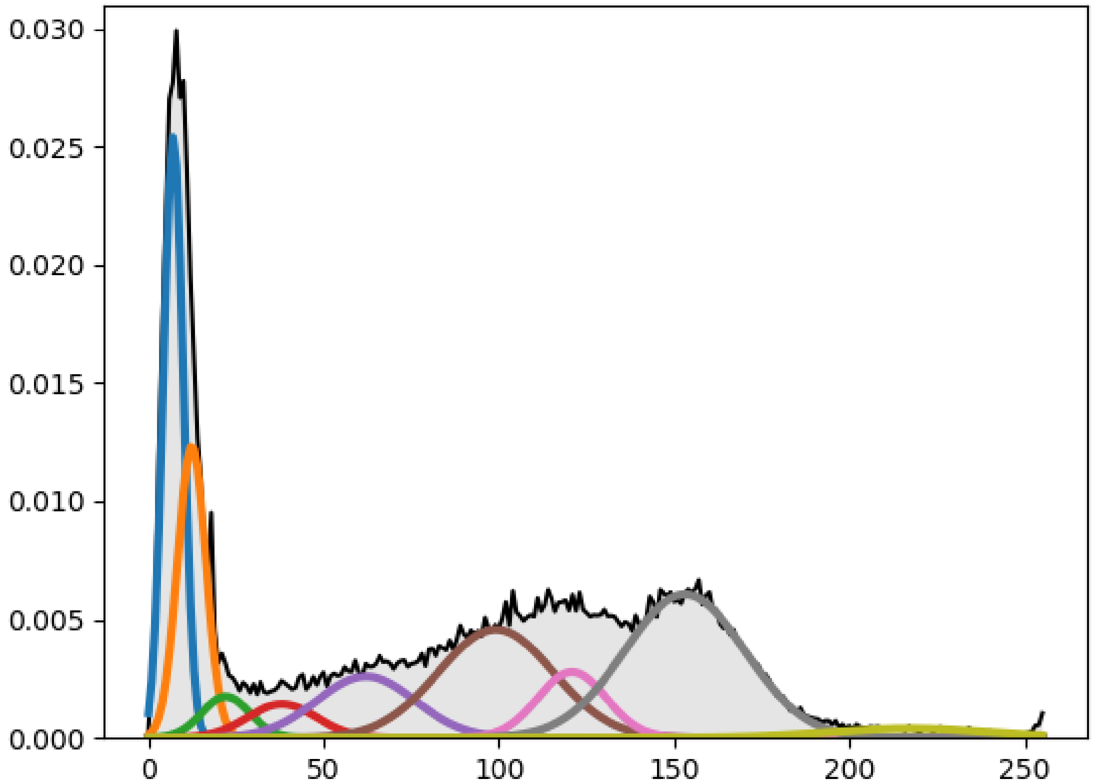




Case Study – Human Head
Solution

$$\lambda_j q_j(y_i|\mu_j,\sigma_j) = \lambda_j N(y_i : \mu_j,\sigma_j)$$

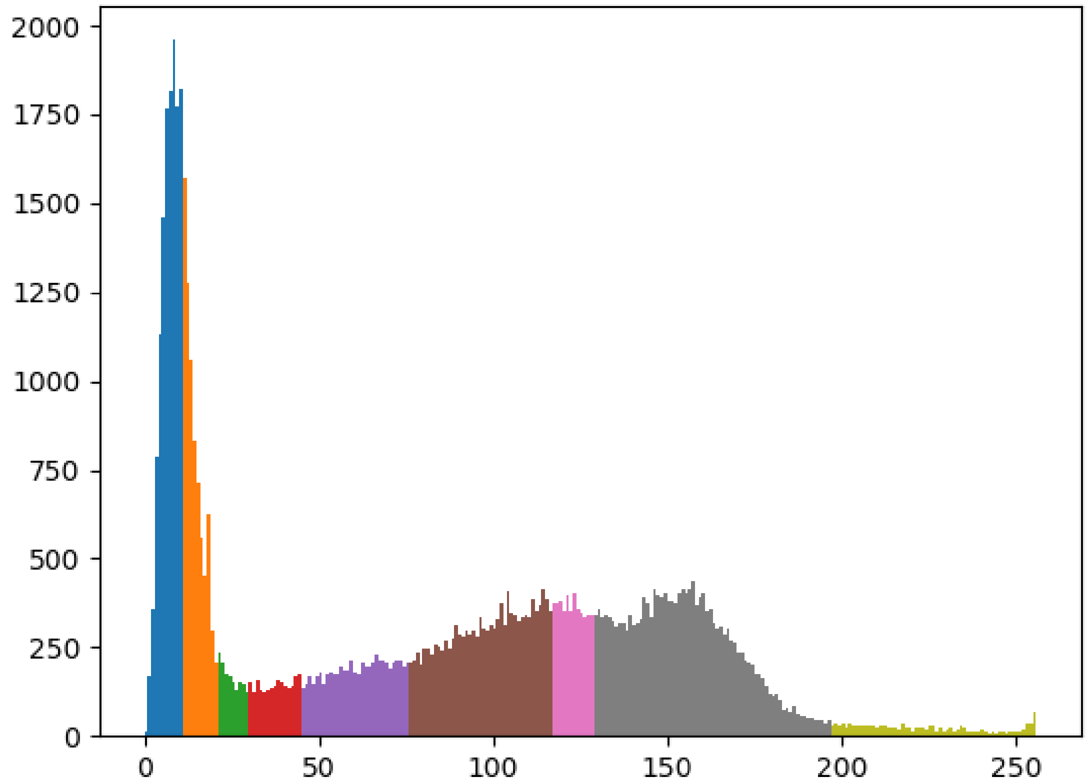
μ_j	σ_j	λ_j
7.1	2.8	0.18
12.4	4.1	0.13
22.1	6.8	0.03
38.1	9.8	0.04
62.1	14.0	0.09
99.1	16.4	0.19
120.8	9.7	0.07
153.0	17.2	0.26
217.2	22.8	0.02



Case Study – Human Head
Solution

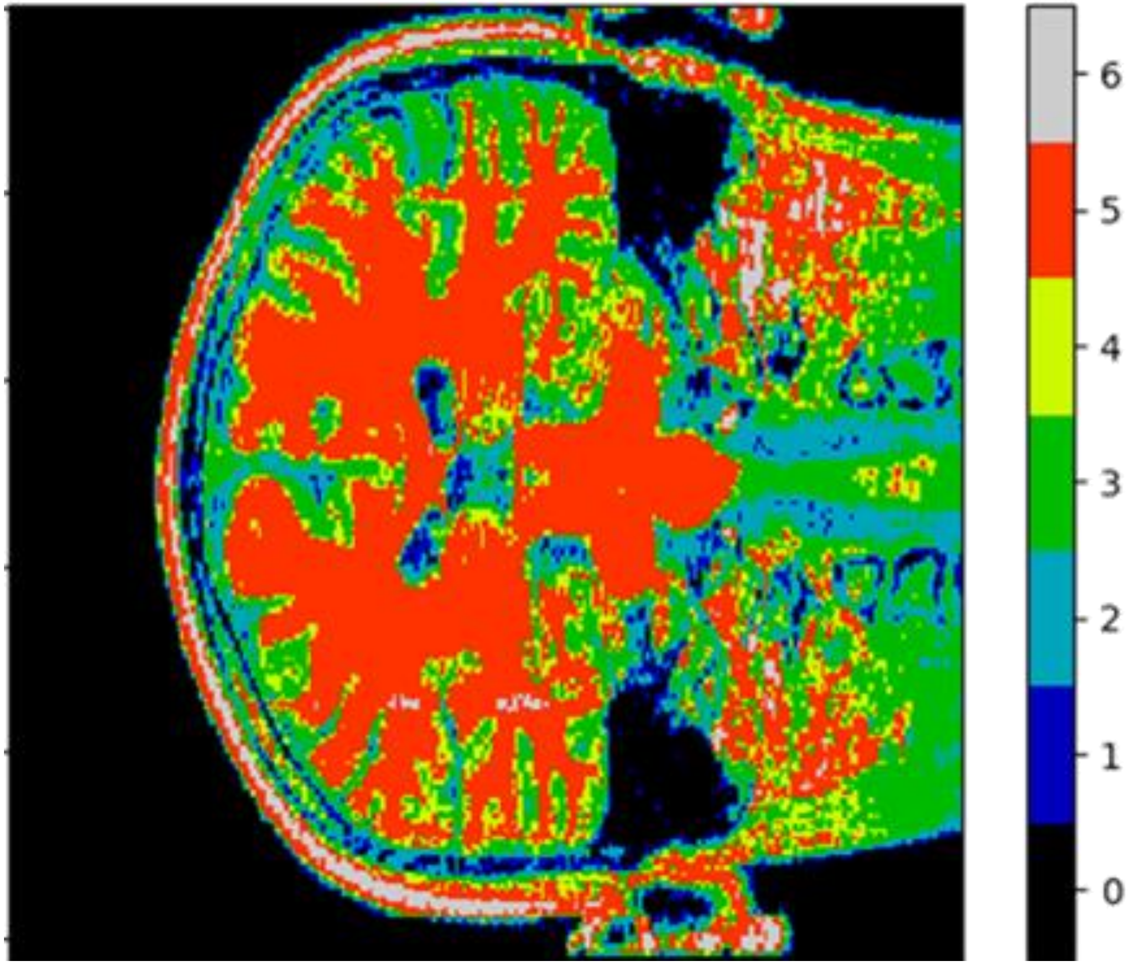
μ_j	σ_j	λ_j
7.1	2.8	0.18
12.4	4.1	0.13
22.1	6.8	0.03
38.1	9.8	0.04
62.1	14.0	0.09
99.1	16.4	0.19
120.8	9.7	0.07
153.0	17.2	0.26
217.2	22.8	0.02

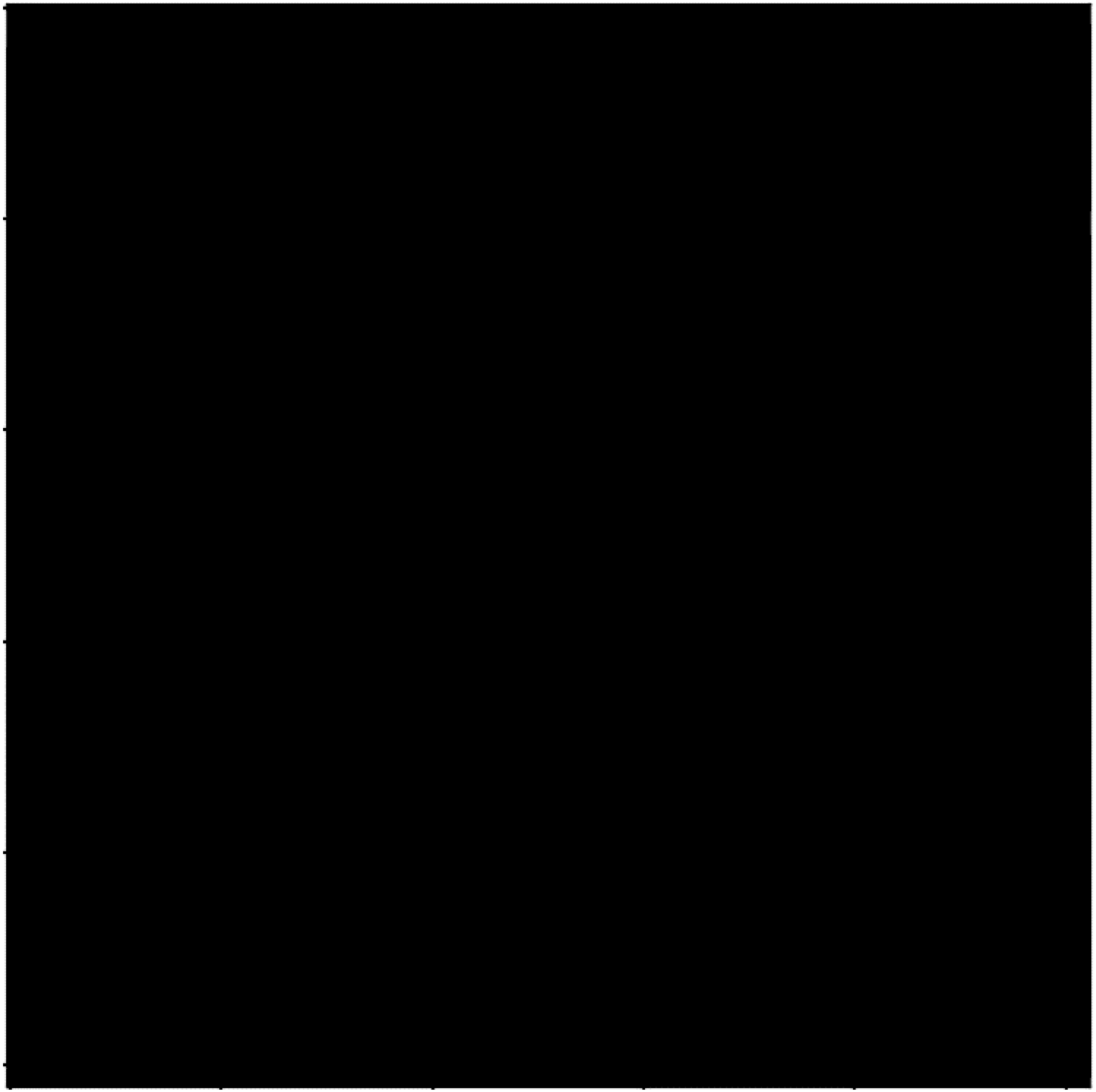
$$\operatorname*{argmax}_j q_{ij}$$



Case Study – Human Head
Solution

μ_j	σ_j	λ_j
7.1	2.8	0.18
12.4	4.1	0.13
22.1	6.8	0.03
38.1	9.8	0.04
62.1	14.0	0.09
99.1	16.4	0.19
120.8	9.7	0.07
153.0	17.2	0.26
217.2	22.8	0.02

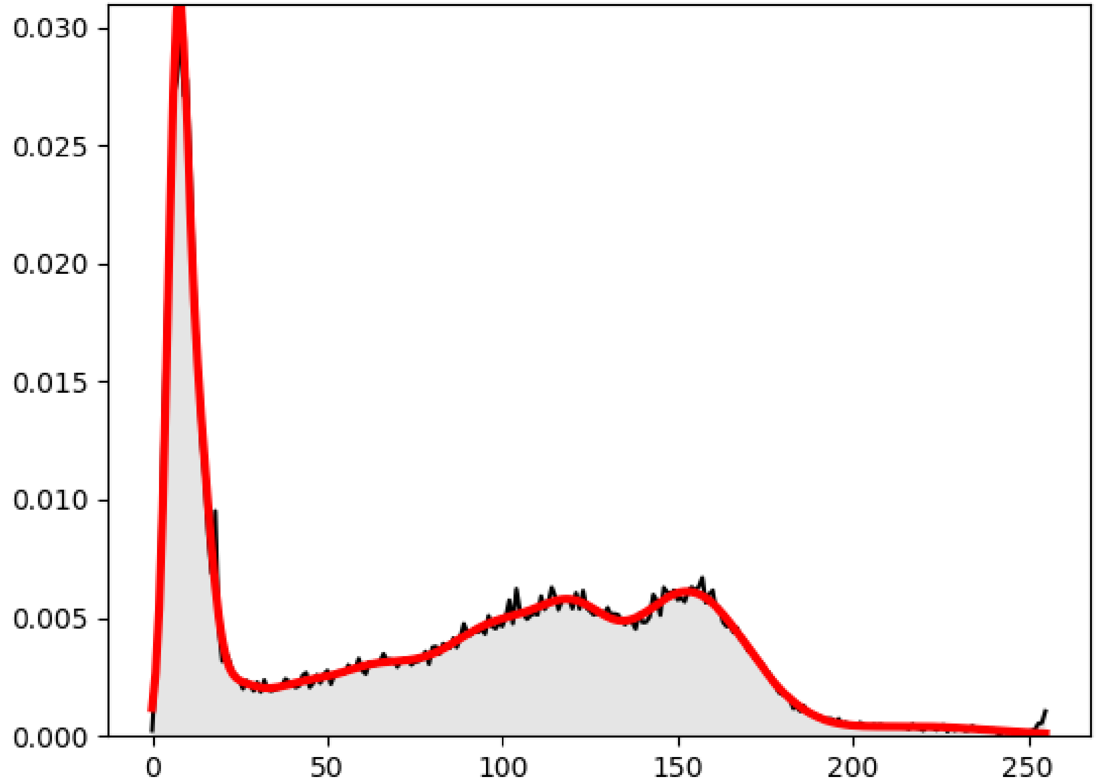




Case Study – Human Head
Solution

$$q(y_i|\mu,\sigma) = \sum_{j=1}^{k} \lambda_j N(y_i: \mu_j, \sigma_j)$$

μ_j	σ_j	λ_j
7.1	2.8	0.18
12.4	4.1	0.13
22.1	6.8	0.03
38.1	9.8	0.04
62.1	14.0	0.09
99.1	16.4	0.19
120.8	9.7	0.07
153.0	17.2	0.26
217.2	22.8	0.02



Introduction

Basic Theory

Mixture Models

Case Studies

Summary

Summary

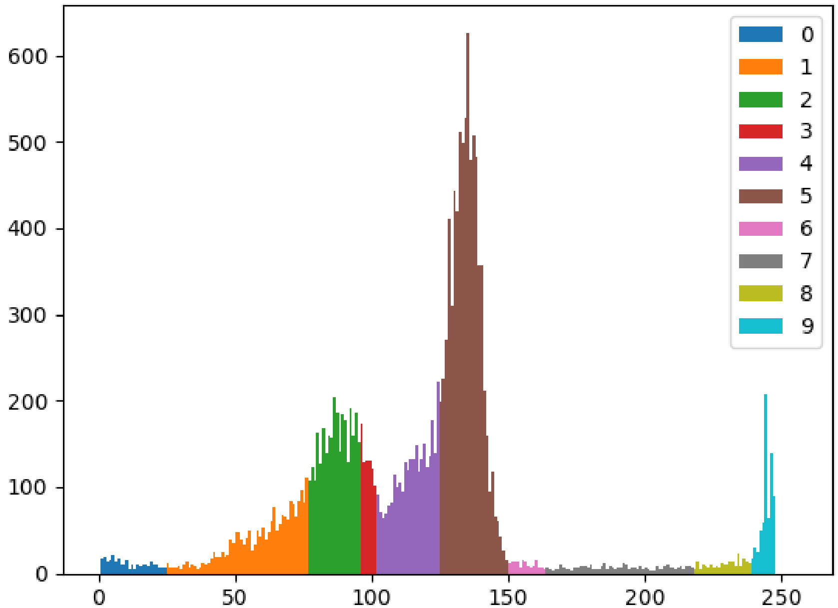
Approximating a complicated posterior by using easy to calculate distributions and minimize the KL divergence

Much faster than MCMC

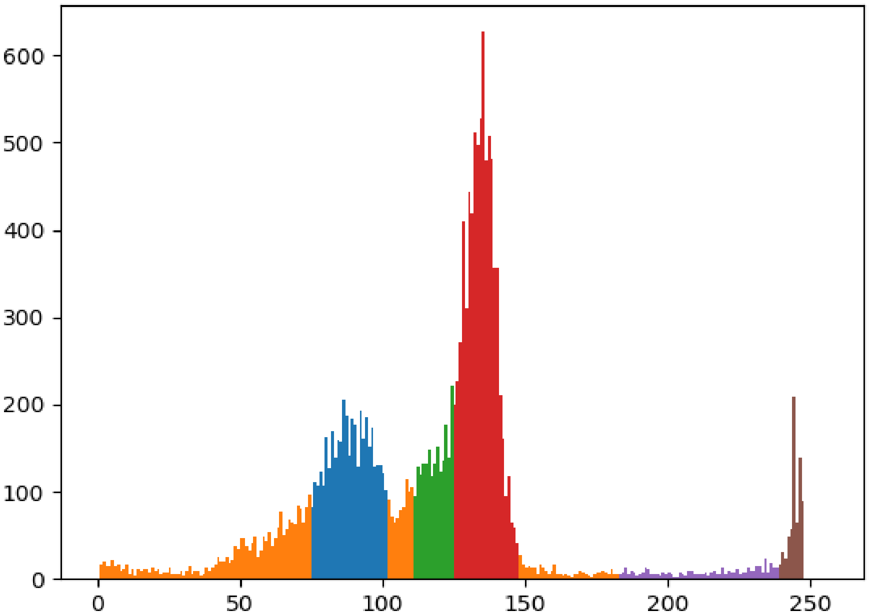
Not as accurate as MCMC

Summary

$$\underset{j}{\operatorname{argmax}}\, q_{ij}$$

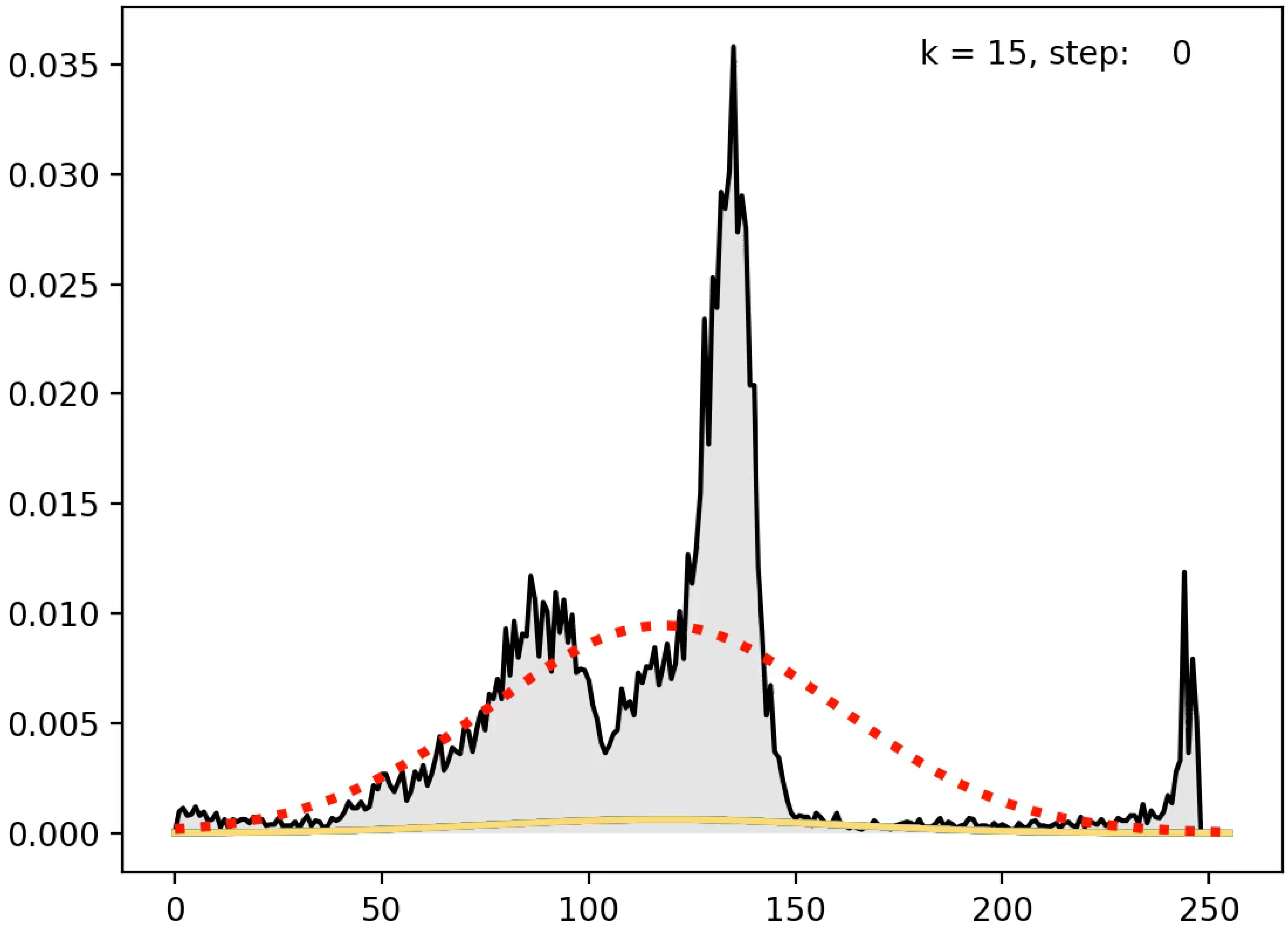


10 final components



6 final components

Thank you for your attention


k = 15, step: 0
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250