Lecture 6:

Variational Bayes, Expectation Maximization

Markus Hohle

University California, Berkeley

Bayesian Data Analysis and Machine Learning for Physical Sciences

Berkeley UNIVERSITY OF CALIFORNIA

Bayesian Data Analysis and Machine Learning for Physical Sciences

Course Map

Berkeley UNIVERSITY OF CALIFORNIA

Variational Bayes, Expectation Maximization

Outline

The Problem

K-means

Actual EM

Variational Bayes

Outline

The Problem

K-means

Actual EM

Variational Bayes

Variational Bayes, Expectation Maximization:

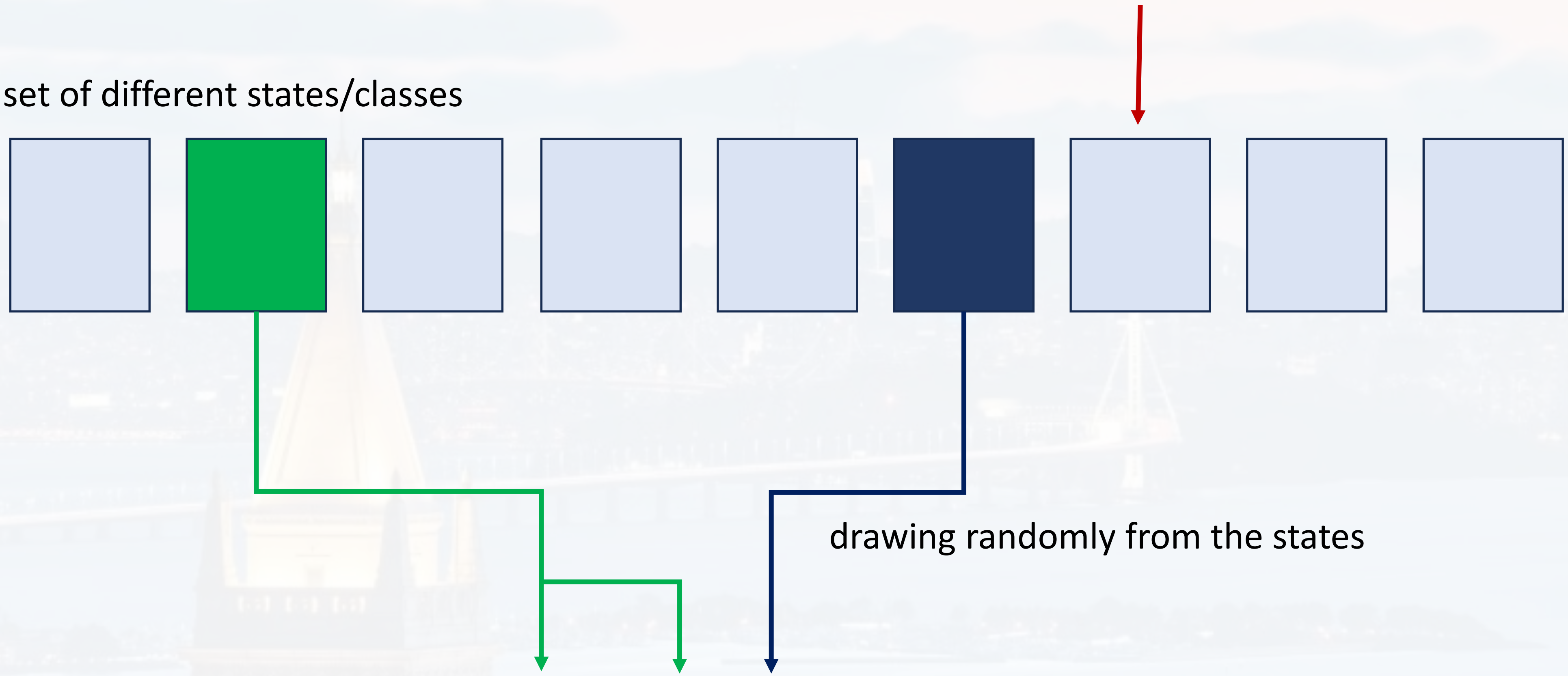
set of different states/classes

a set of N observations $\{x_1, x_2, x_3, \dots x_n, \dots x_N\}$

Variational Bayes, Expectation Maximization:

each state k draws x_n from a distribution $L_k\left(x_{n,k}\middle|\theta\right)$ with parameters θ

set of different states/classes



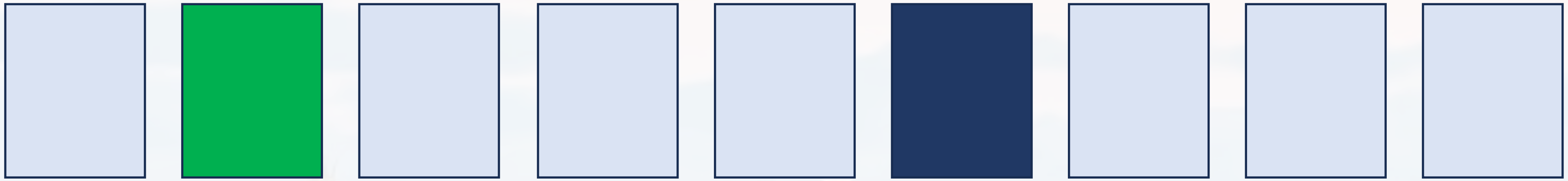
drawing randomly from the states

a set of *N* observations $\{x_1, x_2, x_3, \dots x_n, \dots x_N\}$

Variational Bayes, Expectation Maximization:

The Problem

set of different states/classes $L_k(x_{n,k}|\theta)$



a set of *N* observations $\{x_1, x_2, x_3, \ldots x_n, \ldots x_N\}$

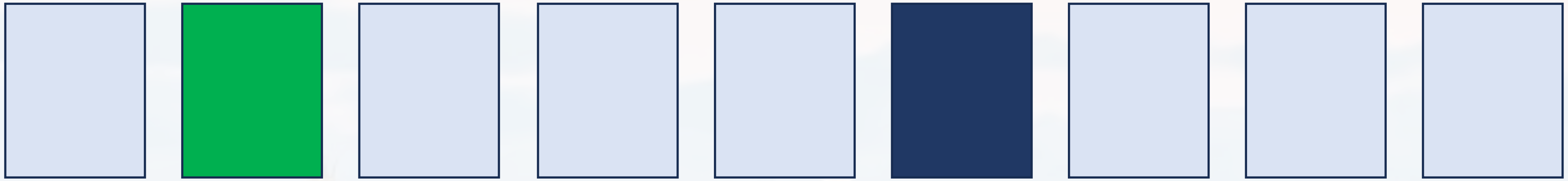
problem: - we have a model of $L_k(x_{n,k}|\theta)$, but
- we don't know θ and
- we don't know from which class/state *k* x_n has been generated

goal: - find an estimator for θ and find the class/state *k* of each x_n

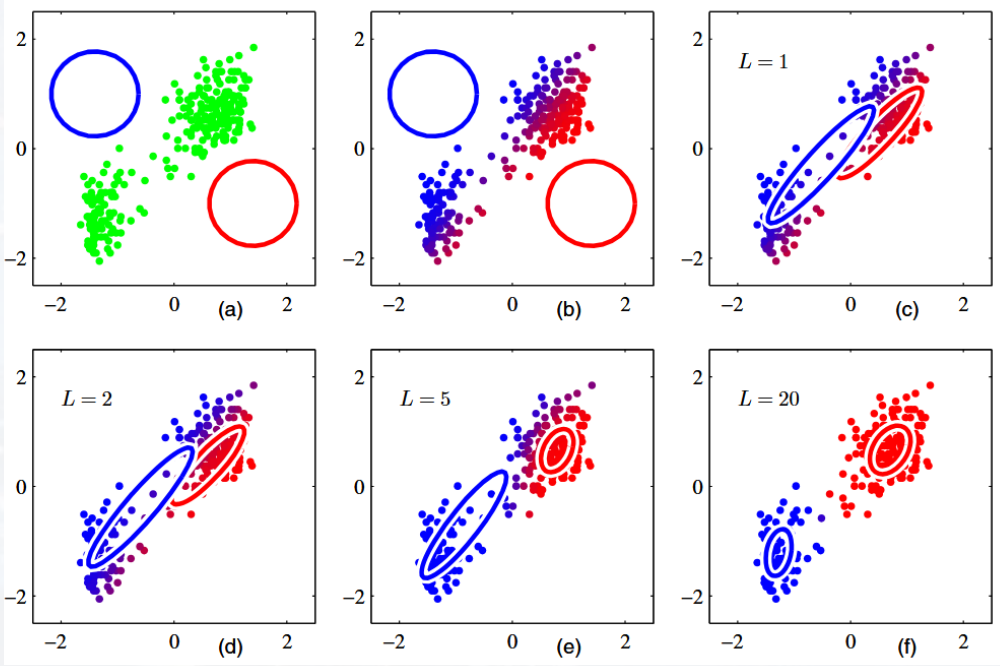
Variational Bayes, Expectation Maximization:

set of different states/classes

$L_k(x_{n,k}|\theta)$



a set of *N* observations $\{x_1, x_2, x_3, \ldots x_n, \ldots x_N\}$



Bishop: “Pattern Recognition”

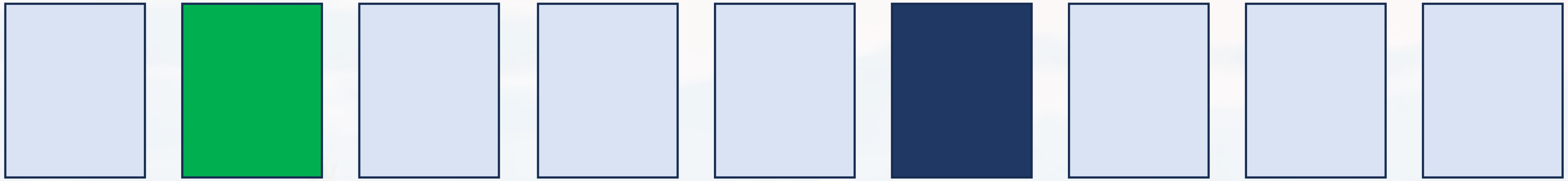
- Gaussian Mixture Models (GMM)
- K-means (clustering, image segmentation)
- HMM

→ unsupervised

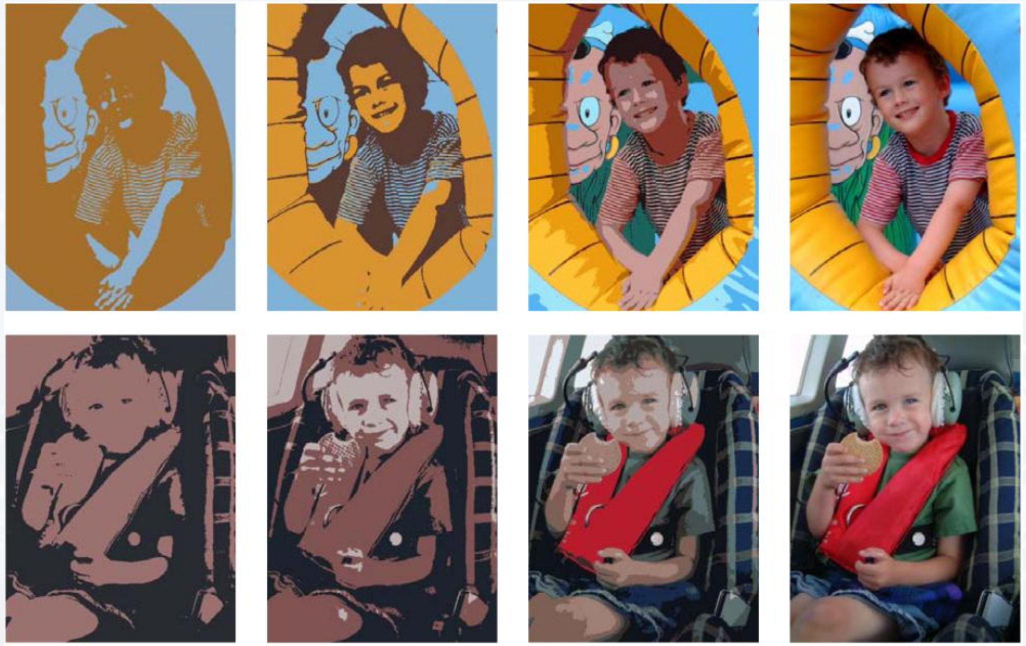
Variational Bayes, Expectation Maximization:

set of different states/classes

$L_k(x_{n,k}|\theta)$



a set of *N* observations $\{x_1, x_2, x_3, \ldots x_n, \ldots x_N\}$



Bishop: “Pattern Recognition”

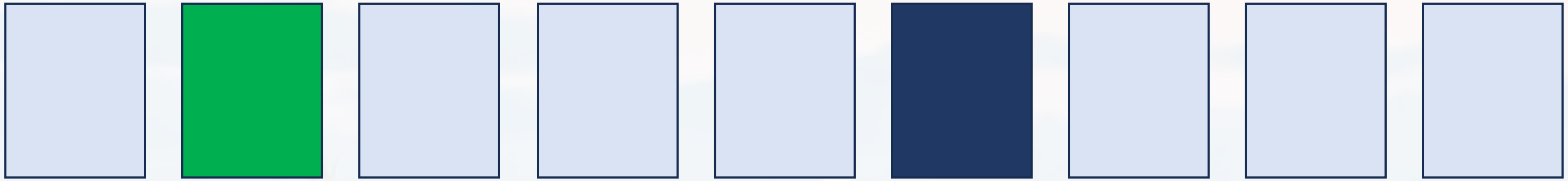
- Gaussian Mixture Models (GMM)
- K-means (clustering, image segmentation)
- HMM

→ unsupervised

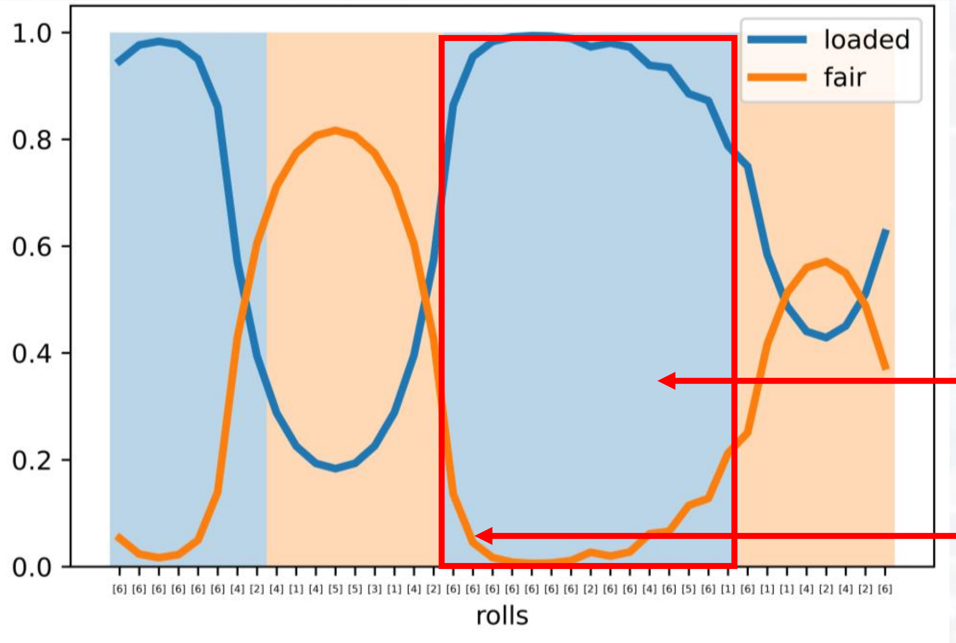
Variational Bayes, Expectation Maximization:

set of different states/classes

$L_k\left(x_{n,k} \middle| \theta\right)$



a set of *N* observations $\{x_1, x_2, x_3, \ldots x_n, \ldots x_N\}$



- Gaussian Mixture Models (GMM)
- K-means (clustering, image segmentation)
- HMM

→ unsupervised

actual (unknown, aka hidden state)

predicted state, given observation

Outline

The Problem

K-means

Actual EM

Variational Bayes

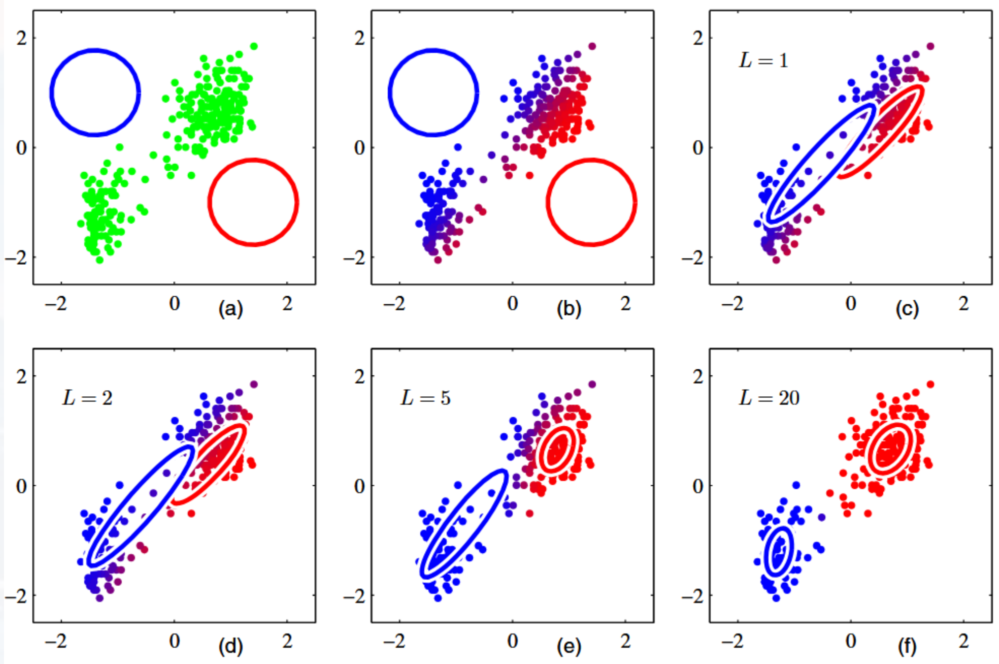
Variational Bayes, Expectation Maximization:

K	: number of cluster
N	: number of observations

indicator function $r_{n,k} \epsilon \{0, 1\}$

goal: minimizing

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \; \|x_n - \mu_k\|^2$$



μ_k: barycenter of cluster k

assigning x_n to its closed mean

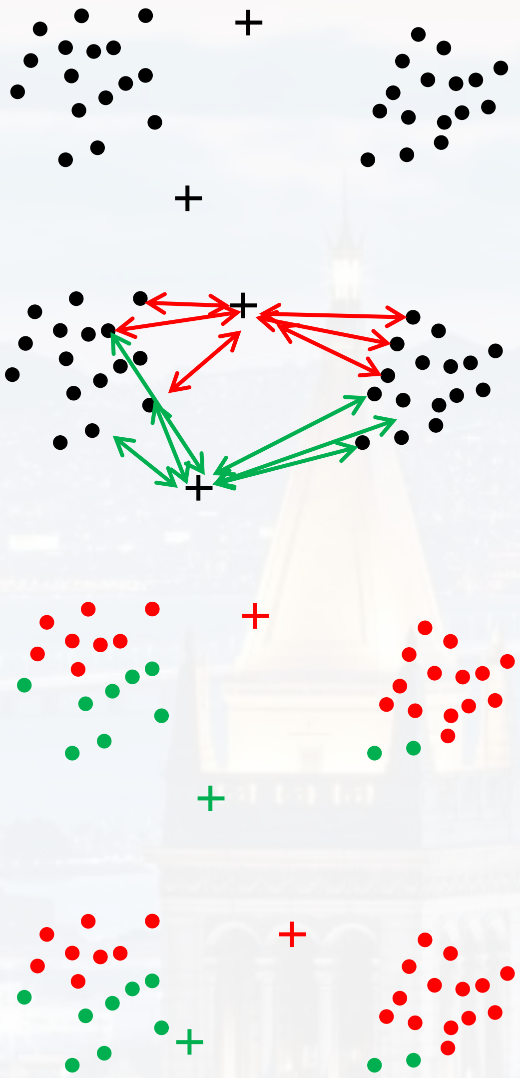
$$r_{n,k} = \begin{cases} 1, & if \; k = \; argmin_j \|x_n - \mu_j\|^2 \\ 0, & else \end{cases}$$

$$\frac{\partial J}{\partial \mu_k} = 0 \qquad \text{MLE} \longrightarrow \qquad \mu_k = \frac{\sum_{n=1}^{N} r_{n,k} x_n}{\sum_{n=1}^{N} r_{n,k}}$$

K – means is an iterative process

Variational Bayes, Expectation Maximization:

K – means is an iterative process



a) assign k means randomly

b) calculate *distance* from each point to each mean

c) assign each point to its closest mean

d) update the means accordingly

K	: number of cluster
N	: number of observations
μ_k	: barycenter of the cluster

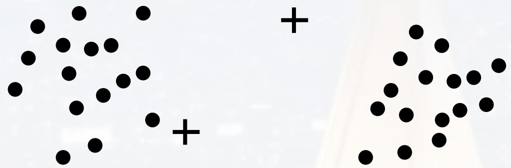
$$r_{n,k} = \begin{cases} 1, & \text{if } k = \operatorname{argmin}_j \|x_n - \mu_j\|^2 \\ 0, & \text{else} \end{cases}$$

$$\boxed{J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \, \|x_n - \mu_k\|^2}$$

Variational Bayes, Expectation Maximization:

K – means is an iterative process

d) update the means accordingly



e) go back to b)

K : number of cluster

N : number of observations

μ_k : barycenter of the cluster

$$r_{n,k} = \begin{cases} 1, & \text{if } k = \operatorname{argmin}_j \|x_n - \mu_j\|^2 \\ 0, & \text{else} \end{cases}$$

$$\boxed{J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \, \|x_n - \mu_k\|^2}$$

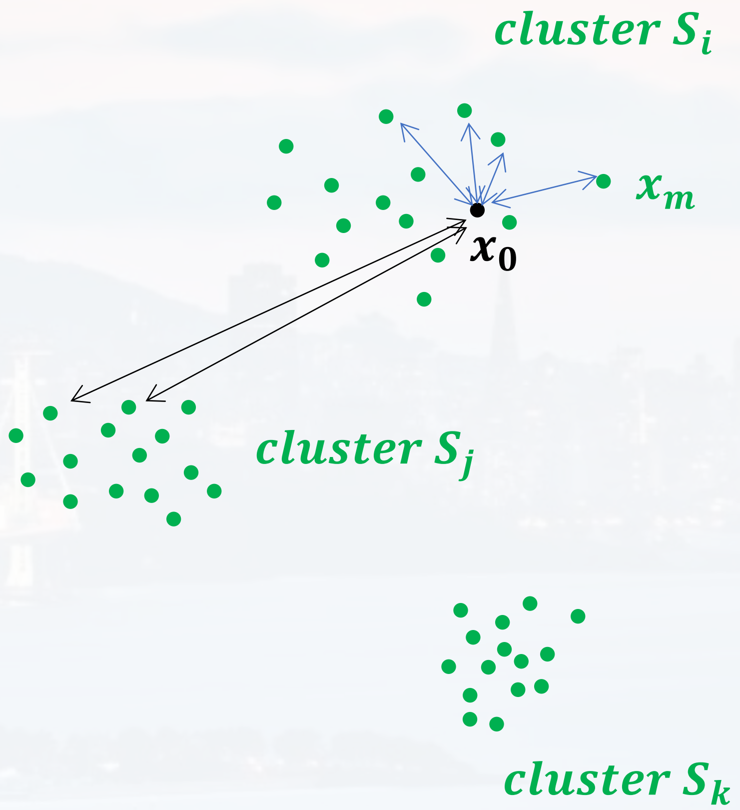
Variational Bayes, Expectation Maximization:

problem: ***K = number of cluster,*** is a hyperparameter. How do I know the correct value for ***K***?

→ silhouette Ψ

- distance $\boldsymbol{d_1}$ of a data point x_0 to ***its assigned cluster*** $\boldsymbol{S_i}$
  vs distance $\boldsymbol{d_2}$ to ***closest cluster (here*** $\boldsymbol{S_j}$***)***

$$\boxed{\Psi(x_0) = \begin{cases} 0 & if\ d_1 = 0 \\ \dfrac{d_2 - \boldsymbol{d_1}}{max[\boldsymbol{d_1}; d_2]} & \end{cases}}$$



- average over all points → ψ_{tot}

if

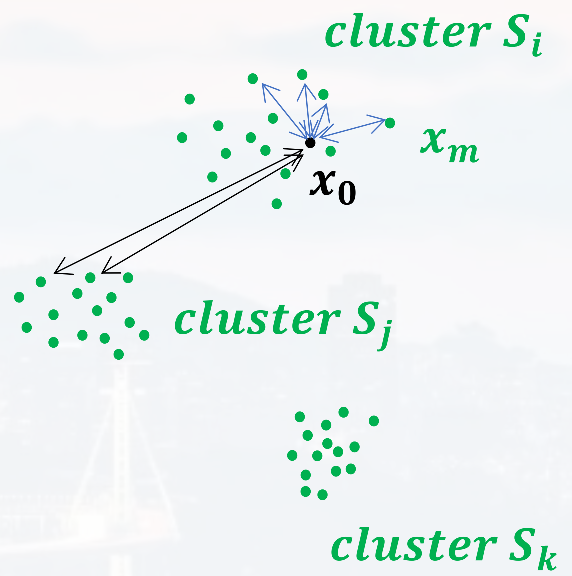
$\psi_{tot} = 0.75\ \ldots 1.00$ → well clustered
$\psi_{tot} = 0.50\ \ldots 0.75$ → medium clustered
$\psi_{tot} = 0.25\ \ldots 0.50$ → poorly clustered
$\psi_{tot} < 0.25$ → data has no structure

Variational Bayes, Expectation Maximization:

problem: ***K* = number of cluster,** is a hyperparameter. How do I know the correct value for ***K***?

→ silhouette Ψ

- distance $\boldsymbol{d_1}$ of a data point x_0 to ***its assigned cluster*** $\boldsymbol{S_i}$
  vs distance $\boldsymbol{d_2}$ to ***closest cluster (here*** $\boldsymbol{S_j}$***)***



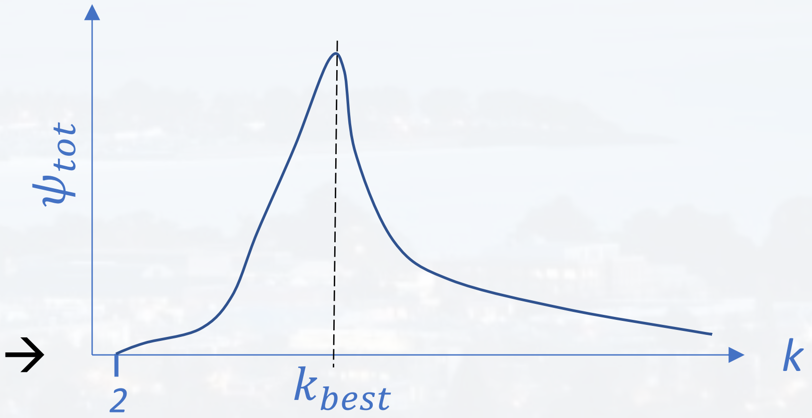
$$\boxed{\Psi(x_0) = \begin{cases} 0 & if\ d_1 = 0 \\ \dfrac{d_2 - \boldsymbol{d_1}}{max[\boldsymbol{d_1}; d_2]} & \end{cases}}$$

- average over all points → ψ_{tot}

if	$\psi_{tot} = 0.75\ \ldots 1.00$	→ well clustered
	$\psi_{tot} = 0.50\ \ldots 0.75$	→ medium clustered
	$\psi_{tot} = 0.25\ \ldots 0.50$	→ poorly clustered
	$\psi_{tot} < 0.25$	→ data has no structure

see `Walk_Through_Kmeans.ipynb`

ideal world →



Variational Bayes, Expectation Maximization:

$\boldsymbol{K}$	: number of cluster
$\boldsymbol{N}$	: number of observations
$\boldsymbol{\mu_k}$	: barycenter of the cluster

$$\boldsymbol{r_{n,k}} = \begin{cases} \boldsymbol{1}, & \boldsymbol{if\ k = argmin_j \|x_n - \mu_j\|^2} \\ \boldsymbol{0}, & \boldsymbol{else} \end{cases}$$

$$\boxed{J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \, \|x_n - \mu_k\|^2}$$

the actual problem:

observation x_n has been drawn from any of the cluster C_k

$$P(x_n) = \sum_{k=1}^{K} P(x_n|C_k)\, P(C_k) \qquad \sum_{k=1}^{K} P(C_k) = 1$$

$P(x_n|C_k)$ likelihood function

$P(C_k)$ mixing coefficient

general:

$$P(x_n) = \sum_{z} P(x_n|z)\, P(z)$$

z: latent variable **(i. e. not observable, but can be inferred from $\{x_n\}$)**

Outline

The Problem

K-means

Actual EM

Variational Bayes

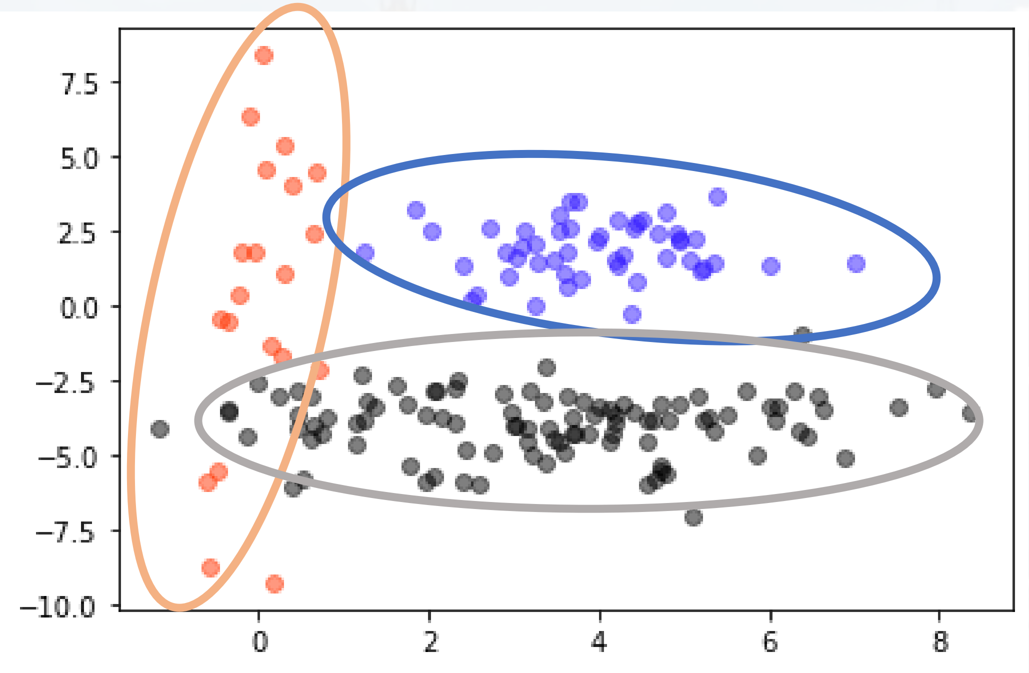
Variational Bayes, Expectation Maximization:

example: **G**aussian **M**ixture **M**odels

K	**: number of cluster**
π_k	**: mixing coefficient**

f features

$$\mathcal{N}_k(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{f/2}\det(\Sigma_k)^{1/2}}\, exp\left[-\frac{1}{2}(x - \mu_k)^T {\Sigma_k}^{-1}(x - \mu_k)\right]$$



two features, $K = 3$ components

$$P(x) = \sum_z P(x|z)\, P(z)$$

$$= \sum_{k=1}^{K} \mathcal{N}_k(x|\mu_k, \Sigma_k)\, \pi_k$$

Variational Bayes, Expectation Maximization:

N	**: number of observations**
K	**: number of cluster**
$\boldsymbol{\pi_k}$	**: mixing coefficient**

example: **G**aussian **M**ixture **M**odels

$$P(x) = \sum_z P(x|z)\, P(z) = \sum_{k=1}^{K} \mathcal{N}_k(x|\mu_k, \Sigma_k)\, \pi_k$$

indicator variable $z_k \, \epsilon \, \{0, 1\}$

goal: $$P(z_k = 1|x) = \frac{P(z_k = 1)\, P(x|z_k = 1)}{P(x)} = \frac{\pi_k\, \mathcal{N}_k(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \mathcal{N}_j\left(x|\mu_j, \Sigma_j\right) \pi_j}$$

via maximizing likelihood by finding best θ

$$L = ln[P(x|\pi, \mu, \Sigma)] = \sum_{n=1}^{N} ln\left\{\sum_{k=1}^{K} \pi_k\, \mathcal{N}_k(x_n|\mu_k, \Sigma_k)\right\}$$

model parameter $\theta = \{\pi, \mu, \Sigma\}$

Variational Bayes, Expectation Maximization:

N	**: number of observations**
K	**: number of cluster**
π_k	**: mixing coefficient**

example: **G**aussian **M**ixture **M**odels

1) initialize $\theta = \{\pi, \mu, \Sigma\}$ and $P(z_k = 1|x_n)$

2) **E**xpectation step *t*:

$$P(z_k = 1|x_n) = \frac{\pi_k\, \mathcal{N}_k(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \mathcal{N}_j\left(x_n|\mu_j, \Sigma_j\right) \pi_j}$$

i.e. evaluate $P(z|x, \theta)$

3) **M**aximization step *t* (for example MLE):

$$\mu_k^t = \frac{1}{N_k} \sum_{n=1}^{N} P(z_k = 1|x_n)\, x_n$$

$$\Sigma_k^t = \frac{1}{N_k} \sum_{n=1}^{N} P(z_k = 1|x_n)\, (x_n - \mu_k^t)\, (x_n - \mu_k^t)^T$$

$$N_k^t = \sum_{n=1}^{N} P(z_k = 1|x_n)$$

$$\pi_k^t = \frac{N_k}{N}$$

$$\theta^t = \underset{\theta}{argmax} \{L(\theta^{t-1})\}$$

Variational Bayes, Expectation Maximization:

example: **G**aussian **M**ixture **M**odels

N	**: number of observations**
K	**: number of cluster**
$\boldsymbol{\pi_k}$	**: mixing coefficient**

1) initialize $\theta = \{\pi, \mu, \Sigma\}$ and $P(z_k = 1|x_n)$

2) **E**xpectation step *t*:

$$P(z_k = 1|x_n) = \frac{\pi_k\, \mathcal{N}_k(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \mathcal{N}_j\left(x_n|\mu_j, \Sigma_j\right) \pi_j}$$

i.e. evaluate $P(z|x, \theta)$

3) **M**aximization step *t* (for example MLE):

$$\mu_k^t = \frac{1}{N_k} \sum_{n=1}^{N} P(z_k = 1|x_n)\, x_n \qquad \Sigma_k^t = \frac{1}{N_k} \sum_{n=1}^{N} P(z_k = 1|x_n)\, (x_n - \mu_k^t)\, (x_n - \mu_k^t)^T \qquad N_k = \sum_{n=1}^{N} P(z_k = 1|x_n) \qquad \pi_k^t = \frac{N_k}{N}$$

$$\theta^t = \underset{\theta}{argmax} \left\{ \sum_z P(z|x, \theta^{t-1})\ ln[P(x, z|\theta^{t-1})] \right\}$$

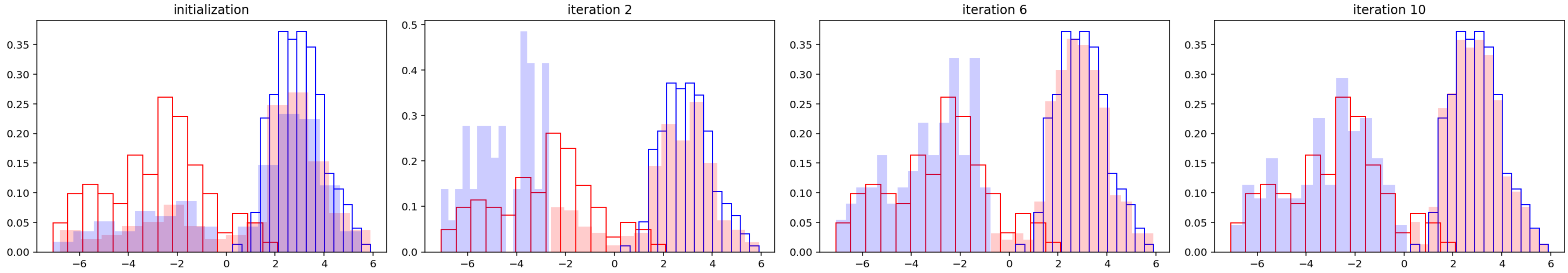
4) evaluate log likelihood

$$ln[P(x|\pi, \mu, \Sigma)] = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k\, \mathcal{N}_k(x_n|\mu_k, \Sigma_k) \right\}$$

Variational Bayes, Expectation Maximization:

note:

- one can show that EM indeed converges and maximizes the likelihood function (Bishop, Sec 9.4)
- no guarantee to find a **global** minimum
- applications: HMM, unsupervised clustering, image segmentation (before CNNs)
- iterative process: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \Delta\theta$ → connection to gradient descent (see later)
- see also `EM__Example.py`



blue and red swapped
→ unsupervised learning

Variational Bayes, Expectation Maximization

Outline

The Problem

K-means

Actual EM

Variational Bayes

Variational Bayes, Expectation Maximization:

EM: - likelihood function given

- parameter via MLE (point estimate)

problem: - integrals can be complicated, calculation takes too long etc ("intractable")

- need **pdf of desired parameter** (see BPE) without ad-hoc constrain

only two assumptions: - 1) maximum entropy

- 2) pdf factorizes wrt its parameters (→ mean field approximation)

D : data set
$Z = \{Z_1, \dots Z_n\}$: set of (latent) parameter

goal: - find an approximation $Q(Z)$ for the posterior $P(Z|D)$ via **max ent**
- the more data, $Q(Z) \rightarrow P(Z|D)$ ("learning")

idea: - $\mathrm{Q}(Z) = \prod_{i=1}^{n} q_i(Z_i|D)$ mean field approximation

Variational Bayes, Expectation Maximization:

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

D	**: data set**
Z	**: set of n (latent) parameter**

KL divergence: tells us how much our information is "off" if we work with the approximation $Q(Z)$

$$KL(P||Q) = -\int P(x)\log\left[\frac{Q(x)}{P(x)}\right]dx \qquad \text{(module 1)}$$

goal: find the $Q(x)$ that **minimizes $\boldsymbol{KL(P||Q)}$** → variational calculus (Euler-Lagrange)

same principle: **maximizing evidence** $P(D)$ → variational calculus (Euler−Lagrange)

we run the 2nd idea → find an equation we know from stat TD
see also reading Paisley, Blei & Jordan

$$ln[P(D)] = ln\left[\int P(D,Z)\,dZ\right] = ln\left[\int P(D,Z)\frac{Q(Z)}{Q(Z)}dZ\right] \geq \int Q(Z)\ln\left[\frac{P(D,Z)}{Q(Z)}\right]dZ$$

"Jensen inequality" → evidence lower bound ("ELBO")

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$ln[P(D)] = ln\left[\int P(D,Z)dZ\right] = ln\left[\int P(D,Z)\frac{Q(Z)}{Q(Z)}dZ\right] \geq \int Q(Z)\ln\left[\frac{P(D,Z)}{Q(Z)}\right]dZ$$

"Jensen inequality"

$$ln[P(D)] \geq \int Q(Z)\ ln[P(D,Z)]\ dZ\ - \boxed{\int Q(Z)\ ln[Q(Z)]\ dZ}$$

entropy of $Q(Z)$

$$ln[P(D)] \geq \int \boxed{\prod_{i=1}^{n} q_i(Z_i|D)}\ ln[P(D,Z)]\ dZ\ - \underbrace{\int \boxed{\prod_{i=1}^{n} q_i(Z_i|D)}\ ln\left[\boxed{\prod_{i=1}^{n} q_i(Z_i|D)}\right]\ dZ}_{S_q}$$

assumption 2

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$S_q = \int \prod_{i=1}^{n} q_i(Z_i|D)\, ln\left[\prod_{i=1}^{n} q_i(Z_i|D)\right] dZ$$

one n-dimensional problem (hard)

$$= \int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_1(Z_1|D)]\, dZ_1 dZ_2 \ldots dZ_n +$$

$$\int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_2(Z_2|D)]\, dZ_1 dZ_2 \ldots dZ_n +$$

$$\ldots +$$

$$\int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_n(Z_n|D)]\, dZ_1 dZ_2 \ldots dZ_n$$

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$S_q = \int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_1(Z_1|D)]\, dZ_1 dZ_2 \ldots dZ_n +$$

$$\int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_2(Z_2|D)]\, dZ_1 dZ_2 \ldots dZ_n + \cdots +$$

$$\int \{q_1(Z_1|D)\, q_2(Z_2|D) \ldots\, q_N(Z_N|D)\}\, ln[q_n(Z_n|D)]\, dZ_1 dZ_2 \ldots dZ_n$$

$$= \int q_1(Z_1|D)\, ln[q_1(Z_1|D)]\, dZ_1 \int q_2(Z_2|D)\, dZ_2 \ldots \int q_n(Z_n|D)\, dZ_n +$$

$$\int q_2(Z_2|D)\, ln[q_2(Z_2|D)]\, dZ_2 \int q_1(Z_1|D)\, dZ_1 \ldots \int q_n(Z_n|D)\, dZ_n + \ldots +$$

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$S_q = \int q_1(Z_1|D)\, ln[q_1(Z_1|D)]\, dZ_1 \int q_2(Z_2|D)\, dZ_2 \ldots \int q_n(Z_n|D)\, dZ_n +$$

$$\int q_2(Z_2|D)\, ln[q_2(Z_2|D)]\, dZ_2 \int q_1(Z_1|D)\, dZ_1 \ldots \int q_n(Z_n|D)\, dZ_n + \cdots +$$

$$= \sum_{i=1}^{n} \int q_i(Z_i|D)\, ln[q_i(Z_i|D)]\, dZ_i \left\{ \prod_{j\neq i}^{n-1} \int q_j(Z_j|D)\, dZ_j \right\}$$

$$= \sum_{i=1}^{n} \langle ln[q_i(Z_i|D)] \rangle \left\{ \prod_{j\neq i}^{n-1} \int q_j(Z_j|D)\, dZ_j \right\}$$

= 1 (the q_j are all a pdf of Z_j)

$$S_q = \sum_{i=1}^{n} \langle ln[q_i(Z_i|D)] \rangle = \sum_{i=1}^{n} \int q_i(Z_i|D)\, ln[q_i(Z_i|D)]\, dZ_i$$

n one-dimensional problems (not so hard)

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$ln[P(D)] \geq \int \prod_{i=1}^{n} q_i(Z_i|D)\ ln[P(D|Z)]\, dZ \ - \underbrace{\int \prod_{i=1}^{n} q_i(Z_i|D)\, ln\left[\prod_{i=1}^{n} q_i(Z_i|D)\right] dZ}_{S_q}$$

$$ln[P(D)] \geq \int \prod_{i=1}^{n} q_i(Z_i|D)\ ln[P(D,Z)]\, dZ \ - \sum_{i=1}^{n} \int q_i(Z_i|D)\ ln[q_i(Z_i|D)]\, dZ_i$$

$$E(D,Z) := ln[P(D,Z)]$$

$$ln[P(D)] \geq \int \prod_{i=1}^{n} q_i(Z_i|D)\, E(D,Z)\, dZ \ - \sum_{i=1}^{n} \int q_i(Z_i|D)\ ln[q_i(Z_i|D)]\, dZ_i$$

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter

find an approximation $Q(Z)$ for the posterior $P(Z|D)$

maximizing evidence $P(D)$

$$ln[P(D)] \geq \int \prod_{i=1}^{n} q_i(Z_i|D)\, E(D,Z)\, dZ \; - \sum_{i=1}^{n} \int q_i(Z_i|D)\, ln[q_i(Z_i|D)]\, dZ_i$$

has the structure of $F = U - TS$, but with $-\mathrm{U} + TS$ term (aka negative variational free energy)

solution:

$$q_i(Z_i|D) = \frac{1}{\mathcal{Z}} \exp\left(\langle E(Z_i, \{Z_{j\neq i}\}, D)\rangle_{\{j\neq i\}}\right)$$

$$\mathcal{Z} = \int \exp\left(\langle E(Z_i, \{Z_{j\neq i}\}, D)\rangle_{\{j\neq i\}}\right) dZ_{\{j\neq i\}}$$

Variational Bayes, Expectation Maximization:

D	: data set
Z	: set of n (latent) parameter
σ^2	: variance
μ	: mean
$\frac{1}{\sigma^2} = \tau$	: precision

example: we can measure $\boldsymbol{\mu}$ and $\frac{1}{\sigma^2} = \boldsymbol{\tau}$ from D

goal: find $q_\mu(\mu|D)$ and $q_\tau(\tau|D)$

$$q_i(Z_i|D) = \frac{1}{\mathcal{Z}} \exp\left(\langle E(Z_i, \{Z_{j\neq i}\}, D)\rangle_{\{j\neq i\}}\right)$$

$$ln\left[q_\mu(\mu|D)\right] = \langle E(\mu, \tau, D)\rangle_\tau - \ln(\mathcal{Z})$$

$$= \langle \ln[P(D|\mu,\tau)P(\mu|\tau)P(\tau)]\rangle_\tau - \ln(\mathcal{Z})$$

$$= \langle \ln[P(D|\mu,\tau)]\rangle_\tau + \langle \ln[P(\mu|\tau)]\rangle_\tau + \langle \ln[P(\tau)]\rangle_\tau - \ln(\mathcal{Z})$$

Variational Bayes, Expectation Maximization:

D	: data set of size K
Z	: set of n (latent) parameter
σ^2	: variance
μ	: mean
$\frac{1}{\sigma^2} = \tau$	: precision

example: we can measure $\boldsymbol{\mu}$ and $\frac{1}{\sigma^2} = \boldsymbol{\tau}$ from D

goal: find $q_\mu(\mu|D)$ and $q_\tau(\tau|D)$

$$ln\left[q_\mu(\mu|D)\right] = \langle \ln[P(D|\mu,\tau)]\rangle_\tau + \langle \ln[P(\mu|\tau)]\rangle_\tau + \langle \ln[P(\tau)]\rangle_\tau - \ln(\mathcal{Z})$$

if no constrain: gaussian

$$P(D|\mu,\tau) = \prod_{k=1}^{K} \mathcal{N}(x_k|\mu,\tau)$$

support is $[0,+\infty)$ → max ent

$$P(\tau) = \Gamma(\tau|a,b) = \frac{b^a\, \tau^{a-1}\, e^{-b\tau}}{\int_0^\infty t^{a-1}\, e^{-t}\, dt}$$

with yet unknown parameter a and b

if μ has support (-∞, + ∞) it is drawn from a gaussian of yet unknown μ_0 and precision τ_0 (= small pos number, max ent)

$$P(\mu|\tau) = \mathcal{N}(\mu|\mu_0,\tau_0)$$

Variational Bayes, Expectation Maximization:

D	: data set of size K
Z	: set of n (latent) parameter
σ^2	: variance
μ	: mean
$\frac{1}{\sigma^2} = \tau$	: precision

example: we can measure $\boldsymbol{\mu}$ and $\frac{1}{\sigma^2} = \boldsymbol{\tau}$ from D

goal: find $q_\mu(\mu|D)$ and $q_\tau(\tau|D)$

$$ln\left[q_\mu(\mu|D)\right] = \langle \ln[P(D|\mu,\tau)] \rangle_\tau + \langle \ln[P(\mu|\tau)] \rangle_\tau + \langle \ln[P(\tau)] \rangle_\tau - \ln(\mathcal{Z})$$

$$P(D|\mu,\tau) = \prod_{k=1}^{K} \mathcal{N}(x_k|\mu,\tau) \qquad P(\mu|\tau) = \mathcal{N}(\mu|\mu_0,\tau_0) \qquad P(\tau) = \Gamma(\tau|a,b) = \frac{b^a\, \tau^{a-1}\, e^{-b\tau}}{\int_0^\infty t^{a-1}\, e^{-t}\, dt}$$

after some (lengthy algebra):

$$ln\left[q_\mu(\mu|D)\right] = -\frac{\langle \tau \rangle_\tau}{2} \left\{ \sum_k (x_k - \mu)^2 + \tau_0 (\mu - \mu_0)^2 \right\} + constant\ terms$$

$$q_\mu(\mu|D) \sim \mathcal{N}(\mu|\mu_K, \lambda_K^{-1})$$

where

$$\mu_K = \frac{\tau_0\, \mu_0 + K\, \bar{x}}{\tau_0 + K}$$

$$\lambda_K = (\tau_0 + K)\, \langle \tau \rangle_\tau$$

$$\bar{x} = \frac{1}{K} \sum_{k=1}^{K} x_k$$

Variational Bayes, Expectation Maximization:

D	: data set of size K
Z	: set of n (latent) parameter
σ^2	: variance
μ	: mean
$\frac{1}{\sigma^2} = \tau$	: precision

example: we can measure $\boldsymbol{\mu}$ and $\frac{1}{\sigma^2} = \boldsymbol{\tau}$ from D

goal: find $q_\mu(\mu|D)$ and $q_\tau(\tau|D)$

$$ln\left[q_\mu(\mu|D)\right] = \langle \ln[P(D|\mu,\tau)]\rangle_\tau + \langle \ln[P(\mu|\tau)]\rangle_\tau + \langle \ln[P(\tau)]\rangle_\tau - \ln(\mathcal{Z})$$

$$P(D|\mu,\tau) = \prod_{k=1}^{K} \mathcal{N}(x_k|\mu,\tau) \qquad P(\mu|\tau) = \mathcal{N}(\mu|\mu_0,\tau_0) \qquad P(\tau) = \Gamma(\tau|a,b) = \frac{b^a\,\tau^{a-1}\,e^{-b\tau}}{\int_0^\infty t^{a-1}\,e^{-t}\,dt}$$

same for $q_\tau(\tau|D)$ (lengthy algebra)

$$q_\tau(\tau|D) \sim \Gamma(\tau|a_K, b_K)$$

where
$$a_K = a + \frac{K+1}{2}$$
$$b_K = b + \frac{1}{2}\left\langle \textstyle\sum_k (x_k-\mu)^2 + \tau_0(\mu-\mu_0)^2\right\rangle_\mu$$

Variational Bayes, Expectation Maximization:

$\boldsymbol{D}$	: data set of size K
$\boldsymbol{Z}$	: set of n (latent) parameter
$\boldsymbol{\sigma^2}$	: variance
$\boldsymbol{\mu}$	: mean
$\frac{1}{\sigma^2} = \tau$	: precision

$$q_\mu(\mu|D) \sim \mathcal{N}(\mu|\mu_K, \lambda_K^{-1})$$

where

$$\mu_K = \frac{\tau_0\, \mu_0 + K\, \bar{x}}{\tau_0 + K}$$
$$\lambda_K = (\tau_0 + K)\, \boxed{\langle\tau\rangle_\tau}$$
$$\bar{x} = \frac{1}{K}\sum_{k=1}^{K} x_k$$

$$q_\tau(\tau|D) \sim \Gamma(\tau|a_K, b_K)$$

where

$$a_K = a + \frac{K+1}{2}$$
$$b_K = b + \frac{1}{2}\langle \textstyle\sum_k (x_k - \mu)^2 + \tau_0(\mu - \mu_0)^2 \rangle_\mu$$

$$\langle\tau\rangle_\tau = \int \tau\, q_\tau(\tau|D)\, d\tau = \frac{a_K}{b_K}$$

We start with setting τ_0, μ_0, a and b to **small positive values** (largest ignorance, broad peaks) and use the circular dependencies (like actual EM)

- from $\langle\tau\rangle_\tau$ (here integral over the gamma distribution) we can calculate $q_\mu(\mu|D)$
- from that we can calculate b_K and $\langle\mu\rangle_\mu = \mu_K$ and $\langle\mu^2\rangle_\mu = \frac{1}{\lambda_K} + \mu_K^2$ etc

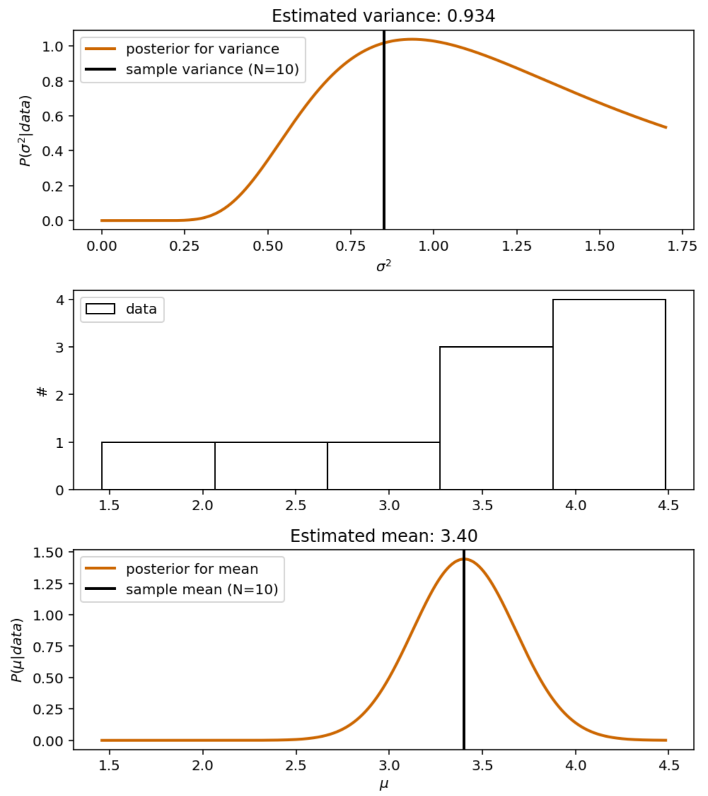
Variational Bayes, Expectation Maximization:

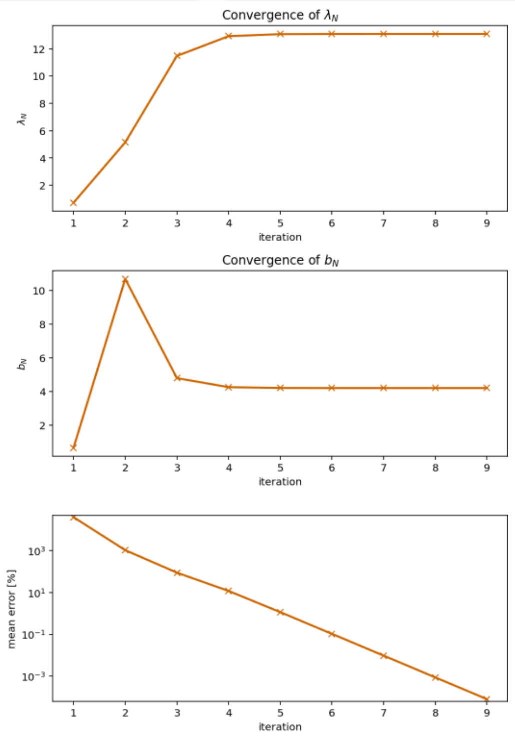
note:

- calculations for $q_i(Z_i|D)$ will lead to an **iterative procedure like for actual EM**
- instead of point estimates for Z_i (MLE) as before, **we get the actual posterior $q_i(Z_i|D)$**
- these distributions get more accurate the larger D → **learning** (see BPE)
- we only use maximum entropy

see `Var_Bayes_Example.py`

```
data = np.random.normal(3, 1, (10,))
Var_Bayes_Example(data)
```





Thank you very much for your attention!