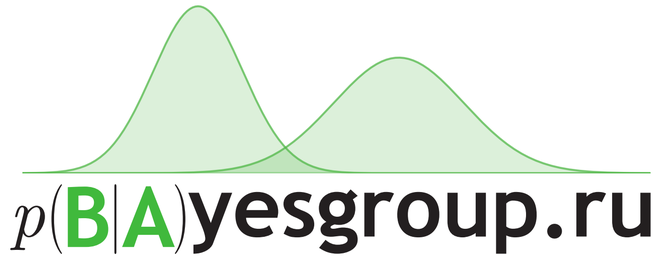
Latent variable models and EM-algorithm

Nadia Chirkova

Higher School of Economics, Samsung-HSE Laboratory
Moscow, Russia

SAMSUNG
Research



Machine learning overview

- Supervised learning: regression and classification
  - Linear models
  - Decision trees and ensembles
  - K Nearest Neighbors
  - Neural networks

- Unsupervised learning: understanding data structure
  - Clustering
  - Dimensionality reduction, e. g. PCA
    or autoencoders

(Bayesian) Machine learning overview

- Supervised learning: regression and classification
  - Linear models — **Bayesian linear regression**
  - Decision trees and ensembles
  - K Nearest Neighbors — **Gaussian processes**
  - Neural networks — **Bayesian neural networks**

- Unsupervised learning: understanding data structure
  - Clustering
  - Dimensionality reduction, e. g. PCA
    or autoencoders

(Bayesian) Machine learning overview

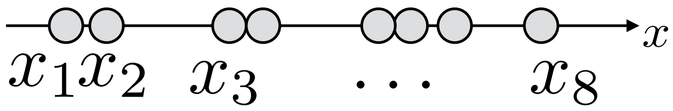
- Supervised learning: regression and classification
  - Linear models — **Bayesian linear regression**
  - Decision trees and ensembles
  - K Nearest Neighbors — **Gaussian processes**
  - Neural networks — **Bayesian neural networks**

- Unsupervised learning: **probabilistic models of data**
  - Clustering — **Gaussian Mixture Models**
  - Dimensionality reduction, e. g. PCA — **Probabilistic PCA**
    or autoencoders — **Variational autoencoders**

Probabilistic models

$$X = \{x_1, \dots, x_N\}, \; x_i \in \mathbb{R}^d \underset{i = 1, \dots, N}{}$$

1-dim data



images


x_1


x_2


x_3

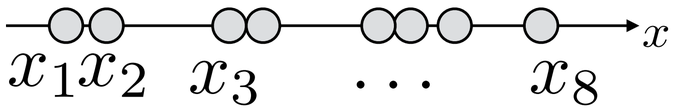
tabular data

	Age	Attrition	BusinessTravel	DailyRate	Department
x_1	41	Yes	Travel_Rarely	1102	Sales
x_2	49	No	Travel_Frequently	279	Research & Development
x_3	37	Yes	Travel_Rarely	1373	Research & Development
x_4	33	No	Travel_Frequently	1392	Research & Development

Probabilistic models

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}^d \underset{i = 1, \dots, N}{}$$

1-dim data



images

x_1

x_2

x_3

$p(x|\theta)$ — probabilistic model

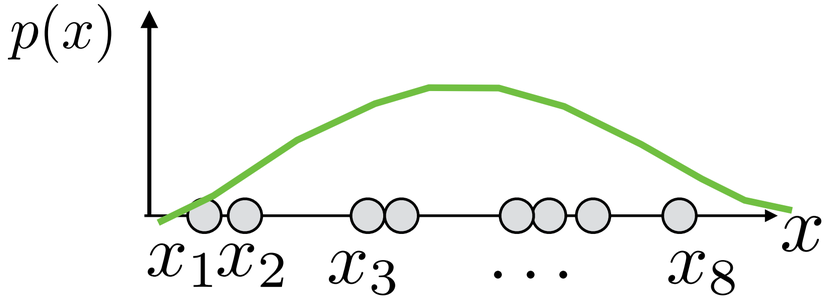
$$\prod_{i=1}^{N} p(x_i|\theta) \to \max_{\theta} \qquad \theta - ?$$

tabular data

	Age	Attrition	BusinessTravel	DailyRate	Department
x_1	41	Yes	Travel_Rarely	1102	Sales
x_2	49	No	Travel_Frequently	279	Research & Development
x_3	37	Yes	Travel_Rarely	1373	Research & Development
x_4	33	No	Travel_Frequently	1392	Research & Development

Probabilistic models

$$X = \{x_1, \ldots, x_N\}, \; x_i \in \mathbb{R}, \; i = 1, \ldots, N$$



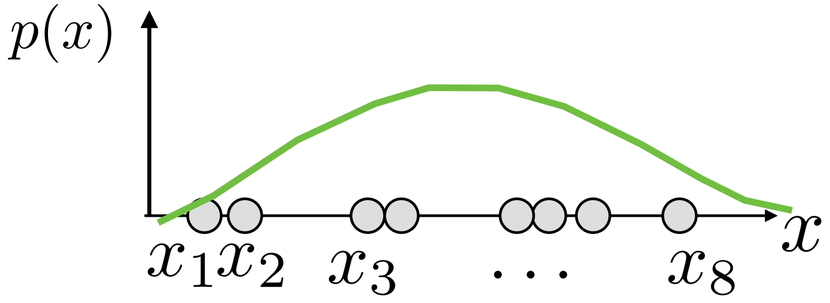
$$p(x|\mu, \sigma) = \mathcal{N}(x|\mu, \sigma^2) =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$\theta = \{\mu, \sigma\}$$

Probabilistic models

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}_{i=1,\dots,N}$$



$$p(x|\mu,\sigma) = \mathcal{N}(x|\mu,\sigma^2) =$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

$$\prod_{i=1}^{N} p(x_i|\mu,\sigma) \to \max_{\mu,\sigma}$$

Solution:

μ — sample mean

σ^2 — sample variance

Only simple models :(

Latent variable models

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables (one per object)

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

$$p(x|\theta) = \int p(x, z|\theta)dz$$

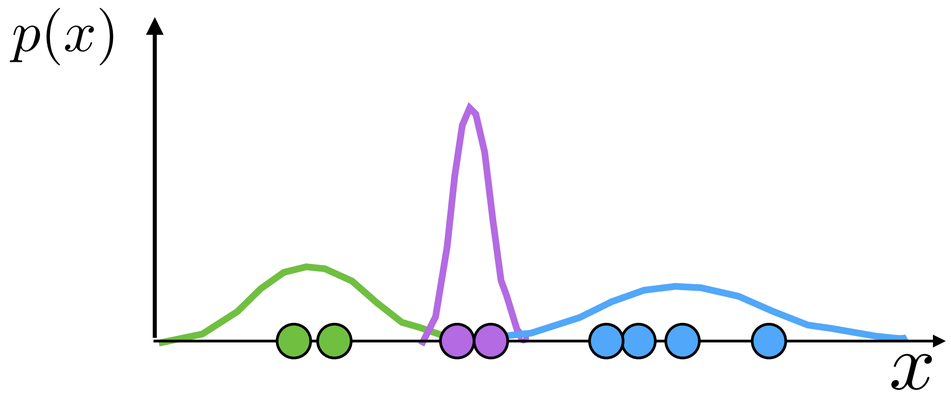
(or sum in case of discrete z)

Allow complex models!

Latent variable models: example 1

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, 2, 3\}$$
$$i = 1, \dots, N$$



Example 1: assume each object comes from one of three Gaussians

Latent variables z:

what Gaussian does an object come from?

clustering!

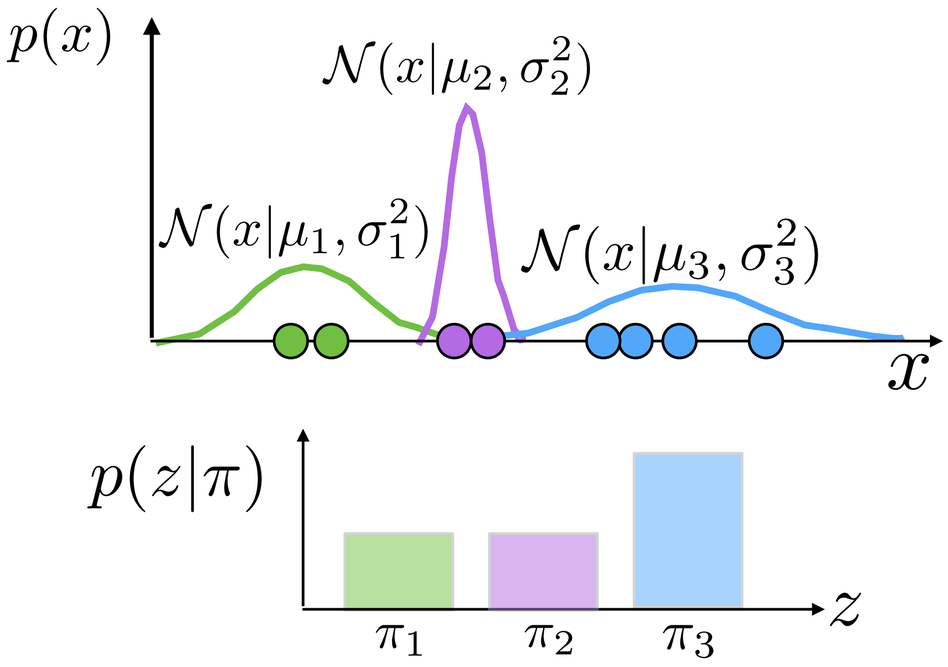
If we knew the values of latent variables, estimating θ would be trivial

Latent variable models: example 1

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, 2, 3\}$$
$$i = 1, \dots, N$$

$$p(x|\underbrace{\pi, \mu, \sigma}_{\theta}) = \sum_{z=1}^{3} p(x|z, \underbrace{\mu, \sigma}_{\theta_1})p(z|\underbrace{\pi}_{\theta_2}) =$$
$$= \sum_{z=1}^{3} \mathcal{N}(x|\mu_z, \sigma_z^2)\pi_z$$



Example 1: assume each object comes from one of three Gaussians

Latent variables z:

what Gaussian does an object come from?

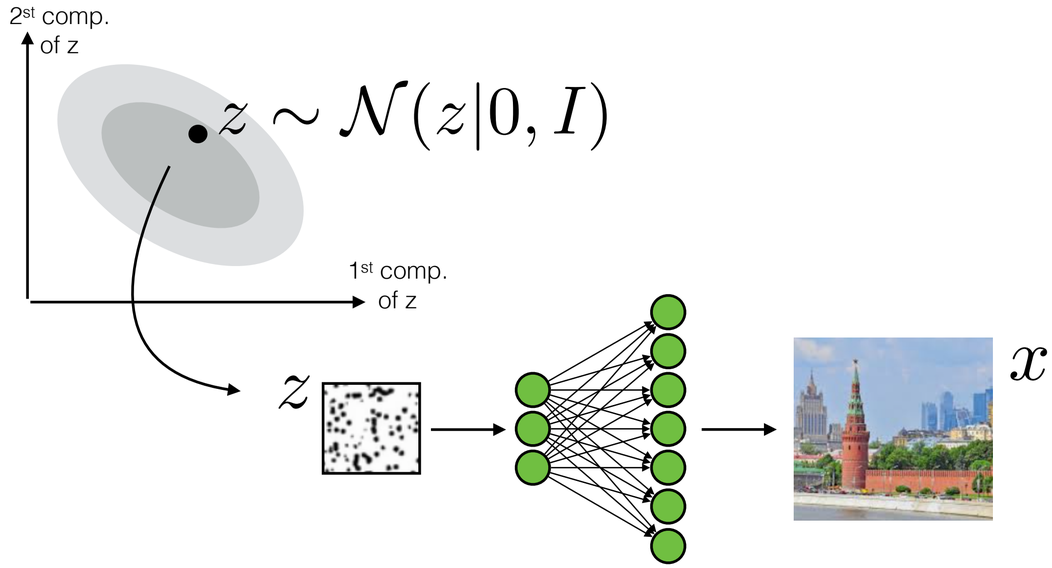
clustering!

If we knew the values of latent variables, estimating θ would be trivial

Latent variable models: example 2

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}}$$
$$i = 1, \ldots, N$$



decoder $p(x|z, \theta, \sigma)$
(linear or non-linear)

Example 2: assume each object is generated by the decoder from an embedding

Latent variables z:
what is an object's embedding?

If we knew the values of latent variables, estimating parameters θ would be trivial.

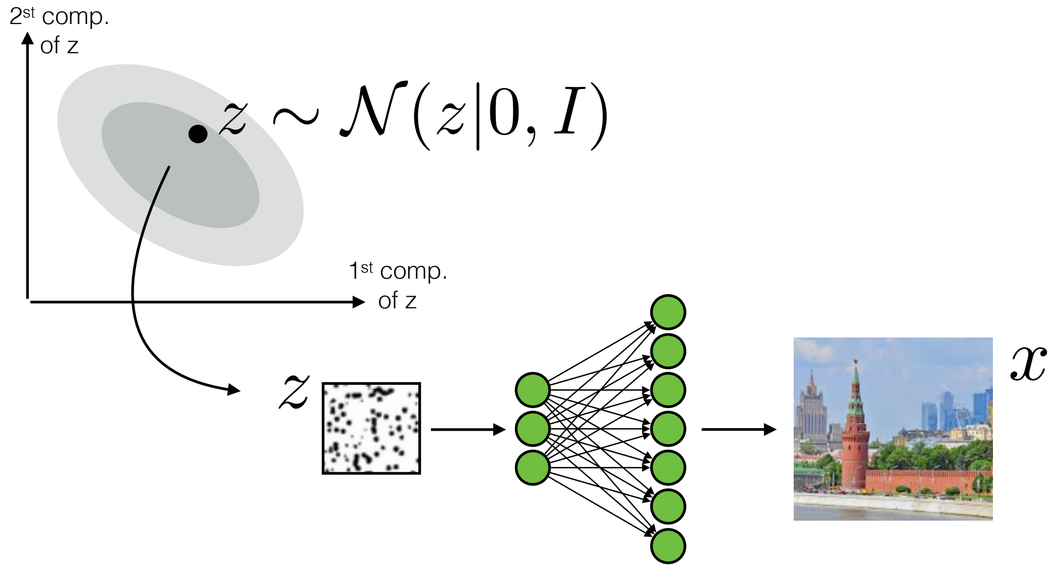
Latent variable models: example 2

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}} \quad i = 1, \ldots, N$$

$$p(x|\theta, \sigma) = \int p(x|z, \theta, \sigma) p(z) dz$$

$$= \int \mathcal{N}(x|\mu_\theta(z), \sigma^2 I) \mathcal{N}(z|0, I) dz$$



decoder $p(x|z, \theta, \sigma)$
(linear or non-linear)

Example 2: assume each object is generated by the decoder from an embedding

Latent variables z:
what is an object's embedding?

If we knew the values of latent variables, estimating parameters θ would be trivial.

Latent variable models

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

$$p(x|\theta) = \int p(x, z|\theta)dz$$

(or sum in case of discrete z)

How can l. v. models be useful?

- Understanding data structure: $p(z|x, \theta) = \frac{p(x|z, \theta_1)p(z|\theta_2)}{\int (x|\tilde{z}, \theta_1)p(\tilde{z}|\theta_2)d\tilde{z}}$

e. g. what cluster does an object come from?

- Generating new objects: $z \sim p(z|\theta_2), \quad x \sim p(x|z, \theta_1)$

e. g. generating new images using decoder

Latent variable models

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables

$$p(x, z|\theta) = \underbrace{p(x|z, \theta_1)}_{\text{likelihood}} \underbrace{p(z|\theta_2)}_{\text{prior}}$$

$$p(x|\theta) = \int p(x, z|\theta) dz$$

(or sum in case of discrete z)

How can l. v. models be useful?

- Understanding data structure: $\underbrace{p(z|x, \theta)}_{\text{posterior over l. v.}} = \frac{p(x|z, \theta_1)p(z|\theta_2)}{\int (x|\tilde{z}, \theta_1)p(\tilde{z}|\theta_2)d\tilde{z}}$
- Generating new objects: $z \sim p(z|\theta_2), \quad x \sim p(x|z, \theta_1)$

Training latent variable models

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

$$p(x|\theta) = \int p(x, z|\theta)dz$$

(or sum in case of discrete z)

θ — ?

$$\log p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \to \max_{\theta}$$

Variational lower bound

$$\log p(X|\theta) = \int q(Z) \log p(X|\theta) dz = \int q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} dZ =$$

$$= \int q(Z) \log \frac{p(X, Z|\theta) q(z)}{p(Z|X, \theta) q(z)} dZ =$$

$$= \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X, \theta)} dZ =$$

$$= \mathcal{L}(q(Z), \theta) + KL(q(Z)||p(Z|X, \theta))$$

(holds for any $q(Z)$)

Training latent variable models

$$\log p(X|\theta) = \int q(Z) \log p(X|\theta) dz = \int q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} dZ =$$

$$= \int q(Z) \log \frac{p(X, Z|\theta) q(z)}{p(Z|X, \theta) q(z)} dZ =$$

$$= \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X, \theta)} dZ =$$

$$= \mathcal{L}(q(Z), \theta) + \underbrace{KL(q(Z)||p(Z|X, \theta))}_{\geqq 0}$$

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \to \max_{q, \theta}$$

Optimization problem for training l. v. models

EM-algorithm

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \to \max_{q,\theta}$$

Block-coordinate ascent: alternate optimizing w. r. t. $q(Z)$ and θ

E-step: (expectation) $\mathcal{L}(q(Z), \theta) \to \max_{q}$

M-step: (maximization) $\mathcal{L}(q(Z), \theta) \to \max_{\theta}$

EM-algorithm

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \rightarrow \max_{q,\theta}$$

Block-coordinate ascent: alternate optimizing w. r. t. $q(Z)$ and θ

E-step:
(expectation)

$$\mathcal{L}(q(Z), \theta) \rightarrow \max_{q}$$

$$\log p(X|\theta) = \mathcal{L}(q(Z), \theta) + KL(q(Z)||p(Z|X, \theta))$$

constant w.r.t.q

$$\rightarrow \max_{q}$$

M-step:
(maximization)

$$\mathcal{L}(q(Z), \theta) \rightarrow \max_{\theta}$$

EM-algorithm

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \to \max_{q,\theta}$$

Block-coordinate ascent: alternate optimizing w. r. t. $q(Z)$ and θ

E-step:
(expectation)

$$\mathcal{L}(q(Z), \theta) \to \max_{q} \quad \Leftrightarrow \quad KL(q(Z)||p(Z|X, \theta)) \to \min_{q}$$

$$q(Z) = p(Z|X, \theta)$$

M-step:
(maximization)

$$\mathcal{L}(q(Z), \theta) \to \max_{\theta}$$

EM-algorithm

$$\log p(X|\theta) \geqslant \mathcal{L}(q(Z), \theta) \to \max_{q,\theta}$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

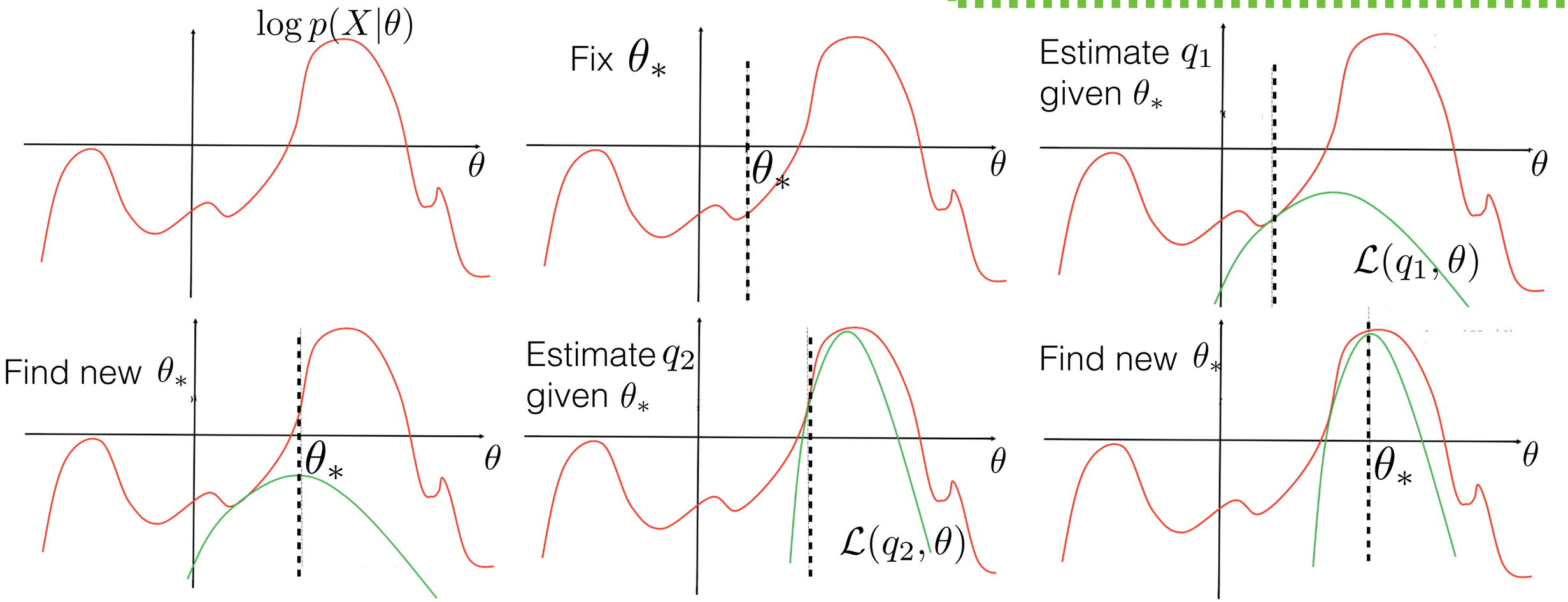
M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

EM-algorithm: visualization

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$



Visualization from Dmitry Vetrov's presentation, deepbayes.ru/2019

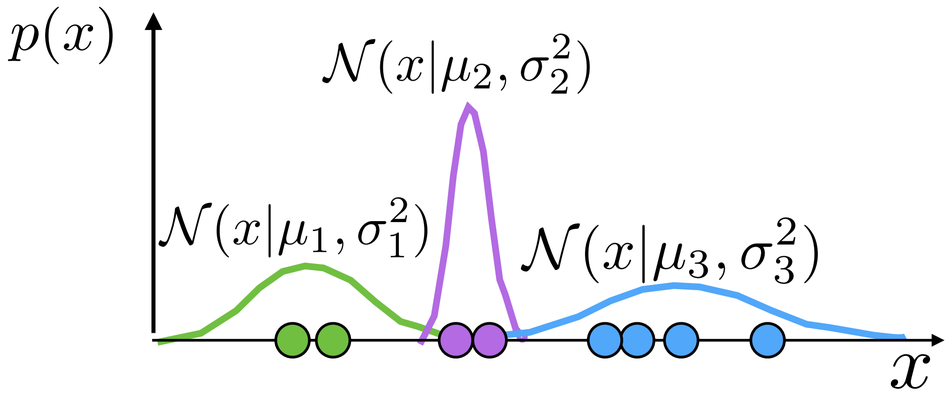
Gaussian mixture model (1-dim case)

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad \underset{i=1,\ldots,N}{}$$

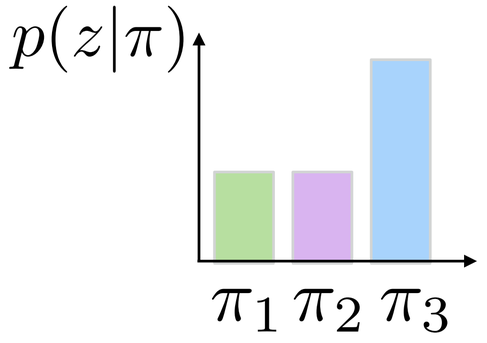
$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\} \quad \underset{i=1,\ldots,N}{}$$

$$p(x|\pi, \mu, \sigma) = \sum_{z=1}^{K} p(x|z, \mu, \sigma)p(z|\pi) =$$

$$= \sum_{z=1}^{K} \mathcal{N}(x|\mu_z, \sigma_z)\pi_z$$



pdfs of Gaussians



prior probabilities of Gaussians

$$\pi = \{\pi_1, \ldots, \pi_K\},$$

$$\pi_k \geqslant 0, \sum_{k=1}^{K} \pi_k = 1$$

Gaussian mixture model

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

E-step: ?

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

Gaussian mixture model

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\}$$
$$i = 1, \ldots, N$$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

E-step: $$p(Z|X, \theta_*) = \frac{p(X|Z, \theta_{*1})p(Z|\theta_{*2})}{\int p(X|\tilde{Z}, \theta_{*1})p(\tilde{Z}|\theta_{*2})d\tilde{Z}} =$$

$$= \prod_{i=1}^{N} \frac{p(x_i|z_i, \theta_{*1})p(z_i|\theta_{*2})}{\int p(x_i|\tilde{z}_i, \theta_{*1})p(\tilde{z}_i|\theta_{*2})d\tilde{z}_i} = \prod_{i=1}^{N} p(z_i|x_i, \theta_*)$$

(or sum in the denominator case of discrete z)

Gaussian mixture model

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

Initialize θ_* and repeat

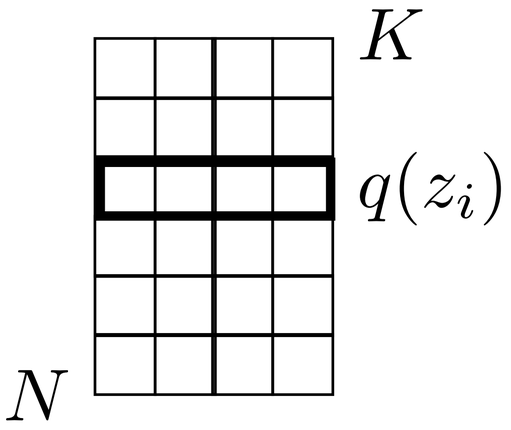
E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$
$$p(z|\pi) = \pi_z$$

E-step:

Matrix q:



$$q_{ik} = q(z_i = k) =$$
$$= p(z_i = k|x_i, \mu, \sigma, \pi) - ?$$

Gaussian mixture model

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

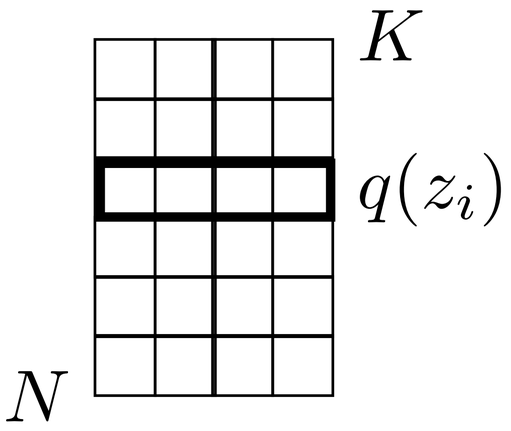
$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\}$$
$$i = 1, \ldots, N$$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$
$$p(z|\pi) = \pi_z$$

E-step:

Matrix q:



$$q_{ik} = p(z_i = k|x_i, \mu, \sigma, \pi) = \frac{p(x_i|z_i = k, \mu, \sigma)p(z_i = k|\pi)}{\sum_{k'=1}^{K} p(x_i|z_i = k', \mu, \sigma)p(z_i = k'|\pi)}$$

Gaussian mixture model

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\}$$
$$i = 1, \ldots, N$$

Initialize θ_* and repeat

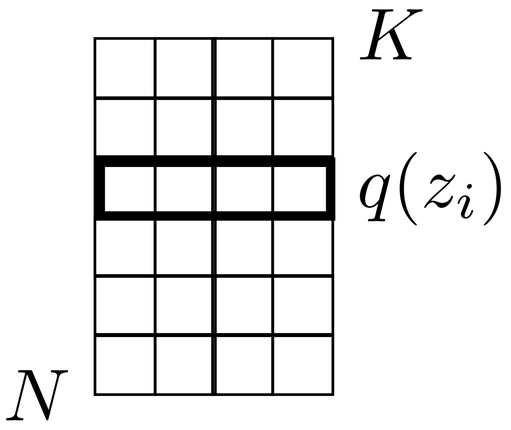
E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$
$$p(z|\pi) = \pi_z$$

E-step:

Matrix q:



$$q_{ik} = \frac{\mathcal{N}(x|\mu_k, \sigma_k)\pi_k}{\sum_{k'=1}^{K} \mathcal{N}(x|\mu_{k'}, \sigma_{k'})\pi_{k'}}$$

Gaussian mixture model

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\}$$
$$i = 1, \ldots, N$$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

M-step - ?

$$\mathcal{L}(q(Z), \theta) = \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ = \int q(Z) \log \frac{p(X|Z, \theta_1) p(Z|\theta_2)}{q(Z)} dZ$$

Gaussian mixture model

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

M-step - ?

$$\mathcal{L}(q(Z), \theta) = \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ = \int q(Z) \log \frac{p(X|Z, \theta_1) p(Z|\theta_2)}{q(Z)} dZ$$

$$= \sum_{i=1}^{N} \int q(z_i) \log \frac{p(x_i|z_i, \theta_1) p(z_i|\theta_2)}{q(z_i)} dz_i = \sum_{i=1}^{N} \mathcal{L}(q(z_i), \theta)$$

Gaussian mixture model

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

M-step - ?

$$\mathcal{L}(q(z_i), \theta) = ?$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

$$\sum_{i=1}^{N} \int q(z_i) \log \frac{p(x_i|z_i, \theta_1) p(z_i|\theta_2)}{q(z_i)} dz_i = \sum_{i=1}^{N} \mathcal{L}(q(z_i), \theta)$$

Gaussian mixture model

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

M-step - ?

$$\mathcal{L}(q(z_i), \mu, \sigma, \pi) = \sum_{k=1}^{K} q_{ik} \log \frac{\mathcal{N}(x_i|\mu_k, \sigma_k)\pi_k}{q_{ik}}$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

Gaussian mixture model

$$X = \{x_1, \dots, x_N\},\ x_i \in \mathbb{R}$$
$$i = 1, \dots, N$$

$$Z = \{z_1, \dots, z_N\},\ z_i \in \{1, \dots, K\}$$
$$i = 1, \dots, N$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, \mu, \sigma) = \mathcal{N}(x|\mu_z, \sigma_z)$$

$$p(z|\pi) = \pi_z$$

M-step:

$$\mathcal{L}(q(z_i), \mu, \sigma, \pi) = \sum_{k=1}^{K} q_{ik} \log \frac{\mathcal{N}(x_i|\mu_k, \sigma_k)\pi_k}{q_{ik}}$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik} \log\big(\mathcal{N}(x_i|\mu_k, \sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\bigl(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\bigr) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. μ_j?

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\left(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\right) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. μ_j?

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\left(-\frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right) + \text{const}$$

$$\frac{\partial}{\partial \mu_j}:$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. μ_j?

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\left(-\frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right) + \text{const}$$

$$\frac{\partial}{\partial \mu_j}: \quad -\sum_{i=1}^{N}\frac{q_{ij}}{\sigma_j^2}(x_i-\mu_j)=0$$

$$\mu_j = \frac{\sum_{i=1}^{N} q_{ij}x_i}{\sum_{i=1}^{N} q_{ij}}$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\bigl(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\bigr) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. σ_j?

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. σ_j?

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\left(-\log\sigma_k - \frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right)+\text{const}$$

$$\frac{\partial}{\partial\sigma_j}:$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. σ_j?

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\left(-\log\sigma_k - \frac{1}{2\sigma_k^2}(x_i-\mu_k)^2\right)+\text{const}$$

$$\frac{\partial}{\partial\sigma_j}: \quad \sum_{i=1}^{N} q_{ij}\left(-\frac{1}{\sigma_j}+\frac{2}{2\sigma_j^3}(x_i-\mu_j)^2\right)=0$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{N} q_{ij}(x_i-\mu_j)^2}{\sum_{i=1}^{N} q_{ij}}$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\bigl(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\bigr) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. π ?

$$\sum_{k=1}^{K}\pi_k = 1$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. π ?

$$\sum_{k=1}^{K}\pi_k = 1$$

Lagrangian!

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. π ?

$$\sum_{k=1}^{K}\pi_k = 1$$

$$L(\pi,\lambda) = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\pi_k + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)$$

$$\frac{\partial L}{\partial \pi_j} =$$

$$\frac{\partial L}{\partial \lambda} =$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. π ?

$$\sum_{k=1}^{K}\pi_k = 1$$

$$L(\pi,\lambda) = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\pi_k + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)$$

$$\frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{N}\frac{q_{ij}}{\pi_j} + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K}\pi_k - 1 = 0$$

$$\pi_j = \frac{\sum_{i=1}^{N} q_{ij}}{\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ij}}$$

Gaussian mixture model: M-step

$$\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\big(\mathcal{N}(x_i|\mu_k,\sigma_k)\pi_k\big) \to \max_{\mu,\sigma,\pi}$$

Maximum w.r.t. π ?

$$\sum_{k=1}^{K}\pi_k = 1$$

$$L(\pi,\lambda) = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\log\pi_k + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)$$

$$\frac{\partial L}{\partial \pi_j} = \sum_{i=1}^{N}\frac{q_{ij}}{\pi_j} + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K}\pi_k - 1 = 0$$

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} q_{ij}$$

Gaussian mixture model: EM-algorithm

Initialize μ, σ, π and repeat

E-step: $$q_{ik} = \frac{\mathcal{N}(x|\mu_k, \sigma_k)\pi_k}{\sum_{k'=1}^{K} \mathcal{N}(x|\mu_{k'}, \sigma_{k'})\pi_{k'}}$$

M-step: $$\mu_k = \frac{\sum_{i=1}^{N} q_{ik} x_i}{\sum_{i=1}^{N} q_{ik}} \qquad \pi_k = \frac{1}{N} \sum_{i=1}^{N} q_{ik}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} q_{ik}(x_i - \mu_k)^2}{\sum_{i=1}^{N} q_{ik}}$$

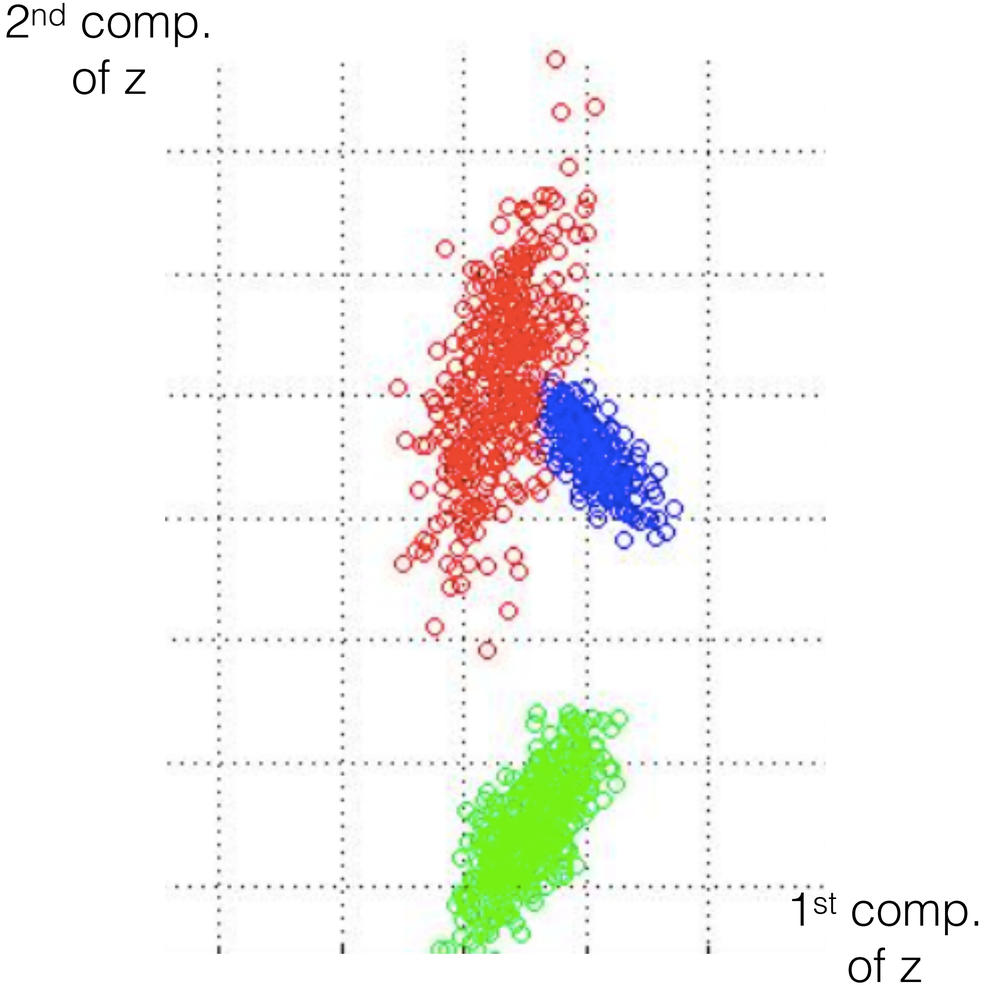
Gaussian mixture model: multi-dim case

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \{1, \ldots, K\} \quad i = 1, \ldots, N$$

$$p(x|\pi, \mu, \Sigma) = \sum_{z=1}^{K} p(x|z, \mu, \Sigma) p(z|\pi)$$

$$= \sum_{z=1}^{K} \mathcal{N}(x|\mu_z, \Sigma_z) \pi_z$$



Gaussian mixture model: EM-algorithm

Initialize μ, Σ, π and repeat

E-step: $$q_{ik} = \frac{\mathcal{N}(x|\mu_k, \Sigma_k)\pi_k}{\sum_{k'=1}^{K} \mathcal{N}(x|\mu_{k'}, \Sigma_{k'})\pi_{k'}}$$

M-step: $$\mu_k = \frac{\sum_{i=1}^{N} q_{ik} x_i}{\sum_{i=1}^{N} q_{ik}} \qquad \pi_k = \frac{1}{N} \sum_{i=1}^{N} q_{ik}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} q_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} q_{ik}}$$

Gaussian mixture model: summary

- GMM provides “soft” clustering of points
- All formulas are easily interpretable
- There are more complicated models that involve priors on GMM parameters

Probabilistic PCA

$$X = \{x_1, \ldots, x_N\},\ x_i \in \mathbb{R}^d \quad i = 1, \ldots, N$$

$$Z = \{z_1, \ldots, z_N\},\ z_i \in \mathbb{R}^{d_{\text{small}}} \quad i = 1, \ldots, N$$

$$p(X, Z|V, \sigma^2) = \prod_{i=1}^{n} p(x_i|z_i, V, \sigma^2) p(z_i) =$$

$$= \prod_{i=1}^{N} \mathcal{N}(x_i|V z_i, \sigma^2 I) \mathcal{N}(z_i|0, I)$$

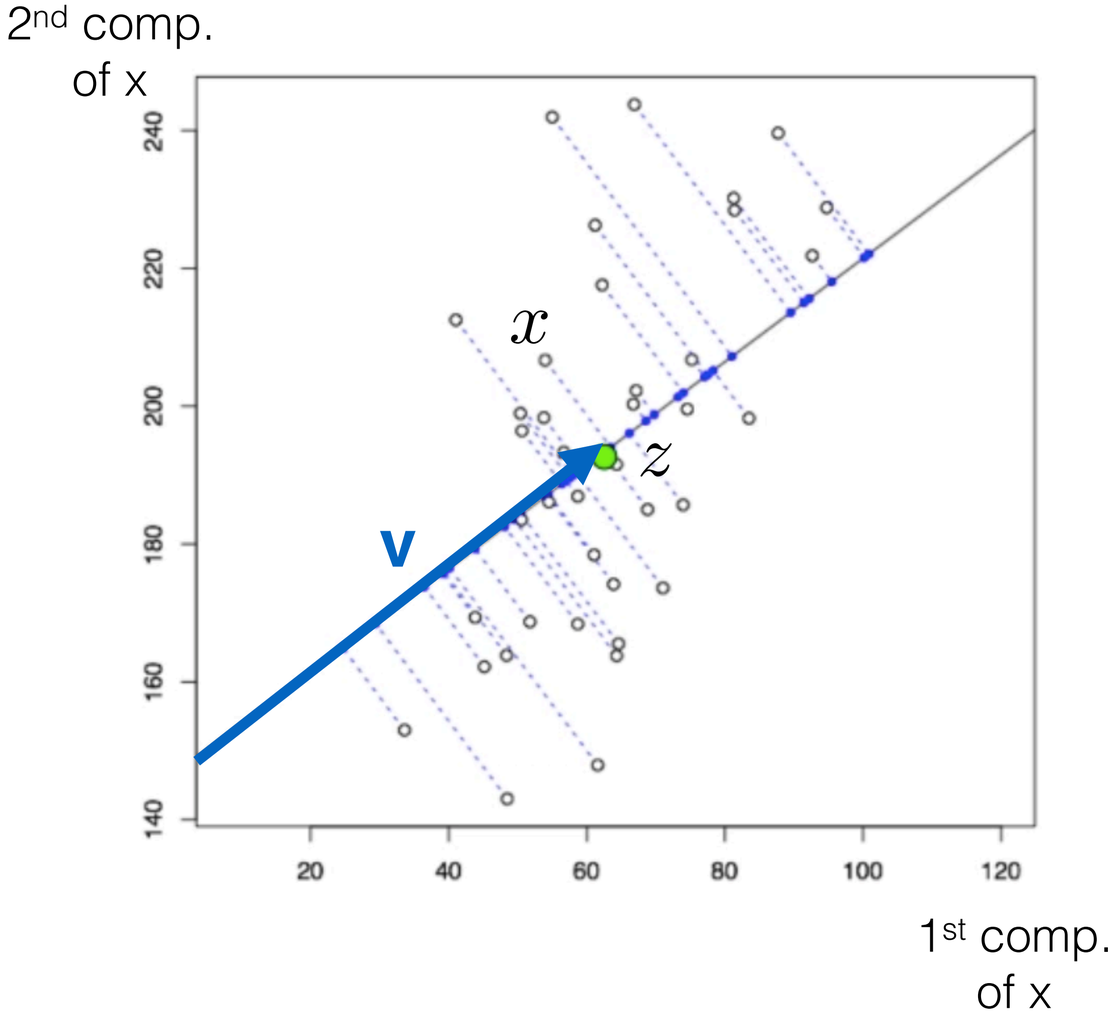


Image from https://liorpachter.wordpress.com/tag/probabilistic-pca/

Probabilistic PCA

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

E-step: $$q(Z) = \prod_{i=1}^{N} q(z_i) = \prod_{i=1}^{N} p(z_i|x_i, V, \sigma^2)$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

Probabilistic PCA

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

E-step: $$q(Z) = \prod_{i=1}^{N} q(z_i) = \prod_{i=1}^{N} p(z_i|x_i, V, \sigma^2)$$

conjugance!

$$q(z_i) \propto \mathcal{N}(x_i|Vz_i, \sigma^2 I)\mathcal{N}(z_i|0, I) =$$

Probabilistic PCA

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

E-step: $$q(Z) = \prod_{i=1}^{N} q(z_i) = \prod_{i=1}^{N} p(z_i|x_i, V, \sigma^2)$$

$$q(z_i) \propto \mathcal{N}(x_i|Vz_i, \sigma^2 I)\mathcal{N}(z_i|0, I) =$$

$$= \mathcal{N}\left(z_i \middle| \underbrace{\left(\tfrac{1}{\sigma^2}V^TV + I\right)^{-1}\left(\textstyle\sum_{i=1}^{N} V^T x_i\right)}_{\mu_i^q}, \underbrace{\left(\tfrac{1}{\sigma^2}V^TV + I\right)^{-1}}_{\Sigma^q}\right)$$

formulas of V, σ and X

Probabilistic PCA

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$\mathcal{L}(q(Z), \theta) = \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ$$

M-step:

$$\sum_{i=1}^{N} \int q(z_i) \log p(x_i, z_i|V, \sigma^2) dz_i = \mathbb{E}_{q(z_i)} \log p(x_i, z_i|V, \sigma^2) \to \max_{V, \sigma^2}$$

Probabilistic PCA

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$\mathcal{L}(q(Z), \theta) = \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ$$

M-step:

$$\sum_{i=1}^{N} \int q(z_i) \log p(x_i, z_i|V, \sigma^2) dz_i = \mathbb{E}_{q(z_i)} \log p(x_i, z_i|V, \sigma^2) \to \max_{V, \sigma^2}$$

1-dim z case:

$$\mathbb{E}_{q(z_i)} \log\big(\mathcal{N}(x_i|vz_i, \sigma^2)\mathcal{N}(z_i|0, 1)\big) = \mathbb{E}_{q(z_i)}\big(-2\log\sigma + \exp\big(0.5\sigma^{-2}||x_i - vz_i||^2\big)$$

$$\mathbb{E}_{q(z_i)} z_i = \mu_i^q \qquad \mathbb{E}_{q(z_i)} z_i^2 = \sigma_i^q + \mu_i^q$$

Probabilistic PCA

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

Initialize θ_* and repeat

E-step: $q(Z) = p(Z|X, \theta_*)$

M-step: $\theta_* = \text{argmax}_\theta \mathcal{L}(q(Z), \theta)$

$$\mathcal{L}(q(Z), \theta) = \int q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} dZ$$

M-step:

$$V = \left(\sum_{i=1}^{N} x_i (\mathbb{E}_{q(z_i)} z_i)^T \right) \left(\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} z_i z_i^T \right)^{-1}$$

$$\sigma^2 = \frac{1}{Nd} \sum_{i=1}^{N} \mathbb{E}_{q(z_i)} ||x_i - V z_i||^2$$

formulas of q params and X

Probabilistic PCA: EM-algorithm

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I) \qquad p(z) = \mathcal{N}(z|0, I)$$

E-step: $$q(z_i) = p(z_i|x_i, V, \sigma^2) = \mathcal{N}(x_i|\mu_i^q, \Sigma^q)$$

formula of V, σ and X

M-step:
$$V = \left(\sum_{i=1}^{N} x_i (\mathbb{E}_{q(z_i)} z_i)^T\right)\left(\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} z_i z_i^T\right)^{-1}$$

$$\sigma^2 = \frac{1}{Nd}\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} ||x_i - Vz_i||^2$$

formulas of q params and X

Probabilistic PCA: EM-algorithm

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I) \qquad p(z) = \mathcal{N}(z|0, I)$$

E-step: $$q(z_i) = p(z_i|x_i, V, \sigma^2) = \mathcal{N}(x_i|\mu_i^q, \Sigma^q)$$

M-step:
$$V = \left(\sum_{i=1}^{N} x_i \overbrace{(\mathbb{E}_{q(z_i)} z_i)^T}^{\mu_i^q}\right)\left(\sum_{i=1}^{N} \overbrace{\mathbb{E}_{q(z_i)} z_i z_i^T}^{\mu_i^q(\mu_i^q)^T + \Sigma^q}\right)^{-1}$$

$$\sigma^2 = \frac{1}{Nd}\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} ||x_i - Vz_i||^2$$

formula of V, σ and X

formulas of q params and X

Probabilistic PCA: pros

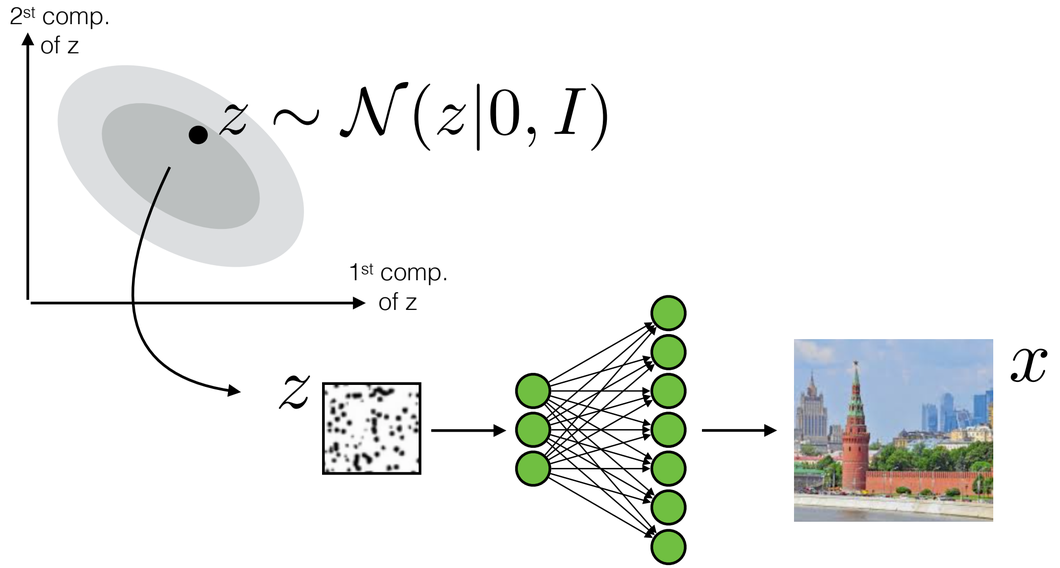
- Probabilistic PCA provides iterative EM-algorithm for finding transformation matrix V that may be more efficient than SVD-decomposition
- EM-algorithm allows combining models, e. g. Mixture of PPCA
- PPCA can determine d_{small} (more complex model)

Next lecture: variational autoencoders

$$p(x|z, V, \sigma^2) = \mathcal{N}(x|Vz, \sigma^2 I)$$

$$p(z) = \mathcal{N}(z|0, I)$$

next lecture:
nonlinear transformation



decoder $p(x|z, \theta, \sigma)$
(linear or non-linear)

Latent variable models

$X = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}^d$, $i = 1, \dots, N$

$Z = \{z_1, \dots, z_N\}$ — latent variables

$$p(x, z|\theta) = p(x|z, \theta_1)p(z|\theta_2)$$

$$p(x|\theta) = \int p(x, z|\theta)dz$$

(or sum in case of discrete z)

How can l. v. models be useful?

- Understanding data structure: $p(z|x, \theta) = \frac{p(x|z, \theta_1)p(z|\theta_2)}{\int (x|\tilde{z}, \theta_1)p(\tilde{z}|\theta_2)d\tilde{z}}$

e. g. what cluster does an object come from?

- Generating new objects: $z \sim p(z|\theta_2), \quad x \sim p(x|z, \theta_1)$

e. g. generating new images using decoder