Probabilistic Machine Learning
Lecture 23
EM

Philipp Hennig

20 July 2023

Eberhard Karls Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

A special option: The EM algorithm

Maximizing model Evidence

<u>The general recipe</u> for hyperparameter inference:
Consider a model with **parameters $\boldsymbol{\theta}$**, **observed data $\boldsymbol{y}$** and **latent variables z**

- Ideally, we would like to maximize the **marginal (log-) likelihood (evidence)**

$$\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \log \left(\int p(\boldsymbol{y}, z \mid \boldsymbol{\theta})\, dz \right) \qquad (\star)$$

- if we can not do this integral, we can try **Laplace**. This is nearly always *possible* (if $\log p(\boldsymbol{y} \mid z)$ is twice differentiable), but it is fundamentally an approximation.
- however, *in some cases*, we may be able to compute the Expectation **of the "complete data" log likelihood** (for a fixed value $\boldsymbol{\theta}_*$)

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = \int p(z \mid \boldsymbol{y}, \boldsymbol{\theta}_*) \log p(\boldsymbol{y}, z \mid \boldsymbol{\theta})\, dz$$

and then Maximize $q(\boldsymbol{\theta}, \boldsymbol{\theta}_*)$ **with respect to $\boldsymbol{\theta}$**. This can be easier than $(\star)$ because the log "simplifies things" (e.g. turns products into sums, thus factors into components).

The EM Algorithm

Dempster, Laird, Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, RSS B **39** (1) 1977, 1–38

Definition: The **Expectation Maximizatiion (EM)** algorithm:
Consider a model with **parameters** $\boldsymbol{\theta}$, **observed data** $\boldsymbol{y}$ and **latent variables** $\boldsymbol{z}$.

```
while not converged, do:
```

E compute the **Expected complete data log-likelihood**

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \int p(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\theta}_t) \log p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, d\boldsymbol{z}$$

M Set θ_{t+1} to **Maximize** $\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} q(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1})$.

EM is an attempt to maximize the evidence $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. Why does it work?

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ for our latent quantity
- For *any* such approximation $q(z)$ (if $q(z) > 0$ wherever $p(x, z \mid \theta) > 0$):

$$\begin{aligned} \log p(x \mid \theta) &= \log \int p(x, z \mid \theta)\, dz && = \log \int q(z) \frac{p(x, z \mid \theta)}{q(z)}\, dz \\ &\geq \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)}\, dz && =: \mathcal{L}(q) \end{aligned}$$

Theorem (Jensen's (1906) inequality)

Let (Ω, A, μ) be a probability space, g be a real-valued, μ-integrable function and ϕ be a convex function on the real line. Then

$$\phi\left(\int_\Omega g\, d\mu\right) \leq \int_\Omega \phi \circ g\, d\mu.$$

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ for our latent quantity
- For *any* such approximation $q(z)$ (if $q(z) > 0$ wherever $p(x, z \mid \theta) > 0$):

$$
\begin{aligned}
\log p(x \mid \theta) = \log \int p(x, z \mid \theta)\, dz \qquad & = \log \int q(z) \frac{p(x, z \mid \theta)}{q(z)}\, dz \\
\geq \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)}\, dz \qquad & =: \mathcal{L}(q)
\end{aligned}
$$

- Thus, by maximizing the RHS in θ in the M-step, we increase a **lower bound on the Evidence**
- $\mathcal{L}(q)$ is thus called the **Evidence Lower Bound** (ELBO)
- But can we be sure that this increases the Evidence? To show that this is the case, we will now establish that the E-step makes the bound *tight* at the local θ.

An Insight

Distance of our bound

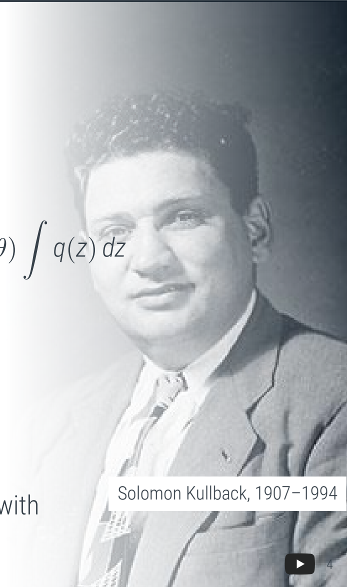
Richard A. Leibler, 1914–2003

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)} \, dz \\
&= \int q(z) \log \frac{p(z \mid x, \theta) \cdot p(x \mid \theta)}{q(z)} \, dz \\
&= \int q(z) \log \frac{p(z \mid x, \theta)}{q(z)} \, dz + \log p(x \mid \theta) \int q(z) \, dz
\end{aligned}
$$

thus

$$
\begin{aligned}
\log p(x \mid \theta) &= \mathcal{L}(q) - \int q(z) \log \frac{p(z \mid x, \theta)}{q(z)} \\
&= \mathcal{L}(q) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))
\end{aligned}
$$

The Kullback-Leibler divergence satisfies $D_{\mathrm{KL}}(q\|p) \geq 0$ with $D_{\mathrm{KL}}(q\|p) = 0 \quad \Leftrightarrow q \equiv p$



Solomon Kullback, 1907–1994

EM maximizes the ELBO / minimizes Free Energy

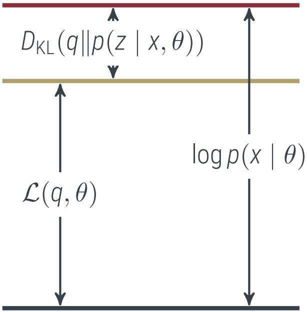
a more general view

exposition based on C.M. Bishop, 2006 §9.4

$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))$$

$$\mathcal{L}(q, \theta) = \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz$$



EM maximizes the ELBO / minimizes Free Energy

a more general view

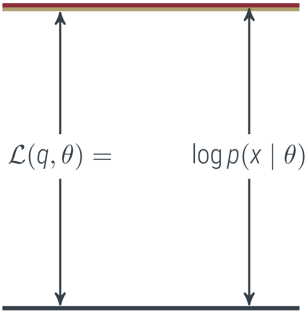
exposition based on C.M. Bishop, 2006 §9.4

$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))$$

$$\mathcal{L}(q, \theta) = \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz$$

E -step: $q(z) = p(z \mid x, \theta_{\text{old}})$, thus $D_{\mathrm{KL}}(q \| p(z \mid x, \theta_i)) = 0$



EM maximizes the ELBO / minimizes Free Energy

a more general view

exposition based on C.M. Bishop, 2006 §9.4

$$\log p(x \mid \theta) = \mathcal{L}(q,\theta) + D_{\mathrm{KL}}(q \| p(z \mid x,\theta))$$

$$\mathcal{L}(q,\theta) = \int q(z) \log\left(\frac{p(x,z \mid \theta)}{q(z)}\right) dz$$

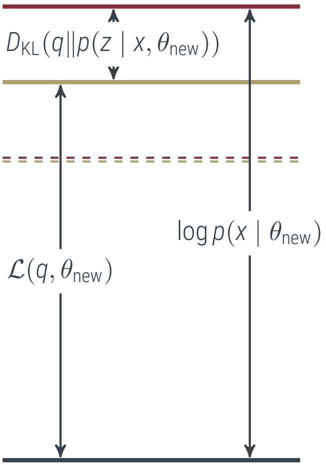
$$D_{\mathrm{KL}}(q \| p(z \mid x,\theta)) = -\int q(z) \log\left(\frac{p(z \mid x,\theta)}{q(z)}\right) dz$$

E -step: $q(z) = p(z \mid x, \theta_{\mathrm{old}})$, thus $D_{\mathrm{KL}}(q \| p(z \mid x, \theta_i)) = 0$

M -step: **Maximize ELBO**

$$\theta_{\mathrm{new}} = \arg\max_\theta \int q(z) \log p(x,z \mid \theta)\, dz$$

$$= \arg\max_\theta \mathcal{L}(q,\theta) + \int q(z) \log q(z)\, dz$$



EM Algorithm — General Form

summary

exposition based on C.M. Bishop, 2006 §9.4

Setting: Want to find *maximum likelihood* (or MAP) estimate for a model involving a **latent** variable

$$\theta_* = \arg\max_\theta \left[\log p(x \mid \theta)\right] = \arg\max_\theta \left[\log\left(\int p(x, z \mid \theta)\, dz\right)\right]$$

Algorithm: Initialize θ_0, then iterate:

1. Compute $q(z) = p(z \mid x, \theta_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(z \mid x, \theta)) = 0$
2. Set θ_{new} to the **Maximize** the **Evidence Lower Bound**

$$\theta_{\text{new}} = \arg\max_\theta \mathcal{L}(q, \theta) = \arg\max_\theta \int q(z) \log\left(\frac{p(x, z \mid \theta)}{q(z)}\right) dz$$

3. Check for convergence of either the log likelihood, or θ.

EM for MAP

dealing with the data is the main challenge, adding a prior is easy

- It is straightforward to extend EM to maximize a **posterior** instead of a likelihood. Just add a log prior for θ:

Initialize θ_0, then iterate between

1. Compute $q(z) = p(z \mid x, \theta_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(z \mid x, \theta)) = 0$
2. Set θ_{new} to the **Maximize** the **Evidence Lower Bound**

$$\theta_{\text{new}} = \arg\max_{\theta} \int q(z) \log \left(\frac{p(x, z \mid \theta) p(\theta)}{q(z)} \right) dz = \arg\max_{\theta} \mathcal{L}(q, \theta) + \log p(\theta)$$

3. Check for convergence of either the log likelihood, or θ.

This maximizes

$$\begin{aligned} \log p(x \mid \theta) + \log p(\theta) &\geq \mathcal{L}(q, \theta) + \log p(\theta) \\ &\triangleq \log p(\theta \mid x) \end{aligned}$$

Another Observation

why is it even useful to build an iterative update in this way?

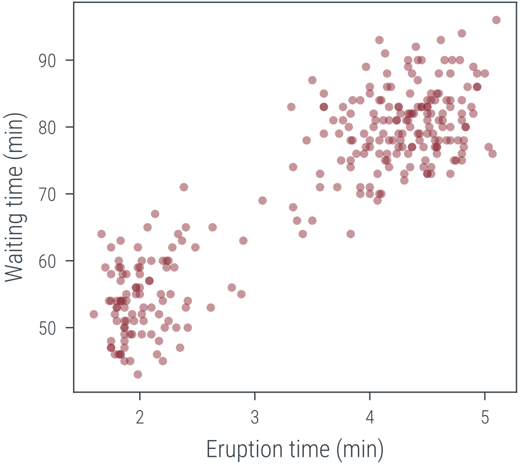
If $p(x, z \mid \theta)$ is an **exponential family** with θ as the natural parameters, then

$$
\begin{aligned}
p(x,z) &= \exp(\phi(x,z)^\intercal\theta - \log Z(\theta)) \\
\mathcal{L}(q(z),\theta) &= \mathbb{E}_{q(z)}(\phi(x,z)^\intercal\theta - \log Z(\theta)) = \mathbb{E}_{q(z)}[\phi(x,z)]^\intercal\theta - \log Z(\theta) \\
\nabla_\theta\mathcal{L}(q(z),\theta) = 0 \quad &\Rightarrow \quad \nabla_\theta \log Z(\theta) = \mathbb{E}_{p(x,z)}[\phi(x,z)] = \mathbb{E}_{q(z)}[\phi(x,z)]
\end{aligned}
$$

and optimization may be analytic (example below: Gaussian Mixture Models).

Example

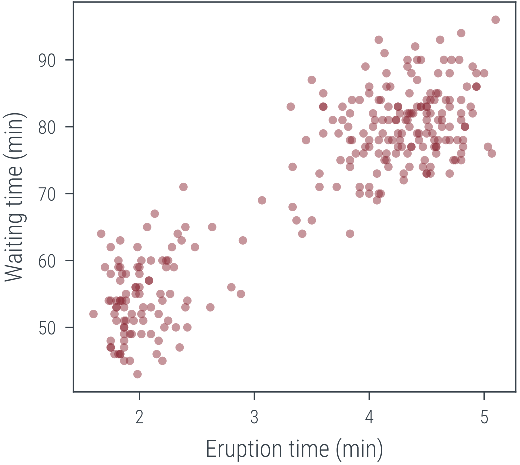
Gaussian Mixture Models



$$p(\boldsymbol{x} \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k),$$

$$\pi_k \in [0, 1], \sum_k \pi_k = 1$$

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \pi, \mu, \Sigma) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$$

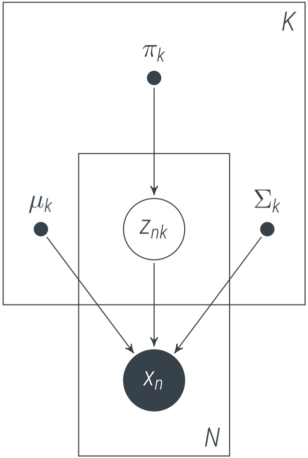


EM for Gaussian Mixtures

introducing a class membership variable z factorizes the likelihood

- Want to maximize, as function of $\theta := (\pi_k, \mu_k, \Sigma_k)_{k=1,\dots,K}$

$$\log p(x \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right)$$



EM for Gaussian Mixtures

introducing a class membership variable z factorizes the likelihood

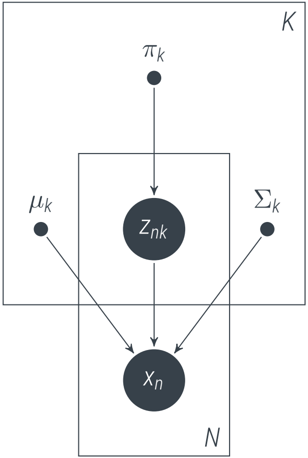
- Want to maximize, as function of $\theta := (\pi_k, \mu_k, \Sigma_k)_{k=1,\dots,K}$

$$\log p(x \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right)$$

- Instead, maximizing the "complete data" likelihood is easier:

$$\log p(x, z \mid \pi, \mu, \Sigma) = \log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n; \mu_k, \Sigma_k)^{z_{nk}}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \underbrace{(\log \pi_k + \log \mathcal{N}(x_n; \mu_k, \Sigma_k))}_{\text{easy to optimize (exponential families!)}}$$



E-Step: Compute $p(z \mid x, \theta)$:

$$p(z_{nk} = 1 \mid x_n, \pi, \mu, \Sigma) = \frac{p(z_{nk} = 1) p(x_n \mid z_{nk} = 1)}{\sum_{k'=1}^K p(z_{nk'} = 1) p(x_n \mid z_{nk'} = 1)}$$

$$= \frac{\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_n; \mu_{k'}, \Sigma_{k'})} =: r_{nk}$$

Note that discrete distributions $q(z_{nk} = 1) = r_{nk}$ have expectation $\mathbb{E}_q[z_{nk}] = r_{nk}$

M-Step: Maximize ELBO

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x},\theta)} \left(\log p(\mathbf{x}, \mathbf{z} \mid \theta) \right) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \left(\log \pi_k + \log \mathcal{N}(x_n; \mu_k, \Sigma_k) \right)$$

EM for Gaussian Mixtures

M-step derivations

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\theta)}\left(\log p(\boldsymbol{x},\boldsymbol{z}\mid\theta)\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_k + \log \mathcal{N}(x_n;\mu_k,\Sigma_k)\right)$$

To maximize w.r.t. μ set gradient of ELBO to 0:

$$\nabla_{\mu_\ell}\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\theta)}\left(\log p(\boldsymbol{x},\boldsymbol{z}\mid\theta)\right) = -\sum_{n=1}^{N} r_{n\ell}\Sigma_\ell^{-1}(x_n-\mu_\ell) \overset{!}{=} 0$$

$$\Rightarrow \quad \mu_\ell = \frac{1}{R_\ell}\sum_{n=1}^{N} r_{n\ell}x_n \qquad R_j := \sum_{n=1}^{N} r_{n\ell}$$

EM for Gaussian Mixtures

M-step derivations

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x},\theta)}\left(\log p(\mathbf{x},\mathbf{z} \mid \theta)\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_k + \log \mathcal{N}(x_n;\mu_k,\Sigma_k)\right)$$

To maximize w.r.t. Σ set gradient of ELBO to 0
(note $\partial|\Sigma|^{-1/2}/\partial\Sigma = -\frac{1}{2}|\Sigma|^{-3/2}|\Sigma|\Sigma^{-1}$ and $\partial(v^\intercal\Sigma^{-1}v)/\partial\Sigma = -\Sigma^{-1}vv^\intercal\Sigma^{-1}$):

$$\nabla_{\Sigma_\ell}\mathbb{E}_{p(\mathbf{z}|\mathbf{x},\theta)}\left(\log p(\mathbf{x},\mathbf{z} \mid \theta)\right) = -\frac{1}{2}\sum_{n=1}^{N} r_{n\ell}\left(\Sigma_\ell^{-1}(x_n-\mu_\ell)(x_n-\mu_\ell)^\intercal\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\right)$$

$$\Rightarrow \quad \Sigma_\ell = \frac{1}{R_\ell}\sum_{n=1}^{N} r_{n\ell}(x_n-\mu_\ell)(x_n-\mu_\ell)^\intercal \qquad R_\ell := \sum_{n=1}^{N} r_{n\ell}$$

EM for Gaussian Mixtures

M-step derivations

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\theta)}\left(\log p(\boldsymbol{x},\boldsymbol{z}\mid\theta)\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_k + \log \mathcal{N}(x_n;\mu_k,\Sigma_k)\right)$$

To maximize w.r.t. π, enforce $\sum_k \pi_k = 1$ by introducing Lagrange multiplier λ and optimize

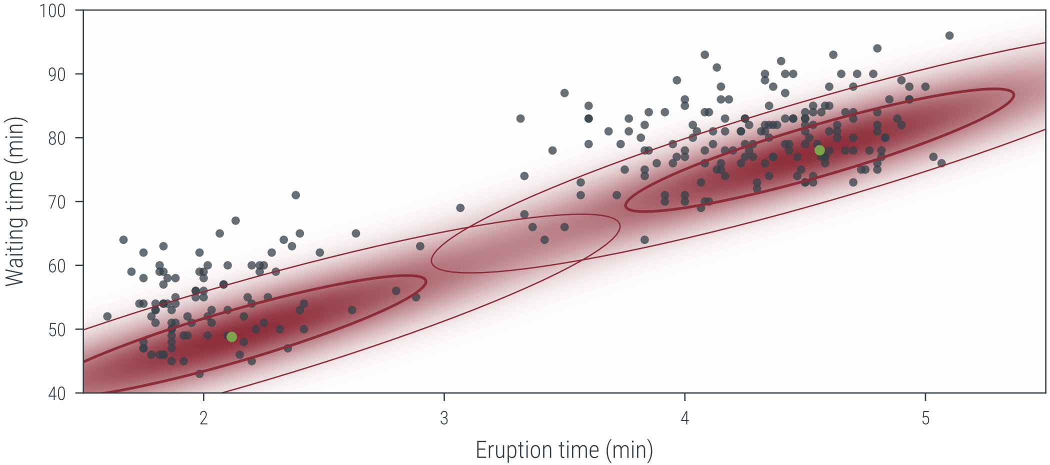
$$\nabla_{\pi_\ell}\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\theta)}\left(\log p(\boldsymbol{x},\boldsymbol{z}\mid\theta)\right) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) = \sum_{n=1}^{N} r_{n\ell}\frac{1}{\pi_\ell} + \lambda \overset{!}{=} 0$$

$$\pi_\ell = -\frac{1}{\lambda}\sum_{n=1}^{N} r_{n\ell} = -\frac{1}{\lambda}R_\ell$$

$$\sum_{k=1}^{K}\pi_k = 1 \Rightarrow \lambda = -N \qquad \text{and} \qquad \pi_\ell = \frac{R_\ell}{N}$$

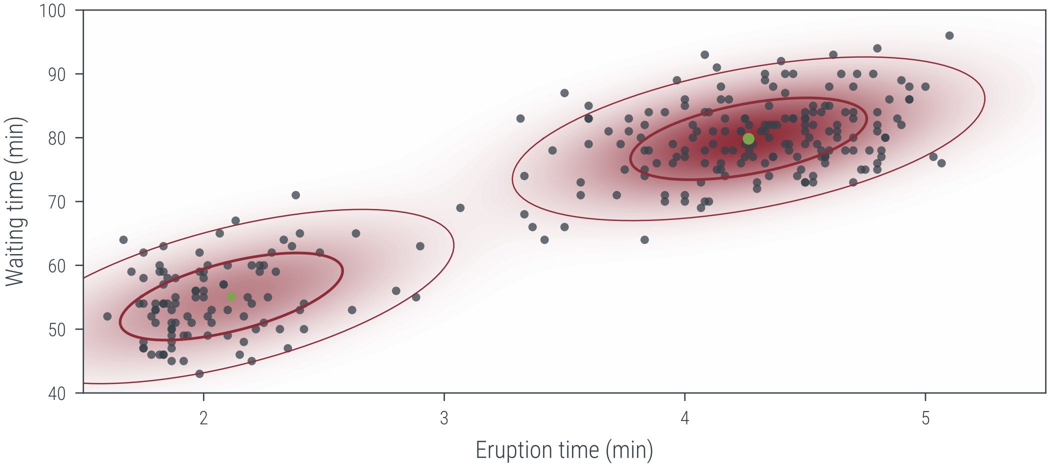
EM for Gaussian Mixtures

example run



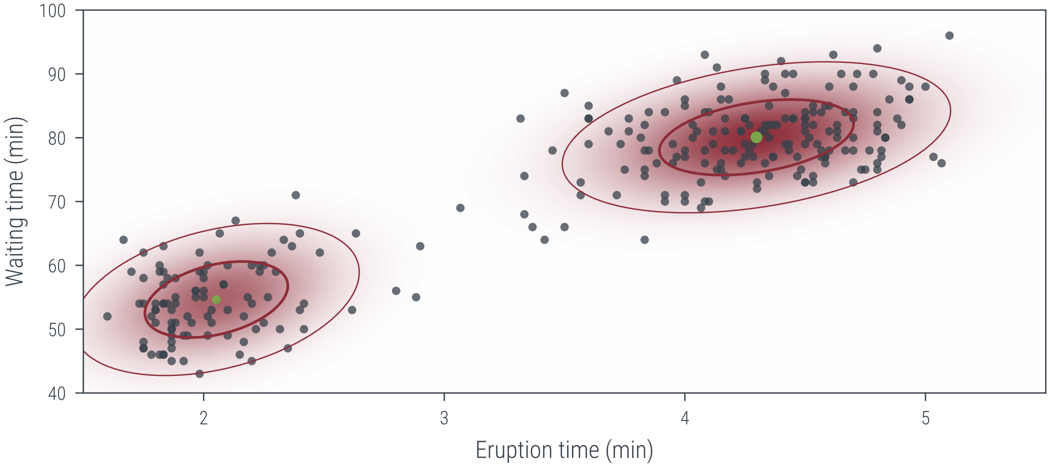
EM for Gaussian Mixtures

example run



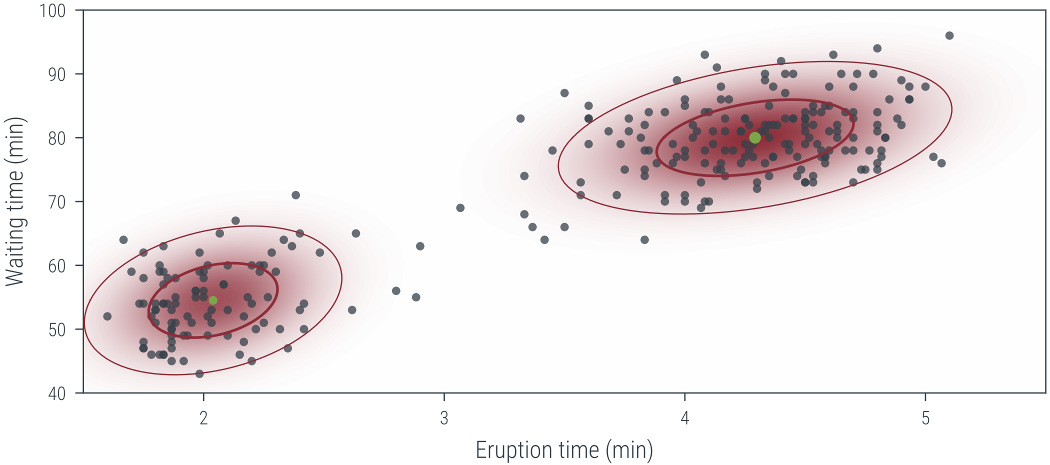
EM for Gaussian Mixtures

example run



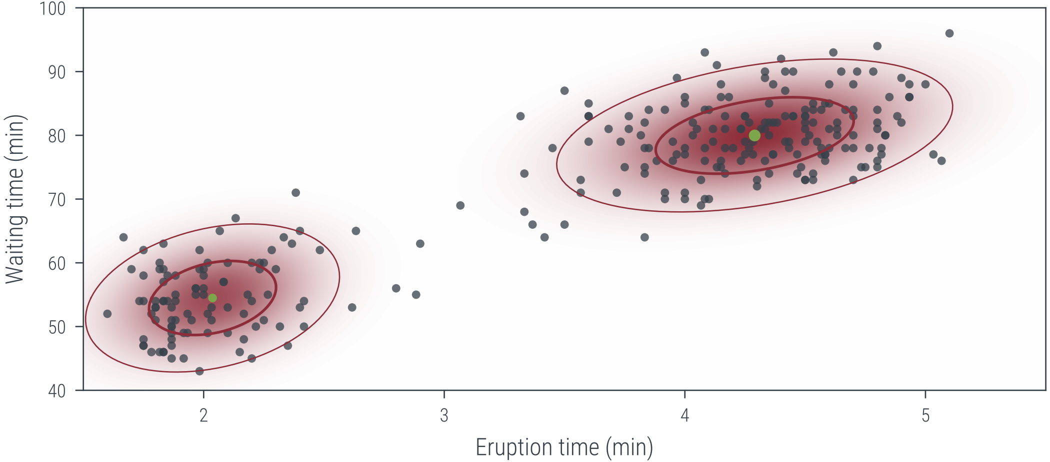
EM for Gaussian Mixtures

example run



EM for Gaussian Mixtures

example run



The EM algorithm:

- to find *maximum likelihood* (or MAP) estimate for a model involving a **latent** variable

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} [\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})] = \arg\max_{\boldsymbol{\theta}} \left[\log \left(\sum_{z} p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \right) \right]$$

- Initialize $\boldsymbol{\theta}_0$, then iterate (checking convergence of either the log likelihood, or $\boldsymbol{\theta}$)

E Compute $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}) = 0$

M Set $\boldsymbol{\theta}_{\text{new}}$ to the **Maximize** the **Expectation Lower Bound**

$$\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{z} q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{z})} \right)$$

Please cite this course, as

```
@techreport{Tuebingen_ProbML23,
    title =
    {Probabilistic Machine Learning},
    author = {Hennig, Philipp},
    series = {Lecture Notes
        in Machine Learning},
    year = {2023},
    institution = {Tübingen AI Center}}
```