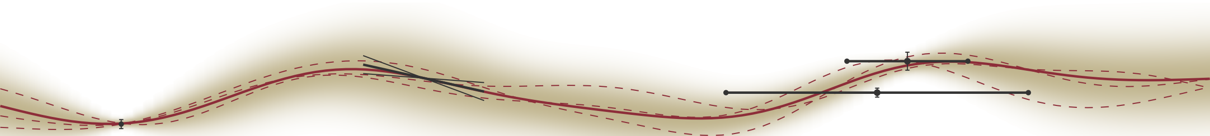
Probabilistic Machine Learning
Lecture 24
Variational Inference

Philipp Hennig

23 July 2023

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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



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

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

Algorithm: Initialize θ_0, then iterate:

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

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

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

EM for Gaussian Mixtures

introducing a class membership variable ***z*** factorizes the likelihood

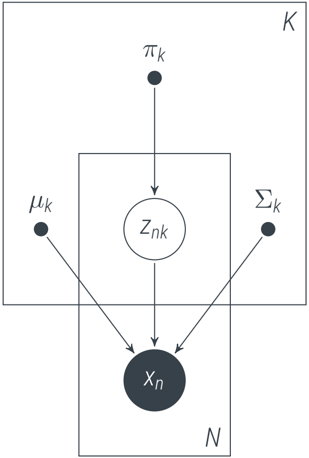
- Want to maximize, as function of $\theta := (\pi_k, \mu_k, \Sigma_k)_{k=1,\dots,K}$

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

- Instead, maximizing the “complete data” likelihood is easier:

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

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



EM for Gaussian Mixtures

EM updates for Exponential Family distributions are usually analytic

```python
K = 2; (N, D) = X.shape

pi = np.ones(K) / K # Initialize parameters
mu = multivariate_normal.rvs(mean=X.mean(axis=0), cov=np.cov(X, rowvar=False), size=K, random_state=1)
sigma = [np.cov(X, rowvar=False) / K for _ in range(K)]
r = np.zeros([N, K])

while True:
    # E-step
    for k in range(K):
        r[:, k] = pi[k] * multivariate_normal.pdf(X, mu[k, :], sigma[k])
    r /= np.sum(r, axis=1, keepdims=True) + 1e-9

    # M-step
    Nk = np.sum(r, axis=0)
    pi = Nk / N
    mu_old = mu.copy()
    mu = np.dot(r.T, X) / Nk[:, None]
    for k in range(K):
        Xc = X - mu[k]
        sigma[k] = np.dot(r[:, k] * Xc.T, Xc) / Nk[k]

    if np.linalg.norm(mu - mu_old) < 1e-3:
        break
```

- What if we can not compute the posterior $p(\boldsymbol{z} \mid \boldsymbol{x}, \theta)$ analytically?
- In Lecture 22 we saw that EM is iteratively optimizing a lower bound on the log likelihood

$$\log p(\boldsymbol{x}) = \underbrace{\int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} \right) \, d\boldsymbol{z}}_{\mathcal{L}(q)} - \underbrace{\int q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \right)}_{D_{\mathrm{KL}}(q \| p(\boldsymbol{z} | \boldsymbol{x}))}$$

  in which we set $D_{\mathrm{KL}}(q \| p(\boldsymbol{z} \mid \boldsymbol{x})) = 0$ at each iteration.
- More generally, we could consider any q and optimize the **Evidence Lower Bound** (ELBO) $\mathcal{L}(q)$
- Because $p(\boldsymbol{x})$ is a constant (assuming fixed model), maximizing the ELBO amounts to minimizing the KL divergence between q and the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$, and thus **finding a good approximation to the posterior**.
- In principle, this is an **optimization in the space of probability distributions** q. It is known as **Variational Inference**.

What's a good approximation?

optimization in function space

- in general, maximizing $\mathcal{L}(q)$ wrt. $q(\mathbf{z})$ is hard, because the extremum is exactly at $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$
- we could choose a **parametric** approximation $q(\mathbf{z} \mid \lambda)$, and optimize wrt. λ. This is restrictive, and does not guarantee success, because we still need to be able to evaluate the integrals

$$\mathcal{L}(q) = \int q(\mathbf{z}) \log \left(\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right) d\mathbf{z} \qquad \text{or} \qquad D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})) = \int q(\mathbf{z}) \log \left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right) d\mathbf{z}$$

- How does one find a *function* (even more, a probability distribution) that minimizes a *functional*?

Factorizing Approximations

A surprisingly subtle approximation with strong implications

- let's assume that $q(\mathbf{z})$ **factorizes**

$$q(\mathbf{z}) = \prod_i q_i(z_i)$$

- then the bound simplifies. Let's focus on one particular variable z_j:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i(z_i) \left(\log p(\mathbf{x}, \mathbf{z}) - \sum_i \log q_i(z_i) \right) \, d\mathbf{z} \\
&= \int q_j(z_j) \left(\int \log p(\mathbf{x}, \mathbf{z}) \prod_{i \neq j} q_i(z_i) \, dz_i \right) dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + \text{const.} \\
&= \int q_j(z_j) \log \tilde{p}(\mathbf{x}, z_j) \, dz_j - \int q_j(z_j) \log q_j(z_j) \, dz_j + \text{const.} \quad \text{where}
\end{aligned}$$

$$\log \tilde{p}(\mathbf{x}, z_j) := \mathbb{E}_{q, i \neq j}[\log p(\mathbf{x}, \mathbf{z})] + \text{const.}$$

this is a **function** of z_j, and a lower bound on $\log \tilde{p}(\mathbf{x}) = \log \int \tilde{p}(\mathbf{x}, z_j) \, dz_j$.

Mean Field Theory

Factorizing variational approximations

Consider a joint distribution $p(\boldsymbol{x}, \boldsymbol{z}) = \tilde{p}(\boldsymbol{x}, \boldsymbol{z}) \cdot \text{const}$. To find a "good" but tractable approximation $q(\boldsymbol{z})$,

- **assume** q **factorizes** $q(\boldsymbol{z}) = \prod_i q_i(\boldsymbol{z}_i)$. Initialize all q_i to some initial *distribution*.
- Iteratively compute

$$\mathcal{L}(q) = \int q_j \log \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)\, d\boldsymbol{z}_j - \int q_j \log q_j \, d\boldsymbol{z}_j + \text{const.} = -D_{\text{KL}}(q_j(\boldsymbol{z}) \| \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j)) + \text{const.}$$

and maximize wrt. q_j. Doing so *minimizes* $D_{\text{KL}}(q(\boldsymbol{z}_j) \| \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j))$, thus the minimum is at q_j^* with

$$\log q_j^*(\boldsymbol{z}_j) = \log \tilde{p}(\boldsymbol{x}, \boldsymbol{z}_j) = \mathbb{E}_{q, i \neq j}(\log p(\boldsymbol{x}, \boldsymbol{z})) + \text{const.} \qquad (\star)$$

this expression identifies a **function** q_j of $\boldsymbol{z}_j$, not some parametric form.

- the optimization converges, because $-\mathcal{L}(q)$ can be shown to be *convex* wrt. q.

In physics, this trick is known as **mean field theory** (because an n-body problem is separated into n separate problems of individual particles who are affected by the "mean field" $\tilde{p}$ summarizing the expected effect of all other particles).

Variational Inference

- ▶ is a general framework to construct approximating **probability distributions** $q(\boldsymbol{z})$ to non-analytic posterior distributions $p(\boldsymbol{z} \mid \boldsymbol{x})$ by minimizing the **functional**

$$q^* = \underset{q \in \mathcal{Q}}{\arg\min}\, D_{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) = \underset{q \in \mathcal{Q}}{\arg\max}\, \mathcal{L}(q)$$

- ▶ the beauty is that we get to *choose* q, so one can nearly always find a tractable approximation.
- ▶ If we impose the *mean field approximation* $q(\boldsymbol{z}) = \prod_i q(z_i)$, get

$$\log q_j^*(z_j) = \mathbb{E}_{q, i \neq j}(\log p(\boldsymbol{x}, \boldsymbol{z})) + \text{const.}.$$

- ▶ for Exponential Family p, we need the expectation under q of the sufficient statistics.

Variational Inference is an extremely flexible and powerful approximation method. Its downside is that constructing the bound and update equations can be tedious. For a quick test, variational inference is often not a good idea. But for a deployed product, it can be the most powerful tool in the box.

- **Variational Inference:** is a general framework to construct approximating **probability distributions** $q(z)$ to non-analytic posterior distributions $p(z \mid x)$ by minimizing the **functional**

$$q^* = \arg\min_{q \in \mathcal{Q}} D_{KL}(q(z) \| p(z \mid x)) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q)$$

- If we impose the *mean field approximation* $q(z) = \prod_i q(z_i)$, the updates reduce to a form of "coordinate ascent in distribution space"

$$\log q_j^*(z_j) = \mathbb{E}_{q, i \neq j}(\log p(x, z)) + \text{const.}$$

- in practice, the construction of the bound involves
  1. write down the log joint $\log p(x, z)$, decide on a factorization $q(z) = \prod_i q(z_i)$
  2. inspect the algebraic form of $\log q_j^*(z_j) = \mathbb{E}_{q, i \neq j}(\log p(x, z)) + \text{const.}$ and identify the distribution q_j
  3. once all q_j^* are identified by type, try to find analytic expressions for $\mathbb{E}_{q, i \neq j}(\log p(x, z))$.

Example: The Gaussian Mixture Model

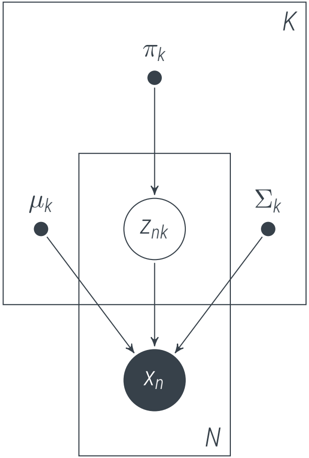
Returning to EM

Exposition after Bishop, PRML 2006, Chapter 10.2

- Remember EM for Gaussian mixtures $\theta := (\pi, \mu, \Sigma)$

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

$$= \prod_{n=1}^{N} p(z_{n:} \mid \pi) \cdot p(x_n \mid z_{n:}, \mu, \Sigma)$$



Example: The Gaussian Mixture Model

Returning to EM

Exposition after Bishop, PRML 2006, Chapter 10.2

- Remember EM for Gaussian mixtures $\theta := (\pi, \mu, \Sigma)$

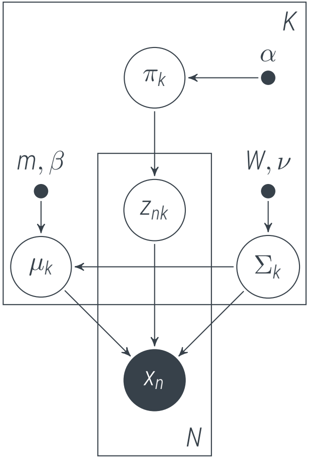
$$p(x, z \mid \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \mathcal{N}(x_n; \mu_k, \Sigma_k)^{z_{nk}}$$

- For Bayesian inference, turn parameters into variables

$$p(x, z, \pi, \mu, \Sigma) = p(x, z \mid \pi, \mu, \Sigma) \cdot p(\pi) \cdot p(\mu \mid \Sigma) \cdot p(\Sigma)$$

$$p(\pi) = \mathcal{D}(\pi \mid \alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

$$p(\mu \mid \Sigma) \cdot p(\Sigma) = \prod_{k=1}^{K} \mathcal{N}(\mu_k; m, \Sigma_k / \beta) \cdot \mathcal{W}(\Sigma_k^{-1}; W, \nu)$$



Constructing the Variational Approximation

Example: The Gaussian Mixture Model

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- We know that the full posterior $p(z, \pi, \mu, \Sigma \mid x)$ is intractable (check the graph!)
- But let's consider an approximation $q(z, \pi, \mu, \Sigma)$ with the factorization

$$q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$$

- from $(\star)$, we have

$$\begin{aligned}
\log q^*(z) &= \mathbb{E}_{q(\pi,\mu,\Sigma)}\left(\log p(x, z, \pi, \mu, \Sigma)\right) + \text{const.} \\
&= \mathbb{E}_{q(\pi)}\left(\log p(z \mid \pi)\right) + \mathbb{E}_{q(\mu,\Sigma)}\left(\log p(x \mid z, \mu, \Sigma)\right) + \text{const.} \\
&= \sum_n^N \sum_k^K z_{nk} \underbrace{\left(\mathbb{E}_{q(\pi)}(\log \pi_k) + \frac{1}{2}\mathbb{E}_{q(\mu,\Sigma)}(\log|\Sigma^{-1}| - (x_n - \mu_k)^\intercal \Sigma_k^{-1}(x - \mu_k))\right)}_{=:\log \rho_{nk}} + \text{const.}
\end{aligned}$$

$q^*(z) \propto \prod_n \prod_k \rho_{nk}^{z_{nk}}$ define $r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}}$, then $q^*(z) = \prod_n \prod_k r_{nk}^{z_{nk}}$ with $\mathbb{E}_{q(z)}[z] = r_{nk}$

Look through the Math!

Identifying the approximation constructed

- a **discrete** or **multinomial** distribution is defined by

$$q(z) = \prod_{k=1}^{K} \rho_k^{z_k} \quad \text{for } z_k \in \{0,1\}, \sum_k z_k = 1$$

It has expectations values $\mathbb{E}[z_k] = \rho_k$ and $\mathbb{E}[z_k z_j] = \rho_k \delta_{kj}$

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- Define some convenient notation:

$$N_k := \sum_{n=1}^{N} r_{nk} \qquad \bar{x}_k := \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \qquad S_k := \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^\intercal$$

- from $(\star)$, we have

$$\begin{aligned}
\log q^*(\pi, \mu, \Sigma) &= \mathbb{E}_{q(z)} \left(\log p(x, z, \pi, \mu, \Sigma)\right) + \text{const.} \\
&= \mathbb{E}_{q(z)} \left(\log p(\pi) + \sum_k \log p(\mu_k, \Sigma_k) + \log p(z \mid \pi) + \sum_n \log p(x_n \mid z, \mu, \Sigma) \right) \\
&= \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)} (\log p(z \mid \pi)) \\
&\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}
\end{aligned}$$

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

$$\log q^*(\pi, \mu, \Sigma) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)}(\log p(z \mid \pi))$$

$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}$$

- The bound exposes an **induced factorization** into $q(\pi, \mu, \Sigma) = q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Sigma_k)$

$$\text{where } \log q(\pi) = \log p(\pi) + \mathbb{E}_{q(z)}(\log p(z \mid \pi)) + \text{const.}$$

$$= (\alpha - 1) \sum_{k} \log \pi_k + \sum_{k} \sum_{n} r_{nk} \log \pi_k + \text{const.}$$

$$q(\pi) = \mathcal{D}(\pi, \alpha_k := \alpha + N_k) \quad \text{with } N_k = \sum_{n} r_{nk}$$

Look through the Math!

Identifying the approximation constructed

- a **Dirichlet** distribution is defined by

$$p(\pi \mid \alpha) = \mathcal{D}(\pi; \alpha) = \frac{\Gamma(\hat{\alpha})}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_k \pi^{\alpha_k - 1} = \frac{1}{B(\alpha)} \prod_k \pi^{\alpha_k - 1} \qquad \hat{\alpha} := \sum_d \alpha_k$$

It satisfies the following identities:

- $\mathbb{E}_p(\pi_k) = \frac{\alpha_k}{\hat{\alpha}}$
- $\mathrm{var}_p(\pi_k) = \frac{\alpha_k(\hat{\alpha}-\alpha_k)}{\hat{\alpha}^2(\hat{\alpha}+1)}$
- $\mathrm{cov}(\pi_k, \pi_i) = -\frac{\alpha_k \alpha_i}{\hat{\alpha}^2(\hat{\alpha}+1)}$
- $\mathrm{mode}(\pi_k) = \frac{\alpha_k - 1}{\hat{\alpha} - D}$
- $\mathbb{E}_p(\log \pi_k) = \digamma(\alpha_k) - \digamma(\hat{\alpha})$
- $\mathbb{H}(p) = -\int p(\pi) \log p(\pi)\, d\pi = -\sum_k (\alpha_k - 1)(\digamma(\alpha_k) - \digamma(\hat{\alpha})) + \log B(\alpha)$

Where $\digamma(z) = \frac{d}{dz} \log \Gamma(z)$ (the “digamma-function”).

`scipy.special.digamma(z)`   `https://dlmf.nist.gov/5`

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

$$\log q^*(\pi, \mu, \Sigma) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)}(\log p(z \mid \pi))$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}$$

- The bound exposes an **induced factorization** into $q(\pi, \mu, \Sigma) = q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Sigma_k)$

where (leaving out some tedious algebra) $q^*(\mu_k, \Sigma_k) = \mathcal{N}(\mu_k; m_k, \Sigma_k/\beta_k)\mathcal{W}(\Sigma_k^{-1}; W_k, \nu_k)$

with $\beta_k := \beta + N_k \qquad m_k := \frac{1}{\beta_k}(\beta m + N_k \bar{x}_k) \qquad \nu_k := \nu + N_k$

$$W_k^{-1} := W^{-1} + N_k S_k + \frac{\beta N_k}{\beta + N_k}(\bar{x}_k - m)(\bar{x}_k - m)^\intercal$$

Constructing the Variational Approximation

closing the loop

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- Recall from above:

$$\log q^*(z) = \sum_n^N \sum_k^K z_{nk} \underbrace{\left(\mathbb{E}_{q(\pi)}(\log \pi_k) + \frac{1}{2}\mathbb{E}_{q(\mu,\Sigma)}(\log |\Sigma^{-1}| - (x_n - \mu_k)^\intercal \Sigma_k^{-1}(x - \mu_k)) \right)}_{=:\log \rho_{nk}} + \text{const.}$$

- now we can evaluate ρ_{nk}, using tabulated identities

$$\log \tilde{\pi}_k := \mathbb{E}_{\mathcal{D}(\pi;\alpha_k)}(\log \pi_k) = F(\alpha_k) - F\left(\sum_k \alpha_k\right)$$

and for the Wishart:

$$\log |\tilde{\Sigma^{-1}}|_k := \mathbb{E}_{\mathcal{W}(\Sigma_k^{-1};W_k,\nu_k)}(\log |\Sigma_k^{-1}|) = \sum_{d=1}^{D} F\left(\nu_k + \frac{1-d}{2}\right) + D \log 2 + \log |W_k|$$

$$\mathbb{E}_{\mathcal{N}(\mu_k;m_k,\Sigma_k/\beta_k)\mathcal{W}(\Sigma_k^{-1};W_k,\nu_k)}\left((x_n - \mu_k)^\intercal \Sigma^{-1}(x_n - \mu_k)\right) = D\beta_k^{-1} + \nu_k(x_n - m_k)^\intercal W_k(x_n - m_k)$$

Constructing the Variational Approximation

connection to EM

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- this yields the update equation

$$\mathbb{E}_q(z_{nk}) = r_{nk} \propto \tilde{\pi}_k |\widetilde{\Sigma^{-1}}|^{1/2} \exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^\intercal W_k (x_n - m_k)\right)$$

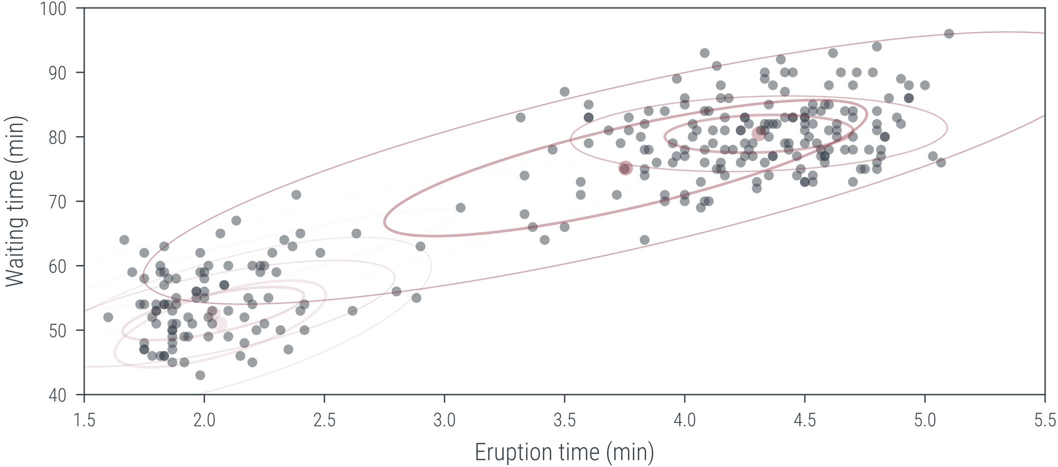
  compare this with the EM-update

$$r_{nk} \propto \pi_k |\Sigma^{-1}|^{1/2} \exp\left(-\frac{1}{2}(x_n - \mu_k)^\intercal \Sigma_k^{-1} (x_n - \mu_k)\right)$$

- Here, variational Inference is the Bayesian version of EM: Instead of maximizing the likelihood for $\theta = (\mu, \Sigma, \pi)$, we maximize a variational bound.
- One advantage of this is that the posterior can actually "decide" to ignore components:

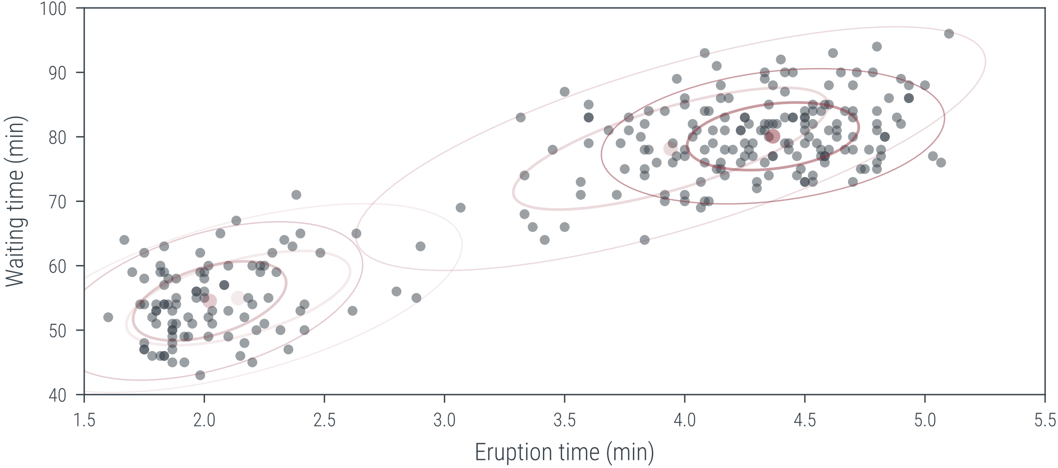
Variational Gaussian Mixture Models

initialized with $K = 6$ components



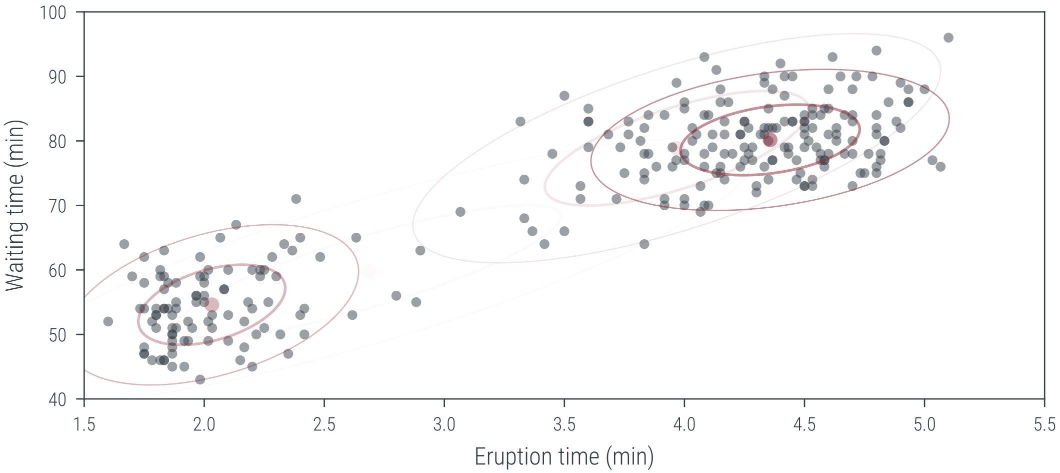
Variational Gaussian Mixture Models

initialized with $K = 6$ components



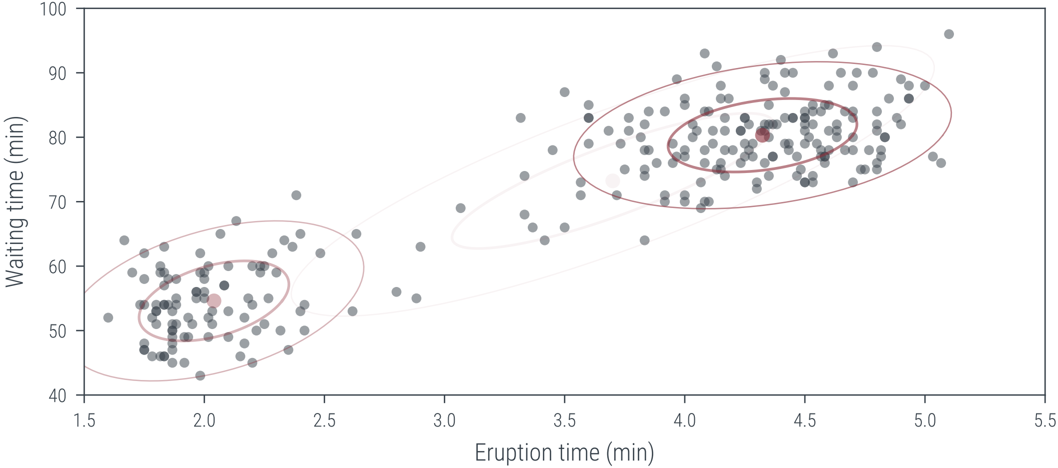
Variational Gaussian Mixture Models

initialized with $K = 6$ components



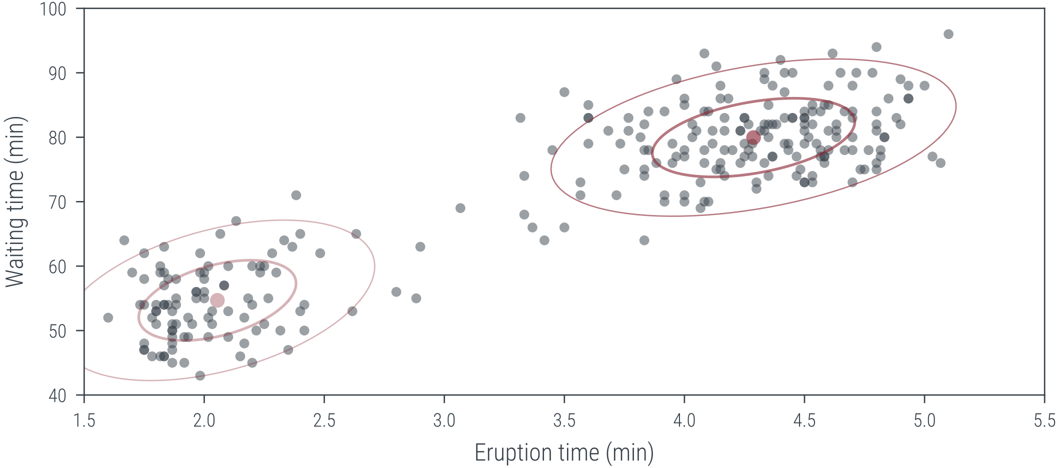
Variational Gaussian Mixture Models

initialized with $K = 6$ components



Variational Gaussian Mixture Models

initialized with $K = 6$ components



- **Variational Inference:** is a general framework to construct approximating **probability distributions** $q(z)$ to non-analytic posterior distributions $p(z \mid x)$ by minimizing the **functional**

$$q^* = \underset{q \in \mathcal{Q}}{\arg\min}\, D_{KL}(q(z) \| p(z \mid x)) = \underset{q \in \mathcal{Q}}{\arg\max}\, \mathcal{L}(q)$$

- If we impose the *mean field approximation* $q(z) = \prod_i q(z_i)$, the updates reduce to a form of "coordinate ascent in distribution space"

$$\log q_j^*(z_j) = \mathbb{E}_{q,i \neq j}(\log p(x,z)) + \text{const.}$$

- in practice, the construction of the bound involves
  1. write down the log joint $\log p(x,z)$, decide on a factorization $q(z) = \prod_i q(z_i)$
  2. inspect the algebraic form of $\log q_j^*(z_j) = \mathbb{E}_{q,i \neq j}(\log p(x,z)) + \text{const.}$ and identify the distribution q_j
  3. once all q_j^* are identified by type, try to find analytic expressions for $\mathbb{E}_{q,i \neq j}(\log p(x,z))$.

Variational Inference is a powerful mathematical tool to construct efficient approximations to intractable *probability distributions* (not just point estimates, but entire distributions). Often, just imposing factorization is enough to make things tractable. The downside of variational inference is that constructing the bound can take significant ELBOw grease. However, the resulting algorithms are often highly efficient compared to tools that require less derivation work, like Monte Carlo.

> “Derive the variational bound while you wait for MCMC to converge.”

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}}
```