02477 – Bayesian Machine Learning: Lecture 10

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

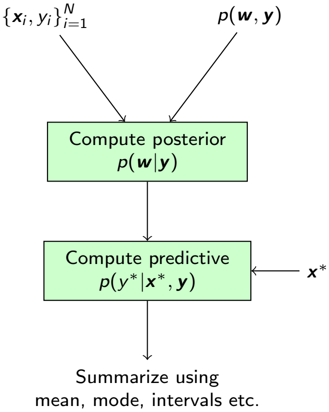
Outline

1. Wrap up: Monte Carlo and sampling methods
2. Variational inference
   - Basic concepts
   - Factorized approximations and CAVI
3. Mixture models

Wrap up: Monte Carlo and sampling methods

Probabilistic machine learning

Probabilistic machine learning = data + model + inference algorithm



$\{\boldsymbol{x}_i, y_i\}_{i=1}^N$

$p(\boldsymbol{w}, \boldsymbol{y})$

Compute posterior
$p(\boldsymbol{w}|\boldsymbol{y})$

Compute predictive
$p(y^*|\boldsymbol{x}^*, \boldsymbol{y})$

$\boldsymbol{x}^*$

Summarize using
mean, mode, intervals etc.

Distributions as Lego blocks: Beta, Binomial, Gaussians, Gamma, Poisson, Categorical, Gaussian processes, Uniform, Dirichlet, ...

Overview of sampling methods

- Ancestral sampling
  - Sampling from joint distributions $p(x, z, w) = p(x|z)p(z|w)p(w)$ or marginals, e.g. $p(x)$, if we can sample from all the conditional distributions
  - Useful for sampling-based posterior inference, e.g. $p(x, z, w|\mathcal{D}) = p(x|z)p(z|w)p(w|\mathcal{D})$

- Gibbs Sampling (MCMC)
  - Usually the first choice for conditionally conjugate models
  - No tuning parameters, acceptance ratio is always 1, but can be very slow for highly correlated distributions
  - Requires model-specific derivations

- Metropolis-Hastings (MCMC)
  - Does not require conjugacy
  - Very flexible
  - Often requires tuning to work well

- Other techniques: Rejection sampling, importance sampling

- Hamiltonian Monte Carlo (MCMC)
  - MH-algorithm that uses Hamiltonian dynamics based on gradient information to generate new proposals
  - Much faster than 'simple' MH
  - Only applies to continuous parameters, where gradients are available

- Probabilistic programming tools supporting HMC/MCMC
  - Stan, PyMC3, Tensorflow Probability, Pyro, BlackJax

The HMC algorithm in a nutshell

- Goal: Generating samples from target distribution $p(\boldsymbol{\theta}) = \frac{1}{Z}\tilde{p}(\boldsymbol{\theta})$
- We augment the target distribution: $p(\boldsymbol{\theta}, \boldsymbol{\nu}) \propto \tilde{p}(\boldsymbol{\theta})\mathcal{N}(\boldsymbol{\nu}|\mathbf{0}, \boldsymbol{\Sigma})$
- Recipe for generating the $k+1$'the sample using HMC
  1. Initialize $\boldsymbol{\theta}'_0 = \boldsymbol{\theta}_k$ and $\boldsymbol{\nu}'_0 \sim \mathcal{N}(\boldsymbol{\nu}|\mathbf{0}, \boldsymbol{\Sigma})$
  2. For $\ell = 1, \ldots, L$

$$\boldsymbol{\nu}'_\ell = \boldsymbol{\nu}'_{\ell-1} + \eta\nabla \log \tilde{p}(\boldsymbol{\theta}_{\ell-1})$$
$$\boldsymbol{\theta}'_\ell = \boldsymbol{\theta}'_{\ell-1} + \eta\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}'_\ell$$

  3. Set $\boldsymbol{\theta}^* = \boldsymbol{\theta}'_L$ and $\boldsymbol{\nu}^* = \boldsymbol{\nu}'_L$
  4. Compute acceptance probability

$$A_{k+1} = \min\left(1, \frac{p(\boldsymbol{\theta}^*, \nu^*)}{p(\boldsymbol{\theta}_k, \nu_k)}\right)$$

  5. Accept proposal $\boldsymbol{\theta}^*$ with probability A_{k+1}, keep $\boldsymbol{\theta}_{k-1}$ otherwise
- Step-size η, covariance $\boldsymbol{\Sigma}$, and the number of integration steps L are parameters of the algorithm.
- Gradient information allows HMC to explore the target distribution much more efficiently
- HMC NUTS (No U-turn Sampler) is state-of-the-art
- Cool visualization of HMC: https://chi-feng.github.io/mcmc-demo/app.html

Inference methods

- Maximum likelihood
  1. Fast and often easy
  2. Prone to overfitting, no model uncertainty, not necessarily well-defined

- Exact Bayesian inference
  1. Extremely limited in choice of models (linear models, conjugate models etc)
  2. Very fast

- Laplace approximations
  1. Very simple to implement and very fast
  2. Limited to continuous distributions
  3. Works well when the exact posterior is close to Gaussian
  4. Can fail horrible for asymmetrical and skewed distributions

- Markov Chain Monte Carlo (MCMC)
  1. Very strong mathematical guarantees, asymptotically exact and Very flexible
  2. Might take forever to converge (literally)
  3. First choice when we have smaller or moderate sized datasets and/or when accuracy and/or uncertainty are prioritized

- Variational inference (VI)
  1. Applies to both continuous and discrete distributions
  2. Can be much faster than MCMC, but without any strict guarantees
  3. Control accuracy vs speed trade-off
  4. Useful for testing different models for very large data sets
  5. Foundation many Bayesian deep learning methods and Variational autoencoders (VAEs)

Variational inference

Variational inference: Basic concepts

Variational inference: big picture

- Our goal is to approximate a posterior distribution of interest

$$p \equiv p(\boldsymbol{z}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{z})p(\boldsymbol{z})}{p(\mathcal{D})}$$

- Variational inference in three steps

  1. Define collection of "simple" approximate probability distributions $\mathcal{Q}$ (*the variational family*)
  2. Define a measure of "distance" between probability distributions $\mathbb{D}\left[q||p\right]$ (*the divergence*)
  3. Search for the distribution $q \in \mathcal{Q}$ that resembles the exact posterior p as close as possible as measured by $\mathbb{D}\left[q||p\right]$ (*optimization*)

- The variational approximation q for the target distribution $p \approx q$ is defined as

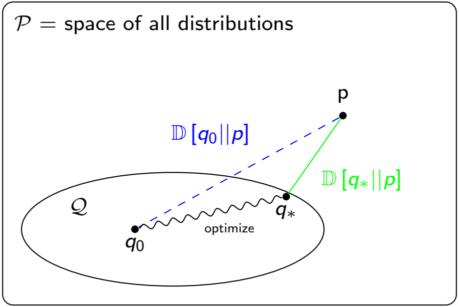
$$q_* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}\left[q||p\right]$$

Variational inference: big picture

- The variational approximation q for target distribution $p \approx q$ is defined as

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}\left[q || p\right]$$

where $\mathcal{Q}$ is a collection of "simple" approximate probability distributions



Example (Bishop p. 464)

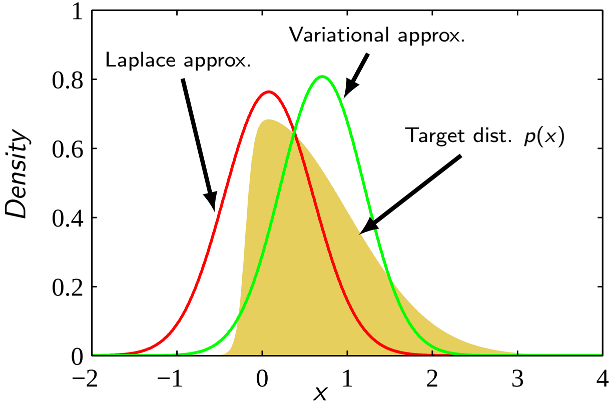
Goal: approximate some target distribution of interest $p(x)$ for $x \in \mathbb{R}$ (yellow).

1. We choose the variational family to be all the Gaussian densities

$$\mathcal{Q} = \{\mathcal{N}(\mu, \sigma^2)\,|\mu \in \mathbb{R}, \sigma^2 > 0\}$$

2. We choose some divergence measure $\mathbb{D}\left[q||p\right]$ and minimize it wrt. q

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathbb{D}\left[q||p\right]$$



The variational family $\mathcal{Q}$

- The *variational family* $\mathcal{Q}$ defines the collection of all possible approximations $q \in \mathcal{Q}$

- $\mathcal{Q}$ is chosen as a compromise between speed/tractability and approximation quality. The larger $\mathcal{Q}$, the smaller approximation error and vice versa.

- Examples of common variational families for $\boldsymbol{z} = [z_1, z_2, \ldots, z_D] \in \mathbb{R}^D$
  - *Full-rank Gaussians*

$$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{m}, \boldsymbol{V})$$

  - *Mean-field Gaussians*

$$q(\boldsymbol{z}) = \prod_{i=1}^{D} \mathcal{N}(z_i|m_i, v_i)$$

  - *Mean-field approximations*

$$q(\boldsymbol{z}) = \prod_{i=1}^{D} q(z_i)$$

Example: $z \in \mathbb{R}^5$: $\quad q(z_1, z_2, z_3, z_4, z_5) = q(z_1)q(z_2)q(z_3)q(z_4)q(z_5)$

  - *Factorized approximations*

$$q(\boldsymbol{z}) = \prod_{j=1}^{J} q(\boldsymbol{z}_j), \quad \text{where} \quad \boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_J] \text{ for } J < D$$

Example: $z \in \mathbb{R}^5$: $\quad q(z_1, z_2, z_3, z_4, z_5) = q(z_1, z_2)q(z_3)q(z_4, z_5)$

Measuring distance between probability distributions

- How do we measure "distance" between probability distributions $\mathbb{D}\left[q, p\right]$?
- The choice of "distance" affects the properties of the approximation
- The *Kullback-Leibler* divergence (for continuous R.V.) is defined as

$$\mathrm{KL}\left[q||p\right] = \int q(\boldsymbol{z}) \ln \left[\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right] \mathrm{d}\boldsymbol{z}$$

- Properties

1. Identity of indiscernibles

$$\mathrm{KL}\left[q||p\right] = 0 \qquad \iff \qquad p = q \quad \text{(a.e)}$$

2. Non-negativity

$$\mathrm{KL}\left[q||p\right] \geq 0$$

3. Asymmetric

$$\mathrm{KL}\left[q||p\right] \neq \mathrm{KL}\left[p||q\right]$$

4. Does not satisfy the triangle inequality, i.e. $\mathrm{KL}\left[q||p\right] \leq \mathrm{KL}\left[q||r\right] + \mathrm{KL}\left[r||p\right]$ does **not** hold in general.

Quiz

Quiz time!

Week 10: Variational inference

Minimizing the KL-divergence

The *variational approximation* q for target distribution $p(\boldsymbol{z}|\mathcal{D}) \approx q$ is defined as

$$q_* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left[q||p\right], \qquad \mathrm{KL}\left[q||p\right] = \int q(\boldsymbol{z}) \ln\left[\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right] \mathrm{d}\boldsymbol{z}$$

- Re-writing the KL-divergence for target posterior distribution $p \equiv p(\boldsymbol{z}|\mathcal{D})$

$$\begin{aligned}
\mathrm{KL}\left[q||p\right] &= \int q(\boldsymbol{z}) \ln\left(\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\mathcal{D})}\right) \mathrm{d}\boldsymbol{z} && \text{(Definition of KL)} \\
&= \mathbb{E}_q\left[\ln \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\mathcal{D})}\right] && \text{(Def. of expectation)} \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{z}|\mathcal{D})\right] && \text{(Linearity of expectations)} \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] - \mathbb{E}_q\left[\ln \frac{p(\mathcal{D}, \boldsymbol{z})}{p(\mathcal{D})}\right] && \text{(Using Bayes' theorem)} \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] - \mathbb{E}_q\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] + \mathbb{E}_q\left[\ln p(\mathcal{D})\right] && \text{(Linearity of expectations)} \\
&= \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] - \mathbb{E}_q\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] + \ln p(\mathcal{D}) && \text{($\mathcal{D}$ is independent of q)}
\end{aligned}$$

Re-arranging

$$\ln p(\mathcal{D}) = \mathbb{E}_q\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] - \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] + \mathrm{KL}\left[q||p\right]$$

The evidence lower bound

- We just derived

$$\begin{aligned}\ln p(\mathcal{D}) &= \mathbb{E}_q\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] - \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right] + \mathrm{KL}\left[q||p\right] \\ &= \mathcal{L}\left[q\right] + \mathrm{KL}\left[q||p\right] \\ &\geq \mathcal{L}\left[q\right]\end{aligned}$$

- We define the *evidence lower bound* (ELBO) as

$$\mathcal{L}\left[q\right] \equiv \mathbb{E}_q\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] - \mathbb{E}_q\left[\ln q(\boldsymbol{z})\right]$$

- Observations

  1. Maximizing $\mathcal{L}$ is equivalent to minimizing $\mathrm{KL}\left[q||p\right]$ since $\mathrm{KL}\left[q||p\right] \geq 0$ and $\ln p(\mathcal{D})$ is const.

  2. We only need to be able to evaluate log joint distribution $p(\boldsymbol{z}, \mathcal{D})$, not the posterior $p(\boldsymbol{z}|\mathcal{D})$

  3. The ELBO $\mathcal{L}[q]$ is a lower bound on the marginal likelihood, i.e.

$$\mathcal{L}\left[q\right] \leq \ln p(\mathcal{D})$$

- Key take away

$$q^* = \arg\min_{q\in\mathcal{Q}} \mathrm{KL}\left[q||p\right] = \arg\max_{q\in\mathcal{Q}} \mathcal{L}\left[q\right]$$

Variational inference: Factorized approximations and CAVI

Minimizing the KL-divergence for factorized distributions I

- Factorized approximations: $q(\boldsymbol{z}) = \prod_{j=1}^{J} q(\boldsymbol{z}_j)$ where $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_J]$

  Example: If $\boldsymbol{z} = [z_1, z_2, z_3, z_4, z_5]$, then we may assume $q(\boldsymbol{z}) = q(z_1)q(z_2, z_3)q(z_4, z_5)$ $(J = 3)$

- The factorization can be in groups of variables or across all variables (mean-field)

- We make *no assumptions on the functional form for each factor*. Hence, often called *free-form* variational inference

- How to identify $q(\boldsymbol{z}_j)$ for $j = 1, \ldots, J$ for a given posterior?

- Minimize KL by substituting the approximation into the ELBO

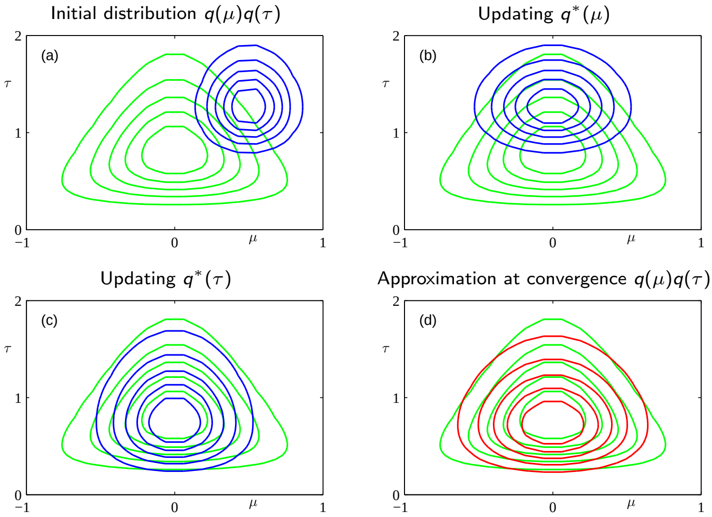
$$\mathcal{L}[q] \equiv \mathbb{E}_q[\ln p(\mathcal{D}, \boldsymbol{z})] - \mathbb{E}_q[\ln q(\boldsymbol{z})] = \int \prod_{j=1}^{J} q(\boldsymbol{z}_j) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int \prod_{j=1}^{J} q(\boldsymbol{z}_j) \ln \prod_{j=1}^{J} q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z}$$

- Optimization strategy: *Coordinate ascent variational inference* (CAVI)

  1. We iterate through all factors, updating one at a time. Starting with the k'th factor
  2. We'll identify all terms that depend on $q(\boldsymbol{z}_k)$ and use that to optimize $\mathcal{L}$.
  3. Repeat for all k and iterate until convergence

Minimizing the KL-divergence for factorized distributions I: Example

Target posterior distribution: $p(\mu, \tau|\mathcal{D})$ (Example from Section 10.1.3 in Bishop)



Initial distribution $q(\mu)q(\tau)$

Updating $q^*(\mu)$

Updating $q^*(\tau)$

Approximation at convergence $q(\mu)q(\tau)$

Minimizing the KL-divergence for factorized distributions IIa

We want to maximize $\mathcal{L}$ wrt. $q(\boldsymbol{z}_k)$

$$\mathcal{L}\left[q\right] = \int \prod_{i=1}^{J} q(\boldsymbol{z}_i) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int \prod_{i=1}^{J} q(\boldsymbol{z}_i) \ln \prod_{j=1}^{J} q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z}$$

Identifying part of the second term that depends on $q(\boldsymbol{z}_k)$

$$\begin{aligned}
\int \prod_{i=1}^{J} q(\boldsymbol{z}_i) \ln \prod_{j=1}^{J} q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z} &= \int \prod_{j=1}^{J} q(\boldsymbol{z}_j) \sum_{j=1}^{J} \ln q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z} && \text{(From product to sums)} \\
&= \sum_{i=1}^{J} \int \prod_{i=1}^{J} q(\boldsymbol{z}_i) \ln q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z} && \text{(Linearity of integrals)} \\
&= \sum_{j=1}^{J} \int q(\boldsymbol{z}_1) q(\boldsymbol{z}_2) \ldots q(\boldsymbol{z}_J) \ln q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z} && \text{(Expand product)} \\
&= \sum_{j=1}^{J} \int q(\boldsymbol{z}_j) \ln q(\boldsymbol{z}_j) \mathrm{d}\boldsymbol{z}_{\boldsymbol{j}} && \text{(Marginalize)} \\
&= \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_{\boldsymbol{k}} + \text{const} && \text{(Dependency on } q(\boldsymbol{z}_k)\text{)}
\end{aligned}$$

Therefore, we can write

$$\mathcal{L}\left[q\right] = \int \prod_{j=1}^{J} q(\boldsymbol{z}_j) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const}$$

Minimizing the KL-divergence for factorized distributions IIb

Simplifying the first term

$$
\begin{aligned}
\mathcal{L}\left[q\right] &= \int \prod_{i=1}^{J} q(\boldsymbol{z}_i) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const} \\
&= \int q(\boldsymbol{z}_1) q(\boldsymbol{z}_2) \ldots q(\boldsymbol{z}_J) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const} && \text{(expand product)} \\
&= \int q(\boldsymbol{z}_k) \left[\int \prod_{i \neq k} q(\boldsymbol{z}_i) \ln p(\mathcal{D}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}_{-k} \right] \mathrm{d}\boldsymbol{z}_k - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const} && \text{(Factor out } q(\boldsymbol{z}_k)\text{)} \\
&= \int q(\boldsymbol{z}_k) \underbrace{\mathbb{E}_{i \neq k} \left[\ln p(\mathcal{D}, \boldsymbol{z})\right]}_{\ln \tilde{p}(\mathcal{D}, \boldsymbol{z}_k)} \mathrm{d}\boldsymbol{z}_k - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const} && \text{(Define } \tilde{p}(\mathcal{D}, \boldsymbol{z}_k)\text{)} \\
&= \int q(\boldsymbol{z}_k) \ln \tilde{p}(\mathcal{D}, \boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k - \int q(\boldsymbol{z}_k) \ln q(\boldsymbol{z}_k) \mathrm{d}\boldsymbol{z}_k + \text{const} && \text{(Use def. of } \tilde{p}\text{)} \\
&= \int q(\boldsymbol{z}_k) \ln \frac{\tilde{p}(\mathcal{D}, \boldsymbol{z}_k)}{q(\boldsymbol{z}_k)} \mathrm{d}\boldsymbol{z}_k + \text{const} && \text{(Linearity of integrals)} \\
&= -\text{KL}\left[q(\boldsymbol{z}_k) || \tilde{p}(\mathcal{D}, \boldsymbol{z}_k)\right] + \text{const} && \text{(Def. of KL)}
\end{aligned}
$$

- Summary: When we consider the ELBO as a function of $q(\boldsymbol{z}_k)$ only, it is equal to the KL-divergence between $q(\boldsymbol{z}_k)$ and $\tilde{p}(\mathcal{D}, \boldsymbol{z}_k)$. When are KL-divergences minimized?

Minimizing the KL-divergence for factorized distributions III

- Goal: We want to minimize the KL-divergence $\mathrm{KL}\left[q||p\right]$ between our approximation q and our target p by maximizing the ELBO $\mathcal{L}$ iterative one factor $q(\boldsymbol{z}_k)$ at time

- We just showed that when we consider $\mathcal{L}$ as a function of $q(\boldsymbol{z}_k)$, then

$$\mathcal{L}\left[q\right] = -\mathrm{KL}\left[q(\boldsymbol{z}_k)||\tilde{p}(\mathcal{D}, \boldsymbol{z}_k)\right] + k$$

- The KL divergence is minimized wrt. $\boldsymbol{z}_k$ when $q(\boldsymbol{z}_k) = \tilde{p}(\mathcal{D}, \boldsymbol{z}_k)$.

- The *optimal choice* for the factor $q(\boldsymbol{z}_k)$ is

$$\ln q^*(\boldsymbol{z}_k) = \ln \tilde{p}(\mathcal{D}, \boldsymbol{z}_k) = \mathbb{E}_{i \neq k}\left[\ln p(\mathcal{D}, \boldsymbol{z})\right] + K$$

- In words: The optimal distribution for $\ln q^*(\boldsymbol{z}_k)$ is obtained by taking the log joint distribution $\ln p(\mathcal{D}, \boldsymbol{z})$ and averaged it wrt. all the other factors, i.e. $q(\boldsymbol{z}_j)$ for $j \neq k$

Coordinate Ascent Variational Inference (CAVI)

Big picture

- We are given a joint distribution for a dataset $\mathcal{D}$ and parameters $\boldsymbol{w} \in \mathbb{R}^D$
$$p(\mathcal{D}, \boldsymbol{w}) = p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$$
- The *variational approximation* q for target distribution s.t. $p \equiv p(\boldsymbol{w}|\mathcal{D}) \approx q$ is defined as
$$q_* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left[q||p\right]$$
- Factorized approximation
$$q(\boldsymbol{w}) = \prod_{j=1}^{J} q(\boldsymbol{w}_j), \qquad \boldsymbol{w} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_J]$$
- CAVI algorithm: Repeat until convergence (or fixed number of iterations)
  1. For $k = 1, \ldots, K$
  $$\ln q^*(\boldsymbol{w}_k) = \mathbb{E}_{\prod_{i \neq k} q(\boldsymbol{w}_i)}\left[\ln p(\mathcal{D}, \boldsymbol{w})\right] + K$$
  2. Compute ELBO $\mathcal{L}\left[q\right]$ (monitoring convergence, model selection)
- Close parallel to Gibbs sampling except we now compute the expectation wrt. $\prod_{i \neq k} q(w_i)$ ("all other parameters")

CAVI Example

- Suppose we are working with a model with two parameters $\boldsymbol{w} \in \mathbb{R}^2$
- The joint distribution of the model is given by

$$\ln p(\boldsymbol{y}, \boldsymbol{w}) = \log p(\boldsymbol{y}|\boldsymbol{w}) + p(\boldsymbol{w}) = -w_1^2 - \frac{1}{2}w_2^2 + w_1 w_2 + 6w_1 - 3w_2$$

- We want to approximate the resulting posterior using a factorized distribution

$$q(\boldsymbol{w}) = q(w_1)q(w_2)$$

- The general CAVI update rule states that

$$\ln q^*(w_k) = \mathbb{E}_{\prod_{i \neq k} q(w_i)} [\ln p(\boldsymbol{y}, \boldsymbol{w})] + K$$

- For this example

$$\ln q^*(w_1) = \mathbb{E}_{q(w_2)} [\ln p(\boldsymbol{w}, \boldsymbol{y})] + K$$
$$\ln q^*(w_2) = \mathbb{E}_{q(w_1)} [\ln p(\boldsymbol{w}, \boldsymbol{y})] + K$$

CAVI Example continued I

The optimal solution for $q(w_1)$ is given by

$$\begin{aligned}\ln q(w_1) &= \mathbb{E}_{q(w_2)}\left[\ln p(\boldsymbol{w}, \boldsymbol{y})\right] + K \\ &= \mathbb{E}_{q(w_2)}\left[-w_1^2 - \frac{1}{2}w_2^2 + w_1 w_2 + 6w_1 - 3w_2\right] + K \\ &= -w_1^2 - \frac{1}{2}\mathbb{E}_{q(w_2)}\left[w_2^2\right] + w_1\mathbb{E}_{q(w_2)}\left[w_2\right] + 6w_1 - 3\mathbb{E}_{q(w_2)}\left[w_2\right] + K \\ &= -w_1^2 + w_1\mathbb{E}_{q(w_2)}\left[w_2\right] + 6w_1 + K' \\ &= -w_1^2 + w_1\left(6 + \mathbb{E}_{q(w_2)}\left[w_2\right]\right) + K'\end{aligned}$$

We recognize the above as a second order polynomial in w_1. Therefore, we conclude that $q(w_1)$ must be a Gaussian distribution and we can identify its mean and variance by matching the coefficients for the first and secord order term as follows

$$v_1^{-1} = 2 \iff v_1 = \frac{1}{2} \qquad \frac{m_1}{v_2} = 6 + \mathbb{E}_{q(w_2)}\left[w_2\right] \iff m_1 = 3 + \frac{1}{2}\mathbb{E}_{q(w_2)}\left[w_2\right]$$

CAVI Example continued II

The optimal solution for $q(w_2)$ is given by

$$\begin{aligned}
\ln q(w_2) &= \mathbb{E}_{q(w_1)}\left[\ln p(\boldsymbol{w}, \boldsymbol{y})\right] + K \\
&= \mathbb{E}_{q(w_1)}\left[-w_1^2 - \frac{1}{2}w_2^2 + w_1 w_2 + 6w_1 - 3w_2\right] + K \\
&= -\mathbb{E}_{q(w_1)}\left[w_1^2\right] - \frac{1}{2}w_2^2 + \mathbb{E}_{q(w_1)}\left[w_1\right] w_2 + 6\mathbb{E}_{q(w_1)}\left[w_1\right] - 3w_2 + K \\
&= -\frac{1}{2}w_2^2 + \mathbb{E}_{q(w_1)}\left[w_1\right] w_2 - 3w_2 + K' \\
&= -\frac{1}{2}w_2^2 + w_2\left(\mathbb{E}_{q(w_1)}\left[w_1\right] - 3\right) + K'
\end{aligned}$$

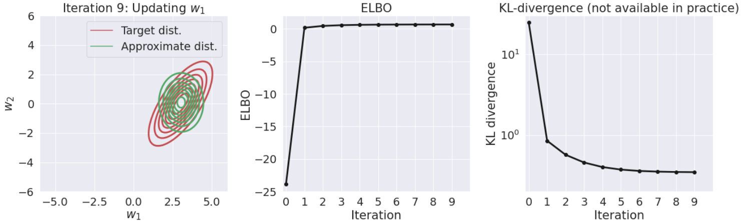
Again, this is a second order polynomial in w_2 and therefore $q(w_2)$ must be Gaussian with parameters

$$v_2^{-1} = 1 \iff v_2 = 1 \qquad \frac{m_2}{v_2} = \mathbb{E}_{q(w_1)}\left[w_1\right] - 3 \iff m_2 = \mathbb{E}_{q(w_1)}\left[w_1\right] - 3$$

CAVI Example continued III

- Initialize variational parameters and iteratively use update equations

$$q(w_1) = \mathcal{N}(w_1|3 + \frac{1}{2}m_2, \frac{1}{2})$$
$$q(w_2) = \mathcal{N}(w_2|m_1 - 3, 1)$$



Mixture models

Unsupervised learning

- Clustering and density estimation as examples of *unsupervised learning*

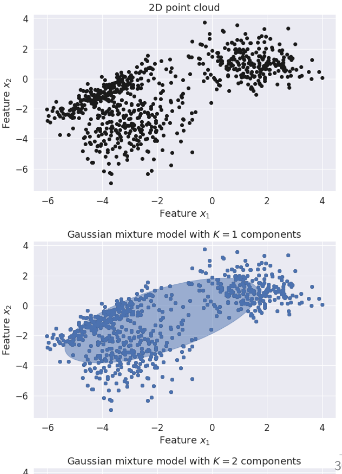
- Dataset $\mathcal{D} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$
    - Input features: $\boldsymbol{x}_i \in \mathbb{R}^D$

- Can we divide the dataset into K groups?

- *Model selection:* How to choose K?

- Common steps
    1. Choose model for the data
    2. Infer parameters of model $\boldsymbol{\theta}$
    3. Use parameters to make predictions for new data, e.g. outlier detection

- Applications
    - Clustering (news articles, songs, ...)
    - Fraud detection
    - ...



2D point cloud

Gaussian mixture model with $K = 1$ components

Gaussian mixture model with $K = 2$ components

The Gaussian Mixture Model

- How to model non-Gaussian data?
- We can construct arbitrary complex distribution by

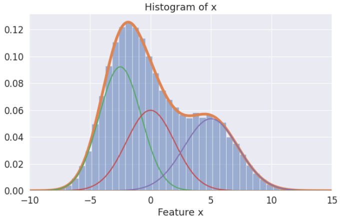
$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

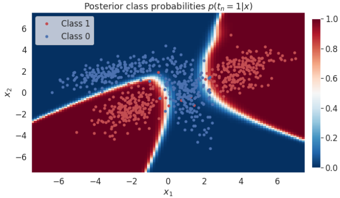
- The *mixing weights* $0 \leq \pi_k \leq 1$ and

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

- Example: Generative classification

$$p(y_n = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x_n}|y_n = k)p(y_n = k)}{p(\boldsymbol{x}_n)}$$





Fitting Gaussian Mixtures using Maximum likelihood

Expectation-maximization algorithm: Maximum likelihood estimation for Gaussian mixture models

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

1. Initialize all parameters: $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ for $k = 1, \ldots, K$

2. Repeat until convergence

   - Expectation-step:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

   - Maximization-step:

$$N_k = \sum_{n=1}^{N} \gamma_{nk}$$

$$\boldsymbol{\pi}_k^* = \frac{N_k}{N}$$

$$\boldsymbol{\mu}_k^* = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k^* = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k^*\right) \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k^*\right)^T$$

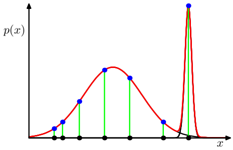
Problems with EM

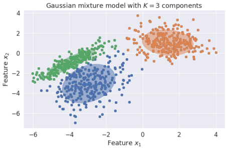
Several issues with the EM algorithm

1. Components can "collapse" onto single data points causing the maximum likelihood to diverge (overfitting due to maximum likelihood)

2. How to determine the number of clusters?

3. Sensitive to initialization



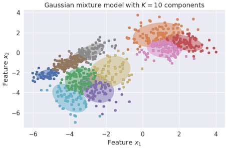


Bayesian approach

1. We can remove problem 1 entirely

2. Problem 2 is non-trivial, but Bayesian methods do have someting to offer

3. The variational approximation we will study is also sensitive to initialization



Bayesian Gaussian Mixture Model I

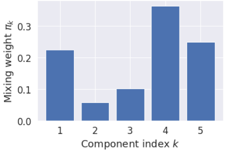
- We follow Bishop and switch to *precision matrix* parametrization $\boldsymbol{\Lambda}_k = \boldsymbol{\Sigma}^{-1}$. The Gaussian Mixture Model (GMM) becomes

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$

- Introducing *binary one-hot encoded latent variables $\boldsymbol{z}$*

$$p(\boldsymbol{x}_n | \boldsymbol{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$$

$$p(\boldsymbol{z}_n) = \text{Categorical}(\boldsymbol{z}_n | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_n}$$



- Example: suppose $K = 5$ and observation n belongs to the 4th cluster

$$\boldsymbol{z}_n = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

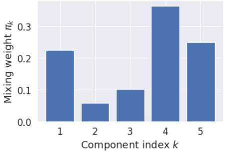
- We can always go back to the original model via the sum rule

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n | z_n = k) p(z_n = k)$$

Bayesian Gaussian Mixture Model II

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$$

$$p(\boldsymbol{z}_n) = \text{Categorical}(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_n}$$



- *Latent variables*: $\boldsymbol{z}_n$ are variable we cannot observe directly
- We need priors for $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ to complete the model

$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\alpha}_0) \\
\boldsymbol{\Lambda}_k &\sim \text{Wishart}(\boldsymbol{W}_0, \nu_0) \\
\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k &\sim \text{Normal}(\boldsymbol{m}_{\mathbf{0}}, (\beta_0\boldsymbol{\Lambda}_k)^{-1}) \\
\boldsymbol{z}_n|\boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}) \\
\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{z}_n &\sim \text{Normal}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Lambda}_{z_n}^{-1}),
\end{aligned}$$

The joint distribution

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \pi) = \prod_{n=1}^{N} p(x_n|\boldsymbol{z_n}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)p(\boldsymbol{\Lambda}_k)$$

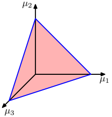
Quiz

Quiz time!

Lecture 10: Mixture models on DTU Learn

The Dirichlet distribution

- A categorial distribution with values $= 1, \ldots, K$ is parametrized by a K-dimensional probability vector $\boldsymbol{\pi}$:



$$z \sim \text{Categorial}(\boldsymbol{\pi})$$

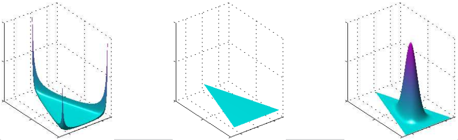
where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

- The *Dirichlet distribution* $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$ is a conjugate prior for the categorical distribution

$$\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \qquad \text{where} \qquad \alpha = \sum_{k=1}^{K} \alpha_k$$

- Hyperparameter: $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_K \end{bmatrix}$, where $\alpha_k > 0$. We often use $\alpha = a \cdot \mathbf{1}_K$ for some $a > 0$.

- Example with $K = 3$ and $a = 0.1$ (left), $a = 1$ (center), $a = 10$ (right)



The Wishart distribution

- For a univariate Gaussian likelihood with unknown mean and precision

$$p(\mathcal{D}|\mu,\tau) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\tau^{-1})$$

- Conjugate prior for the mean and the precision

$$p(\tau) = \mathrm{Gamma}(\tau|a_0,b_0)$$
$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})$$

- High-dimensional equivalent

$$p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}(x_n|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$$

- The *Wishart* prior is a *distribution over precision matrices* and is the high-dimensional equivalent of the Gamma distribution

$$\mathcal{W}(\Lambda|\boldsymbol{W}_0,\nu_0) = B|\Lambda|^{(\nu-D-1)/2}\exp(-\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}\right])$$

- Hyperparameters: $\nu_0 > D-1$ and $\boldsymbol{W}_0 \in \mathbb{R}^{D\times D}$

- Mean: $\mathbb{E}\left[\boldsymbol{\Lambda}\right] = \nu\boldsymbol{W}_0$ and $\mathbb{E}\left[\boldsymbol{\Lambda}^{-1}\right] = \frac{1}{\nu-D-1}\boldsymbol{W}_0^{-1}$

Variational inference for the mixture model

- Our goal is to compute the posterior distribution of all parameters of the mixture model

$$p(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi} | \boldsymbol{X}) = \frac{p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda})}{p(\boldsymbol{X})}$$

- Calculating the evidence requires us to sum over all possible K^N possible assignments

- We use variational inference with a factorized approximation

$$q(\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) = q(\boldsymbol{Z}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi})$$

- This is *the only assumption* we need to make inference feasible!

- Iterative algorithm to minimize the KL divergence

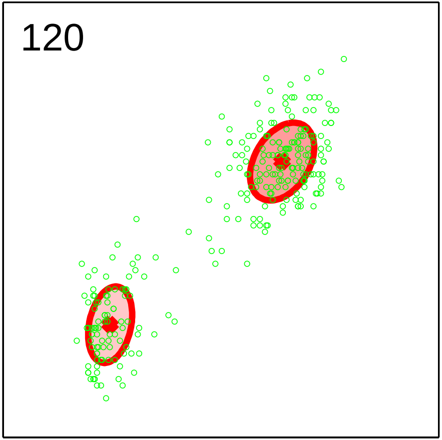
$$\ln q(\boldsymbol{Z}) \propto \mathbb{E}_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi})} [\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi})]$$
$$\ln q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \mathbb{E}_{q(\boldsymbol{Z})} [\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi})]$$

- Resulting approximation

$$\begin{aligned} q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= \underbrace{\prod_{n=1}^{N} \text{Categorial}(z_n | r_n)}_{q(\boldsymbol{Z})} \underbrace{\text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha})}_{q(\boldsymbol{\pi})} \underbrace{\prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{m}_k, \left[\beta_k \boldsymbol{\Lambda}_k^{-1}\right]\right) \mathcal{W}(\boldsymbol{\Lambda}_k | W_k, \nu_k)}_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda})} \end{aligned}$$

Example: Old faithful

- $N = 272$ observations from hydrothermal geyser in Yellowstone National Park
- Feature: x_1 eruption time (minutes), x_2 time until next eruption (minutes)
- Initialize Variational GMM with $K = 6$ clusters



Clustering images

- Initialize Variational GMM using 30 clusters, 10k images for training and 10k test

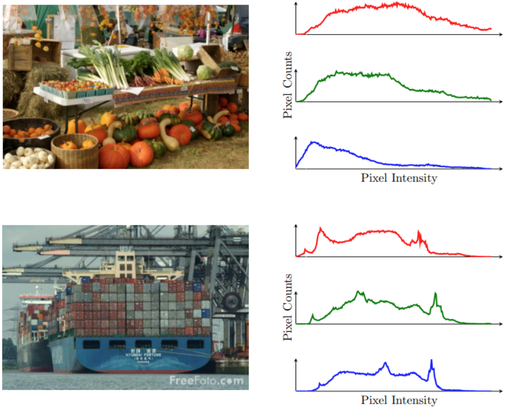
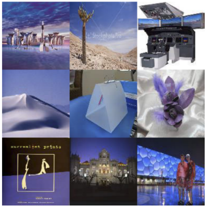


Figure 4: Red, green, and blue channel image histograms for two images from the imageCLEF dataset. The top image lacks blue hues, which is reflected in its blue channel histogram. The bottom image has a few dominant shades of blue and green, as seen in the peaks of its histogram.

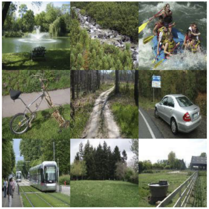
Blei et al: Variational inference: a review for statisticians

Clustering images

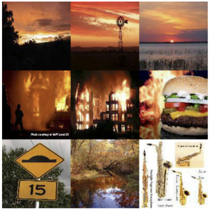
Visualizing 4 of the clusters



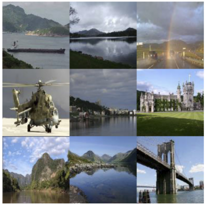
(a) Purple



(b) Green & White



(c) Orange



(d) Grayish Blue

Blei et al: Variational inference: a review for statisticians

Clustering images

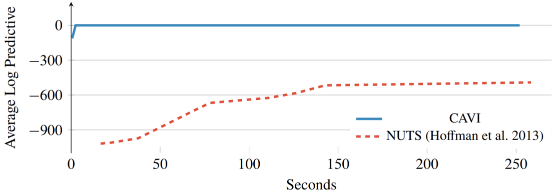


Figure 6: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. CAVI fits a Gaussian mixture model to ten thousand images in less than a minute.

Blei et al: Variational inference: a review for statisticians