Approximate Bayesian Estimation: Lecture 3
Variational Bayesian Inference

Lan Hua

Northwestern Polytechnical University
Key Laboratory of Information Fusion Technology

E-mail: lanhua@nwpu.edu.cn
Tel: 15399035213

October 28, 2019

Lecture Outline

1. What is the VB?

2. Mean Field VB Algorithm

3. Advances in VB

4. VB for Gaussian Mixture Model

5. Applications for JDT of OTHR

What is the VB?

Central task of Bayesian inference is the posterior distribution of latent variables $\mathbf{Z}$ given observations $\mathbf{X}$:

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{\int p(\mathbf{X}, \mathbf{Z})d_{\mathbf{Z}}} \tag{1}$$

In practice, evaluating the posterior is usually difficult because we cannot easily evaluate $p(\mathbf{X})$, especially when:

- analytical solutions are not available (for continuous latent variables)
- numerical integration is too expensive (for discrete latent variables)

What is the VB?

Two kinds of approximation inference methods:

Stochastic approximate inference (in particular sampling)

- design an algorithm that draws samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(m)}$ from $p(\mathbf{Z}|\mathbf{X})$
- inspect sample statistics (e.g., histogram, sample quantiles, ...)
- asymptotically exact but computationally expensive and tricky engineering concerns
- suitable for small-scale inference problem (nonlinear nonGaussian filtering)

Deterministic approximate inference (in particular variational Bayes)

- find an analytical proxy $q(\mathbf{Z})$ that is maximally similar to $p(\mathbf{Z}|\mathbf{X})$
- inspect distribution statistics of $q(\mathbf{Z})$ (e.g., mean, quantiles, ...)
- often insightful and lightning-fast but hard work to derive
- suitable for large-scale inference problem (data associaton)

What is the VB?

Variational Bayesian inference is based on *variational calculus.*

- Standard calculus (Newten, Leibniz, and others)
  - function $f : x \to f(x)$
  - derivatives $\dfrac{df}{dx}$
  - Example: maximize the likelihood expression $p(\mathbf{X}|\mathbf{Z})$ w.r.t $\mathbf{Z}$
- Variational calculus (Euler, Lagrange, and others)
  - functions $F : f \to F(f)$
  - derivatives $\dfrac{dF}{df}$
  - Example: maximize the entropy $\mathrm{H}[p]$ w.r.t $p(x)$.

Tips: The entropy $H[p]$ which takes a probability distribution $p(x)$ as the input and returns the quantity

$$\mathrm{H}[p] = \int p(x) \ln p(x) dx \tag{2}$$

What is the VB?

Variational calculus lends itself nicely to approximate Bayesian inference.

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \ln \frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \\
&= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \left(\ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} + \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right) d\mathbf{Z} \\
&= \underbrace{\int q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}}_{\mathrm{KL}[q||p]\ \text{:divergence}} + \underbrace{\int \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}}_{F(q,X)\ \text{:free energy}}
\end{aligned}
\tag{3}
$$

What is the VB?

In summary, the log model evidence can be expressed as:

$$\ln p(\mathbf{X}) = \underbrace{\mathrm{KL}[q||p]}_{\geq 0,\text{unknown}} + \underbrace{F(q, \mathbf{X})}_{\text{easy to evaluate}} \tag{4}$$

Maximizing $F(q, \mathbf{X})$ is equivalent to:

- minimizing $\mathrm{KL}[q||p]$
- tightening $F(q, \mathbf{X})$ as a lower bound to the log model evidence

Tips: This differs from our discussion of EM only in that the parameters vector $\boldsymbol{\theta}$ no longer appears, because the parameters are now stochastic variables and are absorbed into Z.

What is the VB?

Variational Bayesian inference (VB)

- The goal of VB is to approximate a conditional density of latent variables given observed variables $p(\mathbf{Z}|\mathbf{X})$.
- The key idea is to solve this problem with optimization, e.g., variational calculus.
- We use a family of densities over the latent variables $q(\mathbf{Z})$, parameterized by free "variational parameters" $q(\mathbf{Z}; \boldsymbol{\lambda})$.
- The optimization finds the number of this family, that is, the setting of the parameters, which is closest in KL divergence to the conditional of interest $\boldsymbol{\lambda}^* = \arg\min_{\lambda} \mathrm{KL}[q(\mathbf{Z}; \boldsymbol{\lambda}) || p(\mathbf{Z}|\mathbf{X})]$.
- The fitted variational density then serves as a proxy for the exact conditional density $q(\mathbf{Z}; \boldsymbol{\lambda}^*) \approx p(\mathbf{Z}|\mathbf{X})$.

Mean field VB

How to choose the form of approximated variational family $q(\mathbf{Z})$? The complexity of the family determines the complexity of the optimization; it is more difficult to optimize over a complex family than a simple family.

Mean-field variational family assumes that the latent variables are mutually independent and each governed by a distinct factor in the variational density, i.e.,

$$q(\mathbf{Z};\boldsymbol{\lambda}) = \prod_i q_i(z_i;\lambda_j) \tag{5}$$

Each latent variable z_i is governed by its own variational density $q_i(z_i;\lambda_i)$. In optimization, each variational density is chosen to maximize the ELBO (variational energy free)

Mean field VB

$$
\begin{aligned}
F(q,\mathbf{X}) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int \prod\nolimits_i q_i(z_i) \times (\ln p(\mathbf{X},\mathbf{Z} - \sum\nolimits_i \ln q_i(z_i))d\mathbf{Z} \\
&= \int q_j(z_j) \prod\nolimits_{\setminus j} q_i(z_i)(\ln p(\mathbf{X},\mathbf{Z} - \ln q_j(z_j))d\mathbf{Z} \\
&\quad - \underbrace{\int q_j(z_j) \prod\nolimits_{\setminus j} q_i(z_i) \sum\nolimits_{\setminus j} \ln q_i(z_i) d\mathbf{Z}_{\setminus j} dz_j}_{\text{constant}} \\
&= \int q_j(z_j)(\underbrace{\int \prod\nolimits_{\setminus j} q_i(z_i) \ln p(\mathbf{X},\mathbf{Z}) d\mathbf{Z}_{\setminus j}}_{\mathbb{E}_{q_{\setminus j}}[\ln p(\mathbf{X},\mathbf{Z})]} - \ln q_j(z_j))dz_j + \text{constant} \\
&= \underbrace{\int q_j(z_j) \ln \frac{\exp(\mathbb{E}_{q_{\setminus j}}[\ln p(\mathbf{X},\mathbf{Z})])}{q_j(z_j)} dz_j}_{-\mathrm{KL}[q_j(z_j)||\exp(\mathbb{E}_{q_{\setminus j}}[\ln p(\mathbf{X},\mathbf{Z})])]} + \text{constant}
\end{aligned}
\tag{6}
$$

Mean field VB

In summary:

$$F(q, \mathbf{X}) = -\mathrm{KL}[q_j(z_j) || \exp(\mathbb{E}_{q_{\setminus j}}[\ln p(\mathbf{X}, \mathbf{Z})])] + \text{constant} \tag{7}$$

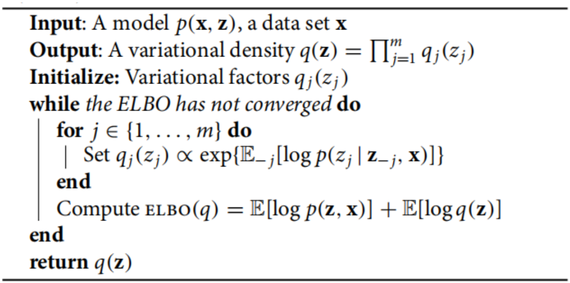
Suppose the densities $q_{\setminus j} := q(\mathbf{Z}_{\setminus j}) = q(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_M)$ are kept fixed. Then the approximate posterior $q(z_j)$ that maximizes $F(q, \mathbf{X})$ is given by

$$\begin{aligned} q_j^*(z_j) &= \arg\max_{q_j(z_j)} F(q, \mathbf{X}) \\ &\propto \exp\left(\mathbb{E}_{q(\mathbf{Z}_{\setminus j})} \ln p(\mathbf{X}, \mathbf{Z}_{\setminus j}, z_j)\right) \end{aligned} \tag{8}$$

Mean field VB

Coordinate ascent variational inference (CAVI) iteratively optimizes each variational density, while holding the others fixed. It climbs the ELBO to a local optimum.

- Initialize all approximate posteriors $q_j(z_j)$, e.g., setting their priors
- Cycle over the parameters, given the current estimates of the others
- Loop until convergence.



```
Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while the ELBO has not converged do
    for $j \in \{1, \ldots, m\}$ do
        Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
    end
    Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$
```

Advances in VB

Beyond Mean-field variational family: Classical MFVI has historically played an important role due to its efficient inference, however, it is limited in multiple ways when it comes to modern applications.

- Scalable variational Bayesian for big datasets.
- Generic variational Bayesian with non-conjugate exponential family
- Accurate variational Bayesian beyond KL and mean field

Advances in VB

Scalable VB: Big datasets raise new challenges for the computational feasibility of Bayesian algorithm, making scalable inference essential.

- Stochastic variational Bayesian
- Collapsed variational Bayesian
- Distributed variational Bayesian

Advances in VB

Stochastic VB: is a *stochastic optimization* algorithm for mean-field VB.

The ELBO of the general model ($q(\boldsymbol{\xi}, \theta) = q(\theta|\gamma) \prod_{i=1}^{M} q(\xi_i|\phi_i)$)

$$\begin{aligned} F(q, \mathbf{X}) = & \mathbb{E}_q \left[\ln p(\theta|\alpha) - \ln q(\theta|\gamma) \right] \\ & + \sum_{i=1}^{M} \mathbb{E}_q \left[\ln p(\xi_i|\theta) + \ln p(x_i|\xi_i, \theta) - \ln q(\xi_i|\phi_i) \right] \end{aligned} \tag{9}$$

- latent variable $\mathbf{Z} = \{\theta, \boldsymbol{\xi}\}$ includes local variables $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_M\}$ and global variable θ
- the variational parameters $\lambda = \{\gamma, \phi\}$ includes global latent variables γ and local latent variables ϕ
- the observation $\mathbf{X}$ and model hyperparameters α
- number of data M and mini-batch size S

Advances in VB

Problem: Eq. (9) can be optimized by coordinate descent on ELBO. Every iteration or gradient step scales with M, and is expensive for large data.

Solution: Stochastic VB solves this problem in the spirit of *stochastic gradient descent*. In every iteration, one randomly selects mini-batches of size S to obtain a stochastic estimate of the ELBO $F(q, \mathbf{X})$.

$$\begin{aligned}\hat{F}(q, \mathbf{X}) =& \mathbb{E}_q \left[\ln p(\theta|\alpha) - \ln q(\theta|\gamma)\right] \\ &+ \frac{M}{S} \sum_{s=1}^{S} \mathbb{E}_q \left[\ln p(\xi_{i_s}|\theta) + \ln p(x_{i_s}|\xi_{i_s}, \theta) - \ln q(\xi_{i_s}|\phi_{i_s})\right]\end{aligned} \tag{10}$$

Advances in VB

Stochastic optimization. Let $f(x)$ be a function to be maximized and $h_t(x)$ be the realization of a random variable $H(x)$ whose expectation is the gradient of $f(x)$. Finally, let ρ_t be a nonnegative scalar. Stochastic optimization updates x at the tth iteration with

$$x_{t+1} \leftarrow x_t + \rho_t h_t(x_t) \tag{11}$$

This converges to a maximum of $f(x)$ when ρ_t, the learning rate, follows the Robbins-Monro conditions,

$$\sum_{t=1}^{\infty} \rho_t = \infty, \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty \tag{12}$$

Because of its simplicity, stochastic optimization is widely used in statistics and machine learning.

Advances in VB

Gradients and probability distributions: The classical gradient method for maximization ELBO $F(q, \mathbf{X})$ by taking steps of size ρ in the direction of the gradient

$$\xi_i^{n+1} = \xi_i^n + \rho \nabla_{\xi_i} F^t(q, \mathbf{X}) \tag{13}$$

Stochastic gradient algorithm: optimizes $\hat{F}(q, \mathbf{X})$ by iteratively following realizations of $B(\boldsymbol{\xi})$ as

$$\xi_i^{n+1} = \xi_i^n + \rho^n b^n(\xi^n) \tag{14}$$

where b_t is an independent draw from the noisy gradient B, and $B(\boldsymbol{\xi})$ is a random function that has expectation equal to the gradient so that $\mathbb{E}_q[B(\boldsymbol{\xi})] = \nabla_{\xi_i} F^t(q, \mathbf{X})$.

Advances in VB

Collapsed VB: relies on the idea of analytically integrating out certain model parameters.

- Due to the reduced number of parameters to be estimated, inference is typically faster
- Constrained in the traditional conjugate exponential families, where the ELBO assumes an analytical form during marginalization
- The computational benefit of collapsed VB depends strongly on the statistics of the collapsed variables

Parallel and Distributed VB: variational inference can be adjusted to distributed computing scenarios, where subsets of the data or parameters are distributed among several machines.

Advances in VB

Generic VB: Beyond the conjugate exponential family variational inference was originally limited to conditionally conjugate models, for which the ELBO could be computed analytically before it was optimized. *Stochastic gradient estimators* of the ELBO are the central tools for non-conjugate models.

- Block box variational inference (BBVI)

Advances in VB

BBVI: The main idea is to represent the gradient as an expectation, and to use Monte Carlo techniques to estimate this expectation.

Consider the ELBO as follows

$$F(q, \mathbf{X}) = \mathbb{E}_{q(\mathbf{Z}|\boldsymbol{\lambda})}[\ln p(\mathbf{X}, \mathbf{Z}) - \ln q(\mathbf{Z}|\boldsymbol{\lambda})] \tag{15}$$

Its gradient can be written as an expectation w.r.t the variational distribution

$$\nabla_{\boldsymbol{\lambda}} F(q, \mathbf{X}) = \mathbb{E}_{q(\mathbf{Z}|\boldsymbol{\lambda})}[\nabla_{\boldsymbol{\lambda}} \ln q(\mathbf{Z}|\boldsymbol{\lambda})(\ln p(\mathbf{X}, \mathbf{Z}) - \ln q(\mathbf{Z}|\boldsymbol{\lambda}))] \tag{16}$$

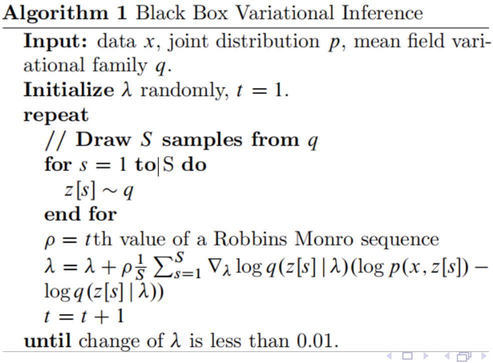
This noisy unbiased gradients of the ELBO can be computed by Monte Carlo samples from the variational distribution:

$$\nabla_{\boldsymbol{\lambda}} F(q, \mathbf{X}) \approx \frac{1}{S} \sum_{s=1}^{S} [\nabla_{\boldsymbol{\lambda}} \ln q(\mathbf{Z_s}|\boldsymbol{\lambda})(\ln p(\mathbf{X}, \mathbf{Z_s}) - \ln q(\mathbf{Z_s}|\boldsymbol{\lambda}))] \tag{17}$$

where $Z_s \sim q(\mathbf{Z}|\boldsymbol{\lambda})$.

Advances in VB

- Score function ln $q(\mathbf{Z_s}|\boldsymbol{\lambda})$ and sampling algorithm depend only on the variational distribution, not the underlying model
- Do not make any assumptions about the form of the model, only that can compute the log of the joint $p(\mathbf{X}, \mathbf{Z_s})$
- This algorithm applies to a wide variety of models
- The variance of the estimated gradient can be too large to be useful



Algorithm 1 Black Box Variational Inference

```
Input: data $x$, joint distribution $p$, mean field variational family $q$.
Initialize $\lambda$ randomly, $t = 1$.
repeat
    // Draw $S$ samples from $q$
    for $s = 1$ to|S do
        $z[s] \sim q$
    end for
    $\rho$ = $t$th value of a Robbins Monro sequence
    $\lambda = \lambda + \rho \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z[s] \mid \lambda)(\log p(x, z[s]) - \log q(z[s] \mid \lambda))$
    $t = t + 1$
until change of $\lambda$ is less than 0.01.
```

Advances in VB

Accurate VB: Beyond KL and Mean Field

The KL divergence often provides a computationally convenient method to measure the distance between two distributions. However, traditional KL-based VB suffers from problems such as underestimating posterior variances. A number of other divergence measures have been proposed

- α-divergence: $D_\alpha^R(p||q) = \dfrac{1}{\alpha - 1} \log \int p(x)^\alpha q(x)^{1-\alpha} dx$ with $\alpha > 0$, and $\alpha \neq 1$. For $\alpha \to 1$, we recover the KL-divergence.
- f-divergence: $D_f(p||q) = \int q(x) f(\dfrac{p(x)}{q(x)}) dx$. where f is a convex function with $f(1) = 0$. The KL-divergence is represented by the f-divergence with $f(r) = r \ln(r)$.
- Stein discrepancy: $D_{\text{stein}}(p, q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{q(z)}[f(z)] - \mathbb{E}_{p(z|x)}[f(z)]|^2$. where $\mathcal{F}$ indicates a set of smooth, real-valued functions.

Advances in VB

The Mean-field family enables efficient inference, but limited by its strong factorization. It cannot capture posterior dependencies between latent variables, dependencies which both improve the fidelity of the approximation and are sometimes of intrinsic interest.

- Structured VB
- Copula VB
- Hierarchical VB

Advances in VB

Structured VB: is an approximation in which the latent variables are not completely factorized, but rather they are related in a structured manner.

- These structured variational distributions are more expressive, but often come at higher computational costs
- Allowing a structured variational distribution to capture dependencies between latent variables is a modeling choice
- different dependencies may be more or less relevant and depend on the model under consideration

Advances in VB

Copula VB assumes the variational family form (Sklars theorem):

$$q(\mathbf{Z}) = \left(\prod_i q(z_i; \lambda_i) \right) c(Q(z_1), \ldots, Q(z_N)), \tag{18}$$

where c is the copula distribution, which is a joint distribution over the marginal cumulative distribution functions $Q(z_1), \ldots, Q(z_N)$. This copula distribution restores the dependencies among the latent variables.

The copula VB decouples the overall inference task into two subtasks: (1) inference of the copula function, which captures the multivariate posterior dependencies; (2) inference of a set of univariate margins, which are allowed to take essentially any form.

Advances in VB

Hierarchical VB Hierarchical variational models are a BBVI framework for structured variational distributions which applies to a broad class of models. In order to capture dependencies between latent variables, one starts with a mean-field variational distribution $\prod_i q(z_i;\lambda_i)$, but instead of estimating the variational parameters $\boldsymbol{\lambda}$, one places a prior $q(\boldsymbol{\lambda};\boldsymbol{\theta})$ over them and marginalizes them out:

$$q(\mathbf{Z};\boldsymbol{\theta}) = \int \left(\prod_i q(z_i;\lambda_i)\right) q(\boldsymbol{\lambda};\boldsymbol{\theta})d\lambda \tag{19}$$

The new variational distribution $q(z;\boldsymbol{\theta})$ thus captures dependencies through the marginalization procedure.

VB for Gaussian Mixture Model

Gaussian Mixture Model can be written as a linear superposition of Gaussian as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{20}$$

- K is the number of Gaussian
- π_k is mixing coefficient: weight for each Gaussian distribution, which satisfies $0 \leq \pi_k \leq 1$ and $\sum_{i=1}^{K} \pi_k = 1$.

VB for Gaussian Mixture Model

Let us introduce a K-dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0. The marginal distribution over z_k is

$$p(z_k = 1) = \pi_k \tag{21}$$

and the joint distribution

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{22}$$

The conditional distribution of $\mathbf{x}$ given z_i is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{23}$$

which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})^{z_k} \tag{24}$$

VB for Gaussian Mixture Model

The marginal distribution of $\mathbf{x}$ is then obtained by summing the joint distribution over all possible states of $\mathbf{z}$

$$p(\mathbf{x}) = \sum_{\mathbf{x}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{25}$$

By introducing latent variable $\mathbf{z}$ to derive the Gaussian mixture, we gain

- Work with the joint distribution $p(\mathbf{x}, \mathbf{z})$ instead of the marginal distribution $p(\mathbf{x})$.
- The conditional probability of $\mathbf{z}$ given $\mathbf{x}$, i.e, $\gamma(z_k) = p(z_k = 1|\mathbf{x})$ will play an important role in the later EM algorithm.
- $\gamma(z_k)$ can be viewed as the *responsibility* that component k takes for 'explaining' the observation $\mathbf{x}$.

Maximum likelihood

Suppose we have a i.i.d data set of observations $\{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, and we wish to model this data using a mixture of Gaussians. Denote the data set as $\mathbf{X} \in \mathcal{R}^{N\times D}$, and the corresponding latent variables $\mathbf{Z} \in \mathcal{R}^{N\times K}$. From (17) the log-likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \right\} \tag{26}$$

Maximizing the log-likelihood function (18) is intractable, the main difficulty arises from the presence of the summation over k that appear insider the logarithm in (18). If we set the derivatives of the log-likelihood to zero, we will no longer obtain a closed form solution.

EM for Gaussian Mixture Model

Now consider the problem of maximizing the likelihood for the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, which takes the form

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\mathbf{k})^{z_{nk}} \tag{27}$$

Taking the logarithm, we obtain

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_\mathbf{k})\} \tag{28}$$

Compared with the log-likelihood function for incomplete data, i.e., (18), the summation over k and the logarithm have been interchanged in (20). The logarithm now acts directly on the Gaussian distribution.

VB for Gaussian Mixture Model

E-step: Evaluate the *responsibilities* using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{j}, \boldsymbol{\Sigma}_\mathbf{j})} \tag{29}$$

M-step: Re-estimate the parameters using the current responsibilities

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n}}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{30}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x_n} - \mu_k)(\mathbf{x_n} - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{31}$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \tag{32}$$

VB for Gaussian Mixture Model

VB for Gaussian Mixture Model Note that the unknown parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ are random variables. First we introduce priors over these parameters. The analysis is considerably simplified if we use conjugate prior distributions. Therefore we choose a Dirichlet distribution for $\boldsymbol{\pi}$, and an independent Gaussian-Wishart prior for $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-\mathbf{1}}$.

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0}) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1} \tag{33}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\mu_k}|\mathbf{m_0}, (\boldsymbol{\beta_0\Lambda_k})^{-\mathbf{1}})\mathcal{W}(\boldsymbol{\Lambda_k}|\mathbf{W_0}, \upsilon_0) \tag{34}$$

Tips: The variational parameters $\theta_0 = \{\boldsymbol{\alpha_0}, \mathbf{m_0}, \beta_0, \mathbf{W_0}, \upsilon_0\}$, also called hyperparameters.

VB for Gaussian Mixture Model

The joint distribution of all the random variables is

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \tag{35}$$

Mean-field approximation

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) \approx & q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ = & q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ = & q(\mathbf{Z})q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}}) \end{aligned} \tag{36}$$

VB for Gaussian Mixture Model

Update for variational distribution $q(\mathbf{Z})$

$$\begin{aligned}\ln q^*(\mathbf{Z}) =& \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{constant} \\ =& \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{constant} \\ =& \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{constant}\end{aligned} \tag{37}$$

where

$$\begin{aligned}\ln \rho_{nk} =& \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_\mathbf{k}|] - \frac{D}{2}\ln(2\pi) \\ & - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_\mathbf{k},\boldsymbol{\Lambda}_\mathbf{k}}[(x_n - \boldsymbol{\mu}_\mathbf{k})^T \boldsymbol{\Lambda}_\mathbf{k}(x_n - \boldsymbol{\mu}_\mathbf{k})]\end{aligned} \tag{38}$$

VB for Gaussian Mixture Model

Taking the exponential of both side of (37), and normalizing the distribution, we obtain

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{39}$$

where $r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nk}}$.

- The optimal solution for $q(\mathbf{Z})$ takes the same functional form as the prior $p(\mathbf{Z}|\boldsymbol{\pi})$
- The optimal solution for $q(\mathbf{Z})$ depends on moments evaluated w.r.t other variables, and so again the variational update equations are coupled and must be solved iteratively
- For the discrete distribution $q^*(\mathbf{Z})$, we have $\mathbb{E}[z_{nk}] = r_{nk}$

VB for Gaussian Mixture Model

Update for variational distribution $q(\boldsymbol{\pi})$

$$\begin{aligned}\ln q^*(\boldsymbol{\pi}) =&\mathbb{E}_{\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{constant} \\ =&\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + p(\boldsymbol{\pi}) + \text{constant} \\ =&(\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln \pi_k + \text{constant}\end{aligned} \tag{40}$$

Taking the exponential of both sides, we recognize $q^*\boldsymbol{\pi}$ as a Dirichlet distribution

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{41}$$

where $\boldsymbol{\alpha}$ has components α_k given by

$$\alpha_k = \alpha_0 + \sum_{n=1}^{N} r_{nk} \tag{42}$$

VB for Gaussian Mixture Model

Update for variational distribution $q(\boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}})$

$$\begin{aligned}\ln q^*(\boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}}) =& \mathbb{E}_{\mathbf{Z},\pi}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{constant} \\ =& \ln p(\boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}}) + \sum_{n=1}^{K} \mathbb{E}[z_{nk}] \ln \mathcal{N}(x_n | \boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}}^{-1}) + \text{constant}\end{aligned} \tag{43}$$

The result, as expected, the variational distribution is a Gaussian-Wishart distribution is given by

$$q^*(\boldsymbol{\mu}_{\mathbf{k}}, \boldsymbol{\Lambda}_{\mathbf{k}}) = \mathcal{N}(\mu_k | \mathbf{m}_{\mathbf{k}}, (\boldsymbol{\beta}_{\mathbf{k}} \boldsymbol{\Lambda}_{\mathbf{k}})^{-1}) \mathcal{W}(\Lambda_k | W_k, \upsilon_k) \tag{44}$$

VB for Gaussian Mixture Model

The variational parameters are updated as follows

$$\beta_k = \beta_0 + N_k \tag{45}$$

$$\mathbf{m_k} = \frac{1}{\beta_k}(\beta_0 \mathbf{m_0} + N_k \bar{\mathbf{x}}_k) \tag{46}$$

$$\mathbf{W_k}^{-1} = \mathbf{W_0}^{-1} + N_k \mathbf{S_k} + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mathbf{m_0})(\bar{\mathbf{x}}_k - \mathbf{m_0})^T \tag{47}$$

$$\upsilon_k = \upsilon_0 + N_k \tag{48}$$

where

$$N_k = \sum_{n=1}^{K} r_{nk}, \bar{\mathbf{x}}_\mathbf{k} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x_n}, \mathbf{S_k} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x_n} - \bar{\mathbf{x}}_\mathbf{k})(\mathbf{x_n} - \bar{\mathbf{x}}_\mathbf{k})^T \tag{49}$$

References

Christopher M.Bishop, Pattern Recognition and Machine Learning, *Spring*, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518): 859-877, 2017.

Zhang Cheng, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

The End