Variational Inference for Bayes vMF Mixture

Hanxiao Liu

September 23, 2014
Lower bound the likelihood

\[ \mathcal{L}(\theta; X) = \mathbb{E}_q \log p(X|\theta) \]

\[ = \mathbb{E}_q \left[ \log \frac{p(X, Z|\theta)}{q(Z)} \right] + \mathbb{E}_q \left[ \log \frac{q(Z)}{p(Z|X, \theta)} \right] \]

\[ \text{VLB}(q, \theta) + \text{D}_{KL}(q(Z)||p(Z|X, \theta)) \]

Raise VLB \((q, \theta)\) by coordinate ascent

1. \[ q^{t+1} = \arg\max_q \text{VLB} \left( q, \theta^t \right) \]
   \[ q = \prod_{i=1}^{M} q_i \]

2. \[ \theta^{t+1} = \arg\max_{\theta} \text{VLB} \left( q^{t+1}, \theta \right) \]
Variational Inference Review

**Goal:** solve \( \arg \max \ VLB (q, \theta^t) \) by coordinate ascent, i.e. sequentially updating a single \( q_i \) in each iteration.

Each coordinate step has a closed-form solution—

\[
VLB \left( q_j; q_{-j}, \theta^t \right) = \mathbb{E}_q \left[ \log \frac{p(X, Z|\theta^t)}{q(Z)} \right] 
\]

\[
= \mathbb{E}_q \log p(X, Z|\theta^t) - \sum_{i=1}^{M} \mathbb{E}_q \log q_i 
\]

\[
= \mathbb{E}_{q_j} \mathbb{E}_{q_{-j}} \log p(X, Z|\theta^t) - \mathbb{E}_{q_j} \log q_j + \text{const}
\]

\[
= \int q_j \log \tilde{q}_j + \text{const} = -D_{KL} (q_j||\tilde{q}_j) + \text{const}
\]

\[
\implies \log q_j^* = \mathbb{E}_{q_{-j}} \log p(X, Z|\theta^t) + \text{const}
\]
Bayes vMF Mixture

[Gopal and Yang, 2014]

\[ \pi \sim \text{Dirichlet} (\cdot | \alpha) \]

\[ \mu_k \sim \text{vMF} (\cdot | \mu_0, C_0) \]

\[ \kappa_k \sim \text{logNormal} (\cdot | m, \sigma^2) \]

\[ z_i \sim \text{Multi} (\cdot | \pi) \]

\[ x_i \sim \text{vMF} (\cdot | \mu_{z_i}, \kappa_{z_i}) \]

\[ q(\pi) \equiv \text{Dirichlet} (\cdot | \rho) \]

\[ q(\mu_k) \equiv \text{vMF} (\cdot | \psi_k, \gamma_k) \]

\[ q(\kappa_k) \equiv \text{logNormal} (\cdot | a_k, b_k) \]

\[ q(z_i) \equiv \text{Multi} (\cdot | \lambda_i) \]
Compute $\log p (X, Z|\theta)$

$$p (X, Z|\theta) = \text{Dirichlet} (\pi|\alpha) \times \prod_{i=1}^{N} \text{Multi} (z_i|\pi) \text{vMF} (x_i|\mu_{z_i}, \kappa_{z_i})$$

$$\times \prod_{k=1}^{K} \text{vMF} (\mu_k|\mu_0, C_0) \log\text{Normal} (\kappa_k|m, \sigma^2$$

$$\log p (X, Z|\theta) = -\log B (\alpha) + \sum_{k=1}^{K} (\alpha - 1) \log \pi_k$$

$$+ \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left( \log C_D (\kappa_k) + \kappa_k x_i^T \mu_k \right)$$

$$+ \sum_{k=1}^{K} \left( \log C_D (C_0) + C_0 \mu_k^T \mu_0 \right)$$

$$+ \sum_{k=1}^{K} \left( -\log \kappa_k - \frac{1}{2} \log (2\pi\sigma^2) - \frac{(\log \kappa_k - m)^2}{2\sigma^2} \right)$$
Updating $q(\pi)$

$q(\pi) \triangleq \text{Dirichlet}(\cdot | \rho)$

$$\log q^*(\pi) = \mathbb{E}_{q\backslash\pi} \log p(X, Z | \theta) + \text{const}$$

$$= \mathbb{E}_{q\backslash\pi} \left[ \sum_{k=1}^{K} (\alpha - 1) \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k \right] + \text{const}$$

$$= \sum_{k=1}^{K} \left( \alpha + \sum_{i=1}^{N} \mathbb{E}_{q}[z_{ik}] - 1 \right) \log \pi_k + \text{const}$$

$$\implies q^*(\pi) \propto \prod_{k=1}^{K} \pi_k^{\alpha + \sum_{i=1}^{N} \mathbb{E}_{q}[z_{ik}] - 1} \sim \text{Dirichlet}$$

$$\implies \rho_k^* = \alpha + \sum_{i=1}^{N} \mathbb{E}_{q}[z_{ik}]$$
Updating $q(z_i)$

$q(z_i) \equiv \text{Multi} (\cdot | \lambda_i)$

$$
\log q^* (z_i) = \mathbb{E}_{q \setminus z_i} \log p (X, Z | \theta) + \text{const}
$$

$$
= \mathbb{E}_{q \setminus z_i} \left[ \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left( \log C_D(\kappa_k) + \kappa_k x_i^\top \mu_k \right) \right] + \text{const}
$$

$$
= \sum_{k=1}^{K} z_{ik} \left( \mathbb{E}_q \log \pi_k + \mathbb{E}_q \log C_D(\kappa_k) + \mathbb{E}_q [\kappa_k] x_i^\top \mathbb{E}_q [\mu_k] \right) + \text{const}
$$

$$
\implies q^* (z_i) \sim \text{Multi}, \ \lambda_{ik}^* \propto e^{\mathbb{E}_q \log \pi_k + \mathbb{E}_q \log C_D(\kappa_k) + \mathbb{E}_q [\kappa_k] x_i^\top \mathbb{E}_q [\mu_k]}
$$

Assume $\mathbb{E}_q \log \pi_k, \mathbb{E}_q \log C_D(\kappa_k), \mathbb{E}_q [\kappa_k]$ and $\mathbb{E}_q [\mu_k]$ are already known. We will explicitly compute them later.
Updating \( q(\mu_k) \)

\[
q(\mu_k) \equiv \text{vMF}(\cdot | \psi_k, \gamma_k)
\]

\[
\log q^*(\mu_k) = \mathbb{E}_{q \setminus \mu_k} \log p(X, Z | \theta) + \text{const}
\]

\[
= \mathbb{E}_{q \setminus \mu_k} \left[ \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \kappa_j \mathbf{x}_i^\top \mu_j + \sum_{j=1}^{K} C_0 \mu_j^\top \mu_0 \right] + \text{const}
\]

\[
= \mathbb{E}_q[\kappa_k] \left( \sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \mathbf{x}_i^\top \mu_k \right) + C_0 \mu_k^\top \mu_0 + \text{const}
\]

\[
\implies q^*(\mu_k) \propto e^{\left[ \mathbb{E}_q[\kappa_k] \left( \sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \mathbf{x}_i \right) + C_0 \mu_0 \right]^\top \mu_k} \sim \text{vMF}
\]

\[
\gamma_k^* = \left\| \mathbb{E}_q[\kappa_k] \left( \sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \mathbf{x}_i \right) + C_0 \mu_0 \right\|, \psi_k^* = \frac{\mathbb{E}_q[\kappa_k] \left( \sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \mathbf{x}_i \right) + C_0 \mu_0}{\gamma_k}
\]
Updating $q(\kappa_k)$

$q(\kappa_k) \equiv \log\text{Normal} (\cdot | a_k, b_k)$

$\log q^*(\kappa_k)$

$= \mathbb{E}_{q \setminus \kappa_k} \log p(X, Z|\theta) + \text{const}$

$= \mathbb{E}_{q \setminus \kappa_k} \left[ \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \left( \log C_D(\kappa_j) + \kappa_j x_i^\top \mu_j \right) + \sum_{j=1}^{K} - \log \kappa_j - \frac{(\log \kappa_j - m)^2}{2\sigma^2} \right] + \text{const}$

$= \mathbb{E}_{q \setminus \kappa_k} \left[ \sum_{i=1}^{N} z_{ik} \left( \log C_D(\kappa_k) + \kappa_k x_i^\top \mu_k \right) - \log \kappa_k - \frac{(\log \kappa_k - m)^2}{2\sigma^2} \right] + \text{const}$

$= \sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \left( \log C_D(\kappa_k) + \kappa_k x_i^\top \mathbb{E}_q[\mu_k] \right) - \log \kappa_k - \frac{(\log \kappa_k - m)^2}{2\sigma^2} + \text{const}$

$\implies q^*(\kappa_k) \not\sim \log\text{Normal} \quad \text{due to the existence of } \log C_D(\kappa_k)$
Intermediate Quantities

Some intermediate quantities are in closed-form

- \( q(z_i) \equiv \text{Multi}(z_i|\lambda_i) \implies E_q[z_{ij}] = \lambda_{ij} \)
- \( q(\pi) \equiv \text{Dirichlet}(\pi|\rho) \implies E_q \log \pi_k = \Psi(\rho_k) - \Psi\left(\sum_j \rho_j\right) \)
- \( q(\mu_k) \equiv \text{vMF}(\mu_k|\psi_k,\gamma_k) \implies E_q[\mu_k] = \frac{I_D(\gamma_k)}{I_D^2 - 1(\gamma_k)} \psi_k \)

[Rothenbuehler, 2005]

Some are not— \( E_q[\kappa_k] \) and \( E_q \log CD(\kappa_k) \)

1. the absence of a good parametric form of \( q(\kappa_k) \)
   - apply sampling
2. even if \( \kappa_k \sim \text{logNormal} \) is assumed, \( E_q \log CD(\kappa_k) \) is still hard to deal with
   - bound \( \log CD(\cdot) \) by some simple functions

\(^1\)can be derived from the characteristic function of \( \text{vMF} \)
Sampling

In principle we can sample $\kappa_k$ from $p(\kappa_k | \mathbf{X}, \theta)$.

Unfortunately, the sampling procedure above requires the samples of $z_i, \mu_k, \pi, \ldots$ which are not maintained by variational inference.

Recall the optimal posterior for $\kappa_k$ satisfies

$$
\log q^* (\kappa_k) = \sum_{i=1}^{N} \mathbb{E} [z_{ik}] (\log C_D (\kappa_k) + \kappa_k \mathbf{x}_i^\top \mathbb{E}_q [\mu_k]) - \log \kappa_k - \frac{(\log \kappa_k - m)^2}{2\sigma^2} + \text{const}
$$

$$
\Rightarrow q^* (\kappa_k) \propto \exp \left( \sum_{i=1}^{N} \mathbb{E} [z_{ik}] (\log C_D (\kappa_k) + \kappa_k \mathbf{x}_i^\top \mathbb{E}_q [\mu_k]) \right) \\
\times \text{logNormal} (\kappa_k | m, \sigma^2)
$$

We can sample from $q^* (\kappa_k)$!

---

$^2$see derivation on p.8
Bounding

Outline

- Assume $q(\kappa_k) \equiv \text{logNormal}(\cdot | a_k, b_k)$
- Lower bound $\mathbb{E}_q \log C_D(\kappa_k)$ in VLB by some simple terms
- To optimize $q(\kappa_k)$, use gradient ascent w.r.t $a_k$ and $b_k$ to raise the VLB

Empirically, sampling outperforms bounding
Empirical Bayes for Hyperparameters

Raise $\text{VLB} \left( q, \theta \right)$ by coordinate ascent

1. $q^{t+1} = \arg\max q = \prod_{i=1}^{M} q_i$ $\text{VLB} \left( q, \theta^t \right)$

2. $\theta^{t+1} = \arg\max_{\theta} \text{VLB} \left( q^{t+1}, \theta \right) = \arg\max_{\theta} \mathbb{E}_{q^{t+1}} \log p \left( X, Z | \theta \right)$

For example, one can use gradient ascent to optimize $\alpha$

$$\max_{\alpha > 0} - \log B \left( \alpha \right) + (\alpha - 1) \sum_{k=1}^{K} \mathbb{E}_{q^{t+1}} \left[ \log \pi_k \right]$$

$m, \sigma^2, \mu_0$ and $C_0$ can be optimized in a similar manner $^3$

$^3$Unlike $\alpha$, their solutions can be written in closed-form
