Variational Inference for Bayes vMF Mixture

Hanxiao Liu

September 23, 2014

Variational Inference Review

Lower bound the likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \mathbb{E}_q \; \log p(\boldsymbol{X}|\boldsymbol{\theta})$$
$$= \underbrace{\mathbb{E}_q\left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right]}_{\mathrm{VLB}(q,\boldsymbol{\theta})} + \underbrace{\mathbb{E}_q\left[\log \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}\right]}_{D_{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}))}$$

Raise $\mathrm{VLB}(q, \boldsymbol{\theta})$ by coordinate ascent

1. $q^{t+1} = \underset{q=\prod_{i=1}^{M} q_i}{\operatorname{argmax}} \; \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$
2. $\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathrm{VLB}\left(q^{t+1}, \boldsymbol{\theta}\right)$

Variational Inference Review

Goal: solve $\underset{q=\prod_{i=1}^{M} q_i}{\operatorname{argmax}} \ \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$ by coordinate ascent, i.e. sequentially updating a single q_i in each iteration.

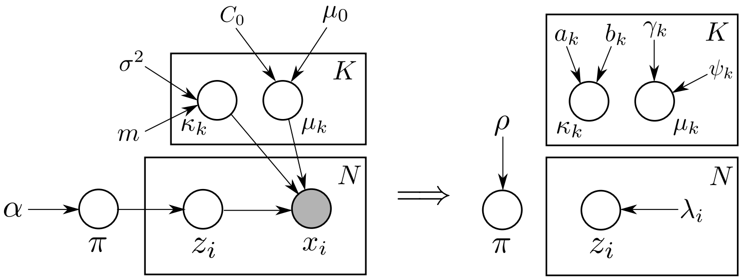
Each coordinate step has a closed-form solution—

$$\begin{aligned}
\mathrm{VLB}\left(q_j; q_{-j}, \boldsymbol{\theta}^t\right) &= \mathbb{E}_q\left[\log \frac{p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^t\right)}{q\left(\boldsymbol{Z}\right)}\right] \\
&= \mathbb{E}_q \log p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^t\right) - \sum_{i=1}^{M} \mathbb{E}_q \log q_i \\
&= \mathbb{E}_{q_j} \underbrace{\mathbb{E}_{q_{-j}} \log p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^t\right)}_{\log \tilde{q}_j + const} - \mathbb{E}_{q_j} \log q_j + const \\
&= \int q_j \log \frac{\tilde{q}_j}{q_j} + const = -D_{KL}\left(q_j || \tilde{q}_j\right) + const
\end{aligned}$$

$$\implies \log q_j^* = \mathbb{E}_{q_{-j}} \log p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^t\right) + const$$

Bayes vMF Mixture

[Gopal and Yang, 2014]



- $\boldsymbol{\pi} \sim \text{Dirichlet}\left(\cdot|\alpha\right)$
- $\boldsymbol{\mu}_k \sim \text{vMF}\left(\cdot|\boldsymbol{\mu}_0, C_0\right)$
- $\kappa_k \sim \text{logNormal}\left(\cdot|m, \sigma^2\right)$
- $\boldsymbol{z}_i \sim \text{Multi}\left(\cdot|\boldsymbol{\pi}\right)$
- $\boldsymbol{x}_i \sim \text{vMF}\left(\cdot|\boldsymbol{\mu}_{z_i}, \kappa_{z_i}\right)$

- $q\left(\boldsymbol{\pi}\right) \stackrel{?}{\equiv} \text{Dirichlet}\left(\cdot|\boldsymbol{\rho}\right)$
- $q\left(\boldsymbol{\mu}_k\right) \stackrel{?}{\equiv} \text{vMF}\left(\cdot|\boldsymbol{\psi}_k, \gamma_k\right)$
- $q\left(\kappa_k\right) \stackrel{?}{\equiv} \text{logNormal}\left(\cdot|a_k, b_k\right)$
- $q\left(\boldsymbol{z}_i\right) \stackrel{?}{\equiv} \text{Multi}\left(\cdot|\boldsymbol{\lambda}_i\right)$

Compute $\log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta})$

$$p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta}) = \text{Dirichlet}(\boldsymbol{\pi} | \alpha) \times \prod_{i=1}^{N} \text{Multi}(z_i | \boldsymbol{\pi}) \, \text{vMF}(\boldsymbol{x}_i | \boldsymbol{\mu}_{z_i}, \kappa_{z_i}) \times \prod_{k=1}^{K} \text{vMF}(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, C_0) \, \text{logNormal}(\kappa_k | m, \sigma^2)$$

$$\begin{aligned}\log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta}) =& - \log B(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \log \pi_k \\ &+ \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left(\log C_D(\kappa_k) + \kappa_k \boldsymbol{x}_i^\top \boldsymbol{\mu}_k \right) \\ &+ \sum_{k=1}^{K} \left(\log C_D(C_0) + C_0 \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_0 \right) \\ &+ \sum_{k=1}^{K} \left(- \log \kappa_k - \frac{1}{2} \log(2\pi\sigma^2) - \frac{(\log \kappa_k - m)^2}{2\sigma^2} \right)\end{aligned}$$

Updating $q(\boldsymbol{\pi})$

$$q(\boldsymbol{\pi}) \stackrel{?}{\equiv} \text{Dirichlet}(\cdot|\boldsymbol{\rho})$$

$$\begin{aligned}
\log q^*(\boldsymbol{\pi}) &= \mathbb{E}_{q\backslash\boldsymbol{\pi}} \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + const \\
&= \mathbb{E}_{q\backslash\boldsymbol{\pi}} \left[\sum_{k=1}^{K} (\alpha - 1) \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k\right] + const \\
&= \sum_{k=1}^{K} \left(\alpha + \sum_{i=1}^{N} \mathbb{E}_q [z_{ik}] - 1\right) \log \pi_k + const
\end{aligned}$$

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

$$\implies \rho_k^* = \alpha + \sum_{i=1}^{N} \mathbb{E}_q [z_{ik}]$$

Updating $q(\boldsymbol{z}_i)$

$$q(\boldsymbol{z}_i) \stackrel{?}{\equiv} \text{Multi}(\cdot|\boldsymbol{\lambda}_i)$$

$$\begin{aligned}
&\log q^*(\boldsymbol{z}_i) \\
&= \mathbb{E}_{q\backslash z_i} \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + const \\
&= \mathbb{E}_{q\backslash z_i} \left[\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left(\log C_D(\kappa_k) + \kappa_k \boldsymbol{x}_i^\top \boldsymbol{\mu}_k \right) \right] + const \\
&= \sum_{k=1}^{K} z_{ik} \left(\mathbb{E}_q \log \pi_k + \mathbb{E}_q \log C_D(\kappa_k) + \mathbb{E}_q[\kappa_k] \boldsymbol{x}_i^\top \mathbb{E}_q[\boldsymbol{\mu}_k] \right) + const
\end{aligned}$$

$$\implies q^*(\boldsymbol{z}_i) \sim \text{Multi}, \ \lambda_{ik}^* \propto e^{\mathbb{E}_q \log \pi_k + \mathbb{E}_q \log C_D(\kappa_k) + \mathbb{E}_q[\kappa_k] \boldsymbol{x}_i^\top \mathbb{E}_q[\boldsymbol{\mu}_k]}$$

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

Updating $q(\boldsymbol{\mu}_k)$

$$q(\boldsymbol{\mu}_k) \overset{?}{\equiv} \mathrm{vMF}(\cdot|\boldsymbol{\psi}_k, \gamma_k)$$

$$\begin{aligned}
\log q^*(\boldsymbol{\mu}_k) &= \mathbb{E}_{q\setminus\boldsymbol{\mu}_k} \log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + const \\
&= \mathbb{E}_{q\setminus\boldsymbol{\mu}_k} \left[\sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij}\kappa_j \boldsymbol{x}_i^\top \boldsymbol{\mu}_j + \sum_{j=1}^{K} C_0 \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_0 \right] + const \\
&= \mathbb{E}_q[\kappa_k] \left(\sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \boldsymbol{x}_i^\top \boldsymbol{\mu}_k \right) + C_0 \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_0 + const
\end{aligned}$$

$$\implies q^*(\boldsymbol{\mu}_k) \propto e^{\left[\mathbb{E}_q[\kappa_k]\left(\sum_{i=1}^{N} \mathbb{E}_q[z_{ik}]\boldsymbol{x}_i\right) + C_0\boldsymbol{\mu}_0\right]^\top \boldsymbol{\mu}_k} \sim \mathrm{vMF}$$

$$\gamma_k^* = \left\| \mathbb{E}_q[\kappa_k] \left(\sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \boldsymbol{x}_i \right) + C_0 \boldsymbol{\mu}_0 \right\|, \quad \boldsymbol{\psi}_k^* = \frac{\mathbb{E}_q[\kappa_k]\left(\sum_{i=1}^{N} \mathbb{E}_q[z_{ik}] \boldsymbol{x}_i\right) + C_0 \boldsymbol{\mu}_0}{\gamma_k}$$

Updating $q(\kappa_k)$

$$q(\kappa_k) \stackrel{?}{\equiv} \text{logNormal}(\cdot | a_k, b_k)$$

$$\begin{aligned}
&\log q^*(\kappa_k) \\
&= \mathbb{E}_{q \setminus \kappa_k} \log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta}) + const \\
&= \mathbb{E}_{q \setminus \kappa_k} \left[\sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \left(\log C_D(\kappa_j) + \kappa_j \boldsymbol{x}_i^\top \boldsymbol{\mu}_j \right) + \sum_{j=1}^{K} - \log \kappa_j - \frac{(\log \kappa_j - m)^2}{2\sigma^2} \right] + const \\
&= \mathbb{E}_{q \setminus \kappa_k} \left[\sum_{i=1}^{N} z_{ik} \left(\log C_D(\kappa_k) + \kappa_k \boldsymbol{x}_i^\top \boldsymbol{\mu}_k \right) - \log \kappa_k - \frac{(\log \kappa_k - m)^2}{2\sigma^2} \right] + const \\
&= \sum_{i=1}^{N} \mathbb{E}_q [z_{ik}] \left(\log C_D(\kappa_k) + \kappa_k \boldsymbol{x}_i^\top \mathbb{E}_q [\boldsymbol{\mu}_k] \right) - \log \kappa_k - \frac{(\log \kappa_k - m)^2}{2\sigma^2} + const
\end{aligned}$$

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

Intermediate Quantities

Some intermediate quantities are in closed-form

- $q(\boldsymbol{z}_i) \equiv \text{Multi}(\boldsymbol{z}_i|\boldsymbol{\lambda}_i) \implies \mathbb{E}_q[z_{ij}] = \lambda_{ij}$
- $q(\boldsymbol{\pi}) \equiv \text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\rho}) \implies \mathbb{E}_q \log \pi_k = \Psi(\rho_k) - \Psi\left(\sum_j \rho_j\right)$
- $q(\boldsymbol{\mu}_k) \equiv \text{vMF}(\boldsymbol{\mu}_k|\boldsymbol{\psi}_k, \gamma_k) \implies \mathbb{E}_q[\boldsymbol{\mu}_k] = \frac{I_{\frac{D}{2}}(\gamma_k)}{I_{\frac{D}{2}-1}(\gamma_k)}\boldsymbol{\psi}_k$ [1] [Rothenbuehler, 2005]

Some are not— $\mathbb{E}_q[\kappa_k]$ and $\mathbb{E}_q \log C_D(\kappa_k)$

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

[1] can be derived from the characteristic function of vMF

Sampling

In principle we can sample κ_k from $p\left(\kappa_k | \boldsymbol{X}, \boldsymbol{\theta}\right)$.

Unfortunately, the sampling procedure above requires the samples of $\boldsymbol{z}_i, \boldsymbol{\mu}_k, \boldsymbol{\pi}, \ldots$ which are not maintained by variational inference.

Recall the optimal posterior for κ_k satisfies [2]

$$\begin{aligned} & \log q^* \left(\kappa_k\right) \\ = & \sum_{i=1}^{N} \mathbb{E}\left[z_{ik}\right] \left(\log C_D \left(\kappa_k\right) + \kappa_k \boldsymbol{x}_i^\top \mathbb{E}_q \left[\boldsymbol{\mu}_k\right]\right) - \log \kappa_k - \frac{\left(\log \kappa_k - m\right)^2}{2\sigma^2} + const \\ \implies & q^* \left(\kappa_k\right) \propto \exp\left(\sum_{i=1}^{N} \mathbb{E}\left[z_{ik}\right] \left(\log C_D \left(\kappa_k\right) + \kappa_k \boldsymbol{x}_i^\top \mathbb{E}_q \left[\boldsymbol{\mu}_k\right]\right)\right) \\ & \qquad \times \mathrm{logNormal}\left(\kappa_k | m, \sigma^2\right) \end{aligned}$$

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

[2] see derivation on p.8

Bounding

Outline

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

Empirically, sampling outperforms bounding

Empirical Bayes for Hyperparameters

Raise $\mathrm{VLB}\,(q, \boldsymbol{\theta})$ by coordinate ascent

1. $q^{t+1} = \underset{q=\prod_{i=1}^{M} q_i}{\operatorname{argmax}} \; \mathrm{VLB}\left(q, \boldsymbol{\theta}^t\right)$
2. $\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} \; \mathrm{VLB}\,(q^{t+1}, \boldsymbol{\theta})$
   $= \operatorname{argmax}_{\boldsymbol{\theta}} \; \mathbb{E}_{q^{t+1}} \log p\,(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta})$

For example, one can use gradient ascent to optimize α

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

m, σ^2, $\boldsymbol{\mu}_0$ and C_0 can be optimized in a similar manner [3]

[3] Unlike α, their solutions can be written in closed-form

Reference I

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005).
Clustering on the unit hypersphere using von mises-fisher distributions.
In *Journal of Machine Learning Research*, pages 1345–1382.

Gopal, S. and Yang, Y. (2014).
Von mises-fisher clustering models.
In *Proceedings of The 31st International Conference on Machine Learning*, pages 154–162.

Rothenbuehler, J. (2005).
Dependence Structures beyond copulas: A new model of a multivariate regular varying distribution based on a finite von Mises-Fisher mixture model.
PhD thesis, Cornell University.