Algorithm for online variational Bayes

November 26, 2023

Here we provide an overview of the model and outline briefly both the theory of behind the online algorithm and the algorithm itself.

The basic model framework

We consider a directed graph $G = (V, E)$, with V being the node set of size n and E the edge set. We assume the graph is fully connected and simple, so that $|E| = n(n-1)$, and we write e_{ij} to denote the edge $i \to j$. Consider the setting in which we have a latent group structure on the network, and write $\mathcal{G} := [K] := \{1, 2, \cdots, d\}$ to denote the set of groups. The group identity of node $i \in V$ is denoted by $z_i \in \{0,1\}^d$, where $z_{ik} = 1$ if node i belongs to groups k, and 0 otherwise, with $z = [z_1 \cdots z_n] \in \{0,1\}^{n \times d}$ the matrix storing these memberships. We will write $\pi = (\pi_1, \cdots, \pi_d) \in [0,1]^d$ to be the vector storing the global group membership probabilities; namely, such that π_k is the global probability of a node belonging to group k, and we clearly have $\sum_{k=1}^d \pi_k = 1$. In this way, we assume that the latent group memberships $\{z_i\}_{i=1}^n$ are independent and identically distributed (iid) according to a categorical distribution with probability vector π:

$$z_i | \pi \overset{iid}{\sim} \text{Cat}(\pi)$$

so that $p(z_i) = \prod_k \pi_k^{z_{ik}}$.

On e_{ij} there lives a point process with rate determined by the group memberships of the enclosing nodes. The point processes are independent of one another given $\{z_i\}_{i=1}^N$. Knowing that $z_{ik} = 1$ and $z_{jm} = 1$, we know that the counting process on e_{ij}, $\{x_{ij}(t), t \in [0, \infty)\}$, writing X_{ij} for short, is distributed as $f(\cdot; \theta_{km}) := f_{km}(\cdot)$, where f_{km} is a probability distribution known up to a finite-dimensional parameter θ_{km}:

$$x_{ij} | z_{ik} z_{jm} = 1 \sim f_{km}(\cdot)$$

For our purposes, we take the point process on e_{ij} to be an homogeneous Poisson process with intensity $\lambda_{z_i z_j}$:

$$x_{ij} | z_{ik} z_{jm} = 1 \sim \mathcal{PP}(\lambda_{km}),$$

where $\mathcal{PP}(\alpha)$ denotes a Poisson process with rate α. We further write $\lambda = [\lambda_{km}]_{k,m=1}^K \in \mathbb{R}_+^{d \times d}$ to be a matrix of rates and $x(t) = [x_{ij}(t)]_{i,j=1}^N \in \mathbb{N}^{n \times n}$ to be the matrix of counting processes, where $x_{ii}(t) = 0$ for all $i = 1, \cdots, n$ and $t \geq 0$ as our network is simple.

A Bayesian implementation

We adopt a Bayesian framework, and place priors $p(\theta)$ and $p(z|\theta)$ on our latent variables. Our goal is then to infer the posterior distribution at time t, namely

$$p(\theta, z | x(t)) \propto p(\theta) \left(\prod_{i=1}^N p(z_i | \theta) \right) \left(\prod_{i \neq j}^N p(x_{ij}(t) | z_i, z_j, \theta) \right)$$

This posterior is intractable, and so we consider a variational approach to approximate, which is referred to as variational Bayes (VB). Here we focus on mean-field variational inference, where we posit a family of factorisable distributions over the latent variables, and look to select the member of this family that is closest to the posterior in the sense of the KL-divergence. Specifically, we posit the family $\mathcal{Q}^{n+d}$, which takes the form

$$\mathcal{Q}^{n+d} = \left\{ q : q(\theta, z) = \prod_{i=1}^{n} q_{z_i}(z_i) \prod_{k=1}^{d} q_{\theta_k}(\theta_k) \right\}$$

and look to compute

$$q^*(\theta, z) = \underset{q(\theta,z)\in\mathcal{Q}^{n+d}}{\arg\min} \, \mathrm{KL}\left(q(\theta, z)||p(\theta, z|x)\right) \tag{1}$$

In applications, this problem is often solved by instead maximising the evidence lower bound (ELBO),

$$\mathrm{ELBO}(q) = \mathbb{E}_q \left\{ \log \frac{p(x, \theta, z)}{q(\theta, z)} \right\}$$

over $q(\theta, z) \in \mathcal{Q}^{n+d}$.

Approximating $q^*(\theta, z)$

Unfortunately, the global minimum of $\mathrm{ELBO}(q)$ is often itself intractable, and thus a further approximation must be made. To address the intractability of $q^*(\theta, z)$, we use the well-known Coordinate Ascent Variational Inference (CAVI) algorithm (David M. Blei and McAuliffe, 2017), which is summarised in Algorithm 1, with the notation $\mathbb{E}_{-\theta_i}$ being used to denote the expectation with respect to all components of θ except θ_i. CAVI is only guaranteed to achieve a global minimum however, and is sensitive to intialisation (Zhang and Zhou, 2020). This provides two layers of approximation, provided by the following diagram:

$$p(\theta, z|x) \overset{\leftarrow}{\approx} q^*(\theta, z) \overset{\leftarrow}{\approx} \hat{q}^{\mathrm{CAVI}}(\theta, z)$$

Algorithm 1 CAVI

Initialise parameter $\phi = (\phi_1, \cdots, \phi_m)$.
while ϕ not converged **do**
  for $i = 1, \cdots, m$ **do**
    Update $q(\phi_i)$ as $q(\phi_i) \propto \exp\left\{\mathbb{E}_{-\phi_i}\left(\log p(x, \phi)\right)\right\}$, holding $\{q(\phi_j)\}_{j\neq i}$ fixed.
  end for
end while

Computing the expectations in Algorithm 1 is an easy task in the problem we are considering if we choose sensible, namely conjugate, priors. Since $z_i|\pi$ is categorical for each i, we specify a Dirichlet prior on π:

$$\pi|n^0 \sim \mathrm{Dirichlet}\left(n^0\right).$$

On λ_{km} we place a gamma prior:

$$\lambda_{km}|\alpha^0_{km}, \beta^0_{km} \sim \mathrm{Gamma}\left(\alpha^0_{km}, \beta^0_{km}\right)$$

independently for each k, m pair.

An online implementation

Suppose that we observe data on an interval $[0, T]$ in an online manner. We will be unable to compute an update upon every arrival, and so we instead consider updating our parameter estimates at every $t = 1$ time units, so that we get a sequence of update points $\mathcal{S} = \{1, 2, \cdots, \lfloor T \rfloor\}$. Consider an update at $t = r \in \mathcal{S}$. The logical approach is to set the priors at this update point to be the posterior of the previous step. As we have chosen conjugate priors, the posteriors from the the run at $t = r - 1$ will be of the same functional form as the initial prior and thus we can run an iterative update. Specifically, using a superscript of $(r-1)$ to signify that the parameters are the parameters from the posterior of $(r-1)$st update, we find that the posterior parameters at $t = r$ take the form given in Equations (2) - (4), where we write $x^{(r)}$ to be the point process on the unit time interval preceeding the rth update.

$$
\begin{aligned}
q(Z_p) &= \text{Cat}\left(\tau_p^{(r)}\right);\ p = 1, \cdots, n \\
&\quad \text{where the } \left\{\tau_i^{(r)}\right\}_{i=1}^{N} \text{ satisfy:} \\
&\quad \tau_{pk}^{(r)} \propto \exp\left\{\psi\left(n_k^{(r)}\right) - \psi\left(\sum_\ell n_\ell^{(r)}\right)\right. \\
&\qquad + \sum_{i \neq p}\sum_m \tau_{im}^{(r)}\left[x_{pi}^{(r)}\left(\psi\left(\alpha_{km}^{(r)}\right) - \log\left(\beta_{km}^{(r)}\right)\right) + x_{ip}^{(r)}\left(\psi\left(\alpha_{mk}^{(r)}\right) - \log\left(\beta_{mk}^{(r)}\right)\right)\right. \\
&\qquad \left.\left. - \left(\frac{\alpha_{km}^{(r)}}{\beta_{km}^{(r)}} + \frac{\alpha_{mk}^{(r)}}{\beta_{mk}^{(r)}}\right)\right]\right\} \qquad (2)
\end{aligned}
$$

$$
q(\pi) = \text{Dirichlet}\left(n^{(r)}\right);\ n_k^{(r)} = n_k^{(r-1)} + \sum_i \tau_{ik}^{(r)} \text{ where } k = 1, \cdots, d \qquad (3)
$$

$$
q(\lambda_{km}) = \text{Gamma}\left(\alpha_{km}^{(r-1)} + \sum_{i \neq j} \tau_{ik}^{(r)}\tau_{jm}^{(r)}x_{ij}^{(r)}, \beta_{km}^{(r-1)} + \sum_{i \neq j}\tau_{ik}^{(r)}\tau_{jm}^{(r)}\right);\ k, m = 1, \cdots, d \qquad (4)
$$

Note that the CAVI algorithm is iterative. In particular, Equations (2) - (4) are cycled through until a convergence criterion is met.

Adapting to include changepoints

Now we suppose that we adjust our model to allow for change points. Specifically, we assume that at random times in our observation window $[0, T]$ node i undergoes a change in latent group membership, i.e. $z_i \mapsto z_i^*$ with $z_i \neq z_i^*$.

As we sequentially run the update procedure, our priors will converge toward the underlying parameter values. However, if we want our algorithm to detect change-points, we need some methodology to down-weight the prior so that we can react quickly to the presence of data that signals a change in the latent structure.

To this end, rather than directly passing the posterior from the update at $t = s_{r-1}$, $p^{(r-1)}(\theta|X)$, as the prior for the update at $t = s_r$, we instead set the prior $p^{(r)}(\theta, z) \propto p^{(r-1)}(\theta, z|x)^\delta$, where $\delta \in (0, 1)$. We think of the posterior from the rth update as the posterior of the data observed between the $(r-1)$st and rth update under a prior that provides a specified "amount" of the information ascertained from the previous runs. Under this new

prior, the updates take the form of Equations (5) - (7).

$$
\begin{aligned}
q(Z_p) &= \text{Cat}\left(\tau_p^{(r)}\right);\ p = 1, \cdots, N \\
&\quad \text{where the } \left\{\tau_i^{(r)}\right\}_{i=1}^{N} \text{ satisfy:} \\
&\quad \tau_{pk}^{(r)} \propto \exp\left\{\delta\left(\psi\left(n_k^{(r)}\right) - \psi\left(\sum_\ell n_\ell^{(r)}\right)\right)\right. \\
&\qquad + \sum_{i\neq p}\sum_m \tau_{im}^{(r)}\left[x_{pi}^{(r)}\left(\psi\left(\alpha_{km}^{(r)}\right) - \log\left(\beta_{km}^{(r)}\right)\right) + x_{ip}^{(r)}\left(\psi\left(\alpha_{mk}^{(r)}\right) - \log\left(\beta_{mk}^{(r)}\right)\right)\right. \\
&\qquad \left.\left. - \left(\frac{\alpha_{km}^{(r)}}{\beta_{km}^{(r)}} + \frac{\alpha_{mk}^{(r)}}{\beta_{mk}^{(r)}}\right)\right]\right\} \qquad (5)
\end{aligned}
$$

$$
q(\pi) = \text{Dirichlet}\left(n^{(r)}\right);\ n_k^{(r)} = \delta\left(n_k^{(r-1)} - 1 + \sum_i \tau_{ik}^{(r)}\right) + 1 \text{ where } k = 1, \cdots, K \qquad (6)
$$

$$
q(\lambda_{km}) = \text{Gamma}\left(\delta\alpha_{km}^{(r-1)} + \sum_{i\neq j}\tau_{ik}^{(r)}\tau_{jm}^{(r)}x_{ij}^{(r)}, \delta\beta_{km}^{(r-1)} + \sum_{i\neq j}\tau_{ik}^{(r)}\tau_{jm}^{(r)}\right);\ k, m = 1, \cdots, K \qquad (7)
$$

References

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

Zhang, A. Y. and Zhou, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *aos*, 48(5):2575–2598.