Structural Expectation Propagation (SEP): Bayesian structure learning for networks with latent variables - supplementary material

Nevena Lazic
Microsoft Research

Christopher M. Bishop
Microsoft Research,
Edinburgh University

John Winn
Microsoft Research

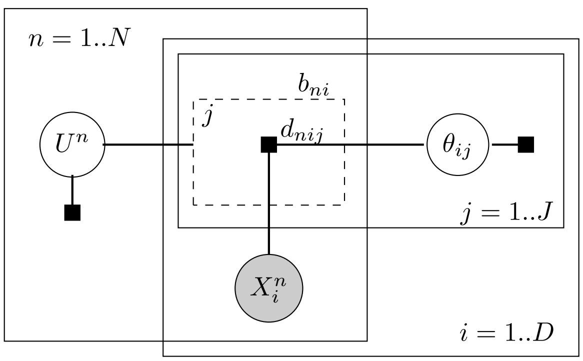

Figure 1: Gated factor graph corresponding to a discrete mixture model.

1 EP update derivations for a discrete mixture model

We first derive EP updates in a discrete mixture model where a discrete latent variable $U \in \{1, ..., J\}$ is the parent of D observed discrete variables $X_1, ..., X_D$. We indicate the values taken on by random variables either by using lower-case symbols, or by writing $X_i = x_i$. Let Θ_{ij} be the multinomial parameters for X_i conditioned on $U = j$, with Dirichlet priors $p(\theta_{ij})$, and and let $\Theta_i = \{\Theta_{i1}, ..., \Theta_{iJ}\}$. Given N observations $\mathcal{D} = \{\mathbf{x}^1, ..., \mathbf{x}^N\}$, the joint distribution is:

$$p(u^{1:N}, \theta_1, ..., \theta_D, \mathbf{x}^{1:N}) = \prod_n \prod_i p(u^n) p(\theta_i) b_{ni}(u^n, \theta_i, x_i^n) \quad (1)$$

$$b_{ni}(u^n, \theta_i m x_i^n) = \prod_j d_{nij}(\theta_{ij}, x_i^n)^{\delta(u^n = j)} \quad (2)$$

$$d_{nij}(\theta_{ij}, x_i^n) = \prod_l {\theta_{ij,l}}^{\delta(x_i^n = l)} \quad (3)$$

Under a factorized approximation, the posterior $q(u^n)$ of each latent variable is the product of the prior and messages $\mu_{ni}(u^n)$ from factors b_{ni}, $i = 1..D$:

$$q(u^n) \propto p(u^n) \prod_i \mu_{ni}(u^n) \quad (4)$$

The posterior $q(\theta_{ij})$ is the product of the prior $p(\theta_{ij})$ and messages $\tau_{nij}(\theta_{ij})$ from factors b_{ni}, $n = 1..N$:

$$q(\theta_{ij}) = p(\theta_{ij}) \prod_n \tau_{nij}(\theta_{ij}) \quad (5)$$

Let $q^{\setminus i}(u^n)$ be the posterior of u^n computed without the message $\mu_{ni}(u^n)$, and let $q^{\setminus n}(\theta_{ij})$ be the posterior of θ_{ij} computed without the message $\tau_{nij}(\theta_{ij})$.

The EP message from a factor b_{ni} to the variable U^n is:

$$\begin{aligned} \mu_{ni}(U^n = j) &\propto \sum_{\theta_i} \left(\prod_{j'} q^{\setminus n}(\theta_{ij'}) \right) d_{nij}(\theta_{ij}, x_i^n) && (6) \\ &\propto \sum_{\theta_{ij}} q^{\setminus n}(\theta_{ij}) d_{nij}(\theta_{ij}, x_i^n) && (7) \\ &\propto E_{q^{\setminus n}(\theta_{ij})}[d_{nij}(\theta_{ij}, x_i^n)] && (8) \end{aligned}$$

Expectations $E_{q(\theta)}[d(x, \theta)]$ can be computed in closed form; when the Dirichlet distribution $q(\theta)$ is parameterized by pseudocounts λ and λ_x is the pseudocount indexed by x, $E_{q(\theta)}[d(x, \theta)]$ evaluates to:

$$E_{q(\theta)}[d(x, \theta)] = \frac{\Gamma(\lambda_0)}{\Gamma(\lambda_0 + 1)} \frac{\Gamma(1 + \lambda_x)}{\Gamma(\lambda_x)} = \frac{\lambda_x}{\lambda_0}. \quad (9)$$

The EP message from b_{ni} to θ_{ij} is:

$$\tau_{nij}(\theta_{ij}) = \frac{\text{proj}[\sum_{j'} q^{\setminus i}(U^n = j') r_{nij'}(\theta_{ij})]}{q^{\setminus n}(\theta_{ij})} \quad (10)$$

The quantities $r_{nij'}(\theta_{ij}$ can be computed by considering the cases $j = j'$ and $j \neq j'$ separately:

$$\begin{aligned} r_{nij'}(\theta_{ij}) &= q^{\setminus n}(\theta_{ij}) \sum_{\theta_i \setminus \theta_{ij}} \left(\prod_{j^* \neq j} q^{\setminus n}(\theta_{ij^*}) \right) d_{nij'}(\theta_{ij'}, x_i^n) \\ &= \begin{cases} q^{\setminus n}(\theta_{ij}) d_{nij}(x_i^n, \theta_{ij}) & \text{if } j = j' \\ q^{\setminus n}(\theta_{ij}) E_{q^{\setminus n}(\theta_{ij'})}[d_{nij'}(\theta_{ij'}, x_i^n)] & \text{if } j \neq j'. \end{cases} \end{aligned} \quad (11)$$

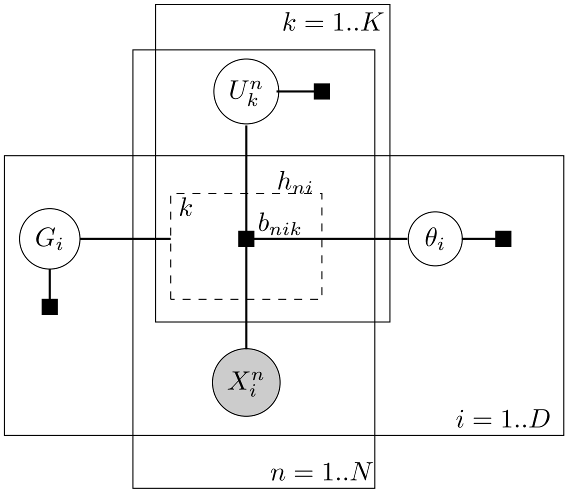

Figure 2: Gated factor graph for learning the structure of networks in which each observed variable X_i is the child of a single latent variable U_k

2 EP update derivations for a single-parent network with latent structure

We now derive EP updates for a network in which there are K latent variables and D observed variables. Each observed variable is the child of a single latent variable (so there are up to K mixture models), but the networkl structure is otherwise unknown. Let $\mathbf{X} = \{X_1, ..., X_D\}$ and $\mathbf{U} = \{U_1, ..., U_K\}$ be the observed and latent variables, respectively. Let $G_i \in \{1, ..., K\}$ be a latent variable indicating the parent of X_i in the graph, and $\mathbf{G} = \{G_1, ..., G_D\}$. Let Θ_{ij} be the parameters for the conditional probability of X_i given that its parent variable takes on the value j. Given N observations $\mathcal{D} = \{\mathbf{x}^1, ..., \mathbf{x}^N\}$, the posterior over the latent structure, variables and parameters is:

$$p(\mathbf{u}^{1:N}, \theta, \mathbf{g}, \mathbf{x}^{1:N}) = \prod_n \prod_i p(\mathbf{u}^n) p(g_i) p(\theta_i) h_{ni}(g_i, \mathbf{u}^n, \theta_i, x_i^n) \quad (12)$$

$$h_{ni}(g_i, \mathbf{u}^n, \theta_i, x_i^n) = \prod_k b_{nik}(u_k^n, \theta_i, x_i^n)^{\delta(g_i=k)} \quad (13)$$

$$b_{nik}(u_k^n, \theta_i, x_i^n) = \prod_j d_{nij}(\theta_{ij}, x_i^n)^{\delta(u_k^n=j)} \quad (14)$$

$$d_{nij}(\theta_{ij}, x_i^n) = \prod_l \theta_{ij,l}{}^{\delta(x_i^n=l)} \quad (15)$$

The model is shown in Figure 2, where we have collapsed the discrete mixture factors $b_{nik}(u_k^n, \theta_i | x_i^n)$ for clarity.

We approximate the posterior by a factorized variational distribution $q(\mathbf{g})q(\mathbf{u})q(\theta)$, and we fit the parameters using expectation propagation. The posterior $q(G_i = k)$ over each structure variable is the product of the prior and messages $\gamma_{ni}(G_i = k)$, $n = 1, ..., N$ from factors h_{ni}:

$$q(G_i = k) \propto p(G_i = k) \prod_n \gamma_{ni}(G_i = k) \quad (16)$$

The posterior of each latent variable $q(u_k^n)$ is the product of the prior and messages $\nu_{nik}(u_k^n)$ from the factors h_{ni}, $i = 1..D$.

$$q(u_k^n) \propto p(u_k^n) \prod_i \nu_{nik}(u_k^n) \quad (17)$$

Each parameter posterior $q(\theta_{ij})$ is the product of the prior and messages $\rho_{nij}(\theta_{ij})$ from factors h_{ni}, $n = 1..N$:

$$q(\theta_{ij}) = p(\theta_{ij}) \prod_n \rho_{nij}(\theta_{ij}) \quad (18)$$

We denote by $q^{\backslash i}(x)$ the approximate posterior of a variable x after removing the message indexed by i.

2.1 Messages from h_{ni} to G_i

The posterior $q(G_i = k)$ over each structure variable is the product of the prior and the messages $\gamma_{ni}(G_i = k)$, $n = 1..N$ from factors h_{ni}:

$$\begin{aligned} &\gamma_{ni}(G_i = k) \\ &\propto \sum_j \sum_{\theta_i} q^{\backslash i}(U_k^n = j) \Big(\prod_{j'} q^{\backslash n}(\theta_{ij'}) \Big) b_{nik}(j, \theta_i, x_i^n) \\ &\propto \sum_j q^{\backslash i}(U_k^n = j) \sum_{\theta_{ij}} q^{\backslash n}(\theta_{ij}) d_{nij}(\theta_{ij}, x_i^n) \quad (19) \\ &\propto \sum_j q^{\backslash i}(U_k^n = j) E_{q^{\backslash n}(\theta_{ij})}[d_{nij}(x_i^n, \theta_{ij})]. \quad (20) \end{aligned}$$

2.2 Messages from h_{ni} to U_k

The posterior of each latent variable $q(u_k^n)$ is the product of the prior and the messages $\nu_{nik}(u_k^n)$ from factors $h_{ni}(g_i, \mathbf{u}^n, \theta_i | x_i^n)$, $i = 1..D$.

$$\nu_{nik}(u_k^n) \propto \frac{\sum_{k'} q^{\backslash n}(G_i = k') r_{nik'}(u_k^n)}{q^{\backslash i}(u_k^n)} \quad (21)$$

$$\begin{aligned} r_{nik'}(u_k^n) = q^{\backslash i}(u_k^n) \sum_{\mathbf{u}^n \backslash u_k^n} \sum_{\theta_i} \Big(\prod_{k^* \neq k} q^{\backslash i}(u_{k^*}^n) \Big) \\ \times \Big(\prod_{j'} q^{\backslash n}(\theta_{ij'}) \Big) b_{nik'}(u_{k'}^n, \theta_i, x_i^n) \quad (22) \end{aligned}$$

When $k' \neq k$,

$$r_{nik'}(U_k^n = j) = q^{\backslash i}(U_k^n = j) \times \sum_{j'} q^{\backslash i}(U_{k'}^n = j') E_{q^{\backslash n}(\theta_{ij'})}[d_{nij'}(\theta_{ij'}|x_i^n)] \quad (23)$$

When $k' = k$,

$$\begin{aligned} r_{nik'}(U_k^n = j) &= q^{\backslash i}(U_k^n = j) \sum_{\theta_i} b_{nik}(j, \theta_i, x_i^n) \\ &= q^{\backslash i}(U_k^n = j) E_{q^{\backslash n}(\theta_{ij})}[d_{nij}(\theta_{ij}, x_i^n)] \end{aligned} \quad (24)$$

2.3 Messages from h_{ni} to θ_{ij}

Each parameter posterior distribution $q(\theta_{ij})$ is computed as the product of the prior and the messages $\rho_{nij}(\theta_{ij})$ from factors $h_{ni}(g_i, \mathbf{u}^n, \theta_i, x_i^n)$, $n = 1..N$:

$$\rho_{nij}(\theta_{ij}) = \frac{\mathrm{proj}[\sum_k q^{\backslash n}(G_i = k) s_{nijk}(\theta_{ij})]}{q^{\backslash n}(\theta_{ij})} \quad (25)$$

The message $\rho_{nij}(\theta_{ij})$ is a weighted average of Dirichlet messages, projected onto a Dirichlet distribution with matching moments. The terms $s_{nijk}(\theta_{ij})$ are EP messages in a discrete mixture model where U_k is the parent of X_i. Each $s_{nijk}(\theta_{ij})$ is a moment-matched weighted average two Dirichlet distributions, for the two cases where $U_k^n = j$ and $U_k^n \neq j$:

$$s_{nijk}(\theta_{ij}) = \mathrm{proj}[\hat{s}_{nijk}(\theta_{ij})] \quad (26)$$

$$\hat{s}_{nijk}(\theta_{ij}) = q^{\backslash n}(\theta_{ij}) \sum_{u_k^n} \sum_{\theta_i \backslash \theta_{ij}} q^{\backslash i}(u_k^n) \times \left(\prod_{j^* \neq j} q^{\backslash n}(\theta_{ij^*}) \right) b_{nik}(u_k^n, \theta_i, x_i^n) \quad (27)$$

$$= q^{\backslash n}(\theta_{ij}) q^{\backslash i}(U_k^n = j) d_{nij}(\theta_{ij}, x_i^n) + q^{\backslash n}(\theta_{ij}) \sum_{j' \neq j} q^{\backslash i}(U_k^n = j') E_{q^{\backslash n}(\theta_{ij'})}[d_{nij'}(\theta_{ij'}, x_i^n)] \quad (28)$$