Bayesian Model Averaging for Causality Estimation and its Approximation based on Gaussian Scale Mixture Distributions

1 Derivation of the analytical form of $p(G|D^N)$ and $p(\theta_G|G, D^N)$

First, we derive $p(\theta_j|G,D^N)$ for a fixed $G \in G$. For $j \in \{1, \ldots, m\}$, let $\text{pa}(X_j) = (X_{j1}, X_{j2}, \ldots, X_{jm_j})$ and $X_j = [x_{j1}, x_{j2}, \ldots, x_{jm_j}] \in \mathbb{R}^{N \times m_j}$, where $x_i \in \mathbb{R}^N$ is the sample of $X_i$. Then, for $\theta_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm_j})$, the likelihood function $p(D^N|G, \theta_j)$ is given by

$$p(D^N|G, \theta_j) = \mathcal{N}(x_j; X_j \theta_j, \tau I_{m_j}) + \text{const.},$$

(1)

where $I_{m_j}$ is the identity matrix of size $m_j$. Since we assumed a conjugate Gaussian prior for $p(\theta_G|D)$, the posterior distribution $p(\theta_j|G,D^N)$ is given by

$$p(\theta_j|G,D^N) = \mathcal{N}(\theta_j; \mu_j, \Sigma_j),$$

(2)

$$\mu_j = s_{\epsilon} \Sigma_j X_j^T x_j,$$

(3)

$$\Sigma_j = (s_{\epsilon} X_j^T X_j + \tau^{-1} I_{m_j})^{-1}.$$  

(4)

Further, we can calculate the likelihood $p(D^N|G)$ as follows.

$$p(D^N|G) = \prod_{j=1}^m p(x_j|X_j),$$

(5)

$$p(x_j|X_j) = \frac{m_j}{2} \ln \tau^{-1} + \frac{N}{2} \ln s_{\epsilon} - E_j - \frac{1}{2} \ln |A_j| - \frac{N}{2} \ln(2\pi),$$

(6)

$$E_j = \frac{s_{\epsilon}}{2} \|x_j - X_j \mu_j\|^2 + \frac{\tau^{-1}}{2} \mu_j^T \mu_j,$$

(7)

$$A_j = \tau^{-1} I_{m_j} + s_{\epsilon} X_j^T X_j.$$  

(8)

We can calculate the posterior probability $p(G|D^N)$ by using the Bayes rule. See [Bishop, 2006] for the derivation of (2) and (6).
2 Derivation of Variational Bayes algorithm

The joint distribution for $x_j, X_j, \theta_j, \tau_j, \alpha_j$ is factorized as

$$p(x_j, X_j, \theta_j, \tau_j, \alpha_j) = p(x_j | X_j, \theta_j)p(\theta_j | \tau_j)p(\tau_j | \alpha_j)p(\alpha_j; \kappa, \nu).$$  \hfill (9)

Let $\xi = (\theta_j, \tau_j, \alpha_j)$. The variational Bayes method finds an approximation distribution $q(\xi)$ that approximates $p(\xi | x_j, X_j)$. The goal is to find $q(\xi)$ that minimizes the Kullback-Leibler divergence $\text{KL}(q(\xi)||p(\xi | x_j, X_j))$:

$$q^*(\xi) = \arg \min_{q(\xi)} \int q(\xi) \ln \frac{q(\xi)}{p(\xi | x_j, X_j)} d\xi$$

$$= \arg \min_{q(\xi)} \int q(\xi) \ln \frac{q(\xi)}{p(\xi | x_j, X_j)} d\xi. \hfill (11)$$

However, it is difficult to minimize (11) for arbitrary distributions. We limit the optimization distributions to $q(\xi)$ that can be factorized as

$$q(\theta_j, \tau_j, \alpha_j) = q(\theta_j)q(\tau_j)q(\alpha_j). \hfill (12)$$

For $\xi_k \in \xi$, the variational Bayes method minimizes (11) by updating $q(\xi_k)$ sequentially. With the distribution $q(\xi \setminus \xi_k)$ of $\xi \setminus \xi_k$ fixed, the update equation of $q(\xi_k)$ is given as follows [Bishop, 2006].

$$\ln q^*(\xi_k) = E_{q(\xi \setminus \xi_k)} [\ln p(\xi, x_j, X_j)] + \text{const.} \hfill (13)$$

In the following, we describe concrete update equation of each $q(\xi_k)$. To keep the description concise, for functions $f(\xi_k)$, the expectation taken by $q(\xi_k)$ at the point is written as $\langle f(\xi_k) \rangle$.

**Update equation of $q(\theta_j)$**

From (13), the update equation of $q(\theta_j)$ is

$$\ln q^*(\theta_j) = E_{q(\tau_j)} [p(x_j | X_j, \theta_j)p(\theta_j | \tau_j)] + \text{const.} \hfill (14)$$

Using the assumption that $p(x_j | X_j, \theta_j)$ and $p(\theta_j | \tau_j)$ are Gaussian distributions, we obtain

$$q^*(\theta_j) = \mathcal{N}(\hat{\theta}_j, \hat{\Sigma}_j),$$

$$\hat{\theta}_j = s_c \hat{\Sigma}_j X_j^T x_j, \hfill (15)$$

$$\hat{\Sigma}_j = (s_c X_j^T X_j + \langle S_{\tau_j} \rangle)^{-1}, \hfill (16)$$

where

$$S_{\tau_j} = \text{diag} \left( \tau_{j,1}^{-1}, \ldots, \tau_{j,m_j}^{-1} \right). \hfill (17)$$

**Update equation of $q(\tau)$**

From (13), the update equation of $q(\tau)$ is

$$\ln q^*(\tau) = E_{q(\theta_j, \alpha_j)} [p(x_j | X_j, \theta_j)p(\theta_j | \tau_j)p(\theta_j | \alpha_j)] + \text{const.} \hfill (19)$$

From the model assumption, without loss of generality, we can assume that $q(\tau_j)$ is decomposed as

$$q(\tau_j) = \prod_{i=1}^{m_j} q(\tau_{j,i}). \hfill (20)$$

By arranging the terms in (19) that include $\tau_{j,i}$, we obtain

$$q^*(\tau_{j,i}) = \mathcal{GIG} \left( \langle \alpha_{j,i} \rangle, \langle \theta_{j,i}^2 \rangle, \frac{1}{2} \right), \hfill (21)$$
where $\mathcal{GI}(a,b,\rho)$ denotes the generalized inverse Gaussian distribution, whose probability density function is given by

$$p(x; a, b, \rho) = \frac{(a/b)^{\rho/2}}{2K_{\rho}(\sqrt{ab})} \rho/2 K_{\rho}(\sqrt{ab}) x^{\rho-1} \exp\left(-\frac{ax + bx^{-1}}{2}\right),$$

where $K_\rho$ is a modified Bessel function of the second kind. To update $q(\theta_j)$ and $q(\alpha_j)$, we need the expected values $\langle \tau_{j,i} \rangle$ and $\langle \tau_{j,i}^{-1} \rangle$. They are given by

$$\langle \tau_{j,i} \rangle = 1 + \frac{\sqrt{\langle \tau_{j,i} \rangle \langle \theta_{j,i}^2 \rangle}}{\alpha_{j,i}},$$

$$\langle \tau_{j,i}^{-1} \rangle = \frac{\sqrt{\langle \alpha_{j,i} \rangle \langle \theta_{j,i}^2 \rangle}}{\theta_{j,i}^2}.$$

**Update equation of $q(\alpha)$**

From (13), the update equation of $q(\alpha)$ is

$$\ln q^*(\alpha) = E_{q(\tau)} [p(\tau|\alpha)p(\alpha; \kappa, \nu)] + \text{const.}$$

As in the case for $\tau_j$, we can assume that $q(\alpha_j)$ is decomposed as

$$q(\alpha_j) = \prod_{i=1}^{m_j} q(\alpha_{j,i}).$$

By arranging the terms in (25) that include $\alpha_{j,i}$, we obtain

$$q^*(\alpha_{j,i}) = \mathcal{GA}(\kappa + 1, \nu + \langle \tau_{j,i} \rangle/2),$$

where $\mathcal{GA}(\kappa, \nu)$ is the gamma distribution, whose probability density function is given by

$$p(x; \kappa, \nu) = \frac{\nu^\kappa}{\Gamma(\kappa)} x^{\kappa-1} e^{-\nu x}.$$

To update $q(\tau)$, we need the expected value $\langle \alpha_{j,i} \rangle$. It is given by

$$\langle \alpha_{j,i} \rangle = (\kappa + 1) \left( \nu + \langle \tau_{j,i} \rangle/2 \right).$$

**References**