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 $p(\boldsymbol{\theta}_G|G, D^N)$

First, we derive $p(\boldsymbol{\theta}_G|G, D^N)$ for a fixed $G \in \mathcal{G}$. For $j \in \{1, \dots, m\}$, let $\mathrm{pa}(X_j) = (X_{j_1}, X_{j_2}, \dots, X_{j_{m_j}})$ and $\boldsymbol{X}_j = [\boldsymbol{x}_{j_1}, \boldsymbol{x}_{j_2}, \dots, \boldsymbol{x}_{j_{m_j}}] \in \mathbb{R}^{N \times m_j}$, where $\boldsymbol{x}_i \in \mathbb{R}^N$ is the sample of X_i. Then, for $\boldsymbol{\theta}_j = (\theta_{j_1 j}, \theta_{j_2 j}, \dots, \theta_{j_{m_j} j})$, the likelihood function $p(D^N|G, \boldsymbol{\theta}_j)$ is given by

$$p(D^N|G, \boldsymbol{\theta}_j) = \mathcal{N}(\boldsymbol{x}_j; \boldsymbol{X}_j \boldsymbol{\theta}_j, \tau \boldsymbol{I}_{m_j}) + \text{const.}, \tag{1}$$

where $\boldsymbol{I}_{m_j}$ is the identity matrix of size m_j. Since we assumed a conjugate Gaussian prior for $p(\boldsymbol{\theta}_G|D)$, the posterior distribution $p(\boldsymbol{\theta}_j|G, D^N)$ is given by

$$p(\boldsymbol{\theta}_j|G, D^N) = \mathcal{N}(\boldsymbol{\theta}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{2}$$

$$\boldsymbol{\mu}_j = s_\epsilon \boldsymbol{\Sigma}_j \boldsymbol{X}_j^T \boldsymbol{x}_j, \tag{3}$$

$$\boldsymbol{\Sigma}_j = \left(s_\epsilon \boldsymbol{X}_j^T \boldsymbol{X}_j + \tau^{-1} \boldsymbol{I}_{m_j} \right)^{-1}. \tag{4}$$

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

$$p(D^N|G) = \prod_{j=1}^{m} p(\boldsymbol{x}_j|\boldsymbol{X}_j), \tag{5}$$

$$p(\boldsymbol{x}_j|\boldsymbol{X}_j) = \frac{m_j}{2} \ln \tau^{-1} + \frac{N}{2} \ln s_\epsilon - E_j - \frac{1}{2} \ln |\boldsymbol{A}_j| - \frac{N}{2} \ln(2\pi), \tag{6}$$

$$E_j = \frac{s_\epsilon}{2} ||\boldsymbol{x}_j - \boldsymbol{X}_j \boldsymbol{\mu}_j||^2 + \frac{\tau^{-1}}{2} \boldsymbol{\mu}_j^T \boldsymbol{\mu}_j, \tag{7}$$

$$\boldsymbol{A}_j = \tau^{-1} \boldsymbol{I}_{m_j} + s_\epsilon \boldsymbol{X}_j^T \boldsymbol{X}_j. \tag{8}$$

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 $\boldsymbol{x}_j, \boldsymbol{X}_j, \boldsymbol{\theta}_j, \boldsymbol{\tau}_j, \boldsymbol{\alpha}_j$ is factorized as

$$p(\boldsymbol{x}_j, \boldsymbol{X}_j, \boldsymbol{\theta}_j, \boldsymbol{\tau}_j, \boldsymbol{\alpha}_j) = p(\boldsymbol{x}_j|\boldsymbol{X}_j, \boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|\boldsymbol{\tau}_j)p(\boldsymbol{\tau}_j|\boldsymbol{\alpha}_j)p(\boldsymbol{\alpha}_j; \kappa, \nu). \tag{9}$$

Let $\boldsymbol{\xi} = (\boldsymbol{\theta}_j, \boldsymbol{\tau}_j, \boldsymbol{\alpha}_j)$. The variational Bayes method finds an approximation distribution $q(\boldsymbol{\xi})$ that approximates $p(\boldsymbol{\xi}|\boldsymbol{x}_j, \boldsymbol{X}_j)$. The goal is to find $q(\boldsymbol{\xi})$ that minimizes the Kullback-Leibler divergence $\mathrm{KL}(q(\boldsymbol{\xi})||p(\boldsymbol{\xi}|\boldsymbol{x}_j, \boldsymbol{X}_j))$:

$$q^*(\boldsymbol{\xi}) = \arg\min_{q(\boldsymbol{\xi})} \int q(\boldsymbol{\xi}) \ln \frac{q(\boldsymbol{\xi})}{p(\boldsymbol{\xi}|\boldsymbol{x}_j, \boldsymbol{X}_j)} \mathrm{d}\boldsymbol{\xi} \tag{10}$$

$$= \arg\min_{q(\boldsymbol{\xi})} \int q(\boldsymbol{\xi}) \ln \frac{q(\boldsymbol{\xi})}{p(\boldsymbol{\xi}, \boldsymbol{x}_j, \boldsymbol{X}_j)} \mathrm{d}\boldsymbol{\xi}. \tag{11}$$

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

$$q(\boldsymbol{\theta}_j, \boldsymbol{\tau}_j, \boldsymbol{\alpha}_j) = q(\boldsymbol{\theta}_j)q(\boldsymbol{\tau}_j)q(\boldsymbol{\alpha}_j). \tag{12}$$

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

$$\ln q^*(\boldsymbol{\xi}_k) = \mathrm{E}_{q(\boldsymbol{\xi}\setminus\boldsymbol{\xi}_k)}\left[\ln p(\boldsymbol{\xi}, \boldsymbol{x}_j, \boldsymbol{X}_j)\right] + \mathrm{const.} \tag{13}$$

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

Update equation of $q(\boldsymbol{\theta}_j)$

From (13), the update equation of $q(\boldsymbol{\theta}_j)$ is

$$\ln q^*(\boldsymbol{\theta}_j) = \mathrm{E}_{q(\boldsymbol{\tau}_j)}\left[p(\boldsymbol{x}_j|\boldsymbol{X}_j, \boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|\boldsymbol{\tau}_j)\right] + \mathrm{const.} \tag{14}$$

Using the assumption that $p(\boldsymbol{x}_j|\boldsymbol{X}_j, \boldsymbol{\theta}_j)$ and $p(\boldsymbol{\theta}_j|\boldsymbol{\tau}_j)$ are Gaussian distributions, we obtain

$$q^*(\boldsymbol{\theta}_j) = \mathcal{N}(\bar{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\Sigma}}_j), \tag{15}$$

$$\bar{\boldsymbol{\theta}}_j = s_\epsilon \tilde{\boldsymbol{\Sigma}}_j \boldsymbol{X}_j^T \boldsymbol{x}_j, \tag{16}$$

$$\tilde{\boldsymbol{\Sigma}}_j = \left(s_\epsilon \boldsymbol{X}_j^T \boldsymbol{X}_j + \left\langle \boldsymbol{S}_{\boldsymbol{\tau}_j} \right\rangle\right)^{-1}, \tag{17}$$

where

$$\boldsymbol{S}_{\boldsymbol{\tau}_j} = \mathrm{diag}\left(\tau_{j,1}^{-1}, \ldots, \tau_{j,m_j}^{-1}\right). \tag{18}$$

Update equation of $q(\boldsymbol{\tau})$

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

$$\ln q^*(\boldsymbol{\tau}) = \mathrm{E}_{q(\boldsymbol{\theta}_j, \boldsymbol{\alpha}_j)}\left[p(\boldsymbol{x}_j|\boldsymbol{X}_j, \boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j|\boldsymbol{\tau}_j)p(\boldsymbol{\theta}_j|\boldsymbol{\alpha}_j)\right] + \mathrm{const.} \tag{19}$$

From the model assumption, without loss of generality, we can assume that $q(\boldsymbol{\tau}_j)$ is decomposed as

$$q(\boldsymbol{\tau}_j) = \prod_{i=1}^{m_j} q(\tau_{j,i}). \tag{20}$$

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

$$q^*(\tau_{j,i}) = \mathcal{GIG}\left(\langle \alpha_{j,i}\rangle, \langle \theta_{j,i}^2\rangle, \frac{1}{2}\right), \tag{21}$$

where $\mathcal{GIG}(a, b, \rho)$ denotes the generalized inverse Gaussian distribution, whose probability density function is given by

$$p(x; a, b, \rho) = \frac{(a/b)^{\rho/2}}{2K_\rho(\sqrt{ab})} x^{\rho-1} \exp\left(-\frac{ax + bx^{-1}}{2}\right), \tag{22}$$

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

$$\langle \tau_{j,i} \rangle = \frac{1 + \sqrt{\langle \tau_{j,i} \rangle \langle \theta_{j,i}^2 \rangle}}{\alpha_{j,i}}, \tag{23}$$

$$\langle \tau_{j,i}^{-1} \rangle = \sqrt{\frac{\langle \alpha_{j,i} \rangle}{\langle \theta_{j,i}^2 \rangle}}. \tag{24}$$

Update equation of $q(\boldsymbol{\alpha})$

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

$$\ln q^*(\boldsymbol{\alpha}) = \mathrm{E}_{q(\boldsymbol{\tau})}\left[p(\boldsymbol{\tau}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}; \kappa, \nu)\right] + \text{const.} \tag{25}$$

As in the case for $\boldsymbol{\tau}_j$, we can assume that $q(\boldsymbol{\alpha}_j)$ is decomposed as

$$q(\boldsymbol{\alpha}_j) = \prod_{i=1}^{m_j} q(\alpha_{j,i}). \tag{26}$$

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

$$q^*(\alpha_{j,i}) = \mathcal{GA}\left(\kappa + 1, \nu + \frac{\langle \tau_{j,i} \rangle}{2}\right), \tag{27}$$

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

$$p(x; \kappa, \nu) = \frac{\nu^\kappa}{\Gamma(\kappa)} x^{\kappa-1} e^{-\nu x}. \tag{28}$$

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

$$\langle \alpha_{j,i} \rangle = (\kappa + 1)\left(\nu + \frac{\langle \tau_{j,i} \rangle}{2}\right). \tag{29}$$

References

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.