Bridging VMP and CEP: Theoretical Insights for Connecting Different Approximate Bayesian Inference Methods

Anonymous authors
Paper under double-blind review

Abstract

Approximate Bayesian inference (ABI) methods have become indispensable tools in modern machine learning and statistics for approximating intractable posterior distributions. Despite the related extensive studies and applications across diverse domains, the theoretical connections among these methods have remained relatively unexplored. This paper takes the first step to uncover the underlying relationships between two widely employed ABI techniques: the variational message passing (VMP) and the conditional expectation propagation (CEP) methods. Through rigorous mathematical analysis, we demonstrate a strong connection between these two approaches under mild conditions, from optimization as well as graphical model perspectives. This newly unveiled connection not only enhances our understanding of the performance and convergence properties of VMP and CEP, but it also facilitates the cross-fertilization of their respective strengths. For instance, we prove the convergence of CEP and enable an online variant of VMP through this connection. Furthermore, our findings provide insights into the underlying relationships and distinctive characteristics of other ABI methods, shedding new light on the understanding and development of more advanced ABI techniques. To validate our theoretical findings, we derive and analyze various ABI methods within the context of Bayesian tensor decomposition, a fundamental tool in machine learning research. Specifically, we show that these two approaches yield the same updates within this context and illustrate how the established connection can be leveraged to construct a streaming version of the VMP-based Bayesian tensor decomposition algorithm.

1 Introduction

Approximating difficult-to-compute posterior distributions is one of the most fundamental challenges in modern machine learning and statistics. To address this challenge, approximate Bayesian inference (ABI) has made significant progress over the years (Blei et al., 2017; Zhang et al., 2019; Theodoridis, 2025; Cheng et al., 2022b; Murphy, 2022), showcasing remarkable performance. These methods have found extensive applications in diverse domains, such as bioinformatics (Daunizeau et al., 2014; Grønbech et al., 2020), computer vision (Chan & Vasconcelos, 2009; Soh & Cho, 2022; Fan et al., 2022), and speech recognition (Cohen & Smith, 2010; Xue et al., 2021). Variational inference (VI) (Jordan et al., 1999; Wainwright & Jordan, 2008) and expectation propagation (EP) (Minka & Picard, 2001), along with their modern variants (Zhang et al., 2019; Broderick et al., 2013; Li et al., 2015; Wang & Zhe, 2020; Vehtari et al., 2020), are two prominent classes of ABI methods widely used in practice, as shown in Fig. 1.

The fundamental principle of VI involves formulating a family of distributions and subsequently finding the member within that family that best approximates the target distribution (Blei et al., 2017; Bishop, 2006; Theodoridis, 2025). The closeness between distributions is typically measured using the Kullback-Leibler (KL) divergence. In the context of the mean-field VI, the variables are assumed to be mutually independent and governed by their respective distributions. By decomposing the model evidence, VI transforms its objective into optimizing the evidence lower bound (ELBO). When analytical expectations can be derived,

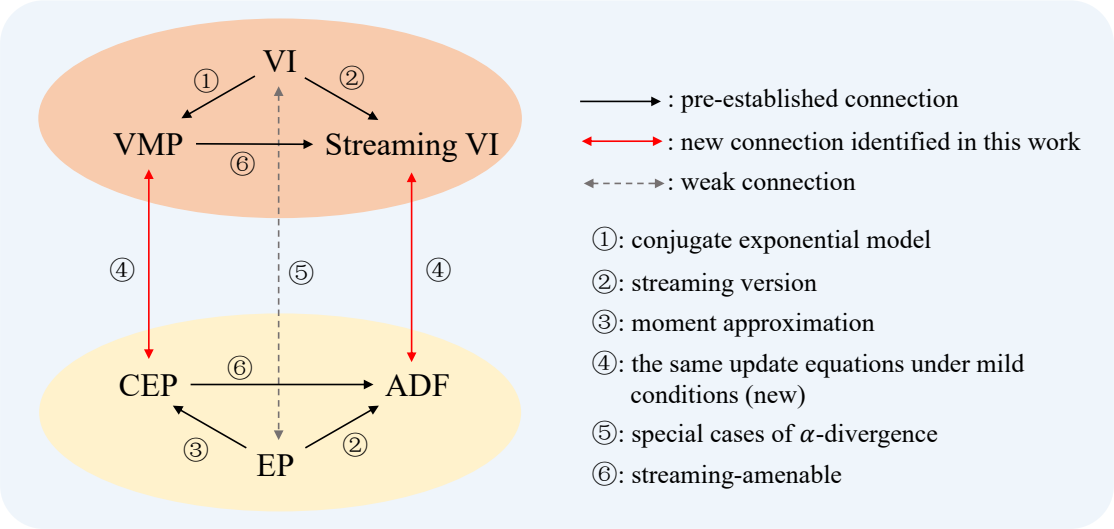

Figure 1: Connections of different ABI methods. The red arrows indicate the new connections established in this paper.

VI demonstrates favorable accuracy and speed. It also guarantees convergence to a local optimum (Beal, 2003). The streaming version of VI has also been developed to handle the streaming data case, but the algorithm design is not straightforward and demands additional effort (Broderick et al., 2013).

When the distributions of variables are restricted to the exponential family (Brown, 1986) and possess conjugate properties, mean-field VI can be implemented based on the convenient and efficient message-passing mechanism. The resulting algorithm is known as the variational message passing (VMP) (Winn et al., 2005). VMP operates by sending messages between nodes in the network and updating posterior beliefs through local operations performed at each node. By introducing additional variational parameters or utilizing approximation methods, VMP can be extended to models containing non-conjugate distributions (Winn et al., 2005; Wang & Blei, 2013). It also guarantees convergence and enables efficient evaluation of the model evidence (Winn et al., 2005).

EP is a generalized message-passing algorithm employed on factor graphs (Minka, 2013), which unifies and extends the concepts of assumed density filtering (ADF) (Maybeck, 1982) and loopy belief propagation (Frey & MacKay, 1997). The ADF can be viewed as a streaming or online version of EP, as shown in Fig. 1. In EP, we construct an approximation of the posterior by iteratively performing simple local computations which refine the factor that represents the contribution of the posterior from each data point. Notably, EP differs from VI in terms of the direction of the KL divergence. In various tasks, such as the clutter problem and mixture weight estimation, EP has shown superior performance compared to VI (Zhou et al., 2023). Additionally, the local computations make EP amenable to parallelized and distributed computation, rendering it well-suited for addressing large-scale problems (Li et al., 2015; Hasenclever et al., 2017; Vehtari et al., 2020). However, applying EP encounters a critical challenge when dealing with models that have complex likelihoods, as the moment matching that is involved in the factor update procedure can become intractable. Additionally, convergence is not guaranteed due to its local optimization nature (Vehtari et al., 2020).

To address the computation barrier in EP, recent advances have introduced alternative approaches for moment computation in EP, such as the Monte Carlo simulations (Li et al., 2018) and the Laplace approximation (Smola et al., 2004). Unfortunately, these approximations often suffer from inefficiency and high computational costs, thereby diminishing the appeal of EP as a fast approximation method. Conditional expectation propagation (CEP) (Wang & Zhe, 2020) has recently emerged as a promising alternative, offering an efficient variant of EP. Instead of directly calculating the moments of the complete distribution, CEP first seeks the tractable and analytical conditional moments and then computes their expectations with respect to the

approximate posterior of the remaining variables. Like EP, CEP's local update nature makes it well-suited for large-scale datasets, but convergence guarantee remains an open question.

Since VMP and CEP are developed from different perspectives (VI and EP, respectively) and have distinct theoretical roots (different directions of the KL divergence), theoretical connections between them remain unexplored. To the best of our knowledge, the most related work is the power EP (Minka, 2004), which unifies the idea of VI and EP by utilizing the α-divergence. By adjusting the value of α, it becomes possible to obtain an intermediate result between VI and EP. While power EP provides a general perspective that connects VI and EP, it does not fully uncover the intrinsic connections and differences between these two methods, nor does it consider the specific cases of CEP and VMP. Another related work is the Bayesian learning rule (Khan & Rue, 2023), which unifies different ABI methods through the natural gradient descent. However, it does not consider the EP algorithm and its variants.

This paper aims to unveil the underlying relationships between VMP and CEP. In particular, we demonstrate a strong connection between these two approaches from optimization as well as graphical model perspectives. This newly identified connection not only deepens our understanding of the performance and convergence properties of these two approaches but, also, it enables the cross-pollination of their respective strengths. Additionally, it provides insights into the underlying relationships and distinct characteristics of other ABI methods, as shown in Fig. 1.

Notably, the established connection provides a guarantee of the convergence of CEP, leveraging the corresponding property enjoyed by VMP. It turns out that the assimilation of the message factors in CEP leads to an increment of ELBO and ensures the attainment of convergence. Additionally, the connection can also provide some insights into the convergence of the standard EP. Furthermore, the parallelized and distributed nature of CEP facilitates the seamless construction of an online or distributed variant of VMP and VI. This adaptability enables the effective handling of large-scale datasets, particularly in scenarios involving continuous data streams or sequential data arrivals.

To corroborate our theoretical analysis, we present an example that showcases the application of VMP and CEP in the context of Bayesian tensor decomposition, which is a powerful tool in machine learning research and finds applications in various real-world scenarios (Cheng et al., 2022b;a; Fang et al., 2021b;a). In this particular context, besides demonstrating the connection between VMP and CEP, we also illustrate how this connection can be leveraged to develop a streaming version of the VMP-based tensor decomposition algorithm.

The remainder of this paper is organized as follows. Section 2 gives a brief review of different ABI methods and provides some useful lemmas. Section 3 contains the main theoretical results and some related extensions. In Section 4, using Bayesian tensor decomposition as an example, different ABI methods are derived and analyzed to validate our theoretical findings. Finally, Section 5 concludes with an overall discussion and suggestions for future research directions.

2 Preliminaries

This section provides a brief review of various ABI methods and presents some useful lemmas. Before delving into the details of each method, we introduce the general problem. Given a set of observations $\mathcal{D} = \{x_1, \cdots, x_N\}$ and a probabilistic model described via a set of latent variables $\boldsymbol{\theta}$, the joint distribution can be expressed as

$$p(\boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta})$ represents the respective prior distribution and $p(\mathcal{D}|\boldsymbol{\theta})$ denotes the data likelihood. The goal is to compute the posterior distribution, $p(\boldsymbol{\theta}|\mathcal{D})$, which can be expressed as

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta}, \mathcal{D})}{\int p(\boldsymbol{\theta}, \mathcal{D}) d\boldsymbol{\theta}},$$

where $p(\mathcal{D})$ denotes the model evidence. For many models of practical interest, it is infeasible to compute the posterior distribution directly due to the analytically intractable integration in the denominator. Therefore,

approximation methods are essential in such cases. In this paper, we primarily focus on VI, EP, and their variants.

2.1 VI and VMP

2.1.1 VI

VI is a technique that approximates the posterior distribution by utilizing a probability distribution with density $q(\boldsymbol{\theta})$ from a tractable family of distributions $\mathcal{Q}$. The aim is to find the best variational approximation, $q^* \in \mathcal{Q}$, by minimizing the KL divergence between $q(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta}|\mathcal{D})$ (Cover, 1999), i.e.,

$$q^* = \min_{q\in\mathcal{Q}} \mathrm{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})) = \min_{q\in\mathcal{Q}} \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta}.$$

This transforms the inference task into an optimization problem, where the flexibility of the family $\mathcal{Q}$ controls the complexity of the optimization process. However, the objective function is not directly computable as it requires the model evidence. To overcome this challenge, VI employs a clever decomposition (e.g., Bishop, 2006; Theodoridis, 2025)

$$\ln p(\mathcal{D}) = \mathcal{L}(q) + \mathrm{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})),$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{1}$$

is the evidence lower bound (ELBO). Since the model evidence is a constant with respect to $\boldsymbol{\theta}$ and the KL divergence is non-negative, minimizing the latter is equivalent to maximizing $\mathcal{L}(q)$.

If there are no restrictions on $\mathcal{L}(q)$, the maximum of the ELBO occurs when $q(\boldsymbol{\theta})$ equals $p(\boldsymbol{\theta}|\mathcal{D})$, which, however, is intractable. Consequently, some restrictions on the functional form of $q(\boldsymbol{\theta})$ are required. In the context of the mean-field VI, the variables are assumed to be mutually independent, and each variable is governed by its own distribution. A typical member of the mean-field variational family can be expressed as (e.g., Blei et al., 2017; Theodoridis, 2025)

$$q(\boldsymbol{\theta}) = \prod_{m=1}^{M} q(\boldsymbol{\theta}_m). \tag{2}$$

Here the elements of $\boldsymbol{\theta}$ are partitioned into M disjoint groups, i.e., $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$. Then, the ELBO $\mathcal{L}(q)$ is optimized by iteratively updating each group in turn. Specifically, the optimal solution for each factor can be obtained by substituting (2) into (1), which gives

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_m q(\boldsymbol{\theta}_m) \left\{ \ln p(\boldsymbol{\theta}, \mathcal{D}) - \sum_m \ln q(\boldsymbol{\theta}_m) \right\} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}_m) \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] d\boldsymbol{\theta}_m - \int q(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m + \text{const} \\
&= -\mathrm{KL}\left(q(\boldsymbol{\theta}_m)\|\tilde{q}(\boldsymbol{\theta}_m)\right) + \text{const},
\end{aligned}$$

where $\boldsymbol{\theta}_{\setminus m}$ represents the set of variables excluding the mth group and

$$\ln \tilde{q}(\boldsymbol{\theta}_m) = \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const}.$$

It can be seen that $\mathcal{L}(q)$ is optimized when the KL divergence equals to zero, which results in

$$\begin{aligned}
\ln q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const} \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})] + \text{const}.
\end{aligned} \tag{3}$$

After taking the exponential of both sides and normalizing, we obtain

$$q^*(\boldsymbol{\theta}_m) = \frac{\exp(\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})])}{\int \exp(\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]) d\boldsymbol{\theta}_m}, \forall m. \tag{4}$$

Although the set of equations in equation 4 provides consistency conditions for maximizing the lower bound, they do not represent an explicit solution. This is because the optimum for each variable group $\boldsymbol{\theta}_m$ depends on the distributions of other groups, $\boldsymbol{\theta}_{\setminus m}$. Therefore, when applying VI, we typically seek a solution by first initializing all of the factors $q(\boldsymbol{\theta}_m)$ appropriately and then iteratively updating each factor, replacing the other variable groups with their current estimates. Convergence is guaranteed because the ELBO is convex with respect to each of the groups (e.g., Bishop, 2006).

2.1.2 VMP

VMP is an implementation of the mean-field VI that operates on the conjugate-exponential model (Winn et al., 2005). In this model, the distribution of variables/nodes, conditioned on their parents, are drawn from the exponential family and are conjugate with respect to the distributions over these parent variables.[1] As a result, each complete conditional is also in the exponential family (e.g., Blei et al., 2017), i.e.,

$$p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D}) = h(\boldsymbol{\theta}_m) \exp\left\{\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D})^T \boldsymbol{\phi}(\boldsymbol{\theta}_m) - Z_m(\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D}))\right\}, \tag{5}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta}_m)$ is the vector of sufficient statistics; $\boldsymbol{\eta}_m$ are the natural parameters; and $Z_m(\cdot)$ is the log partition function. The subscript m indicates that these quantities may vary across different nodes. For simplicity, here, we consider each group $\boldsymbol{\theta}_m$ to contain a single variable.

In the conjugate-exponential model, the update for node m in the mean-field VI becomes significantly simplified. By substituting (5) into (4), the update can be expressed as (e.g., Blei et al., 2017)

$$\begin{aligned} q^*(\boldsymbol{\theta}_m) &\propto \exp(\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]) \\ &= \exp\left\{\ln h(\boldsymbol{\theta}_m) + \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]^T \boldsymbol{\phi}(\boldsymbol{\theta}_m) - \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[Z_m(\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D}))]\right\} \\ &\propto h(\boldsymbol{\theta}_m) \exp\left\{\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]^T \boldsymbol{\phi}(\boldsymbol{\theta}_m)\right\}. \end{aligned} \tag{6}$$

This reveals that the optimal variational distribution for a node has the same functional form as the corresponding prior distribution, indicating that we only need to update the parameters of the corresponding distribution. Furthermore, the updates for each one of the nodes can be implemented locally using the expected values (messages) from the rest of the other nodes. VMP involves the exchange of messages between nodes in the network and iteratively updating the posterior distribution until the convergence is reached (Li et al., 2024). The detailed algorithm for VMP is summarized in Appendix A.

2.2 EP and CEP

2.2.1 EP

EP is a generalized message-passing algorithm that combines and extends the concepts of ADF and loopy belief propagation. Compared to VI, EP also approximates the posterior by minimizing the KL divergence, but in the opposite direction. The detailed algorithm is elucidated as follows.

EP assumes that the joint distribution of the probabilistic model can be expressed in a factorized form, given by

$$p(\boldsymbol{\theta}, \mathcal{D}) = \prod_i f_i(\boldsymbol{\theta}).$$

[1] Equivalently, this can be described as a conjugate-exponential model represented as a Bayesian network (e.g., Theodoridis, 2025).

Particularly, in the case of independently and identically distributed (i.i.d.) observed data, $f_i(\boldsymbol{\theta})$ corresponds to the ith likelihood term $p(\mathbf{x}_i|\boldsymbol{\theta})$, and $f_0(\boldsymbol{\theta})$ represents the respective prior $p(\boldsymbol{\theta})$. Then, the joint distribution is written as

$$p(\boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{i=0}^{N} f_i(\boldsymbol{\theta}). \tag{7}$$

We are interested in evaluating the posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})} = \frac{1}{Z} \prod_i f_i(\boldsymbol{\theta}),$$

where $Z = p(\mathcal{D})$ is the normalization constant, which can be calculated as

$$Z = \int \prod_i f_i(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

EP approximates the posterior by a product of factors[2], given by (Minka, 2013)

$$q(\boldsymbol{\theta}) = \frac{1}{\tilde{Z}} \prod_i \tilde{f}_i(\boldsymbol{\theta}),$$

where $\tilde{f}_i$ is an approximation of f_i that belongs to the exponential family, and $\tilde{Z}$ is the associated normalization constant. Ideally, the determination of the involved factors $\{\tilde{f}_i(\boldsymbol{\theta})\}_{i=1}^N$ involves the minimization of the KL divergence from $p(\boldsymbol{\theta}|\mathcal{D})$ to $q(\boldsymbol{\theta})$, given by

$$\mathrm{KL}(p\|q) = \mathrm{KL}\left(\frac{1}{Z} \prod_i f_i(\boldsymbol{\theta}) \| \frac{1}{\tilde{Z}} \prod_i \tilde{f}_i(\boldsymbol{\theta}) \right).$$

However, this minimization is typically intractable due to the need to compute expectations with respect to the true distribution.

EP provides an approximation approach by iteratively optimizing individual factors while taking into account the influence of the remaining ones. It operates by cycling through the factors and refining them one at a time. To elaborate, EP follows four simple steps in each iteration (Minka, 2013). First, select a factor $\tilde{f}_i$ for updating and remove it from the approximation $q(\boldsymbol{\theta})$ to produce the *calibrating* distribution $q^{\setminus i}(\boldsymbol{\theta})$, defined as $q^{\setminus i}(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/\tilde{f}_i(\boldsymbol{\theta})$. Note that $q^{\setminus i}$ can also be derived from the product of factors $i \neq j$, but in practice, the division is more convenient. Second, the calibrating distribution is combined with the factor $f_i(\boldsymbol{\theta})$ to obtain the *tilted* distribution

$$\hat{p}_i(\boldsymbol{\theta}) = \frac{1}{Z_i} f_i(\boldsymbol{\theta}) q^{\setminus i}(\boldsymbol{\theta}), \tag{8}$$

where Z_i is the associated normalization constant. Third, we obtain an approximation $q^\natural(\boldsymbol{\theta})$ of $\hat{p}_i(\boldsymbol{\theta})$ by minimizing the KL divergence between $\hat{p}_i(\boldsymbol{\theta})$ and $q^\natural(\boldsymbol{\theta})$. If $q^\natural(\boldsymbol{\theta})$ belongs to the exponential family, as it is often the case, the minimum can be obtained by moment matching (Maybeck, 1982), i.e.,

$$\mathbb{E}_{q^\natural(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})] = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})], \tag{9}$$

where $\phi(\boldsymbol{\theta})$ is the sufficient statistics of $q^\natural(\boldsymbol{\theta})$. Note that the natural parameter of $q^\natural(\boldsymbol{\theta})$ is implicitly specified in the moment matching process. For example, if $q^\natural(\boldsymbol{\theta})$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then we can minimize the KL divergence by setting the mean $\boldsymbol{\mu}$ equal to the mean of $\hat{p}_i(\boldsymbol{\theta})$ and the covariance $\boldsymbol{\Sigma}$ equal to the covariance of $\hat{p}_i(\boldsymbol{\theta})$. Finally, the factor $\tilde{f}_i$ is update via $\tilde{f}_i(\boldsymbol{\theta}) \propto q^\natural(\boldsymbol{\theta})/q^{\setminus i}(\boldsymbol{\theta})$.

The rationale behind this update is to ensure that the approximate factor contributes to the posterior in a manner similar to the corresponding data likelihood. Due to the local refinement, the factors can be efficiently calculated in a distributed manner. However, convergence is not guaranteed in general.

[2]Here, we use the term "factor" to maintain consistency with the terminology used in probabilistic graphical models.

2.2.2 CEP

While EP is known for its favorable accuracy and speed on diverse tasks, a significant challenge in its application arises from the computational intractability of the expectations $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})]$ in (9) for models with complex data likelihood. To overcome this limitation, several methods have been proposed. One such method is the CEP, which offers efficient and analytical updates.

In CEP, the approximate factor is assumed to be further factorized with respect to the variable groups $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}$, which can be expressed as

$$\tilde{f}_i(\boldsymbol{\theta}) = \prod_m \tilde{f}_i(\boldsymbol{\theta}_m), \tag{10}$$

where $\{\tilde{f}_i(\boldsymbol{\theta}_m)\}_{m=1}^M$ are constrained to be in the exponential family. As a result, the approximate posterior $q(\boldsymbol{\theta}_m)$ and the calibrating distribution $q^{\setminus i}(\boldsymbol{\theta})$ are both factorized over the variable groups. It is worth noting that the factorized message factors are also widely used in EP algorithms for large-scale applications.

Given the factorized form in (10), the objective is to update each subfactor $\tilde{f}_i(\boldsymbol{\theta}_m)$. By utilizing the law of iterated expectations, the moment $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)]$ required for updating $\tilde{f}_i(\boldsymbol{\theta}_m)$ can be expressed as (Wang & Zhe, 2020)

$$\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}\left[\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})}[\phi(\boldsymbol{\theta}_m)]\right], \tag{11}$$

where $\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})$ is the conditional distribution and $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$ is the marginal distribution. The conditional moment $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})}[\phi(\boldsymbol{\theta}_m)]$ often has an analytical form since the rest of the variables except $\boldsymbol{\theta}_m$ are fixed. More generally, the conditional moment can be represented with a quadrature formula. (Wang & Zhe, 2020)

To compute the moment in (11), EP requires the computation of the expectation of the conditional moment with respect to the marginal posterior $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$. However, this computation is also intractable for models with complex likelihoods. To address this challenge, CEP assumes that $q(\boldsymbol{\theta}_{\setminus m})$ and $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$ are close in high-density regions as their moments are matched (Wang & Zhe, 2020). In the sequel, CEP employs $q(\boldsymbol{\theta}_{\setminus m})$ as a surrogate for $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$ in the respective computation. The goal now becomes to calculate the expectation $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}\left[\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})}[\phi(\boldsymbol{\theta}_m)]\right]$.

Note that the conditional moment $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})}[\phi(\boldsymbol{\theta}_m)]$ is a function of the sufficient statistics of $\boldsymbol{\theta}_{\setminus m}$, denoted as $\mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})}[\phi(\boldsymbol{\theta}_m)] = h(\Phi_m)$, where $\Phi_m = \{\phi(\boldsymbol{\theta}_1), \cdots, \phi(\boldsymbol{\theta}_{m-1}),$
$\phi(\boldsymbol{\theta}_{m+1}), \cdots, \phi(\boldsymbol{\theta}_M)\}$ is the set of sufficient statistics. If the expectation $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[h(\Phi_m)]$ is still intractable, we can approximate it by utilizing the multivariate delta method (Dorfman, 1938; Ver Hoef, 2012), which can be expressed by $h(\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\Phi_m])$. The multivariate delta method can be seen as a first-order Taylor approximation, as detailed in Appendix B. Similar to EP, the messages can be computed in a distributed manner, but the convergence guarantee remains an open question. The detailed algorithm for CEP is summarized in Appendix A.

2.3 ADF

ADF is an online Bayesian inference method that can be seen as a special case of EP. It provides an efficient approach for approximating posterior distributions in a sequential manner. ADF is obtained by initializing all the approximating factors, except the first one, to unity and then updating each factor once in a single pass. The ADF algorithm shares similarities with EP but it simplifies certain aspects of the process. Particularly, in ADF, the removal step, which involves creating a calibrating distribution by removing a factor from the approximation, is ignored. Instead, the calibrating distribution is replaced by the full approximation, given by (Li et al., 2015)

$$q^{\setminus i}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}).$$

Consequently, the tilted distribution in ADF can be expressed as:

$$\hat{p}_i(\boldsymbol{\theta}) \propto f_i(\boldsymbol{\theta}) q(\boldsymbol{\theta}).$$

The subsequent steps in ADF are the same as in EP.

2.4 Lemmas

This subsection presents some useful lemmas to offer a deeper understanding of the exponential family and the KL divergence.

Lemma 1(Minka, 2013): *If $p(\boldsymbol{\theta})$ is an arbitrary fixed distribution and $q(\boldsymbol{\theta})$ is in the exponential family, then minimizing the divergence $KL(p\|q)$ with respect to q gives*

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})] = \mathbb{E}_{p(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta})],$$

where $\phi(\boldsymbol{\theta})$ is the sufficient statistics of $q(\boldsymbol{\theta})$.

Lemma 1, commonly referred to as moment matching or moment projection, reveals that the KL divergence can be minimized by equating expectations of the sufficient statistics of $q(\boldsymbol{\theta})$ to their expectations with respect to $p(\boldsymbol{\theta})$. It is noteworthy that if $p(\boldsymbol{\theta})$ belongs to the exponential family and shares the same sufficient statistics as $q(\boldsymbol{\theta})$ (i.e., they possess the same distributional form), the moment matching procedure guarantees that their natural parameters become identical. As exponential family distributions are uniquely determined by their sufficient statistics and natural parameters, moment matching leads to the equality of $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$. Consequently, the KL divergence between the two distributions is reduced to zero.

Lemma 2(Bishop, 2006): *Assume $p(\boldsymbol{\theta})$ is a fixed distribution and $q(\boldsymbol{\theta})$ factorizes with respect to variable groups, i.e.,*

$$q(\boldsymbol{\theta}) = \prod_m q(\boldsymbol{\theta}_m),$$

then minimizing the divergence $KL(p\|q)$ with respect to q gives

$$q^*(\boldsymbol{\theta}_m) = p(\boldsymbol{\theta}_m), \forall m. \tag{12}$$

Lemma 2 shows that the optimal solution of each factor distribution $q(\boldsymbol{\theta}_m)$ is given by the corresponding marginal distribution of $p(\boldsymbol{\theta})$.

3 Main Results

3.1 Connection between VMP and CEP

To establish the connection between CEP and VMP, we initially present the following lemma.

Lemma 3: *Assume $p(\boldsymbol{\theta})$ is a fixed distribution and $q(\boldsymbol{\theta})$ factorizes with respect to variable groups, i.e.,*

$$q(\boldsymbol{\theta}) = \prod_m q(\boldsymbol{\theta}_m),$$

where each factor $q(\boldsymbol{\theta}_m)$ belongs to the exponential family. Then minimizing the divergence $KL(p\|q)$ with respect to q gives

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{p(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)], \forall m, \tag{13}$$

where $\phi(\boldsymbol{\theta}_m)$ is the sufficient statistics of $q(\boldsymbol{\theta}_m)$.

Proof: See Appendix C.

Lemma 3 can be seen as a combination of Lemma 1 and Lemma 2, establishing a connection between the conditional moment matching and the minimization of KL divergence. In CEP, the optimal factor is given by

$$\tilde{f}_i(\boldsymbol{\theta}_m) \propto q^{\natural}(\boldsymbol{\theta}_m)/q^{\setminus i}(\boldsymbol{\theta}_m),$$

where the tilted distribution $q^{\setminus i}(\boldsymbol{\theta}_m)$ is defined in equation 8 and the variational distribution $q^\natural(\boldsymbol{\theta})$ is obtained through moment matching, satisfying

$$\mathbb{E}_{q^\natural(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)]. \tag{14}$$

To establish the connection between CEP and VMP, it is necessary to derive an analytical form for the factor $\tilde{f}_i(\boldsymbol{\theta}_m)$. Generally, $\tilde{f}_i(\boldsymbol{\theta}_m)$ does not possess an analytical form due to the involvement of moment matching in the computation of $q^\natural(\boldsymbol{\theta})$. However, by comparing (13) and (14), we can show that for conjugate-exponential models under mild conditions, $\tilde{f}_i(\boldsymbol{\theta}_m)$ does indeed have an analytical form.

Lemma 4: *Consider a conjugate-exponential probabilistic model represented as a Bayesian network. If the expectations are approximated using the multivariate delta method, the optimal factor in CEP is expressed as*

$$\tilde{f}_i(\boldsymbol{\theta}_m) \propto \frac{\hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}])}{q^{\setminus i}(\boldsymbol{\theta}_m)}. \tag{15}$$

Proof: See Appendix C.

Based on the analytical form of $\tilde{f}_i(\boldsymbol{\theta}_m)$, we can show the connection between the CEP and VMP, and state the following theorem.

Theorem 1: *Consider a conjugate-exponential probabilistic model represented as a Bayesian network. Suppose the variational distribution follows the mean-field assumption and the observations are i.i.d. Then the CEP and VMP yield the same update equations under the following conditions:*

- *The update in CEP is performed on the variable groups.*
- *The expectations are approximated using the multivariate delta method.*

Proof: To prove Theorem 1, we first give the following lemma.

Lemma 5: *Consider a conjugate-exponential probabilistic model represented as a Bayesian network. Suppose the variational distribution follows the mean-field assumption and the observations are i.i.d. If the update in CEP is performed on the variable groups, then a sufficient condition for the equivalence between the update equations of CEP and VMP is*

$$\ln \tilde{f}_i(\boldsymbol{\theta}_m) = \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})]. \tag{16}$$

Proof: To prove this lemma, we start by considering the logarithm of the optimal distribution in VMP, given by:

$$\begin{aligned} \ln q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\sum_i \ln f_i(\boldsymbol{\theta})] + \text{const} \\ &= \sum_i \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})] + \text{const}. \end{aligned} \tag{17}$$

where $p(\boldsymbol{\theta}, \mathcal{D}) = \prod_i f_i(\boldsymbol{\theta})$ follows from (7) under the i.i.d. assumption. Note that we also use $f_i(\boldsymbol{\theta})$ to represent the likelihood and prior, as in CEP. On the other hand, if the update in CEP is performed on the variable groups, then the optimal distribution for each variable group can be expressed as the product of the approximate factors:[3]

$$q^*(\boldsymbol{\theta}_m) \propto \prod_i \tilde{f}_i(\boldsymbol{\theta}_m).$$

[3]Here, we also use the notation $q^*(\boldsymbol{\theta}_m)$ to denote the optimal variational distribution in CEP.

Taking the logarithm of both sides gives:

$$\ln q^*(\boldsymbol{\theta}_m) = \sum_i \ln \tilde{f}_i(\boldsymbol{\theta}_m) + \text{constant}. \tag{18}$$

By comparing (17) and (18), it can be seen that the updates of CEP and VMP are the same if (16) holds. ■

Then we show that (16) holds if the expectations are approximated using the multivariate delta method. Specifically, from Lemma 4, we can take the logarithm of both sides of (15), which yields

$$\ln \tilde{f}_i(\boldsymbol{\theta}_m) = \ln \hat{p}_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]) - \ln q^{\setminus i}(\boldsymbol{\theta}_m). \tag{19}$$

For simplicity, we omit the constant term. From (17), it can be seen that the optimal variational distribution in VMP consists of some independent terms, which can be expressed as

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})] &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})] + \ln q^{\setminus i}(\boldsymbol{\theta}_m) - \ln q^{\setminus i}(\boldsymbol{\theta}_m) \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta}) q^{\setminus i}(\boldsymbol{\theta}_m)] - \ln q^{\setminus i}(\boldsymbol{\theta}_m) \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln \hat{p}_i(\boldsymbol{\theta}_m | \boldsymbol{\theta}_{\setminus m})] - \ln q^{\setminus i}(\boldsymbol{\theta}_m).
\end{aligned}$$

By utilizing the multivariate delta approximation, the expectation $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})]$ can be expressed as

$$\ln \hat{p}_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]) - \ln q^{\setminus i}(\boldsymbol{\theta}_m) = \ln \tilde{f}_i(\boldsymbol{\theta}_m),$$

which shows that $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})]$ and $\ln \tilde{f}_i(\boldsymbol{\theta}_m)$ are equivalent under the multivariate delta approximation. According to Lemma 5, we can conclude that VMP and CEP yield the same update equations under the specified conditions. ■

In practical applications, these preconditions and conditions are often satisfied, enabling the derivation of analytical updates, as demonstrated in the example provided in Section 4. Below, we delve deeper into these conditions and discuss their respective implications.

The preconditions in Theorem 1 establish a foundation for the effective application of both VMP and CEP, as outlined in the preliminaries. Specifically, the i.i.d. assumption enables the update in VMP to be expressed as a summation of N terms, each corresponding to a factor in the CEP framework. This highlights that the VMP update can be interpreted as the merging of messages sent from the data nodes, aligning with the message-passing nature of VMP, which will be discussed further in the next subsection. Consequently, it becomes easy to derive a streaming version of VMP, which is described in detail in the next section. Additionally, the conjugate-exponential condition ensures that the updates of the factors in CEP can be formulated analytically, avoiding the need for moment matching. When this assumption does not hold, the direct connection between VMP and CEP may break down, as the factor updates in CEP may no longer have analytical solutions.

Regarding the specific conditions, the first ensures that the update of $q(\boldsymbol{\theta}_m)$ in CEP is expressed as the product of a number of factors, enabling efficient parallel or distributed computation and significantly reducing computational costs. The second condition simplifies expectation computations, making the inference process more tractable. Notably, the multivariate delta approximation, also known as the reparameterization trick in standard VMP (Winn et al., 2005), is widely employed in various ABI methods, including CEP.

It is worth noting that the connection between CEP and VMP can be viewed from a more general perspective. To see this, note that a fundamental assumption in CEP is that the message factor $\tilde{f}_i$ factorizes with respect to variable groups, allowing the approximate posterior to be expressed as

$$q(\boldsymbol{\theta}) \propto \prod_i \tilde{f}_i(\boldsymbol{\theta}) = \prod_i \prod_m \tilde{f}_{im}(\boldsymbol{\theta}_m) = \prod_m q(\boldsymbol{\theta}_m),$$

which in fact corresponds to the mean-field assumption. From Lemma 2, we know that the optimal solution of $\min_{q(\boldsymbol{\theta}_m)} \mathrm{KL}(p(\boldsymbol{\theta}|\mathcal{D})\|q(\boldsymbol{\theta}))$ is given by $q^*(\boldsymbol{\theta}_m) = p(\boldsymbol{\theta}_m|\mathcal{D})$, which can be further written as

$$\begin{aligned} q^*(\boldsymbol{\theta}_m) &= p(\boldsymbol{\theta}_m|\mathcal{D}) \\ &= \int p(\boldsymbol{\theta}_m, \boldsymbol{\theta}_{\setminus m}|\mathcal{D}) d\boldsymbol{\theta}_{\setminus m} \\ &= \int p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D}) p(\boldsymbol{\theta}_{\setminus m}|\mathcal{D}) d\boldsymbol{\theta}_{\setminus m} \\ &= \mathbb{E}_p[p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]. \end{aligned}$$

By applying the approximations in CEP, i.e., using $q(\boldsymbol{\theta}_m)$ as a surrogate of $p(\boldsymbol{\theta}_m)$ in moment computation and the delta approximation, the optimal variational distribution can be approximated as

$$q^*(\boldsymbol{\theta}_m) = \mathbb{E}_p[p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})] \approx p(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}], \mathcal{D}). \tag{20}$$

In VMP, again applying the delta approximation, each optimal variational distribution becomes

$$\begin{aligned} \ln q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{constant} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m}, \mathcal{D})] + \text{constant} \\ &\approx \ln p(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}], \mathcal{D}), \end{aligned}$$

which leads to

$$q^*(\boldsymbol{\theta}_m) \approx p(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}], \mathcal{D}). \tag{21}$$

By comparing (20) and (21), it can be seen that the inherent objective of both VMP and CEP is to approximate the conditional marginal distribution $p(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}], \mathcal{D})$. VMP and CEP start from different KL formulations. However, in both cases, the theoretical optimal is the same due to the properties of the KL. This justifies the derived connection.

3.2 Extensions and Implications

The previous section established a strong connection between CEP and VMP. Expanding on this connection, we present new theoretical results regarding the convergence and scalability of several ABI methods.

3.2.1 Convergence of CEP

As previously mentioned, the convergence of EP is not generally guaranteed. To address this issue, some approaches apply energy optimization techniques directly to the associated objective function rather than relying on local updates. For instance, they implement EP based on the convergent double-loop optimization algorithm (Opper et al., 2005; Hasenclever et al., 2017). However, these approaches require additional designs and exhibit increased computational complexities.

Since CEP is developed from EP, its convergence properties also remain an open question. Nevertheless, by leveraging the established connection with VMP, we can demonstrate that CEP is guaranteed to converge under certain mild conditions. Specifically, we present the following corollary.

Corollary 1: *Consider a conjugate-exponential probabilistic model represented as a Bayesian network. Suppose the variational distribution follows the mean-field assumption and the observations are i.i.d. If the conditions in Theorem 1 hold, then CEP updates are guaranteed to converge to a local minimum of the KL divergence.*

Proof: See Appendix C.

From (17), the optimal variational distribution $q^*(\boldsymbol{\theta}_m)$ in VMP is

$$\ln q^*(\boldsymbol{\theta}_m) = \sum_i \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})] + \text{const.}$$

As shown in Section 2, this update increases the ELBO at each iteration, ensuring the convergence property of VMP. According to Lemma 5, we have $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})] = \ln \tilde{f}_i(\boldsymbol{\theta}_m)$. Since $\tilde{f}_i(\boldsymbol{\theta}_m)$ is the message factor, the term $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta})]$ can be interpreted as the message sent from the ith data node. Thus, the update in VMP can be viewed as merging all the messages sent by data nodes. In other words, in CEP, merging the message factors $\tilde{f}_i(\boldsymbol{\theta}_m)$ sent by data nodes increases the ELBO, thereby ensuring convergence.

It is important to note that in the standard implementations of CEP, updates are performed on the factors instead of the variable groups. In other words, $\tilde{f}_i$ is updated sequentially in each iteration. This factor-based update mechanism allows for a more fine-grained local optimization, which might be the reason for its superior performance in various tasks. However, this type of local optimization does not guarantee convergence in general. If the updates in CEP are performed on the factors rather than on the variable groups, the convergence guarantee is lost.

To see this, note that if the updates in CEP are performed on the factors, the increase of ELBO is not guaranteed. Specifically, the ELBO can be expressed as

$$\begin{aligned}
\mathcal{L} &= \int \prod_m q(\boldsymbol{\theta}_m) \left\{ \ln p(\boldsymbol{\theta}, \mathcal{D}) - \sum_m \ln q(\boldsymbol{\theta}_m) \right\} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}_m) \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] d\boldsymbol{\theta}_m - \int q(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
&= \int \prod_i \tilde{f}_i(\boldsymbol{\theta}_m) \sum_i \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[f_i(\boldsymbol{\theta})] d\boldsymbol{\theta}_m + \int \prod_i \tilde{f}_i(\boldsymbol{\theta}_m) \sum_i \ln \tilde{f}_i(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m.
\end{aligned}$$

The optimal factor in CEP can be written as

$$\begin{aligned}
\ln \tilde{f}_i(\boldsymbol{\theta}_m) &= \ln \hat{p}_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]) - \ln q^{\setminus i}(\boldsymbol{\theta}_m) \\
&= \ln q^{\setminus i}(\boldsymbol{\theta}_m) f_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]) - \ln q^{\setminus i}(\boldsymbol{\theta}_m) \\
&= \ln f_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]),
\end{aligned}$$

which leads to $\tilde{f}_i(\boldsymbol{\theta}_m) = f_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}])$. Due to the multiplication and integration involved in ELBO, optimizing $\mathcal{L}$ with respect to $\tilde{f}_i(\boldsymbol{\theta}_m)$ does not yield the same results as in CEP. Therefore, each update does not necessarily increase the ELBO, and CEP may not converge in this scenario. Similarly, this local optimization of message factors is also the reason why standard EP may not converge.

3.2.2 Connections to Streaming Bayes

The concept of EP is developed from ADF, an online Bayesian algorithm designed for streaming data. As CEP is a variant of EP, it can be readily adapted into a streaming version. Furthermore, due to the strong connections between CEP and VMP updates, it is straightforward to construct a streaming version of VMP. The resulting method shares a close connection with streaming variational Bayes, although it is developed from a distinct perspective and offers different interpretations.

In Section II, we observe that ADF differs from EP in the factor removing step. In ADF, the removing step is ignored, and the calibrating distribution is replaced by the full approximation obtained from the previous iteration. The updated approximating posterior is computed by directly multiplying the previous approximation with the newly updated message factor associated with the added data.

Mathematically, assuming the current approximation is denoted as $q(\boldsymbol{\theta})$, the new posterior in ADF is

$$q^*(\boldsymbol{\theta}) = \min_{\hat{q}(\boldsymbol{\theta})} \mathrm{KL}(\hat{p}_i(\boldsymbol{\theta}) \| \hat{q}(\boldsymbol{\theta})),$$

where $\hat{p}_i(\boldsymbol{\theta}) \propto f_i(\boldsymbol{\theta}) q(\boldsymbol{\theta})$. The resulting $q^*(\boldsymbol{\theta})$ is then used as the current approximation in the next iteration. In a conjugate-exponential model with mutually independent variable groups, the update of the posterior for each variable has a closed-form solution, given by

$$q^*(\boldsymbol{\theta}_m) = \hat{p}_i(\boldsymbol{\theta}_m) = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m | \boldsymbol{\theta}_{\setminus m})].$$

Upon the arrival of new data, we can optimize each variable group and multiply their distributions together to obtain the new approximation $q^*(\boldsymbol{\theta})$.

As mentioned in the previous subsection, the variable update in VMP merges all the messages sent from the other nodes simultaneously. If the data arrives in a streaming manner, we can sequentially merge the messages to update the variables. Building upon this insight, we can easily modify the VMP to a streaming version. Specifically, when a new sample $\mathbf{x}_i$ arrives, the updated estimate of the posterior in VMP is

$$\begin{aligned} \ln q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln f_i(\boldsymbol{\theta}) q(\boldsymbol{\theta})] + \text{const} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln \hat{p}_i(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln \hat{p}_i(\boldsymbol{\theta}_m | \boldsymbol{\theta}_{\setminus m})]. \end{aligned} \tag{22}$$

We can exploit the multivariate delta method to approximate the expectation, which leads to $q^*(\boldsymbol{\theta}_m) = \hat{p}_i(\boldsymbol{\theta}_m | \mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}])$. Here, the joint distribution can be expressed as $p(\boldsymbol{\theta}, \mathcal{D}) = \hat{p}_i(\boldsymbol{\theta}) \propto f_i(\boldsymbol{\theta}) q(\boldsymbol{\theta})$. Similar to ADF, we can optimize each variable group through (22) and then multiply the respective distributions together to obtain the new approximate estimate.

It is worth noting that the algorithm can be easily extended to scenarios where data arrive in a batch version. Additionally, standard VI with i.i.d. observations can also be easily modified to a streaming version through this framework. Moreover, it can be seen that the primary difference between streaming VMP and ADF is that the expectations are taken with respect to different distributions. Based on the connection between VMP and CEP, we can present the following corollary.

Corollary 2: *Consider a conjugate-exponential probabilistic model represented as a Bayesian network. Suppose the variational distribution follows the mean-field assumption and the observations are i.i.d. Then, streaming VMP and ADF yield the same update equations under the following conditions:*

- *The current approximation $q(\boldsymbol{\theta}_{\setminus m})$ is used as an surrogate of $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$ in the computation of the expectation in ADF;*
- *The expectations are approximated using the multivariate delta method.*

Proof: See Appendix C.
Since VMP is a special case of VI, it follows that streaming VI also has the same update equations to ADF under these conditions, provided that the underlying probabilistic model is a conjugate-exponential model.

Note that the streaming version of VMP or CEP performs a one-pass update, discarding the data once they are updated, which significantly reduces the storage requirements. Additionally, the variable update in VMP can be implemented in a distributed manner since the messages can also be calculated in parallel. The resulting algorithm is similar to the distributed VMP (Masegosa et al., 2016).

3.3 Interpretation via Graphical Models

Since both VMP and CEP are closely related to graphical models, we can gain further insights into their connection from a graphical model perspective. Specifically, we assume that the model takes the form of a Bayesian network, and the joint distribution can be expressed as[4]

$$p(\mathbf{V}) = \prod_i p(\mathbf{v}_i | \text{pa}_i), \tag{23}$$

where $\mathbf{V} = \{\boldsymbol{\theta}, \mathcal{D}\}$ contains all the visible and hidden variables; pa_i denotes the set of variables corresponding to the parents of node i; and $\mathbf{v}_i$ denotes the variable or group of variables associated with node i.

[4]Here we use a similar notation as in the original VMP paper.

Assume that the variational distribution is fully factorized with respect to the hidden variables, which means each variable group has only one variable. In VMP, the optimized form for each variable is given by

$$\begin{aligned}\ln q^*(\boldsymbol{\theta}_j) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus j})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const} \\ &= \langle \ln p(\mathbf{V}) \rangle_{q(\boldsymbol{\theta}_{\setminus j})} + \text{const},\end{aligned} \tag{24}$$

where $\langle \cdot \rangle_{q(\boldsymbol{\theta}_{\setminus j})}$ denotes the expectation with respect to $q(\boldsymbol{\theta}_{\setminus j})$. Substituting the joint probability distribution (23) into (24) leads to:

$$\ln q^*(\boldsymbol{\theta}_j) = \left\langle \sum_i \ln p(\mathbf{v}_i | \text{pa}_i) \right\rangle_{q(\boldsymbol{\theta}_{\setminus j})} + \text{const}.$$

Here we only need to consider the variables in the Markov blanket of node j since the terms that do not depend on $\boldsymbol{\theta}_j$ are constant under the expectation. Then we have

$$\ln q^*(\boldsymbol{\theta}_j) = \langle \ln p(\boldsymbol{\theta}_j | \text{pa}_j) \rangle_{q(\boldsymbol{\theta}_{\setminus j})} + \sum_{k \in \text{ch}_j} \langle \ln p(\mathbf{v}_k | \text{pa}_k) \rangle_{q(\boldsymbol{\theta}_{\setminus j})} + \text{const}, \tag{25}$$

where ch_j denotes the index set that corresponds to the children of node j. The parent node of $\mathbf{v}_k$ includes the node j and the co-parents cp_j.

In a conjugate-exponential model, we have

$$\ln p(\boldsymbol{\theta}_j | \text{pa}_j) = \boldsymbol{\eta}_j(\text{pa}_j)^T \boldsymbol{\phi}_j(\theta_j) + Z_j(\text{pa}_j) + \ln h_j(\boldsymbol{\theta}_j), \tag{26}$$

and

$$\begin{aligned}\ln p(\mathbf{v}_k | \text{pa}_k) &= \boldsymbol{\eta}_k(\boldsymbol{\theta}_j, \text{cp}_j)^T \boldsymbol{\phi}_k(\mathbf{v}_k) + Z_k(\boldsymbol{\theta}_j, \text{cp}_j) + \ln h_k(\mathbf{v}_k) \\ &= \boldsymbol{\eta}_{kj}(\mathbf{v}_k, \text{cp}_j)^T \boldsymbol{\phi}_j(\boldsymbol{\theta}_j) + \lambda(\mathbf{v}_k, \text{cp}_j),\end{aligned} \tag{27}$$

where λ is a function that contains the terms irrelevant to $\boldsymbol{\eta}_{kj}$ and $\boldsymbol{\phi}_j(\boldsymbol{\theta}_j)$. The second equation holds due to the conjugacy property. Substituting (26) and (27) into (25) will give

$$\ln q^*(\boldsymbol{\theta}_j) = \left[\left\langle \boldsymbol{\eta}_j(\text{pa}_j) \right\rangle_{q(\boldsymbol{\theta}_{\setminus j})} + \sum_{k \in \text{ch}_j} \left\langle \boldsymbol{\eta}_{kj}(\mathbf{v}_k, \text{cp}_j) \right\rangle_{q(\boldsymbol{\theta}_{\setminus j})} \right]^T \boldsymbol{\phi}_j(\boldsymbol{\theta}_j) + \ln h_j(\boldsymbol{\theta}_j) + \text{const}.$$

It follows that the optimal variational distribution $q^*(\boldsymbol{\theta}_j)$ is also an exponential family distribution and has the same form as $p(\boldsymbol{\theta}_j | \text{pa}_j)$, of which the natural parameter is given by

$$\boldsymbol{\eta}_j^* = \left\langle \boldsymbol{\eta}_j(\text{pa}_j) \right\rangle + \sum_{k \in \text{ch}_j} \left\langle \boldsymbol{\eta}_{kj}(\mathbf{v}_k, \text{cp}_j) \right\rangle, \tag{28}$$

where the expectation are with respect to $q(\boldsymbol{\theta}_{\setminus j})$ and we omit it here for notational simplicity. Equation (28) can also be interpreted as merging the messages sent by the nearby nodes. In practice, we usually reparameterise these functions in terms of these expectations to make the computation tractable, which leads to

$$\tilde{\boldsymbol{\eta}}_j^* = \tilde{\boldsymbol{\eta}}_j(\{\langle \boldsymbol{\phi}_s \rangle\}_{s \in \text{pa}_j}) + \sum_{k \in \text{ch}_j} \tilde{\boldsymbol{\eta}}_{kj}(\langle \boldsymbol{\phi}_k \rangle, \{\langle \boldsymbol{\phi}_t \rangle\}_{t \in \text{cp}_k}).$$

To show the connection between VMP and CEP, consider a conjugate-exponential model with i.i.d. observations, with a graphical illustration shown in Fig. 2. For this model, the natural parameter of the optimal variational distribution in VMP is

$$\tilde{\boldsymbol{\eta}}_j^* = \tilde{\boldsymbol{\eta}}_j(\{\langle \boldsymbol{\phi}_s \rangle\}_{s \in \text{pa}_j}) + \sum_{\mathbf{x}_k \in D} \tilde{\boldsymbol{\eta}}_{kj}(\mathbf{x}_k, \{\langle \boldsymbol{\phi}_t \rangle\}_{t \neq j}). \tag{29}$$

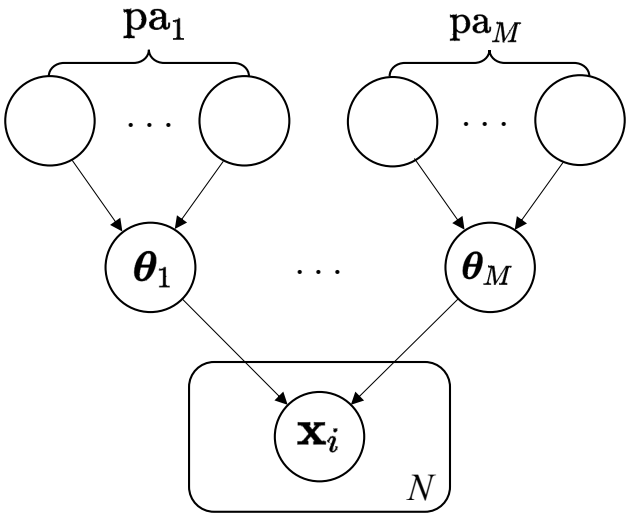

Figure 2: A graphical illustration of the considered conjugate-exponential model with i.i.d. observation.

In CEP, the optimal variational distribution is also in the exponential family and can be expressed as

$$\ln q^*(\boldsymbol{\theta}_j) = \ln p(\boldsymbol{\theta}_j|\mathrm{pa}_j) + \sum_{k=1}^{N} \ln \tilde{f}_k(\boldsymbol{\theta}_j), \tag{30}$$

where $\ln p(\boldsymbol{\theta}_j|\mathrm{pa}_j)$ is given by (26). We need to determine the form of $\ln \tilde{f}_k(\boldsymbol{\theta}_j)$. From Lemma 5, we have

$$\ln \tilde{f}_k(\boldsymbol{\theta}_j) = \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus j})}[\ln f_k(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{\setminus j})]. \tag{31}$$

From (27), the likelihood is also an exponential family distribution and can be expressed as

$$\begin{aligned} \ln f_k(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{\setminus j}) &= \ln p(\mathbf{x}_k|\boldsymbol{\theta}_j, \boldsymbol{\theta}_{\setminus j}) \\ &= \boldsymbol{\eta}_{kj}(\mathbf{v}_k, \mathrm{cp}_j)^T \boldsymbol{\phi}_j(\boldsymbol{\theta}_j) + \lambda(\mathbf{v}_k, \mathrm{cp}_j). \end{aligned} \tag{32}$$

By substituting (32) into (31) and utilizing the reparameterization trick, the approximate factor can be expressed as

$$\ln \tilde{f}_k(\boldsymbol{\theta}_j) = \tilde{\boldsymbol{\eta}}_{kj}(\mathbf{x}_k, \{\langle \boldsymbol{\phi}_t \rangle\}_{t\neq j})^T \boldsymbol{\phi}_j(\boldsymbol{\theta}_j) + \lambda(\mathbf{x}_k, \{\langle \boldsymbol{\phi}_t \rangle\}_{t\neq j})). \tag{33}$$

Substituting it into (30), the resulting distribution shares the same natural parameters as in (29).

Generally, the factors $\{\tilde{f}_k\}$ can be interpreted as the messages sent from the data nodes, after replacing the message sent from the other co-parent nodes with the corresponding moments. Additionally, for a fully factorized model, the standard EP will reduce to loopy belief propagation. More discussions concerning the performance and convergence of LBP can be found in Frey & MacKay (1997); Li et al. (2019); Du et al. (2018a).

3.4 Summary and Practical Suggestions

The previous subsections have unveiled some relationships among various ABI methods, shedding light on their theoretical properties. This subsection presents a brief summary and connections among our established theoretical results (see Fig. 3) and provides recommendations on applying these findings to address practical inference problems.

We begin by illustrating the theoretical connections presented in this study, as depicted in Fig. 3. Lemmas 1 and 2 serve as foundational results leading to Lemma 3, which connects moment matching with the

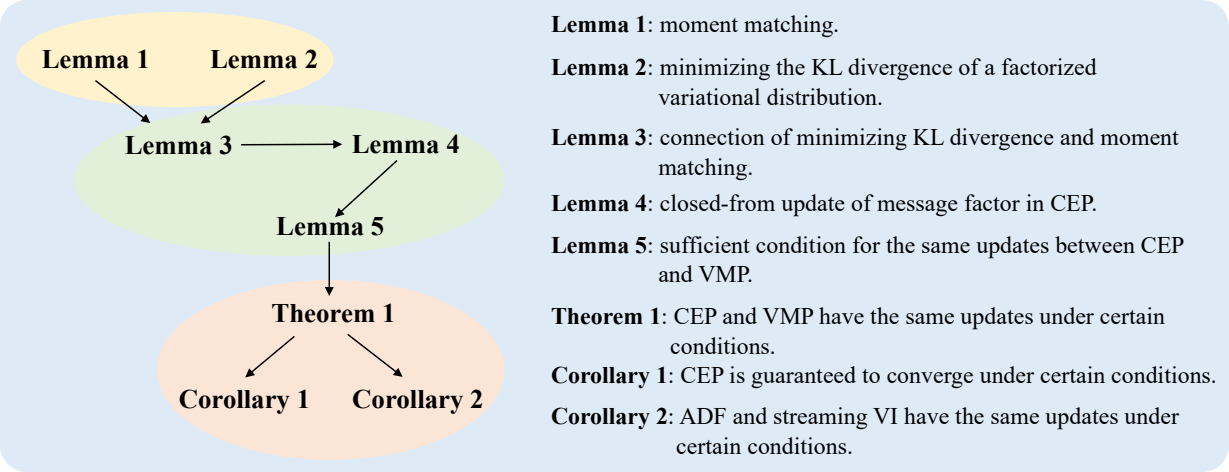

Figure 3: A summary of the theoretical results and their connections in this study.

minimization of KL divergence. This connection enables the derivation of a closed-form message factor in CEP under certain conditions. Based on this closed-form factor and the sufficient condition in Lemma 5, we establish Theorem 1, the main theorem. Following this, Corollaries 1 and 2 provide additional theoretical insights and connections among different ABI methods.

Based on these theoretical results, we offer practical suggestions for applying different ABI methods:

- *Convergent CEP*: For the CEP algorithm, it is guaranteed to converge if the conditions in Theorem 1 are satisfied, as outlined in Corollary 1.

- *Streaming/parallel VMP and VI*: Since the update of VMP or VI can be interpreted as merging messages from other nodes, the resulting algorithms are readily adaptable for streaming data or parallel updates, provided that the conditions in Theorem 1 are satisfied, as shown in (22).

- *Streaming ABI from scratch*: When developing a streaming ABI algorithm, we can first assess whether the model is a conjugate-exponential model with independent variable groups. If so, streaming VI provides closed-form updates for each variable. If not, we could consider employing the moment-matching technique to approximate the posterior distribution with an exponential family distribution.

While these suggestions represent straightforward applications of our findings, the insights we have developed can further pave the way for more advanced Bayesian methods, which is an interesting future research direction.

4 Example

In this section, we demonstrate the strong connections between the updates of VMP and CEP in the context of a Bayesian tensor decomposition model. Our emphasis is on the canonical polyadic decomposition (CPD), which is an essential technique in machine learning and has been used in various real-world applications. Our choice of Bayesian CPD as the illustrative example is motivated by its prominent role in the original CEP paper (Wang & Zhe, 2020), where the inference algorithm is extensively discussed. We start by introducing a probabilistic model for the CPD approach. To this end, we apply both VMP and CEP to infer the associated posterior distribution. Finally, we extend the developed algorithm to a streaming version, enabling it to handle data arriving sequentially.

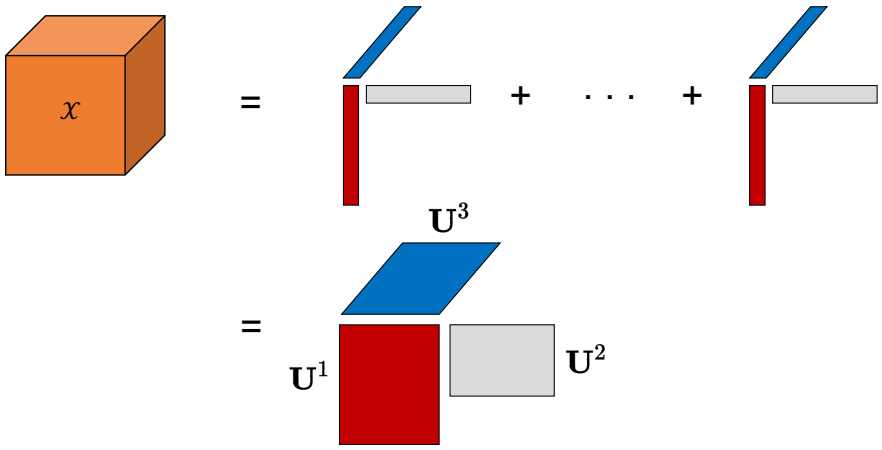

Figure 4: A graphical illustration of the three-dimensional CPD.

4.1 Probabilistic modeling

We denote a K-mode tensor by $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, where d_k is the dimension of the k-th mode. The entry value at location $\mathbf{i} = (i_1, \cdots, i_K)$ is denoted as $x_{\mathbf{i}}$. To perform tensor decomposition, we introduce an R-dimensional embedding vector $\mathbf{u}_j^k$ to represent each object in mode k. Then, a $d_k \times R$ matrix can be constructed by stacking all the embedding vectors in mode k, i.e., $\mathbf{U}^k = [\mathbf{u}_1^k, \cdots, \mathbf{u}_{d_k}^k]^T$. Tensor decomposition aims to find the embedding matrices of all modes $\mathcal{U} = \{\mathbf{U}^1, \cdots, \mathbf{U}^K\}$ from the observed entries.

Mathematically, the CPD of a given tensor $\mathcal{X}$ is written as

$$\mathcal{X} = [\![\mathbf{U}^1, \cdots, \mathbf{U}^K]\!],$$

where $[\![\cdot]\!]$ is the Kruskal operator. A graphical illustration of a three-dimensional CPD ($K = 3$) is shown in Fig. 4. For each entry $x_{\mathbf{i}}$, we have

$$x_{\mathbf{i}} = \sum_{r=1}^{R} \prod_{k=1}^{K} u_{i_k,r}^k = \mathbf{1}^T (\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K),$$

where $\circ$ is the Hadamard product.

Consider a K-mode tensor $\mathcal{Y}$ with N observed entries denoted as $\{y_{\mathbf{i}}\}_{\mathbf{i} \in \mathcal{S}}$. Here, $\mathcal{S}$ represents the index set, and its cardinality is $|\mathcal{S}| = N$. We assume that the observations are contaminated with i.i.d. Gaussian noise. Then the likelihood can be expressed as

$$p(y_{\mathbf{i}} | \mathcal{U}, \tau) = \mathcal{N}(y_{\mathbf{i}} | \mathbf{1}^T (\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K), \tau^{-1}),$$

where τ is the noise precision. We further assign a conjugate Gamma prior over τ, given by

$$p(\tau | a_0, b_0) = \mathrm{Gam}(\tau | a_0, b_0).$$

For each embedding vector $\mathbf{u}_s^k$, we assign a Gaussian prior with mean $\boldsymbol{\beta}_s^k$ and covariance $v\mathbf{I}$, given by

$$p(\mathcal{U}) = \prod_{k=1}^{K} \prod_{s=1}^{d_k} \mathcal{N}(\mathbf{u}_s^k | \boldsymbol{\beta}_s^k, v\mathbf{I}),$$

where $\{\boldsymbol{\beta}_s^k\}$ and v are pre-defined hyperparameters.

Consequently, the joint probability distribution is

$$p(\{y_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{S}},\mathcal{U},\tau) = \mathrm{Gam}(\tau|a_0,b_0)\prod_{k=1}^{K}\prod_{s=1}^{d_k}\mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\beta}_s^k, v\mathbf{I}) \cdot \prod_{\mathbf{i}\in\mathcal{S}}\mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^T(\mathbf{u}_{i_1}^1\circ\cdots\circ\mathbf{u}_{i_K}^K),\tau^{-1}). \tag{34}$$

Note that the prior and likelihood are conjugate and belong to the exponential family, thus the probabilistic model is a conjugate-exponential model. Additionally, the observations are assumed to be i.i.d., therefore satisfying the conditions in Theorem 1.

4.2 VMP

In VMP, the variables are assumed to be mutually independent, allowing us to factorize the variational distribution as

$$q(\mathcal{U},\tau) = q(\tau)\prod_{k=1}^{K}\prod_{s=1}^{d_k}q(\mathbf{u}_s^k).$$

Since the probabilistic model is conjugate-exponential, the variational distribution for each variable is identical to its prior distribution. Consequently, the variational distribution is parameterized by

$$q(\mathcal{U},\tau) = \mathrm{Gam}(\tau|a,b)\prod_{k=1}^{K}\prod_{s=1}^{d_k}\mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^k,\boldsymbol{\Sigma}_s^k).$$

Due to the conjugacy property, we can derive closed-form updates for each variable. Here we present the key steps and leave the detailed derivation in Appendix D. Specifically, the optimal variational distribution of $\mathbf{u}_s^k$ is given by

$$q^*(\mathbf{u}_s^k) = \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^{k^*},\boldsymbol{\Sigma}_s^{k^*}),$$

with the mean $\boldsymbol{\mu}_s^{k^*}$ and covariance $\boldsymbol{\Sigma}_s^{k^*}$ given by

$$\begin{aligned}\boldsymbol{\mu}_s^{k^*} &= \boldsymbol{\Sigma}_s^{k^*}\left(\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S},i_k=s}y_{\mathbf{i}}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}\rangle + v\boldsymbol{\beta}_s^k\right),\\ \boldsymbol{\Sigma}_s^{k^*} &= \left(\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S},i_k=s}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}\mathbf{z}_{\mathbf{i}}^{\backslash k^T}\rangle + v\mathbf{I}\right)^{-1},\end{aligned} \tag{35}$$

where $\langle\cdot\rangle$ denotes the expectation $\mathbb{E}_q[\cdot]$ and

$$\mathbf{z}_{\mathbf{i}}^{\backslash k} = \mathbf{u}_{i_1}^1\circ\cdots\circ\mathbf{u}_{i_{k-1}}^{k-1}\circ\mathbf{u}_{i_{k+1}}^{k+1}\circ\cdots\circ\mathbf{u}_{i_K}^K.$$

The optimal variational distribution of noise precision τ is given by

$$q^*(\tau) = \mathrm{Gam}(\tau|a^*,b^*),$$

with a^* and b^* computed as follows

$$\begin{aligned}a^* &= a_0 + \frac{N}{2},\\ b^* &= b_0 + \frac{1}{2}\sum_{\mathbf{i}\in\mathcal{S}}[y_{\mathbf{i}}^2 - 2y_{\mathbf{i}}\langle\mathbf{1}^T\mathbf{z}_{\mathbf{i}}\rangle + \langle(\mathbf{1}^T\mathbf{z}_{\mathbf{i}})^2\rangle],\end{aligned}$$

where $\mathbf{z}_{\mathbf{i}} = \mathbf{u}_{i_1}^1\circ\cdots\circ\mathbf{u}_{i_K}^K$.

4.3 CEP

In the context of CEP, the approximation factor $\tilde{f}_{\mathbf{i}}$ is assumed to be factorized with variables, given by

$$\tilde{f}_{\mathbf{i}}(\mathcal{U}, \tau) = \tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k),$$

where the factors have the same form as the prior distribution but with different parameters. Specifically, $\tilde{f}_{\mathbf{i}}(\tau) = \mathrm{Gam}(\tau|a_{\mathbf{i}}, b_{\mathbf{i}})$ and $\tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{\mathbf{i}}^k, \mathbf{S}_{\mathbf{i}}^k)$. Consequently, the approximate distribution is given by

$$\begin{aligned} q(\mathcal{U}, \tau) \propto & \mathrm{Gam}(\tau|a_0, b_0) \prod_{k=1}^{K} \prod_{s=1}^{d_k} \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\beta}_s^k, v\mathbf{I}) \\ & \cdot \prod_{\mathbf{i}\in\mathcal{S}} \tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k). \end{aligned}$$

It can be seen that the approximate distribution is factorized over the variables, i.e.,

$$q(\mathcal{U}, \tau) = q(\tau) \prod_{k=1}^{K} \prod_{s=1}^{d_k} q(\mathbf{u}_s^k),$$

where

$$\begin{aligned} q(\tau) &\propto \mathrm{Gam}(\tau|a_0, b_0) \prod_{\mathbf{i}\in\mathcal{S}} \tilde{f}_{\mathbf{i}}(\tau), \\ q(\mathbf{u}_s^k) &\propto \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\beta}_s^k, v\mathbf{I}) \prod_{\mathbf{i}\in\mathcal{S}, i_k=s} \tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k). \end{aligned} \tag{36}$$

To update the variational distribution $q(\mathbf{u}_s^k)$, we need to determine the optimal factor $\tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k)$. The first step is to obtain the calibrating distribution

$$q^{\setminus \mathbf{i}}(\mathcal{U}, \tau) \propto \frac{q(\mathcal{U}, \tau)}{\tilde{f}_{\mathbf{i}}(\tau) \prod_{k=1}^{K} \tilde{f}_{\mathbf{i}}^{k}(\mathbf{u}_{i_k}^k)}.$$

Next, we construct the tilted distribution as

$$\hat{p}_{\mathbf{i}}(\mathcal{U}, \tau) \propto q^{\setminus \mathbf{i}}(\mathcal{U}, \tau) \mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^T(\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K), \tau^{-1}).$$

Since only the moments for the precision τ and the embedding vectors that associate with entry $\mathbf{i}$, $\mathbf{u}_{\mathbf{i}} = \{\mathbf{u}_{i_1}^1, \cdots, \mathbf{u}_{i_k}^K\}$, are needed and the other embeddings vectors will be marginalized out, we can focus on the marginal titled distribution for $\{\mathbf{u}_{\mathbf{i}}, \tau\}$,

$$\hat{p}_{\mathbf{i}}(\mathbf{u}_{\mathbf{i}}, \tau) \propto q^{\setminus \mathbf{i}}(\tau) \prod_{k=1}^{K} q^{\setminus \mathbf{i}}(\mathbf{u}_{i_k}^k) \mathcal{N}(y_{\mathbf{i}}|\mathbf{1}^T(\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K), \tau^{-1}),$$

where

$$q^{\setminus \mathbf{i}}(\tau) = \mathrm{Gam}(\tau|a^{\setminus \mathbf{i}}, b^{\setminus \mathbf{i}}), \quad q^{\setminus \mathbf{i}}(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k|\mathbf{m}_{i_k}^k, \mathbf{S}_{i_k}^k),$$

with

$$
\begin{aligned}
a^{\backslash \mathbf{i}} &= a_0 + \sum_{\mathbf{j}\in\mathcal{S},\mathbf{j}\neq\mathbf{i}} a_{\mathbf{j}} - N + 1, \\
b^{\backslash \mathbf{i}} &= b_0 + \sum_{\mathbf{j}\in\mathcal{S},\mathbf{j}\neq\mathbf{i}} b_{\mathbf{j}}, \\
\mathbf{S}_{i_k}^k &= \left(\sum_{\mathbf{j}\in\mathcal{S},\mathbf{j}\neq\mathbf{i},j_k=i_k} (\mathbf{S}_{\mathbf{j}}^k)^{-1} + v\mathbf{I} \right)^{-1}, \\
\mathbf{m}_{i_k}^k &= \mathbf{S}_{i_k}^k \left(\sum_{\mathbf{j}\in\mathcal{S},\mathbf{j}\neq\mathbf{i},j_k=i_k} (\mathbf{S}_{\mathbf{j}}^k)^{-1}\mathbf{m}_{\mathbf{j}}^k + v\boldsymbol{\beta}_{i_k}^k \right).
\end{aligned}
$$

The next step is to compute conditional moments with respect to the conditional tilted distribution given τ and $\mathbf{u}_{\mathbf{i}}^{\backslash k} = \{\mathbf{u}_{i_1}^1, \cdots, \mathbf{u}_{i_K}^K\}$ fixed, which can be expressed as

$$
\hat{p}_{\mathbf{i}}(\mathbf{u}_{i_k}^k | \mathbf{u}_{\mathbf{i}}^{\backslash k}, \tau) \propto \mathcal{N}(\mathbf{u}_{i_k}^k | \mathbf{m}_{i_k}^k, \mathbf{S}_{i_k}^k) \mathcal{N}(y_{\mathbf{i}} | \mathbf{1}^T (\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K), \tau^{-1}).
$$

It can be observed that this is a Gaussian distribution with covariance and mean given by

$$
\begin{aligned}
\operatorname{cov}(\mathbf{u}_{i_k}^k | \mathbf{u}_{\mathbf{i}}^{\backslash k}, \tau) &= \left[(\mathbf{S}_{i_k}^k)^{-1} + \tau (\mathbf{z}_{\mathbf{i}}^{\backslash k} \mathbf{z}_{\mathbf{i}}^{\backslash k^T}) \right]^{-1}, \\
\mathbb{E}(\mathbf{u}_{i_k}^k | \mathbf{u}_{\mathbf{i}}^{\backslash k}, \tau) &= \operatorname{cov}(\mathbf{u}_{i_k}^k | \mathbf{u}_{\mathbf{i}}^{\backslash k}, \tau) \left[(\mathbf{S}_{i_k}^k)^{-1} \mathbf{m}_{i_k}^k + \tau y_{\mathbf{i}} \mathbf{z}_{\mathbf{i}}^{\backslash k} \right].
\end{aligned}
$$

According to Lemma 4, the optimal factor is given by $\tilde{f}_{\mathbf{i}}^k(\mathbf{u}_{i_k}^k) = \mathcal{N}(\mathbf{u}_{i_k}^k | \mathbf{m}_{\mathbf{i}}^{k^*}, \mathbf{S}_{\mathbf{i}}^{k^*})$ with

$$
\begin{aligned}
\mathbf{S}_{\mathbf{i}}^{k^*} &= \left(\langle \tau \rangle \langle \mathbf{z}_{\mathbf{i}}^{\backslash k} \mathbf{z}_{\mathbf{i}}^{\backslash k^T} \rangle \right)^{-1}, \\
\mathbf{m}_{\mathbf{i}}^{k^*} &= \mathbf{S}_{\mathbf{i}}^{k^*} (y_{\mathbf{i}} \langle \tau \rangle \langle \mathbf{z}_{\mathbf{i}}^{\backslash k} \rangle).
\end{aligned}
$$

It is worth noting that the message factors can be calculated in parallel, which can significantly reduce time consumption. After obtaining all the message factors, we can merge them to obtain the approximation distribution. Based on (36), the optimal variational distribution for $q(\mathbf{u}_s^k)$ is given by $q^*(\mathbf{u}_s^k) = \mathcal{N}(\mathbf{u}_s^k | \boldsymbol{\mu}_s^{k^*}, \boldsymbol{\Sigma}_s^{k^*})$, where

$$
\begin{aligned}
\boldsymbol{\mu}_s^{k^*} &= \boldsymbol{\Sigma}_s^{k^*} \left(\langle \tau \rangle \sum_{\mathbf{i}\in\mathcal{S},i_k=s} y_{\mathbf{i}} \langle \mathbf{z}_{\mathbf{i}}^{\backslash k} \rangle + v\boldsymbol{\beta}_s^k \right), \\
\boldsymbol{\Sigma}_s^{k^*} &= \left(\langle \tau \rangle \sum_{\mathbf{i}\in\mathcal{S},i_k=s} \langle \mathbf{z}_{\mathbf{i}}^{\backslash k} \mathbf{z}_{\mathbf{i}}^{\backslash k^T} \rangle + v\mathbf{I} \right)^{-1}.
\end{aligned} \tag{37}
$$

Comparing (35) and (37), it can be seen that the optimal variational distributions obtained by VMP and CEP are the same.

For noise precision τ, the conditional tilted distribution is given by

$$
\hat{p}_{\mathbf{i}}(\tau | \mathbf{u}_{\mathbf{i}}) = \operatorname{Gam}(\tau | \hat{a}, \hat{b}), \tag{38}
$$

where

$$
\begin{aligned}
\hat{a} &= a^{\backslash \mathbf{i}} + \frac{1}{2}, \\
\hat{b} &= b^{\backslash \mathbf{i}} + \frac{1}{2} [y_{\mathbf{i}} - \mathbf{1}^T \mathbf{z}_{\mathbf{i}}]^2.
\end{aligned} \tag{39}
$$

Then the optimal message factor can be calculated as $\tilde{f}_{\mathbf{i}}(\tau) = \text{Gam}(\tau|a^*_{\mathbf{i}}, b^*_{\mathbf{i}})$ with

$$\begin{aligned} a^*_{\mathbf{i}} &= \frac{1}{2}, \\ b^*_{\mathbf{i}} &= \frac{1}{2}[y^2_{\mathbf{i}} - 2y_{\mathbf{i}}\langle \mathbf{1}^T\mathbf{z}_{\mathbf{i}} + \langle(\mathbf{1}^T\mathbf{z}_{\mathbf{i}})^2\rangle]. \end{aligned} \tag{40}$$

Merging these factors through (36) will leads to $q^*(\tau) = \text{Gam}(\tau|a^*, b^*)$, where

$$\begin{aligned} a^* &= a_0 + \frac{N}{2}, \\ b^* &= b_0 + \frac{1}{2}\sum_{\mathbf{i}\in\mathcal{S}}[y^2_{\mathbf{i}} - 2y_{\mathbf{i}}\langle \mathbf{1}^T\mathbf{z}_{\mathbf{i}}\rangle + \langle(\mathbf{1}^T\mathbf{z}_{\mathbf{i}})^2\rangle], \end{aligned}$$

which are the same as in VMP. Consequently, we can conclude that the update of variables in VMP and CEP are the same. These closed-form updates demonstrate promising accuracy and empirically show a fast convergence in many real-world applications (Wang & Zhe, 2020).

4.4 Streaming VMP

The connection of VMP and CEP enables the algorithm to be easily adapted to a streaming version. In the streaming version, we assume that every time we receive one data point and the current approximation is $q(\mathbf{u}^k_s) = \mathcal{N}(\mathbf{u}^k_s|\boldsymbol{\mu}^k_s, \boldsymbol{\Sigma}^k_s)$ and $q(\tau) = \text{Gam}(\tau|a, b)$. Then based on (22), the posterior update of the streaming version is given by $q^*(\mathbf{u}^k_s) = \hat{p}_{\mathbf{i}}(\mathbf{u}^k_{i_k}|\langle\mathbf{u}^{\setminus k}_{\mathbf{i}}\rangle, \langle\tau\rangle) = \mathcal{N}(\mathbf{u}^k_s|\boldsymbol{\mu}^{k^*}_s, \boldsymbol{\Sigma}^{k^*}_s)$ with

$$\begin{aligned} \boldsymbol{\Sigma}^{k^*}_s &= [(\boldsymbol{\Sigma}^k_s)^{-1} + \langle\tau\rangle(\langle\mathbf{z}^{\setminus k}_{\mathbf{i}}\mathbf{z}^{\setminus k^T}_{\mathbf{i}}\rangle)]^{-1}, \\ \boldsymbol{\mu}^{k^*}_s &= \boldsymbol{\Sigma}^{k^*}_s[(\boldsymbol{\Sigma}^k_s)^{-1}\boldsymbol{\mu}^k_s + \langle\tau\rangle y_{\mathbf{i}}\langle\mathbf{z}^{\setminus k}_{\mathbf{i}}\rangle], \end{aligned} \tag{41}$$

and $q^*(\tau) = \hat{p}_{\mathbf{i}}(\tau|\langle\mathbf{u}_{\mathbf{i}}\rangle) = \text{Gam}(\tau|a^*, b^*)$ with

$$\begin{aligned} a^* &= a + \frac{1}{2}, \\ b^* &= b + \frac{1}{2}[y^2_{\mathbf{i}} - 2y_{\mathbf{i}}\langle \mathbf{1}^T\mathbf{z}_{\mathbf{i}} + \langle(\mathbf{1}^T\mathbf{z}_{\mathbf{i}})^2\rangle]. \end{aligned}$$

It is important to note that the natural parameters used here are different from those used in CEP (see (41) and (37)) since the calibrating distribution is replaced with the full approximation from the previous iteration. Additionally, the developed algorithm can be readily extended to the scenario where data are arrived in a batch version. The resulting algorithm is the same as probabilistic streaming tensor decomposition (POST) (Du et al., 2018b), which is flexible and demonstrates promising performance in this task.

5 Discussions and Future Directions

This paper investigates the theoretical connections among different ABI methods, starting by bridging the VMP and CEP. Specifically, we have demonstrated a strong link between these two methods under mild conditions. This newly identified connection not only guarantees the convergence of CEP but also allows for the seamless construction of a streaming version of the VMP algorithm. The key insight is that the variable updates in VMP and CEP are intrinsically merging the messages sent by all the data points and they share a common objective of approximating the conditional marginal distribution. Additionally, this finding provides insights into the underlying relationships and distinct characteristics of other ABI methods, including the same expressions between ADF and streaming VI updates.

Generally, VI and EP have different properties and performance since they optimize the opposite directions of KL divergence. This work, for the first time, demonstrates that their variants are closely related under certain conditions, which sheds new light on the understanding and development of further advanced ABI

methods. However, our theoretical analysis is restricted to the conjugate-exponential family of models. It would be interesting to explore the application of these connections in other model families or non-conjugate scenarios. We believe that these explorations will open new avenues for future research on efficient and accurate Bayesian learning algorithms, particularly in the context of streaming and large-scale data.

References

Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, 2003.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

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

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. *Advances in neural information processing systems*, 26, 2013.

Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Institute of Mathematical Statistics, 1986.

Antoni B. Chan and Nuno Vasconcelos. Layered dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1862–1879, 2009.

Lei Cheng, Zhongtao Chen, Qingjiang Shi, Yik-Chung Wu, and Sergios Theodoridis. Towards flexible sparsity-aware modeling: Automatic tensor rank learning using the generalized hyperbolic prior. *IEEE Transactions on Signal Processing*, 70:1834–1849, 2022a.

Lei Cheng, Feng Yin, Sergios Theodoridis, Sotirios Chatzis, and Tsung-Hui Chang. Rethinking bayesian learning for data analysis: The art of prior and inference in sparsity-aware modeling. *IEEE Signal Processing Magazine*, 39(6):18–52, 2022b.

Shay B. Cohen and Noah A. Smith. Covariance in unsupervised learning of probabilistic grammars. *Journal of Machine Learning Research*, 11(101):3017–3051, 2010.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Jean Daunizeau, Vincent Adam, and Lionel Rigoux. Vba: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS computational biology*, 10(1):e1003441, 2014.

RA Dorfman. A note on the! d-method for finding variance formulae. *Biometric Bulletin*, 1938.

Jian Du, Shaodan Ma, Yik-Chung Wu, Soummya Kar, and José M. F. Moura. Convergence analysis of distributed inference with vector-valued gaussian belief propagation. *Journal of Machine Learning Research*, 18(172):1–38, 2018a.

Yishuai Du, Yimin Zheng, Kuang-chih Lee, and Shandian Zhe. Probabilistic streaming tensor decomposition. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 99–108, 2018b.

Wentao Fan, Lin Yang, and Nizar Bouguila. Unsupervised grouped axial data modeling via hierarchical bayesian nonparametric models with watson distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9654–9668, 2022.

Shikai Fang, Robert M. Kirby, and Shandian Zhe. Bayesian streaming sparse tucker decomposition. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 558–567. PMLR, 27–30 Jul 2021a.

Shikai Fang, Zheng Wang, Zhimeng Pan, Ji Liu, and Shandian Zhe. Streaming bayesian deep tensor factorization. In *International Conference on Machine Learning*, pp. 3133–3142. PMLR, 2021b.

Brendan J Frey and David MacKay. A revolution: Belief propagation in graphs with cycles. *Advances in neural information processing systems*, 10, 1997.

Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 05 2020.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18(106):1–37, 2017.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An introduction to variational methods for graphical models*, pp. 105âĂŞ161. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262600323.

Mohammad Emtiyaz Khan and Håvard Rue. The bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46, 2023.

Bin Li, Qinliang Su, and Yik-Chung Wu. Fixed points of gaussian belief propagation and relation to convergence. *IEEE Transactions on Signal Processing*, 67(23):6025–6038, 2019.

Bin Li, Nan Wu, and Yik-Chung Wu. Distributed inference with variational message passing in gaussian graphical models: Trade-offs in message schedules and convergence conditions. *IEEE Transactions on Signal Processing*, 2024.

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Wenting Wang. Black-box expectation propagation for bayesian models. In *Proceedings of the 2018 siam international conference on data mining*, pp. 603–611. SIAM, 2018.

Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. Stochastic expectation propagation. *Advances in neural information processing systems*, 28, 2015.

Andres R Masegosa, Ana M Martinez, Helge Langseth, Thomas D Nielsen, Antonio Salmeron, Dario Ramos-Lopez, and Anders L Madsen. d-vmp: Distributed variational message passing. In *Conference on Probabilistic Graphical Models*, pp. 321–332. PMLR, 2016.

P.S. Maybeck. *Stochastic Models, Estimation, and Control*. Mathematics in Science and Engineering. Elsevier Science, 1982. ISBN 9780124807037.

Thomas Minka. Power ep. Technical report, Microsoft Research, Cambridge, 2004.

Thomas Minka. Divergence measures and message passing. Technical Report TR-2005-173, Microsoft Research, 2005.

Thomas P. Minka. Expectation propagation for approximate bayesian inference, 2013.

Thomas P. Minka and Rosalind Picard. *A family of algorithms for approximate Bayesian inference*. PhD thesis, USA, 2001.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

Manfred Opper, Ole Winther, and Michael J Jordan. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(12), 2005.

Alex J Smola, SVN Vishwanathan, and Eleazar Eskin. Laplace propagation. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pp. 441–448, 2004.

Jae Woong Soh and Nam Ik Cho. Variational deep image restoration. *IEEE Transactions on Image Processing*, 31:4363–4376, 2022.

S. Theodoridis. *Machine Learning: From the Classics to Deep Networks, Transformers, and Diffusion Models*. 3nd Ed., Academic Press, 2025.

Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P Cunningham, David Schiminovich, and Christian P Robert. Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21 (17):1–53, 2020.

Jay M Ver Hoef. Who invented the delta method? *The American Statistician*, 66(2):124–127, 2012.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. ISSN 1935-8237.

Chong Wang and David M Blei. Variational inference in non-conjugate models. *Journal of Machine Learning Research*, 2013.

Zheng Wang and Shandian Zhe. Conditional expectation propagation. In *Uncertainty in Artificial Intelligence*, pp. 28–37. PMLR, 2020.

John Winn, Christopher M Bishop, and Tommi Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.

Boyang Xue, Jianwei Yu, Junhao Xu, Shansong Liu, Shoukang Hu, Zi Ye, Mengzhe Geng, Xunying Liu, and Helen M. Meng. Bayesian transformer language models for speech recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7378–7382, 2021.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019.

Jackson Zhou, John T Ormerod, and Clara Grazian. Fast expectation propagation for heteroscedastic, lasso-penalized, and quantile regression. *Journal of Machine Learning Research*, 24(314):1–39, 2023.

A Summary of Algorithms

The procedure of the VMP and CEP methods are summarized in **Algorithm 1** and **Algorithm 2**, respectively.

B The Multivariate Delta Method

In the multivariate delta method, the expectation of a function of a random variable is approximated by the expectation of the function's Taylor expansion. Specifically, given a function $f(\boldsymbol{\theta})$ and a distribution $q(\boldsymbol{\theta})$ with mean $\mathbf{m}$, we can use the first-order Taylor approximation to get,

$$\mathbb{E}_{q(\boldsymbol{\theta})}(f(\boldsymbol{\theta})) \approx \mathbb{E}_q\left[f(\mathbf{m}) + (\boldsymbol{\theta}-\mathbf{m})^T\nabla_{\boldsymbol{\theta}} f(\mathbf{m})\right] \approx f(\mathbf{m}),$$

where ∇ is the differential operator. In CEP, the outer expectation can be approximated by

$$\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[h(\Phi_m)] \approx \mathbb{E}_q\left[h(\mathbb{E}_q(\Phi_m)) + (\Phi_m - \mathbb{E}_q(\Phi_m))^T\nabla h(\mathbb{E}_q(\Phi_m))\right] \approx h(\mathbb{E}_q(\Phi_m)).$$

C Proofs

C.1 Proof of Lemma 3

To prove Lemma 3, we rewrite the KL divergence as

$$\begin{aligned}\mathrm{KL}(p\|q) &= \int p(\boldsymbol{\theta})\ln\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}d\boldsymbol{\theta}\\ &= H[p(\boldsymbol{\theta})] - \int p(\boldsymbol{\theta})\ln q(\boldsymbol{\theta})d\boldsymbol{\theta},\end{aligned}$$

Algorithm 1 Variational Message Passing (VMP)

Input: joint probability distribution $p(\mathcal{D}, \boldsymbol{\theta})$.

1: Initialise each factor distribution $q(\boldsymbol{\theta}_m)$.
2: **while** not converge **do**
3: **for** each variable group **do**
4: Calculate moment of the natural parameters $\mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\boldsymbol{\eta}_m(\boldsymbol{\theta}_{\setminus m}, \mathcal{D})]$ using the messages sent from other nodes.
5: Update the factor distribution $q^*(\boldsymbol{\theta}_m)$ via (6).
6: **end for**
7: **end while**

Output: variational distribution $q(\boldsymbol{\theta}) = \prod_m q^*(\boldsymbol{\theta}_m)$.

Algorithm 2 Conditional Expectation Propagation (CEP)

Input: joint probability distribution $p(\mathcal{D}, \boldsymbol{\theta})$.

1: Initialise each message factor $\tilde{f}_i(\boldsymbol{\theta}_m)$.
2: **while** not converge **do**
3: **for** each variable group **do**
4: **for** each factor $\tilde{f}_i(\boldsymbol{\theta}_m)$ **do**
5: Calculate the calibrating distribution, $q^{\setminus i}(\boldsymbol{\theta}_m) = q(\boldsymbol{\theta}_m)/\tilde{f}_i(\boldsymbol{\theta}_m)$.
6: Derive a new posterior $q^{\natural}(\boldsymbol{\theta}_m)$ via conditional moment matching (11).
7: Update the message factor $\tilde{f}_i(\boldsymbol{\theta}_m) \propto q^{\natural}(\boldsymbol{\theta}_m)/q^{\setminus i}(\boldsymbol{\theta}_m)$.
8: **end for**
9: Merge the message: $q^*(\boldsymbol{\theta}_m) \propto \prod_i \tilde{f}_i(\boldsymbol{\theta}_m)$.
10: **end for**
11: **end while**

Output: variational distribution $q(\boldsymbol{\theta}) = \prod_m q^*(\boldsymbol{\theta}_m) = \frac{1}{Z}\prod_i \prod_m \tilde{f}_i(\boldsymbol{\theta}_m)$.

where $H[\cdot]$ is the entropy. Since the entropy is a constant, minimizing $\mathrm{KL}(p\|q)$ is equivalent to maximizing $\mathcal{L}(q) = \int p(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Exploiting the factorized property, it can be further decomposed as

$$
\begin{aligned}
\mathcal{L}(q) &= \int p(\boldsymbol{\theta}) \sum_m \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta} \\
&= \sum_m \int p(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta} \\
&= \sum_m \int \left(\int p(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\setminus m} \right) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
&= \sum_m \int p(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
&= \sum_m L_m(q(\boldsymbol{\theta}_m)),
\end{aligned}
$$

where we denote $\int p(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m$ as $L_m(q(\boldsymbol{\theta}_m))$. Since the variable groups are mutually independent, maximizing $\mathcal{L}(q)$ with respect to $q(\boldsymbol{\theta})$ is equivalent to maximizing each L_m with respect to $q(\boldsymbol{\theta}_m)$. For each

variable group, the optimum is given by

$$\begin{aligned}\max_{q(\boldsymbol{\theta}_m)} L_m(q(\boldsymbol{\theta}_m)) &= \max_{q(\boldsymbol{\theta}_m)} \int p(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \min_{q(\boldsymbol{\theta}_m)} - \int p(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \min_{q(\boldsymbol{\theta}_m)} H[p(\boldsymbol{\theta}_m)] - \int p(\boldsymbol{\theta}_m) \ln q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \min_{q(\boldsymbol{\theta}_m)} \mathrm{KL}(p(\boldsymbol{\theta}_m) \| q(\boldsymbol{\theta}_m)),\end{aligned}$$

where the third equation holds because the entropy of the marginal distribution $H[p(\boldsymbol{\theta}_m)]$ is irrelevant to $q(\boldsymbol{\theta}_m)$. Using Lemma 1, the optimal solution is achieved by the moment matching

$$\mathbb{E}_{q(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{p(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)].$$

Additionally, the left-hand side of (13) can be expressed as

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] &= \int q(\boldsymbol{\theta}) \phi(\boldsymbol{\theta}_m) d\boldsymbol{\theta} \\ &= \int \left[\int q(\boldsymbol{\theta}_m, \boldsymbol{\theta}_{\setminus m}) d\boldsymbol{\theta}_{\setminus m} \right] \phi(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \int q(\boldsymbol{\theta}_m) \phi(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \mathbb{E}_{q(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)].\end{aligned}$$

Similarly, the right-hand side of (13) can be expressed as $\mathbb{E}_{p(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{p(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)]$. Thus we have

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{q(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{p(\boldsymbol{\theta}_m)}[\phi(\boldsymbol{\theta}_m)] = \mathbb{E}_{p(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}_m)],$$

which completes the proof. ■

C.2 Proof of Lemma 4

From (14) and Lemma 3, it can be seen that the calculation of $q^\natural(\boldsymbol{\theta}_m)$ in CEP is essentially solving the following problem

$$\begin{aligned}&\min_{q(\boldsymbol{\theta}_m)} \mathrm{KL}(\hat{p}_i(\boldsymbol{\theta}) \| q(\boldsymbol{\theta})) \\ &\text{s.t. } q(\boldsymbol{\theta}) = \prod_m q(\boldsymbol{\theta}_m),\end{aligned}$$

where $q(\boldsymbol{\theta}_m)$ belongs to the exponential family. If the $\hat{p}_i(\boldsymbol{\theta}_m)$ is also in the exponential family and has the same form as $q^\natural(\boldsymbol{\theta}_m)$, then the moment matching leads to

$$q^\natural(\boldsymbol{\theta}_m) = \hat{p}_i(\boldsymbol{\theta}_m),$$

where the marginal posterior can be further written as

$$\begin{aligned}\hat{p}_i(\boldsymbol{\theta}_m) &= \int \hat{p}_i(\boldsymbol{\theta}_m, \boldsymbol{\theta}_{\setminus m}) d\boldsymbol{\theta}_{\setminus m} \\ &= \int \hat{p}_i(\boldsymbol{\theta}_{\setminus m}) \hat{p}_i(\boldsymbol{\theta}_m | \boldsymbol{\theta}_{\setminus m}) d\boldsymbol{\theta}_{\setminus m} \\ &= \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m | \boldsymbol{\theta}_{\setminus m})].\end{aligned}$$

In CEP, two approximations are made to derive an analytical form of the update. The first approximation is to use $q(\boldsymbol{\theta}_{\setminus m})$ as a surrogate for $\hat{p}_i(\boldsymbol{\theta}_{\setminus m})$. The second approximation is to use the multivariate delta method

to approximate the expectation of the conditional distribution. Based on these approximations, the optimal approximate posterior $q^\natural(\boldsymbol{\theta}_m)$ can be expressed as

$$\begin{aligned} q^\natural(\boldsymbol{\theta}_m) &= \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})] \\ &\approx \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})] \\ &\approx \hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]). \end{aligned}$$

Generally, $\hat{p}_i(\boldsymbol{\theta}_m)$ is not in the exponential family, so the moment matching step is used to minimize the KL divergence. However, in a conjugate-exponential model, each complete conditional, including $\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})$, is in the exponential family. Additionally, $\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})$ shares the same sufficient statistics as $q^\natural(\boldsymbol{\theta}_m)$ due to the conjugacy property. As a result, $\hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}])$ is used as a surrogate for $q^\natural(\boldsymbol{\theta}_m)$ in CEP. Thus the update of $\tilde{f}_i(\boldsymbol{\theta}_m)$ can be expressed as

$$\tilde{f}_i(\boldsymbol{\theta}_m) \propto \frac{\hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}])}{q^{\setminus i}(\boldsymbol{\theta}_m)},$$

which completes the proof. ■

C.3 Proof of Corollary 1

It has been established in Winn et al. (2005); Minka (2005) that VMP updates are guaranteed to converge to a local minimum of the KL divergence under the conditions stated in Theorem 1. Since CEP follows the same update equations as VMP under these conditions, its convergence property directly follows.

C.4 Proof of Corollary 2

In ADF, the optimal variational distribution in each iteration can be expressed as

$$q^*(\boldsymbol{\theta}_m) = \hat{p}_i(\boldsymbol{\theta}_m) = \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})].$$

With the two conditions in Corollary 2, the optimal distribution can be reformulated as

$$\begin{aligned} q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{\hat{p}_i(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})] \\ &\approx \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})] \\ &\approx \hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]). \end{aligned}$$

Similarly, using the multivariate delta method, the optimal distribution in streaming VMP is given by

$$\begin{aligned} \ln q^*(\boldsymbol{\theta}_m) &= \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln \hat{p}_i(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{\setminus m})] \\ &\approx \ln \hat{p}_i(\boldsymbol{\theta}_m|\mathbb{E}_q[\boldsymbol{\theta}_{\setminus m}]), \end{aligned}$$

which is the same as in ADF. This completes the proof.

D Derivation of the VMP in Tensor Decomposition

In the Bayesian tensor decomposition problem, the unknown parameter set $\boldsymbol{\theta}$ consists of the latent factor matrices $\mathcal{U}$ and hyperparameter τ. The optimal variational distribution for each $\boldsymbol{\theta}_m$ is given by

$$\ln q^*(\boldsymbol{\theta}_m) = \mathbb{E}_{q(\boldsymbol{\theta}_{\setminus m})}[\ln p(\boldsymbol{\theta}, \mathcal{D})] + \text{const.} \tag{42}$$

From (34), the logarithm of the joint density function $\ln p(\boldsymbol{\theta}, \mathcal{D})$ can be expressed as

$$\begin{aligned}\ln p(\boldsymbol{\theta}, \mathcal{D}) &= \ln p(\{y_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{S}}, \mathcal{U}, \tau) \\ &= \frac{N}{2}\ln\tau - \sum_{\mathbf{i}\in\mathcal{S}}\frac{\tau}{2}[y_{\mathbf{i}} - \mathbf{1}^T(\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K)]^2 \\ &\quad - \sum_{k=1}^{K}\sum_{s=1}^{d_k}\frac{v}{2}(\mathbf{u}_s^k - \boldsymbol{\beta}_s^k)^T(\mathbf{u}_s^k - \boldsymbol{\beta}_s^k) \\ &\quad (a_0 - 1)\ln\tau - b_0\tau + \text{const.}\end{aligned} \tag{43}$$

By substituting (43) into (42), we obtain $q^*(\mathbf{u}_s^k)$:

$$\begin{aligned}\ln q^*(\mathbf{u}_s^k) &= \mathbb{E}_q\{-\frac{\tau}{2}\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}[y_{\mathbf{i}} - \mathbf{1}^T(\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K)]^2 \\ &\quad - \frac{v}{2}(\mathbf{u}_s^k - \boldsymbol{\beta}_s^k)^T(\mathbf{u}_s^k - \boldsymbol{\beta}_s^k)\} \\ &= \mathbb{E}_q\{-\frac{\tau}{2}\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}\left[y_{\mathbf{i}}^2 - 2y_{\mathbf{i}}(\mathbf{u}_s^k)^T\mathbf{z}_{\mathbf{i}}^{\backslash k} + (\mathbf{u}_s^k)^T\mathbf{z}_{\mathbf{i}}^{\backslash k}{\mathbf{z}_{\mathbf{i}}^{\backslash k}}^T\mathbf{u}_s^k\right] \\ &\quad - \frac{v}{2}\left[(\mathbf{u}_s^k)^T\mathbf{u}_s^k - 2(\mathbf{u}_s^k)^T\boldsymbol{\beta}_s^k + (\boldsymbol{\beta}_s^k)^T\boldsymbol{\beta}_s^k\right]\} \\ &= -\frac{1}{2}(\mathbf{u}_s^k)^T\left[\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}{\mathbf{z}_{\mathbf{i}}^{\backslash k}}^T\rangle + v\mathbf{I}\right]\mathbf{u}_s^k \\ &\quad + (\mathbf{u}_s^k)^T\left[\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}y_{\mathbf{i}}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}\rangle + v\boldsymbol{\beta}_s^k\right].\end{aligned} \tag{44}$$

We can see from (44) that $\mathbf{u}_s^k$ follows a Gaussian distribution $q^*(\mathbf{u}_s^k) = \mathcal{N}(\mathbf{u}_s^k|\boldsymbol{\mu}_s^{k*}, \boldsymbol{\Sigma}_s^{k*})$, of which the mean and covariance are given by

$$\begin{aligned}\boldsymbol{\mu}_s^{k*} &= \boldsymbol{\Sigma}_s^{k*}\left(\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}y_{\mathbf{i}}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}\rangle + v\boldsymbol{\beta}_s^k\right), \\ \boldsymbol{\Sigma}_s^{k*} &= \left(\langle\tau\rangle\sum_{\mathbf{i}\in\mathcal{S}, i_k=s}\langle\mathbf{z}_{\mathbf{i}}^{\backslash k}{\mathbf{z}_{\mathbf{i}}^{\backslash k}}^T\rangle + v\mathbf{I}\right)^{-1}.\end{aligned}$$

The expression of $q^*(\tau)$ can be found as

$$\begin{aligned}\ln q^*(\tau) = \mathbb{E}_q\{&\frac{N}{2}\ln\tau - \sum_{\mathbf{i}\in\mathcal{S}}\frac{\tau}{2}[y_{\mathbf{i}} - \mathbf{1}^T(\mathbf{u}_{i_1}^1 \circ \cdots \circ \mathbf{u}_{i_K}^K)]^2 \\ &(a_0 - 1)\ln\tau - b_0\tau\},\end{aligned}$$

which is a Gamma distribution $q^*(\tau) = \text{Gam}(\tau|a^*, b^*)$ with a^* and b^* given by

$$\begin{aligned}a^* &= a_0 + \frac{N}{2}, \\ b^* &= b_0 + \frac{1}{2}\sum_{\mathbf{i}\in\mathcal{S}}[y_{\mathbf{i}}^2 - 2y_{\mathbf{i}}\langle\mathbf{1}^T\mathbf{z}_{\mathbf{i}}\rangle + \langle(\mathbf{1}^T\mathbf{z}_{\mathbf{i}})^2\rangle].\end{aligned}$$