Copula Variational Bayes inference via information geometry

Viet Hung Tran

Abstract—Variational Bayes (VB), also known as independent mean-field approximation, has become a popular method for Bayesian network inference in recent years. Its application is vast, e.g. in neural network, compressed sensing, clustering, etc. to name just a few. In this paper, the independence constraint in VB will be relaxed to a conditional constraint class, called copula in statistics. Since a joint probability distribution always belongs to a copula class, the novel copula VB (CVB) approximation is a generalized form of VB. Via information geometry, we will see that CVB algorithm iteratively projects the original joint distribution to a copula constraint space until it reaches a local minimum Kullback-Leibler (KL) divergence. By this way, all mean-field approximations, e.g. iterative VB, Expectation-Maximization (EM), Iterated Conditional Mode (ICM) and k-means algorithms, are special cases of CVB approximation.

For a generic Bayesian network, an augmented hierarchy form of CVB will also be designed. While mean-field algorithms can only return a locally optimal approximation for a correlated network, the augmented CVB network, which is an optimally weighted average of a mixture of simpler network structures, can potentially achieve the globally optimal approximation for the first time. Via simulations of Gaussian mixture clustering, the classification's accuracy of CVB will be shown to be far superior to that of state-of-the-art VB, EM and k-means algorithms.

Index Terms—Copula, Variational Bayes, Bregman divergence, mutual information, k-means, Bayesian network.

I. INTRODUCTION

Originally, the idea of mean-field theory is to approximate an interacting system by a non-interacting system, such that the mean values of system's nodes are kept unchanged [1]. Variational Bayes (VB) is a redefined method of mean-field theory, in which the joint probability distribution f_{θ} of a system is approximated by a free-form independent distribution $\widetilde{f}_{\theta} = \prod_{k=1}^{K} \widetilde{f}_{\theta_k}$, such that the Kullback-Leibler (KL) divergence $\mathrm{KL}_{\widetilde{f}_{\theta}||f_{\theta}}$ is minimized [2], $\theta \triangleq \{\theta_1, \theta_2, \ldots, \theta_K\}$. The term "variational" in VB originates from "calculus of variations" in differential mathematics, which is used to find the derivative of KL divergence over distribution space [3], [4].

The VB approximation is particularly useful for estimating unknown parameters in a complicated system. If the true value of parameters θ is unknown, we assume they follow a probabilistic model a-priori. We then apply Bayesian inference, also called inverse probability in the past [5], [6], to minimizing the expected loss function between true value θ and posterior estimate $\widehat{\theta}(x)$. In practice, the computational

V. H. Tran was with Telecom Paris institute, 75013 Paris, France. He is now with Univeristy of Surrey, GU27XH Surrey, U.K. (e-mails: v.tran@surrey.ac.uk, tranviethung@hcmut.edu.vn).

complexity of posterior estimate $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ often grows exponentially with arriving data \boldsymbol{x} and, hence, yields the curse of dimensionality [7]. For tractable computation, as shown in this paper, the VB algorithm iteratively projects the originally complex distribution into simpler independent class of each unknown parameter θ_k , one by one, until the KL divergence converges to a local minimum. For this reason, the VB algorithm has been used extensively in many fields requiring tractable parameter's inference, e.g. in neural networks [8], compressed sensing [9], data clustering [10], etc. to name just a few.

Nonetheless, the independent class is too strict in practice, particularly in case of highly correlated model [2]. In order to capture the dependence in a probabilistic model, a popular method in statistics is to consider a copula class. The key idea is to separate the dependence structure, namely copula, of a joint distribution from its marginal distributions. In this way, the copula concept is similar to nonnegative compatibility functions over cliques in factor graphs [11], [12], although the compatibility functions are not probability distributions like copula. Indeed, a copula c_{θ} is a joint distribution whose marginals are uniform, as originally proposed in [13]. For example, the copula of a bivariate discrete distribution is a bi-stochastic matrix, whose sum of any row or any column is equal to one [14], [15]. More generally, by Sklar's theorem [13], [16], any joint distribution f_{θ} can always be written in copula form $f_{\theta} = c_{\theta} \prod_{k=1}^{K} f_{\theta_k}$, in which c_{θ} fully describes the inter-dependence of variables in a joint distribution. For independent class, the copula density c_{θ} is merely a constant and equal to one everywhere [14].

In this paper, the novel copula VB (CVB) approximation $\widetilde{f}_{\theta} = \widetilde{c}_{\theta} \prod_{k=1}^{K} \widetilde{f}_{\theta_{k}}$ will extend the independent constraint in VB to a copula class of dependent distributions. After fixing the distributional form of \widetilde{c}_{θ} , the CVB iteratively updates the free-form marginals $\widetilde{f}_{\theta_{k}}$ one by one, similarly to traditional VB, until KL divergence $\mathrm{KL}_{\widetilde{f}_{\theta}||f_{\theta}}$ converges to a local minimum. The CVB approximation will become exact if the form of \widetilde{c}_{θ} is the same as that of original copula c_{θ} . The study of copula form c_{θ} is still an active field in probability theory and statistics [14], owing to its flexibility to modeling the dependence of any joint distribution f_{θ} . Also, because the mutual information f_{θ} is equal to entropy of its copula c_{θ} [17], the copula is currently an interesting topic for information criterions [18], [19].

In information geometry, the KL divergence is a special case of the Bregman divergence, which, in turn, is a generalized concept of distance in Euclidean space [20]. By reinterpreting the KL minimization in VB as the Bregman projection, we

will see that CVB, and its special case VB, iteratively projects the original distribution to a fixed copula constraint space until convergence. Then, similar to the fact that the mean is the point of minimum total distance to data, an augmented CVB approximation will also be designed as a distribution of minimum total Bregman divergence to the original distribution in this paper.

Three popular special cases of VB will also be revisited in this paper, namely Expectation-Maximization (EM) [21], [22], Iterated Conditional Mode (ICM) [23], [24] and k-means algorithms [25], [26]. In literature, the well-known EM algorithm was shown to be a special case of VB [1], [2], in which one of VB's marginal is restricted to a point estimate via Dirac delta function. In this paper, the EM algorithm will be shown that it does not only reach a local minimum KL divergence, but it may also return a local maximum-a-posteriori (MAP) point estimate of the true marginal distribution. This justifies the superiority of EM algorithm to VB in some cases of MAP estimation, since the peaks in VB marginals might not be the same as those of true marginals.

If all VB marginals are restricted to Dirac delta space, the iterative VB algorithm will become ICM algorithm, which returns a locally joint MAP estimate of the original distribution. Also, for standard Normal mixture clustering, the ICM algorithm is equivalent to the well-known k-means algorithm, as shown in this paper. The k-means algorithm is also equivalent to the Lloyd-Max algorithm [25], which has been widely used in quantization context [27].

For illustration, the CVB and its special cases mentioned above will be applied to two canonical models in this paper, namely bivariate Gaussian distribution and Gaussian mixture clustering. By tuning the correlation in these two models, the performance of CVB will be shown to be superior to that of state-of-the-art mean-field methods like VB, EM and k-means algorithm. An augmented CVB form for a generic Bayesian network will also be studied and applied to this Gaussian mixture model.

A. Related works

Although some generalized forms of VB have been proposed in literature, most of them are merely variants of meanfield approximations and, hence, still confined within independent class. For example, in [28], [29], the so-called Conditionally Variational algorithm is an application of traditional VB to a joint conditional distribution $f(\theta|\xi) = \prod_{k=1}^{K} f_k(\theta_k|\xi)$, given a latent variable ξ . Hence, different to CVB above, the approximated marginal f_{ξ} was not updated in their scheme. In [30], the so-called generalized mean-field algorithm is merely to apply the traditional VB method to the independent class of a set of variables, i.e. each θ_k consists of a set of variables. In [31], the so-called Structured Variational inference is the same as the generalized mean-field, except that the dependent structure inside the set θ_k is also specified. In summary, they are different ways of applying traditional VB, without changing the VB's updating formula. In contrast, the CVB in this paper involves new tractable formulae and broader copula constraint class.

The closest form to the CVB of this paper is the so-called Copula Variational inference in [32], which fixes the form of approximated distribution $\tilde{f}_{\theta|\xi} = \tilde{c}_{\theta|\xi} \prod_{i=1}^N \tilde{f}_{\theta_i|\xi}$ and applies gradient decent method upon the latent variable ξ in order to find a local minimum of KL divergence. In contrast, the CVB in this paper is a free-form approximation, i.e. it does not impose any particular form initially, and provides higher-order moment's estimates than a mere point estimate. Hence, the fixed-form constraint class in their Copula Variational inference is much more restricted than the free-form copula constraint class of CVB in this paper. Also, the iterative computation for CVB will be given in closed form with low complexity, rather than relying point estimates of gradient decent methods.

2

B. Contributions and organization

The contributions of this paper are summarized as follows:

- A novel copula VB (CVB) algorithm, which extends the
 independent constraint class of traditional VB to a copula
 constraint class, will be given. The convergence of CVB
 will be proved via three methods: Lagrange multiplier
 method in calculus of variation, Jensen's inequality and
 the Bregman projection in information geometry. The two
 former methods have been used in literature for proof of
 convergence of traditional VB, while the third method is
 new and provides a unified scheme for the former two
 methods.
- The EM, ICM and k-means algorithms will be shown to be special cases of the traditional VB, i.e. they all locally minimize the KL divergence under a fixed-form independent constraint class.
- An augmented form of CVB, namely hierarchical CVB approximation, with linear computational complexity for a generic Bayesian network will also be provided.
- In simulations, the CVB algorithm for Gaussian mixture clustering will be illustrated. The classification's performance of CVB will be shown to be far superior to that of VB, EM and k-means algorithms for this model.

The paper is organized as follows: since the Bregman projection in information geometry is insightful and plays central role to VB method, it will be presented first in section II. The definition and property of copula will then be introduced in section III. The novel copula VB (CVB) method and its special cases will be presented in section IV. The computational flow of CVB for a Bayesian network is studied in section V and will be applied to simulations in section VI. The paper is then concluded in section VII.

Note that, for notational simplicity, the notion of probability density function (p.d.f.) for continuous random variable (r.v.) in this paper will be implicitly understood as the probability mass function (p.m.f) in the case of discrete r.v., when the context is clear.

II. INFORMATION GEOMETRY

In this section, we will revisit a geometric interpretation of one of fundamental measures in information theory, namely Kullback-Leibler (KL) divergence, which is also the central

Figure 1. Illustration of Bregman divergence \mathcal{D} for convex function ϕ . The hyperplane $\mathcal{H}_{\beta}(\alpha) \triangleq \phi(\beta) + \langle \alpha - \beta, \nabla \phi(\beta) \rangle$ is tangent to ϕ at point β . Note that, if $\phi(\alpha)$ is equal to the continuous entropy function $H_{\alpha}(\alpha)$, the hyperplane $\mathcal{H}_{\beta}(\alpha)$ is equal to the cross entropy $H_{\beta}(\alpha)$ from α to β and $\mathcal{D}(\alpha||\beta) = H_{\alpha}(\alpha) - H_{\beta}(\alpha)$ is equal to Kullback-Leibler (KL) divergence (c.f. section II-A2).

part of VB (i.e. mean-field) approximation. For this purpose, the Bregman divergence, which is a generalization of both Euclidean distance and KL divergence, will be defined first. Two important theorems, namely Bregman pythagorean theorem and Bregman variance theorem, will then be presented. These two theorems generalize the concept of Euclidean projection and variance theorem to the probabilistic functional space, respectively. The Bregman divergence is also a key concept in the field of information geometry in literature [20], [33].

A. Bregman divergence for vector space

For simplicity, in this subsection, we will define Bregman divergence for real vector space first, which helps us visualize the Bregman pythagorean theorem later.

Definition 1. (Bregman divergence)

Let $\phi: \mathbb{R}^K \to \mathbb{R}$ be a strictly convex and differentiable function. Given two points $\alpha, \beta \in \mathbb{R}^K$, with $\alpha \triangleq [\alpha_1, \alpha_2, \dots, \alpha_K]^T$ and $\beta \triangleq [\beta_1, \beta_2, \dots, \beta_K]^T$, the Bregman divergence $\mathcal{D}: \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^+$, with $\mathbb{R}^+ \triangleq [0, +\infty)$, is defined as follows:

$$\mathcal{D}(\boldsymbol{\alpha}||\boldsymbol{\beta}) \triangleq \mathcal{H}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) - \mathcal{H}_{\boldsymbol{\beta}}(\boldsymbol{\alpha})$$

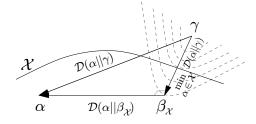
$$= \phi(\boldsymbol{\alpha}) - \phi(\boldsymbol{\beta}) - \langle \boldsymbol{\alpha} - \boldsymbol{\beta}, \nabla \phi(\boldsymbol{\beta}) \rangle,$$
(1)

where ∇ is gradient operator, $\langle \cdot, \cdot \rangle$ denotes inner product and $\mathcal{H}_{\beta}(\alpha) \triangleq \phi(\beta) + \langle \alpha - \beta, \nabla \phi(\beta) \rangle$ is hyperplane tangent to ϕ at point β , as illustrated in Fig. 1.

For simplicity, the notations \mathcal{D} and \mathcal{D}_{ϕ} are used interchangeably in this paper when the context is clear. Some well-known properties of Bregman divergence (1) are summarized below:

Proposition 2. (Bregman divergence's properties)

- 1) Non-negativity: $\mathcal{D}(\boldsymbol{\alpha}||\boldsymbol{\beta}) \geq 0$.
- 2) Equality: $\mathcal{D}(\boldsymbol{\alpha}||\boldsymbol{\beta}) = 0 \Leftrightarrow \boldsymbol{\alpha} = \boldsymbol{\beta}.$
- 3) Asymmetry: $\mathcal{D}(\alpha||\beta) \neq \mathcal{D}(\beta||\alpha)$ in general.
- 4) Convexity: $\mathcal{D}(\alpha||\beta)$ is convex over α , but not over β in general.
- 5) Gradient: $\nabla_{\alpha} \mathcal{D}(\alpha||\beta) = \nabla \phi(\alpha) \nabla \phi(\beta)$ and $\nabla_{\beta} \mathcal{D}(\alpha||\beta) = \nabla^2 \phi(\beta)[\beta \alpha]$.



3

Figure 2. Illustration of Bregman pythagorean inequality over closed convex set $\alpha \in \mathcal{X}$. The point $\beta_{\mathcal{X}} \in \mathcal{X}$ is called the Bregman projection of $\gamma \in \mathbb{R}^K$ onto $\mathcal{X} \subset \mathbb{R}^K$. The dashed contours represent the convexity of $\mathcal{D}(\boldsymbol{\beta}||\gamma)$ over arbitrary point $\boldsymbol{\beta} \in \mathbb{R}^K$ in general.

- 6) Affine equivalence class: $\mathcal{D}_{\phi}(\boldsymbol{\alpha}||\boldsymbol{\beta}) = \mathcal{D}_{\widetilde{\phi}}(\boldsymbol{\alpha}||\boldsymbol{\beta})$ if $\widetilde{\phi}(\boldsymbol{x}) = \phi(\boldsymbol{x}) + \langle \boldsymbol{\gamma}, \boldsymbol{x} \rangle + c$, e.g. $\widetilde{\phi}(\boldsymbol{x}) = \mathcal{D}_{\phi}(\boldsymbol{x}||\boldsymbol{\beta})$.
- 7) Three-point property:

$$\mathcal{D}(\boldsymbol{\alpha}||\boldsymbol{\beta}) + \mathcal{D}(\boldsymbol{\beta}||\boldsymbol{\gamma}) - \mathcal{D}(\boldsymbol{\alpha}||\boldsymbol{\gamma}) = \left\langle \boldsymbol{\beta} - \boldsymbol{\alpha}, \underbrace{\nabla \phi(\boldsymbol{\beta}) - \nabla \phi(\boldsymbol{\gamma})}_{\nabla_{\boldsymbol{\beta}} \mathcal{D}(\boldsymbol{\beta}||\boldsymbol{\gamma})} \right\rangle$$

The points $\{\alpha, \beta, \gamma\}$ in (2) are called *Bregman orthogonal* at point β if $\langle \beta - \alpha, \nabla \phi(\beta) - \nabla \phi(\gamma) \rangle = 0$.

Proof: All properties 1-7 are direct consequence of Bregman definition (1). The derivation of well-known properties 1-4 and 6-7 can be found in [20], [34] and [35], [36], respectively, for any $x, \gamma \in \mathbb{R}^K$, $c \in \mathbb{R}$. In property 6, since $\mathcal{D}_{\phi}(x||\beta)$ is both convex and affine over x, as defined in (1), we can assign $\widetilde{\phi}(x) = \mathcal{D}_{\phi}(x||\beta)$. In property 7, the ∇_{β} form is a consequence of gradient property. The gradient property, i.e. the property 5, can be derived from definition (1) as follows: $\nabla_{\alpha}\mathcal{D}_{\phi}(\alpha||\beta) = \nabla_{\alpha}\phi(\alpha) - \nabla_{\alpha}\langle\alpha - \beta, \nabla\phi(\beta)\rangle = \nabla_{\alpha}\phi(\alpha) - \nabla\phi(\beta)$. Similarly, from (1), we have $\nabla_{\beta}\mathcal{D}_{\phi}(\alpha||\beta) = -\nabla_{\beta}\phi(\beta) - \nabla_{\beta}\langle\alpha - \beta, \nabla\phi(\beta)\rangle = -\nabla_{\beta}\phi(\beta) - (-\nabla\phi(\beta) + \nabla^2\phi(\beta)[\alpha - \beta]) = \nabla^2\phi(\beta)[\beta - \alpha]$, in which ∇^2 denotes Hessian matrix operator.

Remark 3. The gradient property gives us some insight on Bregman divergence. For example, from gradient property, we can see that $\alpha = \beta$ is the stationary and minimum point of $\mathcal{D}(\alpha||\beta)$. Also, $\mathcal{D}(\alpha||\beta)$ is convex over α but not over β since $\phi(\cdot)$ is a convex function, as shown intuitively in the form of $\nabla_{\alpha}\mathcal{D}(\alpha||\beta)$ and $\nabla_{\beta}\mathcal{D}(\alpha||\beta)$, respectively. The gradient form $\nabla_{\beta}\mathcal{D}(\beta||\gamma)$ in (2) represents the changing value of $\mathcal{D}(\beta||\gamma)$ over β and, hence, explains the three-point property intuitively, as illustrated in Fig. 2.

Let us now consider the most important property of Bregman divergence in this paper, namely Bregman pythagorean inequality, which defines the Bregman projection over a closed convex subset $\mathcal{X} \subset \mathbb{R}^K$.

Theorem 4. (Bregman pythagorean inequality)

Let \mathcal{X} be a closed convex subset in \mathbb{R}^K . For any points $\alpha \in \mathcal{X}$ and $\gamma \in \mathbb{R}^K$, we have:

$$\mathcal{D}(\alpha||\beta_{\chi}) + \mathcal{D}(\beta_{\chi}||\gamma) \le \mathcal{D}(\alpha||\gamma), \tag{3}$$

where the unique point β_{χ} is called the Bayesian projection

Figure 3. Illustration of equivalence between Jensen's inequality (left) and Bregman variance theorem (right). Similar to Fig. 1 and Fig. 2, the dashed contours on the right represent the convexity of $\mathcal{D}_{\phi}(\boldsymbol{x}||\tilde{\boldsymbol{x}})$ over \boldsymbol{x} , which, in turn, can be regarded as another convex function $\tilde{\phi}$ for Jensen's inequality on the left.

of γ onto \mathcal{X} and defined as follows:

$$\beta_{\mathcal{X}} \triangleq \arg\min_{\alpha \in \mathcal{X}} \mathcal{D}(\alpha||\gamma). \tag{4}$$

From three-point property (2), we can see that the Bregman pythagorean inequality in (3) becomes equality for all $\alpha \in \mathcal{X}$ if and only if \mathcal{X} is an affine set (i.e. the triple points $\{\alpha, \beta_{\mathcal{X}}, \gamma\}$ are Bregman orthogonal at $\beta_{\mathcal{X}}, \forall \alpha \in \mathcal{X}$).

Proof: Note that $\beta_{\mathcal{X}}$, as defined in (4), is not necessarily unique if \mathcal{X} is not convex [20]. The uniqueness of $\beta_{\mathcal{X}}$ (4) for convex set \mathcal{X} can be proved either via contradiction [34] or via convexity of \mathcal{X} in three-point property (2), c.f. [20], [37]. Substituting $\beta_{\mathcal{X}}$ in (4) to three-point property (2) yields the Bregman pythagorean inequality (3).

Owing to Bregman divergence, we also have a geometrical interpretation of probabilistic variance, as shown in the following theorem on Jensen's inequality:

Theorem 5. (Bregman variance theorem - Jensen's inequality)

Let $x \in \mathbb{R}^K$ be a r.v. with mean $\mathbb{E}[x]$ and variance Var[x]. The Bregman variance $Var_{\phi}[x]$ is defined as follows:

$$Var_{\phi}[\mathbf{x}] \triangleq \mathbb{E}[\mathcal{D}_{\phi}(\mathbf{x}||\mathbb{E}[\mathbf{x}])] = \mathbb{E}[\phi(\mathbf{x})] - \phi(\mathbb{E}[\mathbf{x}]) \ge 0.$$
 (5)

Equivalently, we have:

$$Var_{\phi}[\boldsymbol{x}] \triangleq \mathbb{E}[\mathcal{D}_{\phi}(\boldsymbol{x}||\mathbb{E}[\boldsymbol{x}])] = \mathbb{E}[D_{\phi}(\boldsymbol{x}||\widetilde{\boldsymbol{x}})] - \mathcal{D}_{\phi}(\mathbb{E}[\boldsymbol{x}]||\widetilde{\boldsymbol{x}}) \ge 0$$
(6)

for any fixed point $\tilde{x} \in \mathbb{R}^K$. The right hand side (r.h.s.) of (5) is called Jensen's inequality in literature, i.e. $\mathbb{E}(\phi(x)) \geq \phi(\mathbb{E}(x))$, for any convex function ϕ [38]. Also, from (6), we have:

$$x_0 \triangleq \mathbb{E}[x] = \operatorname*{arg\,min}_{\widetilde{x}} \mathbb{E}[D(x||\widetilde{x})],$$
 (7)

as illustrated in Fig. 3.

Proof: Let us show the proof in reverse way. Firstly, the mean property (7) is a consequence of (6), i.e. we have: $\mathbb{E}[D(\boldsymbol{x}||\widetilde{\boldsymbol{x}})] = \mathbb{E}[\mathcal{D}(\boldsymbol{x}||\mathbb{E}[\boldsymbol{x}])] + \mathcal{D}(\mathbb{E}[\boldsymbol{x}]||\widetilde{\boldsymbol{x}})$ and $\mathcal{D}(\mathbb{E}[\boldsymbol{x}]||\widetilde{\boldsymbol{x}}) = 0 \Leftrightarrow \widetilde{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}]$. Secondly, by replacing $\phi(\boldsymbol{x})$ in (5) with $\widetilde{\phi}(\boldsymbol{x}) = D_{\phi}(\boldsymbol{x}||\widetilde{\boldsymbol{x}})$, the form (6) is equivalent to (5), owing to the affine equivalence property in Proposition 2. Lastly, the form (5) is a direct derivation from Bregman definition (1), with $\alpha = \boldsymbol{x}$ and $\beta = \mathbb{E}[\boldsymbol{x}]$, as follows: $\mathcal{D}(\boldsymbol{x}||\mathbb{E}[\boldsymbol{x}]) = \phi(\boldsymbol{x}) - \phi(\mathbb{E}[\boldsymbol{x}]) - \langle \boldsymbol{x} - \mathbb{E}[\boldsymbol{x}], \nabla \phi(\mathbb{E}[\boldsymbol{x}]) \rangle$ and, hence, $\mathbb{E}[\mathcal{D}(\boldsymbol{x}||\mathbb{E}[\boldsymbol{x}])] =$

$$\mathbb{E}[\phi(\boldsymbol{x})] - \mathbb{E}[\phi(\mathbb{E}[\boldsymbol{x}])] - \left\langle \underbrace{\mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{x}]}_{=0}, \nabla \phi(\mathbb{E}[\boldsymbol{x}]) \right\rangle.$$

Remark 6. Although we have $Var[x] \neq Var_{\phi}[x]$ in general, the mean $\mathbb{E}[x]$ is the same minimum point for any expected Bregman divergence, as shown in (7). This notable property of the mean has been exploited extensively for Bregman k-means algorithms in literature [34], [35].

A list of Bregman divergences, corresponding to different functional forms of $\phi(x)$, can be found feasibly in literature, e.g. in [20], [39]. Let us recall two most popular forms below.

1) Euclidean distance: A special case of Bregman divergence is squared Euclidean distance [35]:

$$\mathcal{D}_{\phi_E}(\boldsymbol{\alpha}||\boldsymbol{\beta}) = ||\boldsymbol{\alpha} - \boldsymbol{\beta}||^2, \text{ with } \phi_E(\boldsymbol{x}) \triangleq ||\boldsymbol{x}||^2,$$
 (8)

where $||\cdot||$ denotes \mathcal{L}_2 -norm for elements of a vector or matrix. In this case, the Bregman pythagorean theorem (3) becomes the traditional Pythagorean theorem and the Bregman variance (5) becomes the traditional variance theorem, i.e. $\operatorname{Var}_{\phi_E}[x] = \operatorname{Var}[x] = \mathbb{E}[||x||^2] - ||\mathbb{E}[x]||^2$.

2) Kullback-Leibler (KL) divergence: Another popular case of Bregman divergence is the KL divergence [35]:

$$\mathrm{KL}(\boldsymbol{\alpha}||\boldsymbol{\beta}) \triangleq \mathcal{D}_{\mathrm{KL}}(\boldsymbol{\alpha}||\boldsymbol{\beta}) = \sum_{k=1}^{K} \alpha_k \log \frac{\alpha_k}{\beta_k} - \sum_{k=1}^{K} \alpha_k + \sum_{k=1}^{K} \beta_k,$$

with $\phi_{KL}(\boldsymbol{x}) \triangleq \sum_{k=1}^{K} x_k \log x_k$, $\forall x_k \in \mathbb{R}^+$. More generally, it can be shown that [39]:

$$\mathrm{KL}_{\widetilde{f}||f} \triangleq \mathcal{D}_{\mathrm{KL}}(\widetilde{f}||f) = \mathbb{E}_{\widetilde{f}(\theta)} \log \frac{\widetilde{f}(\theta)}{f(\theta)},$$
 (9)

where $\phi_{\text{KL}}(f(\theta)) \triangleq H(\theta) = \mathbb{E}_{f(\theta)} \log f(\theta)$ is the continuous entropy and $\mathcal{D}_{\text{KL}}(\widetilde{f}||f)$ is the Bregman divergence between two density distributions $\widetilde{f}(\theta)$ and $f(\theta)$, as presented below.

B. Bregman divergence for functional space

In the calculus of variations, the Bregman divergence for vector space is a special case of the Bregman divergence for functional space, defined as follows:

Definition 7. (Bregman divergence for functional space) [33] Let $\phi: \mathcal{L}_p(\theta) \to \mathbb{R}$ be a strictly convex and twice Fréchet-differentiable functional over \mathcal{L}_p -normed space. The Bregman divergence $\mathcal{D}: \mathcal{L}_p(\theta) \times \mathcal{L}_p(\theta) \to \mathbb{R}^+$ between two functions $f, g \in \mathcal{L}_p(\theta)$ is defined as follows:

$$\mathcal{D}(f||g) \triangleq \phi(f) - \phi(g) - \delta\phi(f - g; g), \tag{10}$$

where $\delta\phi(\cdot;q)$ is Fréchet derivative of ϕ at q.

Apart from gradient form, all well-known properties of Bregman divergence in Proposition 2 are also valid for functional space [33], [40]. Hence, we can feasibly derive the Bregman variance theorem for probabilistic functional space, as follows:

Proposition 8. (Bregman variance theorem for functions) Let functional point $f(\theta)$ be a r.v. drawn from the functional

space $\mathcal{L}_p(\theta)$ with functional mean $\mathbb{E}[f] \triangleq \mathbb{E}[f(\theta)]$ and functional variance $Var[f] \triangleq \mathbb{E}[||f(\theta) - \mathbb{E}[f]||^2]$. Then we have:

$$Var_{\phi}[f] \triangleq \mathbb{E} \left[\mathcal{D}(f || \mathbb{E}(f)) \right] = \mathbb{E} [\phi(f)] - \phi(\mathbb{E}[f]) > 0.$$

Equivalently, we have:

$$\mathit{Var}_{\phi}[f] \triangleq \mathbb{E}\left[\mathcal{D}(f||\mathbb{E}(f))\right] = \mathbb{E}[D(f||\widetilde{f})] - \mathcal{D}(\mathbb{E}[f]||\widetilde{f}) \geq 0,$$

for any functional point $\widetilde{f} \triangleq \widetilde{f}(\theta) \in \mathcal{L}_p(\theta)$ and:

$$f_0 \triangleq \mathbb{E}[f] = \underset{\widetilde{f}}{\operatorname{arg\,min}} \, \mathbb{E}[D(f||\widetilde{f})].$$
 (11)

Proof: Because the Fréchet derivative in (10) is a linear operator like gradient in (1), we can derive the above results in the same manner of the proof of Theorem 5.

Remark 9. From Proposition 8, we have $\mathrm{Var}[f] = \mathrm{Var}_{\phi_E}[f]$ for Euclidean case $\phi_E(f) = ||f(\theta) - \mathbb{E}[f]||^2$, but $\mathrm{Var}[f] \neq \mathrm{Var}_{\phi}[f]$ in general. The functional mean $f_0 \triangleq \mathbb{E}[f]$ is also the same minimum function for any expected Bregman divergence, similarly to Remark 6.

For later use, let us apply Proposition 8 and show here the Bregman variance for a probabilistic mixture:

Corollary 10. (Bregman variance theorem for mixture) Let functional point $f(\theta)$ be a r.v. drawn from a functional set $\mathbf{f} \triangleq \{f_{i_1}(\theta), \dots, f_{i_N}(\theta)\}$ of N distributions over θ , with probabilities $p_i \in \mathbb{I} \triangleq [0,1]$, $\sum_{i=1}^N p_i = 1$. The functional mean (11) is then regarded as a mixture, as follows:

$$f_0(\theta) \triangleq \mathbb{E}[f] = \sum_{i=1}^K p_i f_i(\theta),$$
 (12)

with variance $Var[f] = \sum_{i=1}^K p_i ||f(\theta) - \bar{f}(\theta)||^2$. The Bregman variance is then:

$$Var_{\phi}[f] = \sum_{i=1}^{K} p_i D(f_i||f_0) = \sum_{i=1}^{K} p_i D(f_i||\widetilde{f}) - D(f_0||\widetilde{f}) \ge 0,$$
(13)

for any distribution $\widetilde{f} \triangleq \widetilde{f}(\theta)$ and, hence, from (11-12), we have:

$$f_0(\theta) = \mathbb{E}[f] = \sum_{i=1}^{K} p_i f_i(\theta) = \arg\min_{\tilde{f}} \sum_{i=1}^{K} p_i D(f_i || \tilde{f}).$$
 (14)

Proof: This case is a consequence of Proposition 8. The case of KL divergence, which is a special case of Bregman variance with $\phi = \phi_{KL}$ in (13), is illustrated in Fig. 4.

Remark 11. The computation of KL variance via (13) for a mixture is often more feasible than the computation of Euclidean variance in practice. Indeed, the KL form corresponds to geometric mean [39], which can yield linearly computational complexity over exponential coordinates (particularly for exponential family [20], [39]), while the Euclidean form corresponds to arithmetic mean, which would yield exponentially computational complexity for exponential family distributions over Euclidean coordinates in general, as shown in section IV-B3.

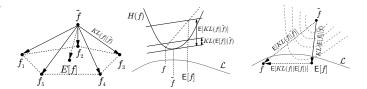


Figure 4. Application of Bregman variance theorem (13) to KL divergence in distribution space $f \in \mathcal{L}$, with the same convention in Fig. 3. As an example, the mixture $f_0(\theta) \triangleq \mathbb{E}[f] = \sum_{i=1}^5 p_i f_i(\theta)$ in (12) must lie inside the polytope $\mathcal{L} = \{f_1(\theta), \dots, f_5(\theta)\}$. In middle sub-figure, H(f) denotes the continuous entropy over p.d.f. f. The mixture $\widetilde{f} = f_0 = \mathbb{E}[f]$ is then the minimum functional point of $\mathbb{E}[\mathrm{KL}(f||\widetilde{f})]$, which is also an upper bound of $\mathrm{KL}(\mathbb{E}[f]||\widetilde{f})$ over $\widetilde{f} \in \mathcal{L}$, as shown in (13-14).

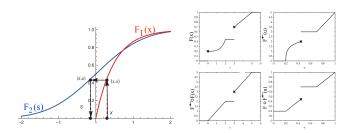


Figure 5. Illustration of variable transformation from x to s in the case of continuous c.d.f. [43] (left), together with pseudo-inverse $F^{\leftarrow}(u)$ of a non-continuous c.d.f. F(x) and their concatenations $F^{\leftarrow} \circ F(x)$, $F \circ F^{\leftarrow}(u)$ [42] (right). We can see that the uniqueness of copula requires the continuous property of c.d.f., since non-continuous c.d.f. does not preserve the inverse transformation.

III. COPULA THEORY

The copula concept was firstly defined in [13], although it was also defined under different names such as "uniform representation" or "dependence function" [14]. The copula has been studied intensively in many decades in statistics, particularly for finances [41], [42]. Yet the application of copula in information theory is still limited at the moment. In this section, we will review the basic concept of copula and its direct connection to mutual information of a system. The KL divergence for copula, which is the nutshell of CVB approximation in next section, will be provided at the end of this section.

A. Sklar's theorem

Because the Sklar's paper [13] is the beginning of copula's history, let us recall the Sklar's theorem first.

Definition 12. (Pseudo-inverse function)

Let $F: \mathbb{R} \to \mathbb{I}$ be a cumulative distributional function (c.d.f.) of a r.v. $\theta \in \mathbb{R}$. Since $F(\theta)$ is not strictly increasing in general, as illustrated in Fig. 5, a pseudo-inverse function (also called quantile function) $F^{\leftarrow}: \mathbb{I} \to \mathbb{R}$ is defined as follows:

$$F^{\leftarrow}(u) \triangleq \inf\{\theta \in \mathbb{R} : F(\theta) > u\}, \ u \in \mathbb{I}.$$

Note that, the quasi-inverse F^{\leftarrow} coincides with the inverse function F^{-1} if $F(\theta)$ is continuous and strictly increasing, as illustrated in Fig. 5.

Theorem 13. (Sklar's theorem) [13], [16] For any r.v. $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_K]^T \in \mathbb{R}^K$ with joint c.d.f. $F(\boldsymbol{\theta})$

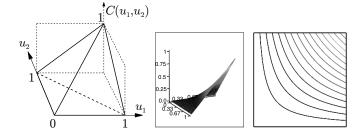


Figure 6. All bivariate copulas must lie inside the pyramid of Fréchet-Hoeffding bound. Both marginal c.d.f. $C(u_1,1)$ and $C(1,u_2)$ must be uniform over [0,1] by definition and, hence, plotted in the left sub-figure as straight lines. The two sub-figures on the right illustrates the contours of independent copula $C(u_1,u_2)=u_1u_2$ [41].

and marginal c.d.f. $F_k(\theta_k)$, $\forall k \in \{1, 2, ..., K\}$, there always exists an equivalent joint c.d.f., namely copula C, whose all marginal c.d.f. $C_k(F_k(\theta_k))$ are uniform over \mathbb{I} as follows:

$$F(\boldsymbol{\theta}) = C(F_1(\theta_1), \dots, F_K(\theta_K)) \tag{15}$$

In general, the copula form C of a joint c.d.f. F is not unique, but its value on the range $u \in Range\{F_1\} \times ... \times Range\{F_K\} \subseteq \mathbb{I}^K$ is always unique, as follows:

$$C(\boldsymbol{u}) = F(F_1^{\leftarrow}(u_1), \dots, F_K^{\leftarrow}(u_K)) \tag{16}$$

with $\mathbf{u} \triangleq [u_1, u_2, \dots, u_K]^T$ and $u_k \triangleq F_k(\theta_k) : \mathbb{R} \to \mathbb{I}$, $\forall k$. If all marginals $F_1, \dots F_K$ are continuous, the copula C in (15) is uniquely determined by quantile transformation (16), in which F^{\leftarrow} coincides with the inverse function F^{-1} .

1) Bound of copula: For rough visualization of copula, let us recall the Fréchet-Hoeffding bound of copula [14], [42]:

$$\max\{0, 1 - K + \sum_{k=1}^{K} u_k\} \le C(u_1, \dots, u_K) \le \min\{u_1, \dots, u_K\}$$

where $u_k \triangleq F_k(\theta_k) : \mathbb{R} \to \mathbb{I}$. This bound is illustrated in Fig. 6 for the case of two dimensions.

2) Discrete copula: The pseudo-inverse form (16) is often called sub-copula in literature [14], since its values are only defined on a possibly subset of \mathbb{I}^K . This mostly happens in the case of discrete distributions, where marginal $F(\theta)$ is not continuous, as illustrated in Fig. 5. Hence, there are possibly more than one continuous copula (15) satisfying the discrete sub-copula form (15) at specific values $u \in \text{Range}\{F_1\} \times \ldots \times \text{Range}\{F_K\}$.

As illustrated in Fig. 5, the Sklar's theorem only guarantees the uniqueness of copula form C for a strictly increasing continuous F (c.f. [14] for examples of non-unique copulas C associated with a discrete F). Nonetheless, the quantile function in (16) is still useful to compute copula values in the discrete range of F. For example, in [14], [15], the copula form of any discrete bivariate distribution was shown to be equivalent to a bi-stochastic non-negative matrix, whose sum of any row or column is equal to one.

3) Continuous copula: For simplicity, let us focus on copula form of continuous c.d.f. F, although the results in this paper can be extended to discrete case via pseudo-

inverse function in (16). For continuous case, the quantile transformation (16) yields the density form of copula C, as follows:

Corollary 14. (Copula density function) [14]

If all marginals $F_1, ..., F_K$ are absolutely continuous with respect to Lebesgue measure on \mathbb{R}^K , the density $c(\boldsymbol{u}) \triangleq \frac{\partial C(\boldsymbol{u})}{\partial u_1...\partial u_K}$ of copula C in (15) is given by:

$$f(\boldsymbol{\theta}) = c(\boldsymbol{u}(\boldsymbol{\theta})) \prod_{k=1}^{K} f_k(\theta_k)$$
 (17)

where f is density of c.d.f. F and $\mathbf{u} \triangleq \mathbf{u}(\boldsymbol{\theta}) \in \mathbb{I}^K$, with $u_k \triangleq F_k(\theta_k), \forall k \in \{1, 2, ..., K\}$.

Proof: By chain rule, we have $f(\boldsymbol{\theta}) = \frac{\partial F(\boldsymbol{\theta})}{\partial \theta_1...\partial \theta_K} = \frac{\partial C(\boldsymbol{u})}{\partial u_1...\partial u_K} \prod_{k=1}^K \frac{\partial u_k}{\partial \theta_k} = c(\boldsymbol{u}) \prod_{k=1}^K f_k(\theta_k).$ The density (17) shows that a joint p.d.f. can be factorized

The density (17) shows that a joint p.d.f. can be factorized into two parts: the dependent part represented by copula and the independent part represented by product of its marginals. Hence, the copula fully extracts all dependent relationships between r.v. θ_k , $k \in \{1, 2, ..., K\}$, from joint p.d.f. $f(\theta)$.

Remark 15. Note that, since copula C is essentially a c.d.f. by definition (15), the copula C(u) of independent c.d.f. $F(\theta) = \prod_{k=1}^K F_k(\theta_k)$ is also factorable, i.e. $C(u) = \prod_{k=1}^K u_k$, and, hence, c(u) = 1 by (17), as illustrated in Fig. 6.

B. Copula's invariant transformations

Let us focus on continuous copula and its useful transformation's properties in this subsection. These properties are also satisfied with discrete copulas via pseudo-inverse function (16).

1) Copula's rescaling transformation: By copula's density definition (17), we can see that a copula $c(u(\theta))$ is merely a rescaled coordinate form of original joint p.d.f. $f(\theta)$, as follows:

Corollary 16. (Copula's rescaling property)

$$1 = \int_{\boldsymbol{u}(\boldsymbol{\theta}) \in \mathbb{T}^K} c(\boldsymbol{u}) d\boldsymbol{u} = \int_{\boldsymbol{\theta} \in \mathbb{R}^K} f(\boldsymbol{\theta}) d\boldsymbol{\theta}$$
 (18)

Proof: By definition in (17), we have $u_k \triangleq F_k(\theta_k)$ and $d\theta \triangleq \prod_{k=1}^K d\theta_k$, which yields: $d\mathbf{u} = \prod_{k=1}^K du_k = \prod_{k=1}^K f_k(\theta_k) d\theta = \frac{f(\theta)}{c(\mathbf{u}(\theta))} d\theta$. Q.E.D.

The rescaling property (18) will be useful later when we wish to change the integrated variables from θ to u in copula's manipulation.

2) Copula's monotone transformation: Under generally monotonic transformation, which is not necessarily strictly increasing, the copula is linearly invariant (c.f. [14] for details). In this paper, let us recall here the useful rank-invariant property of copula under increasing transformation, as follows:

Theorem 17. (Copula's rank-invariant property) [14], [42] Let $\widetilde{\boldsymbol{\theta}} \triangleq [\widetilde{\theta}_1, \widetilde{\theta}_2, \dots, \widetilde{\theta}_K]^T \in \mathbb{R}^K$, in which $\widetilde{\theta}_k \triangleq \varphi_k(\theta_k)$ is a strictly increasing function of r.v. $\theta_k \in \mathbb{R}$, $\forall k \in \{1, 2, \dots, K\}$. Then the density copulas \widetilde{c} and c of $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$, respectively, have the same form, i.e. $\widetilde{c}(\boldsymbol{u}) = c(\boldsymbol{u})$, $\forall \boldsymbol{u} \in \mathbb{I}^K$.

Intuitively, the copula's rank-invariant property is merely a consequence of natural rank-invariant property of marginal c.d.f. under increasing transformation, as implied by definition of copula (15) and illustrated in Fig. 5.

3) Copula's marginal transformation: For later use, let us emphasize a very special case of rank-invariant property, namely marginal transformation. By definition (17), we can see that copula separates the dependence part of joint p.d.f. from its marginals. Hence, we can freely replace any marginal F_k with new marginal \widetilde{F}_k , $\forall k \in \{1, 2, \dots, K\}$, without changing the form of copula, as shown below:

Corollary 18. (Copula's marginal-invariant property) Let $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta}) \triangleq [\theta_1, \dots, \widetilde{\theta}_k(\theta_k), \dots, \theta_K]^T \in \mathbb{R}^K$, in which r.v. θ_k in $\boldsymbol{\theta}$ is replaced by a continuously transformed r.v. $\widetilde{\theta}_k(\theta_k) \triangleq \widetilde{F}_k^{\leftarrow}(F_k(\theta_k))$, for any $k \in \{1, 2, \dots, K\}$. Then the density copulas \widetilde{c}_k and c of $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$, respectively, have the same form, i.e. $\widetilde{c}_k(\boldsymbol{u}) = c(\boldsymbol{u}), \ \forall \boldsymbol{u} \in \mathbb{I}^K$.

Proof: This corollary is a direct consequence of the copula's rank-invariant property, since the continuous c.d.f. functions $\widetilde{F}_k^{\leftarrow}$ and F_k are both strictly increasing function for continuous variables.

The marginal-invariant property shows that when we replace the marginal distribution $f_k(\theta_k)$ of joint p.d.f. $f(\theta)$ in (17) by another marginal distribution $\widetilde{f}_k(\theta_k)$, the resulted joint distribution $\widetilde{f}(\theta)$ does not change its original copula form, i.e.:

$$\begin{cases}
f(\boldsymbol{\theta}) &= f(\theta_{\backslash k} | \theta_k) f_k(\theta_k) \\
\widetilde{f}(\boldsymbol{\theta}) &= f(\theta_{\backslash k} | \theta_k) \widetilde{f}_k(\theta_k)
\end{cases} \Rightarrow \widetilde{c}(\boldsymbol{u}) = c(\boldsymbol{u}), \forall \boldsymbol{u} \in \mathbb{I}^K$$
(19)

Indeed, by Corollary 18, we have $f(\theta) = \widetilde{f}(\widetilde{\theta}(\theta))$, i.e. the distribution $\widetilde{f}(\theta)$ is merely a marginally rescaling form of $f(\theta)$ and, hence, does not change the form of copula.

C. Copula's divergence

Because the copula is essentially a distribution itself, the KL divergence (9) can be applied directly to any two copulas. Let us show the relationship between joint p.d.f. and its copula via KL divergence in this subsection.

1) Mutual information: Because all dependencies in a joint p.d.f. f in (17) is captured by its copula, it is natural that the mutual information of joint p.d.f. f can also be computed via its copula form c in (17), as shown below.

Proposition 19. (Mutual information)

For continuous copula c in (17), the mutual information $I(\theta)$ of joint p.d.f. $f(\theta)$ is equal to continuous entropy H of copula density $c(u(\theta))$, as follows:

$$I(\boldsymbol{\theta}) = H(c(\boldsymbol{u})). \tag{20}$$

Proof: The proof is straight-forward from definition of KL divergence (9) and copula density (17), as follows: $I(\theta) \triangleq \mathrm{KL}(f(\theta)||\prod_{k=1}^K f_k(\theta_k)) = \mathbb{E}_{f(\theta)}\log\frac{f(\theta)}{\prod_{k=1}^K f_k(\theta_k)} = \mathbb{E}_{f(\theta)}\log c(u(\theta)) = \mathbb{E}_{c(u)}\log c(u) = \mathrm{H}(c(u)),$ in which θ was transformed to u via rescaling property (18). For a special case of bivariate copula density, another proof was given in [17].

2) KL divergence (KLD): In literature, the below copulabased KL divergence for a joint p.d.f. was already given for a special case of conditional structure [44]. For later use, let us recall their proof here in a slightly more generally form, via pseudo-inverse (16) and rescaling property (18).

Proposition 20. (Copula's divergence) [44]

The KLD of two joint p.d.f. f, \widetilde{f} in (17) is the sum of KLD of their copulas c, \widetilde{c} and KLDs of their marginals f_k , \widetilde{f}_k , as follows:

$$\mathit{KL}_{\widetilde{f}||f} = \mathit{KL}(\widetilde{c}(\boldsymbol{u})||c(F(\widetilde{F}^{\leftarrow}(\boldsymbol{u})))) + \sum_{k=1}^{K} \mathit{KL}_{\widetilde{f}_{k}||f_{k}}$$
 (21)

in which the copula \widetilde{c} of \widetilde{f} was rescaled back to marginal coordinates of f, i.e. $\widetilde{c}(\widetilde{F}(F^{\leftarrow}(\boldsymbol{u}))) \triangleq \widetilde{c}(\widetilde{F}_1(F_1^{\leftarrow}(u_1)), \ldots, \widetilde{F}_K(F_K^{\leftarrow}(u_K))).$

Proof: By definition of KLD (9) and copula density (17), we have: $\mathrm{KL}(f(\theta)||\widetilde{f}(\theta)) = \mathbb{E}_{f(\theta)}\log\frac{f(\theta)}{\widetilde{f}(\theta)} = \mathbb{E}_{f(\theta)}\log\frac{c(u(\theta))}{\widetilde{c}(\widetilde{u}(\theta))} + \sum_{k=1}^K \mathbb{E}_{f(\theta)}\log\frac{f_k(\theta_k)}{\widetilde{f}_k(\theta_k)}$, of which the second term in r.h.s. is actually KLDs of marginal, i.e. $\mathbb{E}_{f(\theta)}\log\frac{f_k(\theta_k)}{\widetilde{f}_k(\theta_k)} = \mathbb{E}_{f_k(\theta_k)}\log\frac{f_k(\theta_k)}{\widetilde{f}_k(\theta_k)} = \mathrm{KL}(f_k(\theta_k))||\widetilde{f}_k(\theta_k))$ and the first term in r.h.s. is actually KLD of copulas, via rescaling property (18), as follows: $\mathbb{E}_{f(\theta)}\log\frac{c(u(\theta))}{\widetilde{c}(\widetilde{u}(\theta))} = \mathbb{E}_{f(\theta)}\log\frac{c(F_1(\theta_1),\dots,F_K(\theta_K))}{\widetilde{c}(\widetilde{F}_1(\theta_1),\dots,\widetilde{F}_K(\theta_K))} = \mathbb{E}_{c(u)}\log\frac{c(u)}{\widetilde{c}(\widetilde{F}(F^{\leftarrow}(u)))} = \mathrm{KL}(c(u)||\widetilde{c}(F(F^{\leftarrow}(u)))).$

Note that, by copula's marginal- and rank-invariant properties in section III-B, we can see that the marginal rescaling form $\widetilde{c}(\widetilde{F}(F^{\leftarrow}(\boldsymbol{u})))$ of \widetilde{c} in (21) does not change the original form of copula \widetilde{c} .

Remark 21. If all \widetilde{F}_k are exact marginals of $F(\theta)$, i.e. $\widetilde{F}_k = F_k$ in (21), $\forall k \in \{1, 2, \dots, K\}$, we have $\mathrm{KL}(f(\theta)||\widetilde{f}(\theta)) = \mathrm{KL}(c(\boldsymbol{u})||\widetilde{c}(\boldsymbol{u}))$. Furthermore, if $\widetilde{c}(\boldsymbol{u})$ is also an independent copula, as noted in Remark 15, the KL divergence in (21) will be equal to mutual information $I(\theta)$ in (20).

IV. COPULA VARIATIONAL BAYES APPROXIMATION

As shown in (21), the KL divergence between any two distributions can always be factorized as the sum of KL divergence of their copulas and KL divergences of their marginals. Exploiting this property, we will design a novel iterative copula VB (CVB) algorithm in this section, such that the CVB distribution is closest to the true distribution in terms of KL divergence, under constraint of initially approximated copula's form. The mean-field approximations, which are special cases of CVB, will also be revisited later in this section.

A. Motivation of marginal approximation

Let us now consider a joint p.d.f. $f(\theta)$, of which the true marginals $f_k(\theta_k) = \int_{\theta \setminus k} f(\theta) d\theta_{\setminus k}, \ k \in \{1,2,\ldots,K\}$, are either unknown or too complicated to compute. A natural approximation of $f_k(\theta_k)$ is then to seek a closed form distribution $\widetilde{f}_k(\theta_k)$ such that their KL divergences $\sum_{k=1}^K \mathrm{KL}_{f_k||\widetilde{f}_k}$ in (21) is minimized. This direct approach is, however, not

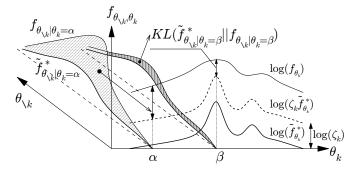


Figure 7. Illustration of Conditionally Variational approximation (CVA), as defined in (23). The lower KL divergence, the better approximation. Given initially a conditional form $\widetilde{f}^*_{\theta \setminus k}|\theta_k$ for $\widetilde{f}_\theta = \widetilde{f}^*_{\theta \setminus k}|\theta_k \widetilde{f}_{\theta_k}$, the optimally approximated marginal $\widetilde{f}^*_{\theta k}$ minimizing KL($\widetilde{f}_\theta || f_\theta)$ is proportional to the true marginal $f_{\theta k}$ in $f_\theta = f_{\theta \setminus k}|\theta_k f_{\theta_k}$ by a fraction of normalized conditional divergence $\zeta_k \exp \mathrm{KL}(\widetilde{f}^*_{\theta \setminus k}|\theta_k||f_{\theta \setminus k}|\theta_k)$, where ζ_k is the normalizing constant. In traditional VB approximation (29), we simply set $\widetilde{f}^*_{\theta \setminus k}|\theta_k = \widetilde{f}^*_{\theta \setminus k}$, which is independent of θ_k .

feasible if the integration for true marginal $f_k(\theta_k)$ is very hard to compute at the beginning.

A popular approach in literature is to find an approximation $f(\theta)$ of the joint distribution $f(\theta)$ such that their KL divergence $\mathrm{KL}_{\widetilde{f}||f} \triangleq \mathrm{KL}(\widetilde{f}(\pmb{\theta})||f(\pmb{\theta}))$ can be minimized. This indirect approach is more feasible since it circumvents the explicit form of $f_k(\theta_k)$. Also, since $\mathrm{KL}_{\widetilde{f}||f}$ is the upper bound of $\sum_{k=1}^{K} \mathrm{KL}_{\widetilde{f}_k||f_k}$, as shown in (21), it would yield good approximated marginals $\widetilde{f}_k(\theta_k)$ if $\mathrm{KL}_{\widetilde{f}||f}$ could be set low enough. This is the objective of CVB algorithm in this section. Remark 22. Another approximation approach is find $f(\theta)$ such that the copula's KL divergence $KL(\widetilde{c}(\boldsymbol{u})||c(F(F^{\leftarrow}(\boldsymbol{u}))))$ in (21) is as close as possible to $KL(c(u)||\widetilde{c}(u))$, which is equivalent to the exact case $f_k = f_k, \forall k \in \{1, 2, \dots, K\}$. This copula's analysis approach is promising, since the original copula form can be extracted from mutual dependence part of the original $f(\theta)$, without the need of marginal's normalization, as shown in [44] for a simple case of a Gaussian copula function. However, this approach would generally involve copula's explicit analysis, which is not a focus of this paper and will be left for future work.

B. Copula Variational approximation

Since the CVB algorithm is actually an iterative procedure of many Conditionally Variational approximation (CVA) steps, let us define the CVA step first, which is also illustrated in Fig. 7.

1) Conditionally Variational approximation (CVA): For a good approximation \widetilde{f}_k of f_k , let us initially pick a closed form p.d.f. $\widetilde{f}(\boldsymbol{\theta}) = \widetilde{f}^*(\theta_{\backslash k}|\theta_k)\widetilde{f}_k(\theta_k)$, in which the conditional distribution $\widetilde{f}_{\backslash k|k}^* \triangleq \widetilde{f}^*(\theta_{\backslash k}|\theta_k)$ is fixed and given. The optimal approximation $\widetilde{f}_k^* \triangleq \widetilde{f}_k^*(\theta_k)$ is then found by the following Theorem, which is also the foundational idea of this paper:

Theorem 23. (Conditionally Variational approximation) Let $\widetilde{f} = \widetilde{f}^*_{\backslash k|k} \widetilde{f}_k$ be a family of distributions with fixed-form conditional $\widetilde{f}_{\backslash k|k}^*$. Then \widetilde{f} is convex over marginals \widetilde{f}_k , which yields:

$$KL_{\widetilde{f}||f} = KL_{\widetilde{f}||\widetilde{f}^*} + KL_{\widetilde{f}^*||f} \ge KL_{\widetilde{f}^*||f} = \log\frac{1}{\zeta_L}$$
 (22)

owing to Bregman pythagorean property (3) for functional space (9-10). The distribution $\tilde{f} = \tilde{f}^*$ minimizing $KL_{\tilde{f}||f}$ and the value ζ_k in (22) are given as follows:

$$\widetilde{f}_{k}^{*}(\theta_{k}) = \frac{f_{k}(\theta_{k})}{\zeta_{k} \exp(KL_{\widetilde{f}_{\backslash k|k}^{*}||f_{\backslash k|k}})}$$

$$= \frac{1}{\zeta_{k}} \exp \mathbb{E}_{\widetilde{f}^{*}(\theta_{\backslash k}|\theta_{k})} \log \frac{f(\boldsymbol{\theta})}{\widetilde{f}^{*}(\theta_{\backslash k}|\theta_{k})}$$
(23)

in which ζ_k is the normalizing constant of \widetilde{f}_k^* in (23) and $KL_{\widetilde{f}_{\backslash k|k}^*||f_{\backslash k|k}} \triangleq KL(\widetilde{f}^*(\theta_{\backslash k}|\theta_k)||f(\theta_{\backslash k}|\theta_k)).$

Note that, if the marginal $\widetilde{f}_k = \widetilde{f}_k$ is initially fixed instead, \widetilde{f} is then convex over $\widetilde{f}_{\backslash k|k}$ and, hence, the conditional $\widetilde{f}_{\backslash k|k}^*$ minimizing $KL_{\widetilde{f}||f}$ in (22) is the true conditional distribution $f_{\backslash k|k}$, i.e. $\widetilde{f}_{\backslash k|k}^* = f_{\backslash k|k}$.

Proof: Firstly, we note that, for any mixture $\widetilde{f}_k(\theta_k) = p_1\widetilde{f}_1(\theta_k) + p_2\widetilde{f}_2(\theta_k)$, we always have $\widetilde{f}(\theta) = p_1\widetilde{f}_1(\theta) + p_2\widetilde{f}_2(\theta)$. Hence, \widetilde{f} is convex over \widetilde{f}_k with fixed $\widetilde{f}_{\setminus k|k}$ and satisfies the Bregman pythagorean equality (22), since KL divergence is a special case of Bregman divergence (9). We can also verify the pythagorean equality (22) directly, similarly to the proof of copula's KL divergence (21), as follows:

$$\begin{split} \mathrm{KL}_{\widetilde{f}||f} &= \mathbb{E}_{\widetilde{f}_{k}} \mathrm{KL}_{\widetilde{f}_{\backslash k|k}^{*}||f_{\backslash k|k}} + \mathrm{KL}_{\widetilde{f}_{k}||f_{k}} \\ &= \mathbb{E}_{\widetilde{f}_{k}} \log \frac{\widetilde{f}_{k}}{\frac{1}{\zeta_{k}} \frac{f_{k}}{\exp(\mathrm{KL}_{\widetilde{f}_{\backslash k|k}^{*}||f_{\backslash k|k}})}} + \mathbb{E}_{\widetilde{f}_{k}} \log \frac{1}{\zeta_{k}} \\ &= \underbrace{\mathrm{KL}_{\widetilde{f}_{k}||\widetilde{f}_{k}^{*}}}_{\mathrm{KL}_{\widetilde{f}||\widetilde{f}^{*}}} + \underbrace{\log \frac{1}{\zeta_{k}}}_{\mathrm{KL}_{\widetilde{f}^{*}||f}} \end{split}$$
(24)

in which the form \widetilde{f}_k^* is defined in (23) and ζ_k is independent of θ_k . Also, we have $\mathrm{KL}_{\widetilde{f}||\widetilde{f}^*} = \mathrm{KL}_{\widetilde{f}_k||\widetilde{f}_k^*}$ in the first term of r.h.s. of (24) since \widetilde{f} and \widetilde{f}^* only differ in marginals \widetilde{f}_k , \widetilde{f}_k^* . For the second term, by definition (23), we have $\mathrm{KL}_{\widetilde{f}_k^*|k||f_{\backslash k|k}} = \log\frac{1}{\zeta_k}\frac{f_k(\theta_k)}{\widetilde{f}_k^*(\theta_k)}$, which yields: $\mathbb{E}_{\widetilde{f}_k^*}\mathrm{KL}_{\widetilde{f}_{\backslash k|k}^*||f_{\backslash k|k}} = \log\frac{1}{\zeta_k} - \mathrm{KL}_{\widetilde{f}_k^*||f_k} \Leftrightarrow \mathrm{KL}_{\widetilde{f}^*||f} = \log\frac{1}{\zeta_k}$ in (22) and (24). Lastly, the second equality in (23) is given as follows: $f_k(\theta_k)/\exp\mathrm{KL}_{\widetilde{f}_{\backslash k|k}^*||f_{\backslash k|k}} = f_k(\theta_k)\exp\mathbb{E}_{\widetilde{f}_{\backslash k|k}^*}\log\frac{f_{\backslash k|k}}{\widetilde{f}_{\backslash k|k}^*} = \exp\mathbb{E}_{\widetilde{f}_{\backslash k|k}^*}\log\frac{f(\theta)}{\widetilde{f}_{\backslash k|k}^*}$.

If $\widetilde{f}_k = \widetilde{f}_k^*$ is fixed instead, \widetilde{f} is then convex over a mixture of $\widetilde{f}_{\backslash k|k}$ as shown above. Then, $\mathrm{KL}_{\widetilde{f}||f}$ in (24) is minimum at $\widetilde{f}_{\backslash k|k}^* = f_{\backslash k|k}$, since the term $\mathrm{KL}_{\widetilde{f}_k||f_k} = \mathrm{KL}_{\widetilde{f}_k^*||f_k}$ in (24) is now fixed and the term $\mathbb{E}_{\widetilde{f}_k^*}\mathrm{KL}_{\widetilde{f}_{\backslash k|k}^*||f_{\backslash k|k}}$ is minimum at zero with $\widetilde{f}_{\backslash k|k}^* = f_{\backslash k|k}$.

In Theorem 23, the conditional $\widetilde{f}_{\backslash k|k}^*$ is fixed beforehand and \widetilde{f}_k^* is found in a free-form variational space, hence the name

Conditionally Variational approximation (CVA). The case of fixed marginal \widetilde{f}_k^* is not interesting, since $\mathrm{KL}_{\widetilde{f}||f}$ in this case is only minimized at the true conditional $f_{\backslash k|k}$, which is often unknown initially.

Remark 24. The CVA form above is a generalized form of the so-called Conditional Variational Bayesian inference [28] or Conditional mean-field [29] in literature, which are merely applications of mean-field approximations to a conditionally independent structure, i.e. $\widetilde{f}(\boldsymbol{\theta}|\boldsymbol{\xi}) = \prod_{k=1}^K \widetilde{f}_k(\theta_k|\boldsymbol{\xi})$, given a latent variable $\boldsymbol{\xi}$ in this case.

2) Copula Variational algorithm: In CVA form above, we can only find one approximated marginal $\widetilde{f}_k^*(\theta_k)$, given conditional form $\widetilde{f}_{\backslash k|k}^*(\theta_{\backslash k}|\theta_k)$. In the iterative form below, we will iteratively multiply $\widetilde{f}_k^*(\theta_k)$ back to $\widetilde{f}_{\backslash k|k}^*(\theta_{\backslash k}|\theta_k)$ in order to find the reverse conditional $\widetilde{f}_{k|\backslash k}^*(\theta_k|\theta_{\backslash k})$ for the next $\widetilde{f}_{\backslash k}^*(\theta_{\backslash k})$ via (23). At convergence, we can find a set of approximations \widetilde{f}_k , $\forall k \in \{1,2,\ldots,K\}$, such that the $\mathrm{KL}_{\widetilde{f}||f}$ is locally minimized, as follows:

Corollary 25. (Copula Variational approximation) Let $\widetilde{f} = \widetilde{f}_{\backslash k|k}^{[0]} \widetilde{f}_k$ be the initial approximation with initial form $\widetilde{f}_{\backslash k|k}^{[0]}$. At iteration $\nu \in \{1,2,\ldots,\nu_c\}$, the approximation $\widetilde{f}^{[\nu]} = \widetilde{f}_{\backslash k|k}^{[\nu-1]} \widetilde{f}_k^{[\nu]} = \widetilde{f}_{k|\lambda}^{[\nu]} \widetilde{f}_{\backslash k}^{[\nu]}$ is given by (23), as follows:

$$\widetilde{f}_k^{[\nu]}(\theta_k) = \frac{f_k(\theta_k)}{\zeta_k^{[\nu]} \exp(KL_{\widetilde{f}_{k|k}^{[\nu-1]}||f_{k|k}})}$$
(25)

in which the reverse conditional is $\widetilde{f}_{k|k}^{[\nu]} = \frac{\widehat{f}_{k|k}^{[\nu-1]}\widehat{f}_{k}^{[\nu]}}{\widehat{f}_{k}^{[\nu]}}$ and $\widetilde{f}_{k}^{[\nu]} \triangleq \int_{\theta_k} \widetilde{f}_{k|k}^{[\nu-1]} \widetilde{f}_{k}^{[\nu]}$, $\forall k \in \{1,2,\ldots,K\}$. Then, the value $KL_{\widetilde{f}^{[\nu]}||f} = \log \frac{1}{\zeta_k^{[\nu]}}$ in (22), where $\zeta_k^{[\nu]}$ is the normalizing constant of marginal $\widetilde{f}_k^{[\nu]}$, monotonically decreases to a local minimum at convergence $\nu = \nu_c$, as illustrated in Fig. 8.

Note that, by copula's marginal-invariant property (19), the copula's form of the iterative joint distribution $\tilde{f}^{[\nu]}(\boldsymbol{\theta})$ is invariant with any updated marginals $\tilde{f}_k^{[\nu]}(\theta_k)$, $\forall k$, hence the name Copula Variational approximation.

Proof: Since the calculation of reverse form $\widetilde{f}_{k|\setminus k}^{[\nu]}$ does not change $\widetilde{f}^{[\nu]}(\boldsymbol{\theta})$, the value $\mathrm{KL}_{\widetilde{f}^{[\nu]}||f}$ only decreases with marginal update $\widetilde{f}_k^{[\nu]}$ via (22-23) and, hence, converges monotonically.

If the initial form $\widetilde{f}_{\backslash k|k}^{[0]}$ belongs to the independent space, i.e. $\widetilde{f}_{\backslash k|k}^{[0]} = \widetilde{f}_{\backslash k}^{[0]}$, the copula of the joint $\widetilde{f}_{\boldsymbol{\theta}}^{[0]}$ will have independent form, as noted in Remark 15, and cannot leave this independence space via dual iterations of (25). Hence, for a binary partition $\boldsymbol{\theta} = \{\theta_{\backslash k}, \theta_k\}$, an initially independent copula will lead to a mean-field approximation.

Nonetheless, this is not true in general for ternary partition $\boldsymbol{\theta} = \{\theta_k, \theta_j, \theta_m\}$ or for a generic network of parameters, since the iterative CVA (25) can be implemented with different partitions of a network at any iteration, without changing the joint network's copula or increasing the joint KL divergence $\mathrm{KL}_{\widetilde{f}^{[\nu]}||f}$.

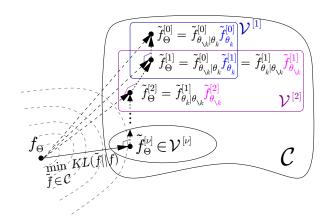


Figure 8. Venn diagram for iterative Copula Variational approximation (CVA), given in (25). The dashed contours represent the convexity of $\mathrm{KL}(\widetilde{f}_{\theta}||f_{\theta})$ over distributional points \widetilde{f}_{θ} . The set \mathcal{C} , possibly nonconvex, denotes a class of distributions with the same copula form. Given initial form $\widetilde{f}_{\theta \setminus k}^{[0]}|_{\theta_k}$, the joint distributions $\widetilde{f}_{\theta}^{[0]}$ and $\widetilde{f}_{\theta}^{[1]}$ belong to the same convex set $\mathcal{V}^{[1]} \subseteq \mathcal{C}$, by Theorem 23 and Corollary 25. The CVA $\widetilde{f}_{\theta}^{[1]}$ is the Bregman projection of the true distribution f_{θ} onto $\mathcal{V}^{[1]}$, with $\widetilde{f}_{\theta_k}^{[1]} = \arg\min_{\widetilde{f}_{\theta_k} \in \mathcal{V}^{[1]}} \mathrm{KL}(\widetilde{f}_{\theta}||f_{\theta})$, as shown in (22) and illustrated in Fig. 2. By interchanging the role of $\theta_{\setminus k}$ and θ_k , the $\mathrm{KL}(\widetilde{f}_{\theta}^{[\nu]}||f_{\theta})$ never increases over iterations ν and, hence, converges to a local minimum inside copula set \mathcal{C} . In traditional VB algorithm, we set $\widetilde{f}_{\theta_{\setminus k}}^{[\nu]}|_{\theta_k} = \widetilde{f}_{\theta_{\setminus k}}^{[\nu]}$, which belongs to the independent copula class at all iterations ν .

For example, in ternary partition, even if we initially set $\widetilde{f}_{k|\backslash k}^{[0]} = \widetilde{f}_k^{[0]}$ independent of $\theta_{\backslash k} = \{\theta_j, \theta_m\}$ and yield the updated $\widetilde{f}_{\backslash k}^{[1]} = \widetilde{f}_{j}^{[1]}(\theta_j, \theta_m)$ for $\widetilde{f}_{j}^{[1]} = \widetilde{f}_{j}^{[0]}\widetilde{f}_{j}^{[1]} = \widetilde{f}_{m|j}^{[1]}\widetilde{f}_{j}^{[1]}$ via (25), the reverse form $\widetilde{f}_{m|j}^{[1]}$ yields $\widetilde{f}_{j}^{(2)} \triangleq \widetilde{f}_{j}^{[2]}(\theta_k, \theta_j)$ via (25) again and, hence $\widetilde{f}_{k|\backslash k}^{[2]} = \widetilde{f}_{j}^{[2]}(\theta_k|\theta_j)$ dependent on θ_j again, which does not yield a mean-field approximation in subsequent iterations of (25). This ternary partition scheme will be implemented in (59) and clarified further in Remark 42.

3) Conditionally exponential family (CEF) approximation: The computation in above approximations will be linearly tractable, if the true joint $f(\theta)$ and the approximated conditional $\tilde{f}_{\backslash k|k}$ can be linearly factorized with respect to logoperator in (23) and (25). The distributions satisfying this property belong to a special class of distributions, namely CEF, defined as follows:

Definition 26. (Conditionally Exponential Family)

A joint distribution $f(\theta)$ is a member of CEF if it has the following form:

$$f(\boldsymbol{\theta}) \propto \exp \left\langle \boldsymbol{g}_k(\theta_k), \boldsymbol{g}_{\setminus k}(\theta_{\setminus k}) \right\rangle$$
 (26)

where g_k , $g_{\setminus k}$ are vectors dependent on θ_k , $\theta_{\setminus k}$ element-wise, respectively. Note that, the form (26) is similar to the well-known Exponential Family in literature [2], [45], hence the name CEF.

From (26), the marginal of a joint CEF distribution is:

$$f(\theta_k) \propto \int_{\theta_{\setminus k}} \exp\left\langle \boldsymbol{g}_k(\theta_k), \boldsymbol{g}_{\setminus k}(\theta_{\setminus k}) \right\rangle d\theta_{\setminus k}$$
 (27)

which may not be tractable, since the CEF form is not factorable in general. In contrast, the CVA (23) for CEF distributions (26) is more tractable, as follows:

Corollary 27. (CEF approximation)

Let $\widetilde{f} = \widetilde{f}_{\backslash k|k}^* \widetilde{f}_k$ be a distribution with $\widetilde{f}_{\backslash k|k}^* = \exp \langle h_k(\theta_k), h_{\backslash k}(\theta_{\backslash k}) \rangle$ given by CEF form in (26). If the true distribution $f(\theta)$ also takes the CEF form (26), the approximation \widetilde{f}^* minimizing $KL_{\widetilde{f}||f}$ in (22), as given by (23), also belongs to CEF, as follows:

$$\widetilde{f}_k^*(\theta_k) \propto \exp\left\langle \boldsymbol{\eta}_k(\theta_k), \boldsymbol{\eta}_{\backslash k}^*(\theta_k) \right\rangle$$
 (28)

where $\eta_{\backslash k}^*(\theta_k) \triangleq \mathbb{E}_{\widetilde{f}_{\backslash k|k}^*} \eta_{\backslash k}(\theta_{\backslash k})$, with $\eta_k \triangleq g_k - h_k$ and $\eta_{\backslash k} \triangleq g_{\backslash k} - h_{\backslash k}$.

Proof: The form (28) is a direct consequence of (23), since both $\widetilde{f}_{\backslash k|k}^*$ and $f(\theta)$ in (23) now have CEF form (26).

From (27-28), we can see that the integral in (27) has moved inside the non-linear exp operator in (28) and, hence, become linear and numerically tractable. Then, substituting (28) into iterative CVA (25), we can see that the iterative CVA for CEF is also tractable, since we only have to update the parameters of CEF iteratively in (28) until convergence.

Remark 28. In the nutshell, the key advantage of KL divergence is to approximate the originally intractable arithmetic mean (27) by the tractable geometric mean in exponential domain (28), as noted in Remark 11.

4) Backward KLD and minimum-risk (MR) approximation: In above approximations, we have used the forward $\mathrm{KL}_{\widetilde{f}||f}$ (22) as the approximation criterion, since the Bregman pythagorean property (3) is only valid for forward $\mathrm{KL}_{\widetilde{f}||f}$. Moreover, the approximation via backward $\mathrm{KL}_{f||\widetilde{f}}$ is not interesting since the minimum is only achieved with the true distributions, as shown below:

Corollary 29. (Conditionally minimum-risk approximation) The approximation $\widetilde{f}^* = \widetilde{f}_{\backslash k|k}\widetilde{f}_k$ minimizing backward $KL_{f||\widetilde{f}}$ is either $\widetilde{f}^* = \widetilde{f}_{\backslash k|k}^{MR}f_k$ or $\widetilde{f}^* = f_{\backslash k|k}\widetilde{f}_k^{MR}$ for fixed $\widetilde{f}_{\backslash k|k}^{MR}$ or fixed \widetilde{f}_k^{MR} , respectively, where f_k and $f_{\backslash k|k}$ are the true marginal and conditional distributions.

Proof: Similar to proof of Theorem 23, the backward form is $\mathrm{KL}_{f||\widetilde{f}} = \mathbb{E}_{f_k} \mathrm{KL}_{f_{\backslash k|k}||\widetilde{f}_{\backslash k|k}} + \mathrm{KL}_{f_k||\widetilde{f}_k}$. Hence, $\mathrm{KL}_{f||\widetilde{f}^*}$ is minimum at $\widetilde{f}_k^{\mathrm{MR}} = f_k$ for fixed $\mathrm{KL}_{f_{\backslash k|k}||\widetilde{f}_{\backslash k|k}^{\mathrm{MR}}}$ and minimum at $\widetilde{f}_{\backslash k|k}^{\mathrm{MR}} = f_{\backslash k|k}$ for fixed $\mathrm{KL}_{f_k||\widetilde{f}_k^{\mathrm{MR}}}$.

**Remark 30. The Corollary 29 is the generalized form of the minimum-risk approximation in [2], which minimizes

Remark 30. The Corollary 29 is the generalized form of the minimum-risk approximation in [2], which minimizes backward KL divergence in the context of VB approximation in mean-field theory. The name "minimum-risk" refers to the fact that the true distribution always yields minimum-risk estimation in Bayesian theory (c.f. Appendix A).

C. Mean-field approximations

If we confine the conditional form $\widetilde{f}=\widetilde{f}_{\backslash k|k}\widetilde{f}_k$ in above approximations by independent form, i.e. $\widetilde{f}=\widetilde{f}_{\backslash k}\widetilde{f}_k$, we will

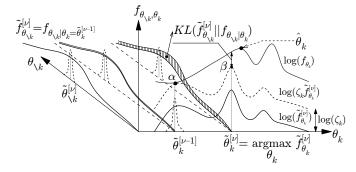


Figure 9. Illustration of Expectation-Maximization (EM) algorithm (30) as a special case of VB approximation. The lower KL divergence, the better approximation. Given restricted form $\widetilde{f}_{\theta \backslash_k}^{[\nu]} = f(\theta \backslash_k |\widetilde{\theta}_k^{[\nu-1]})$ at iteration ν , the approximated $\widetilde{f}_{\theta \backslash_k}^{[\nu]}$ minimizing $\mathrm{KL}(\widetilde{f}_{\theta \backslash_k}^{[\nu]}|f_{\theta})$ is proportional to the true marginal $f_{\theta k}$ by a fraction of conditional divergence, similar to Fig. 7. Note that, $\widetilde{\theta}_k^{[\nu]}$ might fail to converge to a local mode $\widehat{\theta}_k$ of the true marginal $f_{\theta k}$, if the peak β is lower than point α . For ICM algorithm (31), we further restrict $\widetilde{f}_{\theta \backslash_k}^{[\nu]}$ to a Dirac delta distribution concentrating around its mode $\widetilde{\theta}_k^{[\nu]}$ and, hence, $\widetilde{\theta}_k^{[\nu]} = \{\widetilde{\theta}_k^{[\nu]}, \widetilde{\theta}_k^{[\nu]}\}$ always converges to a joint local mode $\widehat{\theta}$ of the true distribution f_{θ} .

recover the so-called mean-field approximations in literature. Four cases of them, namely VB, EM, ICM and k-means algorithms, will be presented below.

1) Variational Bayes (VB) algorithm: From CVA (23), the VB algorithm is given as follows:

Corollary 31. (VB approximation)

The independent distribution $\tilde{f}^* = \tilde{f}_{\backslash k}^* \tilde{f}_k^*$ minimizing $KL_{\tilde{f}||f}$ in (22) is given by (23), as follows:

$$\widetilde{f}_k^*(\theta_k) \propto \frac{f_k(\theta_k)}{\exp(\mathit{KL}_{\widetilde{f}_{\setminus k}^*||f_{\setminus k}|k})} \propto \exp \mathbb{E}_{\widetilde{f}_{\setminus k}^*(\theta_{\setminus k})} \log f(\boldsymbol{\theta}), \quad (29)$$

 $\forall k \in \{1, 2, \dots, K\}$, as illustrated in Fig. 7.

Proof: Since $\widetilde{f}_{\backslash k|k} = \widetilde{f}_{\backslash k}$ does not depends on θ_k in this case, substituting $\widetilde{f}_{\backslash k|k} = \widetilde{f}_{\backslash k}$ into (23) yields (29).

Since there is no conditional form $\widetilde{f}_{\backslash k|k}$ to be updated, the iterative VB algorithm simply updates (29) iteratively for all marginals \widetilde{f}_k and $\widetilde{f}_{\backslash k}$, similar to (25), until convergence. Hence, VB algorithm is a special case of Copula Variational algorithm in Corollary 25, in which the approximated copula is of independent form, as noted in Remark 15.

2) Expectation-Maximization (EM) algorithm: If we restrict the independent form $\widetilde{f} = \widetilde{f}_{\setminus k}\widetilde{f}_k$ in VB algorithm with Dirac delta form $\widetilde{f}_{\rm EM} \triangleq \widetilde{f}_{\setminus k}\widetilde{\delta}_k$, where $\widetilde{\delta}_k \triangleq \delta(\theta_k - \widetilde{\theta}_k)$, we will recover the EM algorithm, as follows:

Corollary 32. (EM algorithm)

At iteration $\nu \in \{1, 2, \dots, \nu_c\}$, the EM approximation of $f(\boldsymbol{\theta})$ is $\widetilde{f}_{EM}^{[\nu]} \triangleq \widetilde{f}_{\backslash k}^{[\nu]} \widetilde{\delta}_k^{[\nu]}$, in which $\widetilde{f}_{\backslash k}^{[\nu]} = f(\theta_{\backslash k} | \widetilde{\theta}_k^{[\nu]})$ and $\widetilde{\delta}_k^{[\nu]} \triangleq \delta(\theta_k - \widetilde{\theta}_k^{[\nu]})$, as given by (29):

$$\begin{aligned} \widehat{\theta}_{k}^{[\nu]} &\triangleq \underset{\theta_{k}}{\operatorname{arg\,max}} \, \widehat{f}_{k}^{[\nu]}(\theta_{k}) = \underset{\theta_{k}}{\operatorname{arg\,max}} \, \mathbb{E}_{f(\theta_{\backslash k} | \widetilde{\theta}_{k}^{[\nu-1]})} \log f(\boldsymbol{\theta}) \\ &= \underset{\theta_{k}}{\operatorname{arg\,max}} \, \frac{f_{k}(\theta_{k})}{\exp(KL_{f(\theta_{\backslash k} | \widetilde{\theta}_{k}^{[\nu-1]}) | | f(\theta_{\backslash k} | \theta_{k})})}. \end{aligned}$$
(30)

If $\widetilde{\theta}_k^{[\nu]}$ converges to a true local maximum $\widehat{\theta}_k$ of the original marginal $f_k(\theta_k)$, as illustrated in Fig. 9, then $KL_{\widetilde{f}_{\rm EM}^{[\nu]}||f} = -\log f_k(\widetilde{\theta}_k^{[\nu]})$ converges to a local minimum.

Proof: Substituting the Dirac delta function $\widetilde{f}_k^*(\theta_k) = \delta(\theta_k - \widetilde{\theta}_k^{[\nu-1]})$ to VB approximation (29), we have $\widetilde{f}_{\setminus k}^{[\nu-1]}(\theta_{\setminus k}) = f(\theta_{\setminus k}|\widetilde{\theta}_k^{[\nu-1]})$, which yields (30) owing to (29). Since the KL value in (30) is never negative, we have $g(\theta_k) \triangleq f_k(\theta_k)/\exp(\mathrm{KL}_{f(\theta_{\setminus k}|\widetilde{\theta}_k^{[\nu-1]})||f(\theta_{\setminus k}|\theta_k)}) \leq f_k(\theta_k)$ and the equality $g(\widetilde{\theta}_k^{[\nu]}) = f_k(\widetilde{\theta}_k^{[\nu]})$ happens at $\theta_k = \widetilde{\theta}_k^{[\nu]}$, which means: $f_k(\widetilde{\theta}_k^{[\nu-1]}) = g(\widetilde{\theta}_k^{[\nu-1]}) \leq g(\widetilde{\theta}_k^{[\nu]}) = \max_k g(\theta_k) \leq \max_k f_k(\theta_k)$. Then, as illustrated in Fig. 9, if $\widetilde{f}_k^{[\nu]}(\widetilde{\theta}_k^{[\nu]})$ strictly increases over ν , $\widetilde{\theta}_k^{[\nu]}$ will converge to a local mode $\widehat{\theta}_k$ of $f_k(\theta_k)$, owing to majorization-maximization (MM) principle [21], [22]. Otherwise, $\widetilde{\theta}_k^{[\nu]}$ might fail to converge to $\widehat{\theta}_k$.

Lastly, from (24), we have:
$$\mathrm{KL}_{\widetilde{f}_{\mathrm{EM}}^{[\nu]}||f} = \mathbb{E}_{\widetilde{\delta}_{k}^{[\nu]}}\mathrm{KL}_{f(\theta_{\backslash k}|\widetilde{\theta}_{k}^{[\nu]})||f(\theta_{\backslash k}|\theta_{k})} + \mathrm{KL}_{\widetilde{\delta}_{k}^{[\nu]}||f_{k}} = \mathrm{KL}_{\widetilde{\delta}_{k}^{[\nu]}||f_{k}} = -\log f_{k}(\widetilde{\theta}_{k}^{[\nu]})$$
 by sifting property of Dirac delta function.

From (29-30), we can see that EM algorithm is a special case of VB algorithm. Both of them minimizes the KL divergence within the independent distribution space, namely mean-field space.

Since EM algorithm is a fixed-form approximation, it has low computational complexity. Nonetheless, as illustrated in Fig. 9, the point estimate $\widetilde{\theta}_k^{[\nu]}$ in EM algorithm (30) might fail to converge to a local mode $\widehat{\theta}_k$ of true marginal $f_k(\theta_k)$ in practice. In contrast, VB approximation is a free-form distribution and capable of approximating higher-order moments of true marginal $f_k(\theta_k)$.

Remark 33. Note that, EM algorithm is also a special case of Copula Variational algorithm (25) in conditional space. Indeed, if the marginal \widetilde{f}_k of $\widetilde{f} = \widetilde{f}_{\backslash k|k}\widetilde{f}_k$ in (25) is restricted to Dirac delta form, i.e. $\widetilde{f}_k = \widetilde{\delta}_k$, the joint \widetilde{f} will become a degenerated independent distribution, i.e. $\widetilde{f} = \widetilde{f}(\theta_{\backslash k}|\theta_k = \widetilde{\theta}_k)\delta(\theta_k - \widetilde{\theta}_k) = \widetilde{f}_{\backslash k}\widetilde{\delta}_k$, owing to sifting property of Dirac delta. Hence, EM algorithm is a very special approximation, since it belongs to both mean-field and copula-field approximations.

3) Iterative conditional mode (ICM) algorithm: If we further restrict the independent form $\widetilde{f} = \widetilde{f}_{\setminus k} \widetilde{f}_k$ in VB algorithm fully to Dirac delta form $\widetilde{f}_{\rm ICM} \triangleq \widetilde{\delta}_{\setminus k} \widetilde{\delta}_k$, we will recover the iterative plug-in algorithm, also called Iterative Conditional Mode (ICM) in literature [23], [24], as follows:

Corollary 34. (ICM algorithm)

At iteration $\nu \in \{1, 2, \dots, \nu_c\}$, the ICM approximation of $f(\theta)$ is $\widetilde{f}_{ICM}^{[\nu]} \triangleq \widetilde{f}_{\backslash k}^{[\nu]} \widetilde{f}_k^{[\nu]} = \widetilde{\delta}_{\backslash k}^{[\nu]} \widetilde{\delta}_k^{[\nu]}$, where $\widetilde{\delta}_{\backslash k}^{[\nu]} \triangleq \delta(\theta_{\backslash k} - \widetilde{\theta}_{\backslash k}^{[\nu]})$ and $\widetilde{\delta}_k^{[\nu]} \triangleq \delta(\theta_k - \widetilde{\theta}_k^{[\nu]})$ is given by (29), as follows:

$$\widetilde{\theta}_{k}^{[\nu]} \triangleq \underset{\theta_{k}}{\operatorname{arg\,max}} f(\theta_{k} | \widetilde{\theta}_{\backslash k}^{[\nu-1]}) = \underset{\theta_{k}}{\operatorname{arg\,max}} f(\theta_{k}, \theta_{\backslash k} = \widetilde{\theta}_{\backslash k}^{[\nu-1]}) \\
\widetilde{\theta}_{\backslash k}^{[\nu]} \triangleq \underset{\theta_{\backslash k}}{\operatorname{arg\,max}} f(\theta_{\backslash k} | \widetilde{\theta}_{k}^{[\nu]}) = \underset{\theta_{\backslash k}}{\operatorname{arg\,max}} f(\theta_{k} = \widetilde{\theta}_{k}^{[\nu]}, \theta_{\backslash k})$$
(31)

From (31), we can see that $\tilde{\boldsymbol{\theta}}^{[\nu]} = \{\tilde{\boldsymbol{\theta}}_{\backslash k}^{[\nu]}, \tilde{\boldsymbol{\theta}}_{k}^{[\nu]}\}$ iteratively converges to a local maximum $\hat{\boldsymbol{\theta}}$ of the true distribution $f(\boldsymbol{\theta})$ and, hence, $KL_{\widetilde{f}_{ICM}^{[\nu]}||f} = -\log f(\tilde{\boldsymbol{\theta}}^{[\nu]})$ converges to a local minimum.

Proof: The proof is a straight-forward derivation from either VB (29) or EM (30) algorithms, by sifting property of Dirac delta forms $\tilde{f}_{\backslash k}^{[\nu]} = \delta(\theta_{\backslash k} - \tilde{\theta}_{\backslash k}^{[\nu]})$ and $\tilde{f}_{k}^{[\nu]} = \delta(\theta_{k} - \tilde{\theta}_{k}^{[\nu]})$.

Since we merely plug the value $\{\widetilde{\theta}_{\backslash k}^{[\nu]}, \widetilde{\theta}_k^{[\nu]}\}$ into the true distribution $f(\theta)$ iteratively in (31) until it reaches a local maximum, the performance of this naive hit-or-miss approach is strongly influenced by the initial points $\{\widetilde{\theta}_{\backslash k}^{[0]}, \widetilde{\theta}_k^{[0]}\}$. Hence it is often used in practice when very low computational complexity is required or when the true distribution $f(\theta)$ does not have tractable CEF form (26).

Remark 35. Similar to the Remark 33, we can see that ICM is a degenerated form of VB, EM and Copula Variational approximations, owing to its very simple form (31).

4) K-means algorithm: In section VI-B1, we will show that the popular k-means algorithm is equivalent to ICM algorithm being applied to a mixture of independent Gaussian distributions. Hence, k-means is also a member of mean-field approximations.

D. Copula Variational Bayes (CVB) approximation

In a model with unknown multi-parameters $\boldsymbol{\theta} = \{\theta_k, \theta_{\backslash k}\}$, the minimum-risk estimation of θ_k can be evaluated from the marginal posterior $f(\theta_k|\boldsymbol{x}) = \int_{\theta_{\backslash k}} f(\boldsymbol{\theta}|\boldsymbol{x}) d\theta_{\backslash k}$ (c.f. Appendix A), in which the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{x})$ is then given via Bayes' rule: $f(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x},\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$. In practice, however, the computational complexity of the normalizing constant of $f(\boldsymbol{\theta}|\boldsymbol{x})$ involves all possible values of $\boldsymbol{\theta}$ and typically grows exponentially with number of data's dimension, which is termed the curse of dimensionality [7]. Then, without normalizing constant of $f(\boldsymbol{\theta}|\boldsymbol{x})$, the computation of moments of $f(\theta_k|\boldsymbol{x})$ is also intractable.

In this subsection, we will apply both copula-field and mean-field approximations to the joint posterior distribution $f(\theta|x) \propto f(x,\theta)$ and, then, return all marginal approximations $\tilde{f}(\theta_k|x)$ directly from the joint model $f(x,\theta)$, without computing the normalizing constant of $f(\theta|x)$, as explained below.

Corollary 36. (Copula Variational Bayes algorithm) At iteration $\nu \in \{1, 2, ..., \nu_c\}$, the CVB approximation $\widetilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x}) = \widetilde{f}^{[\nu-1]}(\boldsymbol{\theta}_{\backslash k}|\boldsymbol{\theta}_k, \boldsymbol{x})\widetilde{f}_k^{[\nu]}(\boldsymbol{\theta}_k|\boldsymbol{x})$ for the joint posterior $f(\boldsymbol{\theta}|\boldsymbol{x})$ is given by (23) and (25), as follows:

$$\widetilde{f}_{k}^{[\nu]}(\theta_{k}|\boldsymbol{x}) = \frac{1}{\zeta_{k}^{[\nu]}(\boldsymbol{x})} \frac{f(\theta_{k}|\boldsymbol{x})}{KL(\widetilde{f}^{[\nu-1]}(\theta_{\backslash k}|\theta_{k},\boldsymbol{x})||f(\theta_{\backslash k}|\theta_{k},\boldsymbol{x}))} (32)$$

$$= \frac{1}{\zeta_{k}^{[\nu]}(\boldsymbol{x})} \exp \mathbb{E}_{\widetilde{f}^{[\nu-1]}(\theta_{\backslash k}|\theta_{k},\boldsymbol{x})} \log \frac{f(\boldsymbol{x},\boldsymbol{\theta})}{\widetilde{f}^{[\nu-1]}(\theta_{\backslash k}|\theta_{k},\boldsymbol{x})}$$

in which $\widetilde{f}^{[\nu]}(\theta_k|\theta_{\backslash k}, \boldsymbol{x}) = \widetilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x})/\widetilde{f}^{[\nu]}_{\backslash k}(\theta_{\backslash k}|\boldsymbol{x}), \ \forall k \in \{1, 2, \dots, K\}.$ For stopping rule, the evidence lower bound (ELBO) for CVB is defined similarly to (22), as follows: $KL_{\widetilde{f}^{[\nu]}||f} = -ELBO^{[\nu]} + \log f(\boldsymbol{x}) \geq 0$, i.e. we have:

$$\log f(\boldsymbol{x}) \ge ELBO^{[\nu]} \triangleq -KL_{\widetilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x})||f(\boldsymbol{x},\boldsymbol{\theta})} = \log \zeta_k^{[\nu]}(\boldsymbol{x})$$
(33)

Since the evidence f(x) is a constant, $KL_{\widetilde{f}^{[\nu]}||f} \triangleq KL_{\widetilde{f}^{[\nu]}(\theta|x)||f(\theta|x)}$ monotonically decreases to a local minimum, while the marginal normalizing constant $\zeta_k^{[\nu]}(x)$ in (32) and $ELBO^{[\nu]}$ in (33) monotonically increase to a local maximum at convergence $\nu = \nu_c$.

Note that, the copula's form of the iterative CVB $\tilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x})$ is invariant with any updated marginal $\tilde{f}_k^{[\nu]}(\theta_k|\boldsymbol{x}), \ \forall k \in \{1,2,\ldots,K\}$, as shown in (19), hence the name Copula Variational Bayes approximation.

Proof: Firstly, we have $\mathrm{KL}_{\widetilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x})||f(\boldsymbol{x},\boldsymbol{\theta})} = \mathrm{KL}_{\widetilde{f}^{[\nu]}||f} - \log f(\boldsymbol{x})$ by definition of KL divergence (9), hence the definition of $\mathrm{ELBO}^{[\nu]}$ in (33). Then, similar to (24), the value $\mathrm{KL}_{\widetilde{f}||f} \triangleq \mathrm{KL}_{\widetilde{f}(\boldsymbol{\theta}|\boldsymbol{x})||f(\boldsymbol{\theta}|\boldsymbol{x})}$ for arbitrary \widetilde{f} in this case is:

$$KL_{\widetilde{f}||f} = \underbrace{KL_{\widetilde{f}_{k}||\widetilde{f}_{k}^{[\nu]}} + \log \frac{1}{\zeta_{k}^{[\nu]}(\boldsymbol{x})}}_{-\text{ELBO}} + \log f(\boldsymbol{x}), \quad (34)$$

in which $\widetilde{f}_k^{[\nu]}$ is defined in 32, the form $f(\theta)$ in (24) is now replaced by $f(\theta|\boldsymbol{x}) = f(\boldsymbol{x},\theta)/f(\boldsymbol{x})$, hence the term $f(\boldsymbol{x},\theta)$ in (32) and the constant evidence $\log f(\boldsymbol{x})$ in (34). Since $\mathrm{KL}_{\widetilde{f}_k||\widetilde{f}_k^{[\nu]}|} = 0$ for the case $\widetilde{f}_k = \widetilde{f}_k^{[\nu]}$, the value ELBO in (34) is equal to $\log \zeta_k^{[\nu]}(\boldsymbol{x})$, which yields (33). The rest of proof is similar to the proof of Corollary 25.

Note that, CVB algorithm (32) is essentially the same as the Copula Variational algorithm in (25). The key difference is that the former is applied to a joint posterior $f(\theta|x)$, while the latter is applied to a joint distribution $f(\theta)$. Hence, in CVB, the joint model $f(x,\theta)$ and ELBO (33) are preferred, since the evidence $\log f(x)$ is often hard to compute in practice. Nevertheless, for notational simplicity, let us call both of them CVB hereafter. By this way, the name CVA (23) also implies that it is the first step of CVB algorithm.

Remark 37. Although the iterative CVB form (32) is novel, the definition of ELBO via KL divergence in (33) was recently proposed in [32]. Nevertheless, the value $\log \zeta_k^{[\nu]}(x)$ of ELBO in (33) was not given therein. Also, the so-called copula variational inference in [32] was to locally minimize ELBO (33) via a sampling-based stochastic-gradient decent for copula's parameters, rather than via a deterministic expectation operator

in (32). No explicit CVB's marginal form at convergence was given in [32].

1) Conditionally Exponential Family (CEF) for posterior distribution: Similar to (28), the computation of CVB algorithm (32) will be linearly tractable if the true posterior $f(\theta|x)$ belongs to CEF (26), as follows:

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{Z} \exp \left\langle \boldsymbol{g}_k(\theta_k, \boldsymbol{x}), \boldsymbol{g}_{\setminus k}(\theta_{\setminus k}, \boldsymbol{x}) \right\rangle.$$
 (35)

Since Z is merely a normalizing constant in (35), we can also replace $f(x, \theta)$ in CVB algorithm (32) by its unnormalized form $q(x, \theta) \triangleq \exp \left\langle g_k, g_{\setminus k} \right\rangle$ in (35). Since the parameters θ_k and $\theta_{\setminus k}$ in (35) are separable, the CVB form (28) is tractable and conjugate to the original distribution (35). For this reason, the CEF form (35) was also called the conditionally conjugate model for exponential family [46], the conjugate-exponential class [3] or the separable-in-parameter family [2] in mean-field context.

2) Mean-field approximations for posterior distribution: Similar to CVB (32), the mean-field algorithms in section IV-C can be applied to the posterior $f(\theta|x)$, except that the original joint distribution $f(\theta)$ in those mean-field algorithms is now replaced by the joint model $f(x, \theta)$. By this way, the EM and ICM algorithms are also able to return a local maximum-aposteriori (MAP) estimate of the true marginal $f(\theta_k|x)$ and the true joint $f(\theta|x)$, respectively, either directly from joint model $f(x, \theta)$ or indirectly from its unnormalized form $g(x, \theta)$.

In literature, there are three main approaches for proof of VB approximation (29) when applied to the joint model $f(x, \theta)$ in (32), as briefly summarized below. All VB's proofs were, however, confined within independent space $\widetilde{f} = \widetilde{f}_{\setminus k} \widetilde{f}_k$ and, hence, did not yield the CVB form (32):

- The first approach (e.g. in [2], [47], [48]) is to expand $KL_{\widetilde{f}||f}$ directly, i.e. similar to CVA's proof (24).
- The second approach (e.g. in [11], [49]) is to start with Jensen's inequality for the so-called energy [12], [50]: $\log Z(x) = \log \mathbb{E}_{\widetilde{f}(\theta)} \frac{q(x,\theta)}{\widetilde{f}(\theta)} \geq \mathbb{E}_{\widetilde{f}(\theta)} \log \frac{q(x,\theta)}{\widetilde{f}(\theta)}$, which is equivalent to the ELBO's inequality in (33), since the term $Z(x) \triangleq \int_{\theta} q(x,\theta) d\theta$ is proportional to f(x), i.e. $f(x) = \int_{\theta} f(x,\theta) d\theta = \frac{1}{Z} \int_{\theta} q(x,\theta) d\theta = \frac{Z(x)}{Z}$, owing to (35). Note that, the Jensen's inequality is merely a consequence of Bregman variance theorem, of which KL divergence is a special case, as shown in Theorem 5.
- The third approach (e.g. in [3], [4]) is to derive the functional derivative of KL_{f||f|} via Lagrange multiplier in calculus of variations (hence the name "variational" in VB). In this paper, however, the Bregman pythagorean projection for functional space (3, 10) was applied instead and it gave a simpler proof for CVA (22) and VB (29), since the gradient form of Bregman divergence in (2) is more concise than traditional functional derivative.

In practice, since the evidence f(x) is hard to compute, the ELBO term in (33) was originally defined as a feasible stopping rule for iterative VB algorithm [46]. The ELBO for CVB in (33), computed via conditional form $\widetilde{f}_{\langle k|k}^{[\nu-1]}$ in (32), can also be used as a stopping rule for CVB algorithm.

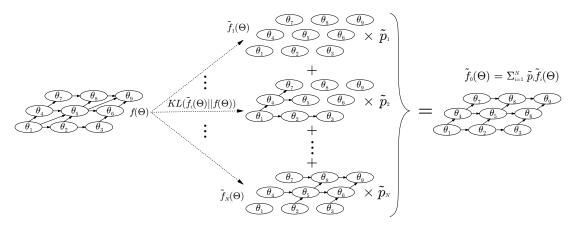


Figure 10. Augmented CVB approximation $\widetilde{f}_0(\theta)$ for a complicated joint distribution $f(\theta)$, illustrated via directed acyclic graphs (DAG). Each $\widetilde{f}_i(\theta)$ is a converged CVB approximation of $f(\theta)$ with simpler structure. The weight vector $\widetilde{p} \triangleq [\widetilde{p}_1, \widetilde{p}_2, \dots, \widetilde{p}_N]^T$, with $\sum_{i=1}^N \widetilde{p}_i = 1$, is then calculated via (38) and yields the optimal mixture $\widetilde{f}_0(\theta) \triangleq \sum_{i=1}^N \widetilde{p}_i \widetilde{f}_i(\theta)$ minimizing the upper bounds (39-40) of $\mathrm{KL}_{\widetilde{f}_0||f}$. Since $\mathrm{KL}_{\widetilde{f}_0||f}$ is convex over \widetilde{f}_0 , the mixture \widetilde{f}_0 would be close to the original f, if we can design a set of \widetilde{f}_i such that f stays inside a polytope bounded by vertices \widetilde{f}_i , as illustrated in Fig. 4. Hence, a good choice of \widetilde{f}_i might be a set of overlapped sectors of the original network f, such that its mixture would have a similar structure of f, as illustrated in above DAGs.

V. HIERARCHICAL CVB FOR BAYESIAN NETWORK

In this section, let us apply the CVB approximation to a joint posterior $f(\theta|x)$ of a generic Bayesian network. Since the network structure of $f(\theta|x)$ is often complicated in practice, an intuitive approach is to approximate $f(\theta|x)$ with a simpler CEF structure $\tilde{f}(\theta|x)$, such that the $\mathrm{KL}_{\tilde{f}||f}$ can be locally minimized via iterative CVB algorithm.

Nevertheless, since CVB approximation $\widetilde{f}^{[\nu]}(\boldsymbol{\theta}|\boldsymbol{x})$ in (32) cannot change its copula form at any iteration ν , a natural approach is to design initially a set of simple network structures $\widetilde{f}_i^{[0]}$, $i \in \{1,2,\ldots,N\}$, and then combine them into a more complex structure with lowest $\mathrm{KL}_{\widetilde{f}^{[\nu_c]}||f}$, or equivalently, highest ELBO (33) at convergence $\nu = \nu_c$. An augmented hierarchy method for merging potential CVB's structures, as illustrated in Fig. 10, will be studied below.

For simplicity, let us consider the case of joint distribution $f(\theta)$ first, before applying the augmented approach to joint posterior $f(\theta|x)$.

A. Augmented CVB for mixture model

Let us firstly consider a mixture model, which is the simplest structure of a hierarchical network. The traditional mixture $f(\boldsymbol{\theta}|\boldsymbol{p}) = \sum_{i=1}^N p_i f_i(\boldsymbol{\theta}) = \sum_{\boldsymbol{l}} f(\boldsymbol{\theta}, \boldsymbol{l}|\boldsymbol{p})$ and its approximation $\widetilde{f}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}) = \sum_{i=1}^N \widetilde{p}_i \widetilde{f}_i(\boldsymbol{\theta}) = \sum_{\boldsymbol{l}} \widetilde{f}(\boldsymbol{\theta}, \boldsymbol{l}|\widetilde{\boldsymbol{p}})$ can be written in augmented form via a boolean label vector $\boldsymbol{l} \triangleq [l_1, l_2, \dots, l_N]^T \in \mathbb{I}^N$, as follows:

$$f(\boldsymbol{\theta}, \boldsymbol{l}|\boldsymbol{p}) = f(\boldsymbol{\theta}|\boldsymbol{l})f(\boldsymbol{l}|\boldsymbol{p}) = \prod_{i=1}^{N} f_{i}^{l_{i}}(\boldsymbol{\theta})p_{i}^{l_{i}}, \qquad (36)$$
$$\widetilde{f}(\boldsymbol{\theta}, \boldsymbol{l}|\widetilde{\boldsymbol{p}}) = \widetilde{f}(\boldsymbol{\theta}|\boldsymbol{l})\widetilde{f}(\boldsymbol{l}|\widetilde{\boldsymbol{p}}) = \prod_{i=1}^{N} \widetilde{f}_{i}^{l_{i}}(\boldsymbol{\theta})\widetilde{p}_{i}^{l_{i}},$$

where $\mathbf{l} \in \{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_N\}$ and $\boldsymbol{\epsilon}_i \triangleq [0, \dots, 1, \dots 0]^T$ is a $N \times 1$ element vector with all zero elements except the unit value at i-th position, $\forall i \in \{1, 2, \dots, N\}$. Each \widetilde{f}_i is then

assumed to be the converged CVB approximation of each original component f_i .

Ideally, our aim is to pick the weight vector $\widetilde{\boldsymbol{p}} \triangleq [\widetilde{p}_1, \widetilde{p}_2, \dots, \widetilde{p}_N]^T$ such that $\mathrm{KL}(\widetilde{f}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}})||f(\boldsymbol{\theta}|\boldsymbol{p}))$ is minimized. Nevertheless, it is not feasible to directly factorize the mixture form $f(\boldsymbol{\theta}|\boldsymbol{p})$ and $\widetilde{f}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}})$ via non-linear form of KL divergence. Instead, let us minimize the KL divergence of their augmented forms in (36), as follows:

$$\widetilde{p}^* \triangleq \underset{\widetilde{p}}{\operatorname{arg\,min}} \operatorname{KL}(\widetilde{f}(\boldsymbol{\theta}, \boldsymbol{l}|\widetilde{p})||f(\boldsymbol{\theta}, \boldsymbol{l}|\boldsymbol{p})),$$
 (37)

which is also an upper bound of $KL(f(\theta|\tilde{p})||f(\theta|p))$, as shown in (21). The solution for (37) can be found via CVA (23), as follows:

Corollary 38. (CVA for mixture model)

Applying CVA (23) to (37), we can compute the optimal weight $\widetilde{\boldsymbol{p}}^* \triangleq [\widetilde{p}^*_1, \widetilde{p}^*_2, \dots, \widetilde{p}^*_N]^T$ minimizing (37), as follows:

$$\widetilde{p}_i^* \propto \frac{p_i}{\exp(\mathit{KL}_{\widetilde{f}_i||f_i})}, \ \forall i \in \{1, 2, \dots, N\}.$$
 (38)

From (24), the minimum value of (37) is then:

$$KL_{\widetilde{\boldsymbol{p}}^*} \triangleq \sum_{i=1}^{N} \widetilde{p}_i^* KL_{\widetilde{f}_i||f_i} + \sum_{i=1}^{N} \widetilde{p}_i^* \log \frac{p_i}{\widetilde{p}_i^*}$$
(39)

Proof: From CVA (23), the marginal $\widetilde{f}(\boldsymbol{l}|\widetilde{\boldsymbol{p}})$ minimizing (37) is $\widetilde{f}(\boldsymbol{l}|\widetilde{\boldsymbol{p}}) \propto f(\boldsymbol{l}|\boldsymbol{p})/\exp(\mathrm{KL}(f(\boldsymbol{\theta}|\boldsymbol{l})||f(\boldsymbol{\theta}|\boldsymbol{l})),$ which yields (38), since $\mathrm{KL}(\widetilde{f}(\boldsymbol{\theta}|\boldsymbol{l})||f(\boldsymbol{\theta}|\boldsymbol{l})) = \sum_{i=1}^{N} l_i \mathrm{KL}(\widetilde{f}_i(\boldsymbol{\theta})||f_i(\boldsymbol{\theta})).$

B. Augmented CVB for Bayesian network

Let us now apply the above approach to a generic network $f(\theta)$. In (36), let us set $f_i(\theta) = f(\theta)$, $\forall i$, together with uniform weight $\mathbf{p} = \bar{\mathbf{p}} \triangleq [\bar{p}_1, \bar{p}_2, \dots, \bar{p}_N]^T = [\frac{1}{N}, \dots, \frac{1}{N}]^T$. Each \tilde{f}_i in (38) is now a CVB approximation, with possibly simpler structures, of the same original network $f(\theta)$, as illustrated in Fig. 10.

Owing to Bregman's property 4 in Proposition 2, $\mathrm{KL}_{\widetilde{f}||f}$ is convex over \widetilde{f} . Hence, there exists a linear mixture $\widetilde{f}_0(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}) = \sum_{i=1}^N \widetilde{p}_i \widetilde{f}_i(\boldsymbol{\theta})$, such that:

$$\mathrm{KL}_{\widetilde{f}_{0}||f} \leq \mathrm{KL}_{\epsilon_{i^{*}}} \triangleq \min_{i \in \{1, 2, \dots, N\}} \mathrm{KL}_{\widetilde{f}_{i}||f}$$
 (40)

in which the equality is reached if we set $\tilde{p} = \epsilon_{i^*}$, with $i^* \triangleq \arg\min_i \mathrm{KL}_{\widetilde{f}_i || f}$.

Since minimizing $KL_{\widetilde{f}_0||f}$ directly is not feasible, as explained above, we can firstly minimize $KL_{\widetilde{f}_i||f}$ in (40) via iterative CVB algorithm for each approximated structure \widetilde{f}_i . We then compute the optimal weights \widetilde{p}^* in (37, 38) for the minimum upper bound $KL_{\widetilde{p}^*}$ of $KL_{\widetilde{f}_0||f}$. Note that $KL_{\widetilde{p}^*}$ in (39) and $KL_{\epsilon_{i^*}}$ in (40) are two different upper bounds of $KL_{\widetilde{f}_0||f}$ and may not yield the global minimum solution for $KL_{\widetilde{f}_0||f}$ in general. The choice $\widetilde{p}=\epsilon_{i^*}$ might yield lower $KL_{\widetilde{f}_0||f}$ than $\widetilde{p}=\widetilde{p}^*$, even when we have $KL_{\epsilon_{i^*}}>KL_{\widetilde{p}^*}$.

Although we can only find the minimum upper bound solution for the mixture \widetilde{f}_0 in this paper, the key advantage of the mixture form is that the moments of \widetilde{f}_0 are simply a mixture of moments of \widetilde{f}_i , i.e.:

$$\widehat{\boldsymbol{\theta}}_0 = \mathbb{E}_{\widetilde{f}_0}(\boldsymbol{\theta}) = \sum_{i=1}^N \widetilde{p}_i \mathbb{E}_{\widetilde{f}_i}(\boldsymbol{\theta}) = \sum_{i=1}^N \widetilde{p}_i \widehat{\boldsymbol{\theta}}_i. \tag{41}$$

By this way, the true moments $\widehat{\boldsymbol{\theta}}$ of complicated network $f(\boldsymbol{\theta})$ can be approximated by a mixture of moments $\widehat{\boldsymbol{\theta}}_i$ of simpler CVB's network structure $\widetilde{f}_i(\boldsymbol{\theta})$.

Another advantage of mixture form is that the optimal weight vector $\tilde{\boldsymbol{p}}$ can be evaluated tractably, without the need of normalizing constant of $f(\boldsymbol{\theta}|\boldsymbol{x})$ in Bayesian context. Indeed, for a posterior Bayesian network $f(\boldsymbol{\theta}|\boldsymbol{x})$, we can simply replace the value $\mathrm{KL}_{\tilde{f}_i||f}$ in (38-40) by ELBO's value in (33), since the evidence $f(\boldsymbol{x})$ is a constant.

C. Hierarchical CVB approximation

In principle, if we keep augmenting the above CVB's augmented mixture, it is possible to establish an m-order hierarchical CVB approximation $\widetilde{f}^{\{m\}}(\boldsymbol{\theta})$ for a complicated network $f(\boldsymbol{\theta}), \ \forall m \in \{0,1,\ldots,M\}$. For example, each zero-order mixture $\widetilde{f}_i^{\{0\}}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}_i^*) = \sum_{m=1}^M \widetilde{p}_{i,m}^* \widetilde{f}_{i,m}(\boldsymbol{\theta}) = \sum_{l_i} \widetilde{f}(\boldsymbol{\theta}, \boldsymbol{l}_i|\widetilde{\boldsymbol{p}}_i^*), \ \forall i \in \{1,2,\ldots,N\}, \ \text{can be considered as a component of the first-order mixture } \widetilde{f}_0^{\{1\}}(\boldsymbol{\theta}|\widetilde{\boldsymbol{q}},\widetilde{\boldsymbol{P}}^*) = \sum_{i=1}^N \widetilde{q}_i \widetilde{f}_i^{\{0\}}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}_i^*), \ \text{where } \widetilde{\boldsymbol{P}}^* \triangleq [\widetilde{\boldsymbol{p}}^*_i,\widetilde{\boldsymbol{p}}^*_2,\ldots,\widetilde{\boldsymbol{p}}_N^*] \ \text{and } \widetilde{\boldsymbol{q}} \triangleq [\widetilde{q}_1,\widetilde{q}_2,\ldots,\widetilde{q}_N]^T.$

If $\widetilde{f}_{i,m}(\theta)$ are all tractable CVB's approximations with simpler and possibly overlapped sectors of the network $f(\theta)$, the optimal vectors \widetilde{p}_i^* can be evaluated feasibly via $\mathrm{KL}_{\widetilde{f}_{i,m}||f}$ in (38). Nonetheless, the computation of the optimal vector \widetilde{q}^* via $\mathrm{KL}_{\widetilde{f}_i^{\{0\}}||f}$ in (38) might be intractable in practice, because $\mathrm{KL}_{\widetilde{f}_i^{\{0\}}||f}$ is a KL divergence of a mixture of distributions and, hence, it is difficult to evaluate $\mathrm{KL}_{\widetilde{f}_i^{\{0\}}||f}$ directly in closed form.

An intuitive solution for this issue might be to apply CVB again to the augmented form $\mathrm{KL}(\widetilde{f}(\boldsymbol{\theta},\boldsymbol{l}_i|\widetilde{\boldsymbol{p}}_i)||f(\boldsymbol{\theta},\boldsymbol{l}_i|\overline{\boldsymbol{p}}))$, similar to (37). By this way, we could avoid the mixture form $\widetilde{f}_i^{\{0\}}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}_i) = \sum_{\boldsymbol{l}_i} \widetilde{f}(\boldsymbol{\theta},\boldsymbol{l}_i|\widetilde{\boldsymbol{p}}_i^*)$ and directly derive a CVB's

closed form for $\widetilde{f}_i^{\{0\}}(\boldsymbol{\theta}|\widetilde{\boldsymbol{p}}_i)$. This hierarchical CVB approach is, however, outside the scope of this paper and will be left for future work.

Remark 39. In literature, the idea of augmented hierarchy was mentioned briefly in [51], [52], in which the potential approximations \tilde{f}_i are confined to a set of mean-field approximations and the prior $\tilde{f}(l|\tilde{p})$ is extended from a mixture to a latent Markovian model. Nevertheless, the ELBO minimization in [51], [52] was implemented via stochastic-gradient decent methods and did not yield an explicit form for the mixture's weights in (38).

VI. CASE STUDY

In this section, let us illustrate the superior performance of CVB to mean-field approximations for two canonical scenarios in practice: the bivariate Gaussian distribution and Gaussian mixture clustering. These two cases belong to CEF class (26) and, hence, their CVB approximation is tractable, as shown below.

A. Bivariate Gaussian distribution

In this subsection, let us approximate a bivariate Gaussian distribution $f(\theta) = \mathcal{N}_{\theta}(0, \Sigma)$ with zero mean and covariance matrix $\Sigma \triangleq \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}$. The purpose is then to illustrate the performance of CVB and VB approximations for $f(\theta)$ with different values of correlation coefficient $\rho \in [-1, 1]$.

For simple notation, let us denote the marginal and conditional distributions of $f(\boldsymbol{\theta})$ by $f_1 = \mathcal{N}_{\theta_1}(0, \sigma_1)$ and $f_{2|1} = \mathcal{N}_{\theta_2}(\beta_{2|1}\theta_1, \sigma_{2|1})$, respectively, in which $\beta_{2|1} \triangleq \rho \frac{\sigma_2}{\sigma_1}$ and $\sigma_{2|1} \triangleq \sigma_2 \sqrt{1 - \rho^2}$.

1) CVB approximation: Since Gaussian distribution belongs to CEF class (26), the CVB form $\widetilde{f}_{\text{CVB}}^{[1]} = \widetilde{f}_{2|1}^{[0]} \widetilde{f}_{1}^{[1]} = \widetilde{f}_{1|2}^{[1]} \widetilde{f}_{2}^{[1]}$ in (25) is also Gaussian, as shown in (28). Then, given initial values $\widetilde{\beta}_{2|1}^{[0]} \triangleq \widetilde{\rho}_{[0]} \widetilde{\overline{\sigma}_{2}^{[0]}}$ and $\widetilde{\sigma}_{2|1}^{[0]} \triangleq \widetilde{\sigma}_{2}^{[0]} \sqrt{1 - \widetilde{\rho}_{[0]}^{2}}$, we have $\widetilde{f}_{2|1}^{[0]} = \mathcal{N}_{\theta_{2}}(\widetilde{\beta}_{2|1}^{[0]}\theta_{1}, \widetilde{\sigma}_{2|1}^{[0]})$. At iteration $\nu = 1$, the CVA form (23) yields:

$$\begin{split} \widetilde{f}_{1}^{[1]} &= \frac{1}{\zeta_{1}^{[1]}} \frac{f_{1}}{\exp(\text{KL}_{\widetilde{f}_{2|1}^{[0]}||f_{2|1}})} \\ &= \frac{1}{\zeta_{1}^{[1]}} \frac{\frac{1}{\sigma_{1}\sqrt{2\pi}} \exp{-\frac{\theta_{1}^{2}}{2\sigma_{1}^{2}}}}{\frac{\sigma_{2|1}}{\overline{\sigma}_{2|1}^{[0]}} \exp{\frac{1}{2}} \left[\frac{\left(\widetilde{\beta}_{2|1}^{[0]} - \beta_{2|1}\right)^{2} \theta_{1}^{2} + (\widetilde{\sigma}_{2|1}^{[0]})^{2}}{\sigma_{2|1}^{2}} - 1 \right] \\ &= \mathcal{N}_{\theta_{1}}(0, \widetilde{\sigma}_{1}^{[1]}), \end{split}$$

in which $\mathrm{KL}_{\widetilde{f}_{2|1}^{[0]}||f_{2|1}}$ is KL divergence between Gaussian distributions and:

$$\widetilde{\sigma}_{1}^{[1]} = \frac{1}{\sqrt{\frac{1}{\sigma_{1}^{2}} + \frac{\left(\widetilde{\beta}_{2|1}^{[0]} - \beta_{2|1}\right)^{2}}{\sigma_{2}^{2}(1 - \rho^{2})}}}, \ \zeta_{1}^{[1]} = \frac{\widetilde{\sigma}_{1}^{[1]}}{\sigma_{1}} \frac{\widetilde{\sigma}_{2|1}^{[0]}}{\sigma_{2|1}} \exp \frac{\sigma_{2|1}^{2} - (\widetilde{\sigma}_{2|1}^{[0]})^{2}}{2\sigma_{2|1}^{2}}.$$

$$(42)$$

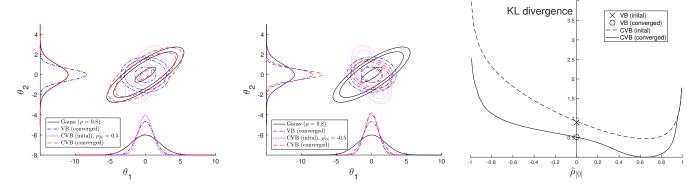


Figure 11. CVB and VB approximations \tilde{f}_{θ} for a zero-mean bivariate Gaussian distribution f_{θ} , with true variances $\sigma_1^2=4$, $\sigma_2^2=1$ and correlation coefficient $\rho=0.8$. The initial guess values for CVB and VB are $\tilde{\sigma}_1^{[0]}=\tilde{\sigma}_2^{[0]}=1$, together with various $\tilde{\rho}_{[0]}\in(-1,1)$ for CVB. The cases $\tilde{\rho}_{[0]}=0.5$ and $\tilde{\rho}_{[0]}=-0.5$ are shown on the left and middle panel, respectively. The marginal distributions, which are also Gaussian, are plotted on two axes in these two panels. The lower KL divergence $\mathrm{KL}(\tilde{f}_{\theta}||f_{\theta})$ on the right panel, the better approximation, as illustrated in Fig. 7, 9. The CVB will be exact, i.e. $\mathrm{KL}(\tilde{f}_{\theta}||f_{\theta})\approx0$ at convergence, if the initial guess values $\tilde{\rho}_{[0]}$ are in range $\tilde{\rho}_{[0]}\in[0.6,0.7]$, which is close to the true value $\rho=0.8$. If $\tilde{\rho}_{[0]}=0$, the CVB is equivalent to VB approximation in independent class. The number ν_c of iterations until convergence for VB and CVB are, respectively, 8 and 11.1 ± 5.2 , averaged over all cases of $\tilde{\rho}_{[0]}\in(-1,1)$ for CVB. Only one marginal is updated per iteration.

Then, in order to derive the reverse form $\widetilde{f}_{1|2}^{[1]}\widetilde{f}_{2}^{[1]}=\widetilde{f}_{2|1}^{[0]}\widetilde{f}_{1}^{[1]}$, let us firstly note that $\widetilde{\beta}_{2|1}^{[0]}=\widetilde{\beta}_{2|1}^{[1]}$ and $\widetilde{\sigma}_{2|1}^{[0]}=\widetilde{\sigma}_{2|1}^{[1]}$, since the conditional form $\widetilde{f}_{2|1}$ of two distributions $\widetilde{f}_{2|1}^{[0]}\widetilde{f}_{1}^{[0]}$ and $\widetilde{f}_{2|1}^{[0]}\widetilde{f}_{1}^{[1]}=\widetilde{f}_{1|2}^{[1]}\widetilde{f}_{2}^{[1]}$ are still the same. Then, the updated parameters are:

$$\begin{cases} \tilde{\beta}_{2|1}^{[0]} &= \tilde{\beta}_{2|1}^{[1]} \\ \tilde{\sigma}_{2|1}^{[0]} &= \tilde{\sigma}_{2|1}^{[1]} \end{cases} \Leftrightarrow \begin{cases} \tilde{\rho}_{[0]} \frac{\tilde{\sigma}_{2}^{[0]}}{\tilde{\sigma}_{1}^{[0]}} &= \tilde{\rho}_{[1]} \frac{\tilde{\sigma}_{2}^{[1]}}{\tilde{\sigma}_{1}^{[1]}} \\ \tilde{\sigma}_{2}^{[0]} \sqrt{1 - \tilde{\rho}_{[0]}^{2}}) &= \tilde{\sigma}_{2}^{[1]} \sqrt{1 - \tilde{\rho}_{[1]}^{2}} \end{cases}$$

which, by solving for $\widetilde{\rho}_{[1]}$ and $\widetilde{\sigma}_2^{[1]},$ yields:

$$\begin{split} \widetilde{\rho}_{[1]}^2 &= \frac{\widetilde{\rho}_{[0]}^2}{\widetilde{\rho}_{[0]}^2 + \left(\frac{\widetilde{\sigma}_{1}^{[0]}}{\widetilde{\sigma}_{1}^{[1]}}\right)^2 (1 - \widetilde{\rho}_{[0]}^2)}, \\ \widetilde{\sigma}_{2}^{[1]} &= \widetilde{\sigma}_{2}^{[0]} \sqrt{\widetilde{\rho}_{[0]}^2 \left(\frac{\widetilde{\sigma}_{1}^{[1]}}{\widetilde{\sigma}_{1}^{[0]}}\right)^2 + (1 - \widetilde{\rho}_{[0]}^2))}. \end{split}$$

Hence, we have $\tilde{\beta}_{1|2}^{[1]} = \widetilde{\rho}_{[1]} \frac{\widetilde{\sigma}_{1}^{[1]}}{\widetilde{\sigma}_{2}^{[1]}}$ and $\widetilde{\sigma}_{1|2}^{[1]} = \widetilde{\sigma}_{1}^{[1]} \sqrt{1 - \widetilde{\rho}_{[1]}^{2}}$, which yield the updated forms $\widetilde{f}_{2}^{[1]} = \mathcal{N}_{\theta_{1}}(0, \widetilde{\sigma}_{2}^{[1]})$ and $\widetilde{f}_{1|2}^{[1]} = \mathcal{N}_{\theta_{1}}(\widetilde{\beta}_{1|2}^{[1]}\theta_{2}, \widetilde{\sigma}_{1|2}^{[1]})$. Reversing the role of θ_{1} with θ_{2} and repeating the above steps for iteration $\nu > 1$, we will achieve the CVB approximation at convergence $\nu = \nu_{c}$, with $\mathrm{KL}_{\widetilde{f}_{\mathrm{CVB}}^{[\nu]}||f} = \log \frac{1}{\zeta_{1}^{[\nu]}}$.

The CVB approximation will be exact if its conditional mean and variance are exact, i.e. $\tilde{\beta}_{2|1}^{[\nu_c]} = \beta_{2|1}$ and $\tilde{\sigma}_{2|1}^{[\nu_c]} = \sigma_{2|1}$, since we have $\mathrm{KL}_{\tilde{f}_{\mathrm{CVB}}^{[\nu_c]}||f} = \log \frac{1}{\zeta_1^{[\nu_c]}} = 0$ in this case, as shown in (42).

- 2) VB approximation: Since VB is a special case of CVB in independence space, we can simply set $\rho=0$ in above CVB algorithm and the result will be VB approximation.
- 3) Simulation's results: The CVB and VB approximations for the case of $f(\theta) = \mathcal{N}_{\theta}(0, \Sigma)$ are illustrated in Fig. 11. Since $\mathrm{KL}_{\widetilde{f}_{\theta}^{[\nu]}||f_{\theta}}$ monotonically decreases with iteration ν , the

right panel shows the value of KL divergence at initialization $\nu=0$ and at convergence $\nu=\nu_c$, with $0 \leq \mathrm{KL}(\widetilde{f}_{\theta}^{[\nu_c-1]}||f_{\theta}) - \mathrm{KL}(\widetilde{f}_{\theta}^{[\nu_c]}||f_{\theta}) \leq 0.01$. We can see that VB is a mean-field approximation and, hence, cannot accurately approximate a correlated Gaussian distribution. In contrast, the CVB belongs to a conditional copula class and, hence, can yield higher accuracy. In this sense, CVB can potentially return a globally optimal approximation for a correlated distribution, while VB can only return a locally optimal approximation.

Nevertheless, since the iterative CVB cannot escape its initialized copula class, its accuracy depends heavily on initialization. A solution for this issue is to initialize CVB with some information of original distribution. For example, merely setting the initial sign of $\widetilde{\rho}_{[0]}$ equal to the sign of true value ρ would gain tremendously higher accuracy for CVB at convergence, as shown in the left and middle panel of Fig. 11.

Another solution for CVB's initialization issue is to generate a lot of potential structures initially and take the average of the results at convergence. This CVB's mixture-scheme will be illustrated in the next subsection.

B. Gaussian mixture clustering

In this subsection, let us illustrate the performance of CVB for a simple bivariate Gaussian mixture model. For this purpose, let us consider clusters of bivariate observation data $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{2 \times N}$, such that $\boldsymbol{x}_i = [x_{1,i}, x_{2,i}]^T \in \mathbb{R}^2$ at each time $i \in \{1, 2, \dots, N\}$ randomly belongs to one of K bivariate independent Gaussian clusters $\mathcal{N}_{\boldsymbol{x}_i}(\boldsymbol{\mu}, \mathbf{I}_2)$ with equal probability $\boldsymbol{p} \triangleq [p_1, p_2, \dots, p_K]^T$, i.e. $p_k = \frac{1}{K}, \ \forall k \in \{1, 2, \dots, K\}$, at unknown means $\boldsymbol{\Upsilon} \triangleq [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K] \in \mathbb{R}^{2 \times K}$. \mathbf{I}_2 denotes the 2×2 identity covariance matrix.

Let us also define a temporal matrix $\boldsymbol{L} \triangleq [\boldsymbol{l}_1, \boldsymbol{l}_2, \dots, \boldsymbol{l}_N] \in \mathbb{I}^{K \times N}$ of categorical vector labels $\boldsymbol{l}_i = [l_{1,i}, l_{2,i}, \dots, l_{K,i}]^T \in \{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_K\}$, where $\boldsymbol{\epsilon}_k = [0, \dots, 1, \dots 0]^T \in \mathbb{I}^K$ denotes the boolean vector with k-th non-zero element. By this way, we set $\boldsymbol{l}_i = \boldsymbol{\epsilon}_k$ if \boldsymbol{x}_i belongs to k-th cluster.

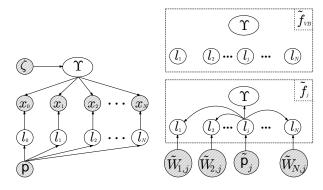


Figure 12. Directed acyclic graphs (DAG) for Gaussian clustering model with uniform hyper-parameters ζ , \boldsymbol{p} (left), the VB approximation with independent structure (upper right) and the CVB approximations (lower right). All variables in shaded nodes are known, while the others are random variables. Each \widetilde{f}_j is a ternary structure centered around l_j , $j \in \{1,2,\ldots,N\}$. The augmented CVB approximation $\widetilde{f}_0 = \sum_{j=1}^N q_j^* \widetilde{f}_j$ is designed in (70) and illustrated in Fig. 10.

Then, by probability chain rule, our model is a Gaussian mixture $f(X, \Theta) = f(X|\Theta)f(\Theta)$, in which $\Theta \triangleq [\Upsilon, L]$ are unknown parameters, as follows:

$$f(X, \Upsilon, L) = f(X|\Upsilon, L)f(\Upsilon, L)$$

$$= f(\Upsilon|L, X)f(L, X)$$

$$= f(L|\Upsilon, X)f(\Upsilon, X).$$
(43)

In the first line of (43), the distributions are:

$$f(\boldsymbol{X}|\boldsymbol{\Upsilon}, \boldsymbol{L}) = \prod_{i=1}^{N} f(\boldsymbol{x}_{i}|\boldsymbol{\Upsilon}, \boldsymbol{l}_{i}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{x}_{i}}^{l_{k,i}}(\boldsymbol{\mu}_{k}, \boldsymbol{I}_{2}),$$
$$f(\boldsymbol{\Upsilon}, \boldsymbol{L}) = f(\boldsymbol{\Upsilon})f(\boldsymbol{L}) = \frac{1}{\zeta K^{N}}, \tag{44}$$

in which the prior $f(\boldsymbol{L}) = \prod_{i=1}^N \boldsymbol{l}_i^T \boldsymbol{p} = \frac{1}{K^N}$ is uniform by default and $f(\boldsymbol{\Upsilon})$ is the non-informative prior over $\mathbb{R}^{2\times K}$, i.e. $f(\boldsymbol{\Upsilon}) = \frac{1}{\zeta}$, with constant ζ being set as high as possible (ideally $\zeta \to \infty$).

The second line of (43) can be written as follows:

$$f(\mathbf{\Upsilon}|\mathbf{L}, \mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{\mu}_{k}} \left(\overline{\boldsymbol{\mu}}_{k}(\mathbf{L}), \overline{\sigma}_{k}(\mathbf{L}) \mathbf{I}_{2} \right), \qquad (45)$$

$$f(\mathbf{L}, \mathbf{X}) = \frac{1}{\zeta K^{N}} \prod_{k=1}^{K} \gamma_{k}(\mathbf{L}),$$

with $\overline{\mu}_k(L)$ and $\overline{\sigma}_k(L)$ denoting posterior mean and standard deviation of μ_k , respectively, and $\gamma_k(L)$ denoting the updated form of weight's probability $p_k = \frac{1}{K}$, as follows:

$$\overline{\boldsymbol{\mu}}_{k}(\boldsymbol{L}) \triangleq \frac{\sum_{i=1}^{N} l_{k,i} \boldsymbol{x}_{i}}{\sum_{i=1}^{N} l_{k,i}}, \ \overline{\sigma}_{k}(\boldsymbol{L}) \triangleq \frac{1}{\sqrt{\sum_{i=1}^{N} l_{k,i}}},$$

$$\gamma_{k}(\boldsymbol{L}) \triangleq 2\pi \overline{\sigma}_{k}^{2}(\boldsymbol{L}) \prod_{i=1}^{N} \mathcal{N}_{\boldsymbol{x}_{i}}^{l_{k,i}}(\overline{\boldsymbol{\mu}}_{k}(\boldsymbol{L}), \mathbf{I}_{2}),$$

$$(46)$$

Note that, the first form (44) is equivalent to the second form

(45-46) since we have:

$$\frac{\sum_{i=1}^{N} l_{k,i} ||\boldsymbol{x}_{i} - \boldsymbol{\mu}_{k}||^{2}}{\sum_{i=1}^{N} l_{k,i}} = \frac{\sum_{i=1}^{N} l_{k,i} ||\boldsymbol{x}_{i} - \overline{\boldsymbol{\mu}}_{k}(\boldsymbol{L})||^{2}}{\sum_{i=1}^{N} l_{k,i}} + ||\boldsymbol{\mu}_{k} - \overline{\boldsymbol{\mu}}_{k}(\boldsymbol{L})||^{2}, \tag{47}$$

owing to Bregman variance theorem in (6), (13).

Similarly, the third line in (43) can be derived from (44), as follows:

$$f(\boldsymbol{X}, \boldsymbol{\Upsilon}) = \sum_{\boldsymbol{L}} f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L}) = \frac{\prod_{i=1}^{N} \sum_{k=1}^{K} \mathcal{N}_{\boldsymbol{x}_{i}}(\boldsymbol{\mu}_{k}, \boldsymbol{I}_{2})}{\zeta K^{N}},$$

$$f(\boldsymbol{L}|\boldsymbol{\Upsilon}, \boldsymbol{X}) = \frac{f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L})}{f(\boldsymbol{\Upsilon}, \boldsymbol{X})} = \prod_{i=1}^{N} \underbrace{\prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{x}_{i}}^{l_{k,i}}(\boldsymbol{\mu}_{k}, \boldsymbol{I}_{2})}_{f(\boldsymbol{l}_{i}|\boldsymbol{\Upsilon}, \boldsymbol{x}_{i})}.$$

$$(48)$$

Note that, the model without labels $f(X, \Upsilon) = f(X|\Upsilon)f(\Upsilon)$ in (48) is a mixture of K^N Gaussian components with unknown means Υ , since we have augmented the model $f(X|\Upsilon)$ with label's form $f(X|\Upsilon, L)$ above. The posterior form $f(\Upsilon|X) \propto f(X,\Upsilon) = \sum_L f(X,\Upsilon, L)$ in this case is intractable, since its normalization's complexity $\mathcal{O}(K^N)$ grows exponentially with number of data N, hence the curse of dimensionality.

1) ICM and k-means algorithms: From (45-46), we can see that the conditional mean $\overline{\mu}_k(L)$ is actually the k-th clustering sample's mean of μ_k , given all possible boolean values of $l_{k,i} \in \mathbb{I} = \{0,1\}$ over time $i \in \{1,2,\ldots,N\}$. The probability of categorical label $f(L|X) \propto f(L,X)$ in (45) is, in turn, calculated as the distance of all observation x_i to sample's mean $\overline{\mu}_k$ of each cluster $k \in \{1,2,\ldots,K\}$ via $\gamma_k(L)$ in (46). Nevertheless, since the weights $\gamma_k(L)$ in (45-46) are not factorable over L, the posterior probability f(L|X) needs to be computed brute-forcedly over all K^N possible values of label matrix L as a whole and, hence, yields the curse of dimensionality.

A popular solution for this case is the k-means algorithm, which is merely an application of iteratively conditional mode (ICM) algorithm (31) to above clustering mixture (45), (48), as follows:

$$\widehat{\boldsymbol{\Upsilon}}^{[\nu]} = \underset{\boldsymbol{\Upsilon}}{\arg\max} f(\boldsymbol{\Upsilon}|\widehat{\boldsymbol{L}}^{[\nu-1]}, \boldsymbol{X}), \tag{49}$$

$$\widehat{\boldsymbol{L}}^{[\nu]} = \underset{\boldsymbol{L}}{\arg\max} f(\boldsymbol{L}|\widehat{\boldsymbol{\mu}}^{[\nu]}, \boldsymbol{X}).$$

where $\widehat{\mathbf{Y}}^{[\nu]} \triangleq [\widehat{\boldsymbol{\mu}}_1^{[\nu]}, \widehat{\boldsymbol{\mu}}_2^{[\nu]}, \ldots, \widehat{\boldsymbol{\mu}}_K^{[\nu]}]$ and $\widehat{\boldsymbol{L}}^{[\nu]} \triangleq [\widehat{\boldsymbol{l}}_1^{[\nu]}, \widehat{\boldsymbol{l}}_2^{[\nu]}, \ldots, \widehat{\boldsymbol{l}}_K^{[\nu]}]$. Since the mode of Gaussian distribution is also its mean value, let us substitute (49) to $f(\mathbf{\Upsilon}|\boldsymbol{L}, \boldsymbol{X})$ in (45-46) and $f(\boldsymbol{L}|\mathbf{\Upsilon}, \boldsymbol{X})$ in (48), as follows:

$$\widehat{\boldsymbol{\mu}}_{k}^{[\nu]} = \overline{\boldsymbol{\mu}}_{k}(\widehat{\boldsymbol{L}}^{[\nu-1]}) = \frac{\sum_{i=1}^{N} \widehat{l}_{k,i}^{[\nu-1]} \boldsymbol{x}_{i}}{\sum_{i=1}^{N} \widehat{l}_{k,i}^{[\nu-1]}},$$

$$\widehat{k}_{i}^{[\nu]} = \arg\max_{k} \mathcal{N}_{\boldsymbol{x}_{i}}(\overline{\boldsymbol{\mu}}_{k}^{[\nu]}, \mathbf{I}_{2})$$

$$= \arg\min_{k} ||\boldsymbol{x}_{i} - \overline{\boldsymbol{\mu}}_{k}^{[\nu]}||^{2},$$
(50)

in which the form of $\overline{\mu}_k$ is given in (46), $\widehat{l}_i^{[\nu]} \triangleq [\widehat{l}_{1,i}^{[\nu]}, \widehat{l}_{2,i}^{[\nu]}, \dots, \widehat{l}_{K,i}^{[\nu]}]^T$ and $\widehat{l}_{k,i}^{[\nu]} = \delta[k - \widehat{k}_i^{[\nu]}]$, with $\delta[\cdot]$ denoting the Kronecker delta function, $\forall i \in \{1, 2, \dots, N\}$. By convention, we keep $\widehat{\mu}_k^{[\nu]} = \widehat{\mu}_k^{[\nu-1]}$ unchanged if $\sum_{i=1}^N \widehat{l}_{k,i}^{[\nu-1]} = 0$, since no update for k-th cluster is found in this case.

From (50), we can see that the algorithm starts with K initial mean values $\overline{\mu}_k^{[0]}$, $\forall k \in \{1,2,\ldots,K\}$, then assigns categorical labels to clusters via minimum Euclidean distance in (50), which, in turn, yields K new cluster's means $\overline{\mu}_k^{[1]}$, $\forall k \in \{1,2,\ldots,K\}$, and so forth. Hence it is called the k-means algorithm in literature [25], [26].

At convergence $\nu=\nu_c$, the k-means algorithm returns a locally joint MAP value $\widehat{\Theta}^{[\nu_c]}=[\widehat{\Upsilon}^{[\nu_c]},\widehat{L}^{[\nu_c]}]$, which depends on initial guess value $\widehat{\Theta}^{[0]}$.

From Corollary 34, the convergence of ELBO can be used as a stopping rule, as follows:

$$\begin{split} \text{ELBO}_{\text{ICM}}^{[\nu]} &= \log f(\boldsymbol{X}, \widehat{\boldsymbol{\Upsilon}}^{[\nu]}, \widehat{\boldsymbol{L}}^{[\nu]}) \\ &= \log \frac{\prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{x}_{i}}^{\widehat{l}_{i}^{[\nu]}}(\widehat{\boldsymbol{\mu}}_{k}^{[\nu]}, \mathbf{I}_{2})}{CK^{N}}, \end{split}$$

since $\mathrm{KL}_{\widetilde{f}_{\mathrm{ICM}}^{[\nu]}||f} = -\mathrm{ELBO}_{\mathrm{ICM}}^{[\nu]} + \log f(\boldsymbol{X}),$ as shown in (34).

2) EM algorithms: Let us now derive two EM approximations for true posterior distribution $f(\Upsilon, L|X)$ via (30), as follows:

$$\widetilde{f}_{\text{EM}_1}(\Upsilon, L|X) = f(\Upsilon|\widehat{L}_{\text{EM}_1}, X)\delta[L - \widehat{L}_{\text{EM}_1}],
\widetilde{f}_{\text{EM}_2}(\Upsilon, L|X) = f(L|\widehat{\Upsilon}_{\text{EM}_2}, X)\delta(\Upsilon - \widehat{\Upsilon}_{\text{EM}_2}).$$
(51)

Since our joint model $f(\Upsilon, L, X)$ in (43-48) is of CEF form (26), the EM forms (51) can be feasibly identified via (28), as follows:

$$\widehat{\boldsymbol{L}}_{\mathrm{EM}_{1}}^{[\nu]} = \arg\max_{\boldsymbol{L}} \mathbb{E}_{f(\boldsymbol{\Upsilon}|\widehat{\boldsymbol{L}}_{\mathrm{EM}_{1}}^{[\nu-1]}, \boldsymbol{X})} \log f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L}), \quad (52)$$

$$\widehat{\boldsymbol{\Upsilon}}_{\mathrm{EM}_{2}}^{[\nu]} = \arg\max_{\boldsymbol{\Upsilon}} \mathbb{E}_{f(\boldsymbol{L}|\widehat{\boldsymbol{\Upsilon}}_{\mathrm{EM}_{2}}^{[\nu-1]}, \boldsymbol{X})} \log f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L}), \quad (53)$$

where $f(\mathbf{\Upsilon}|\widehat{\mathbf{L}}_{\mathrm{EM}_1}^{[\nu-1]}, \mathbf{X}) = \prod_{k=1}^K \mathcal{N}_{\mu_k}(\widetilde{\boldsymbol{\mu}}_k^{[\nu]}, \widetilde{\boldsymbol{\sigma}}_k^{[\nu]} \mathbf{I}_2)$ and $f(\mathbf{L}|\widehat{\mathbf{\Upsilon}}_{\mathrm{EM}_2}^{[\nu-1]}, \mathbf{X}) = \prod_{i=1}^N Mu_{l_i}(\widetilde{\boldsymbol{p}}_i^{[\nu-1]}).$

Replacing Υ and L in $f(X, \Upsilon, L)$ in (52) and (53) with $\widetilde{\Upsilon}^{[\nu-1]} \triangleq \mathbb{E}_{f(\Upsilon|\widehat{L}_{\mathrm{EM}_{1}}^{[\nu-1]}, X)}(\Upsilon)$ and $\widetilde{P}^{[\nu-1]} \triangleq \mathbb{E}_{f(L|\widehat{\Upsilon}_{\mathrm{EM}_{2}}^{[\nu-1]}, X)}(L)$, we then have, respectively:

$$\widetilde{\boldsymbol{\mu}}_{k}^{[\nu]} = \overline{\boldsymbol{\mu}}_{k}(\widehat{\boldsymbol{L}}_{\text{EM}_{1}}^{[\nu-1]}), \ \widetilde{\boldsymbol{\sigma}}_{k}^{[\nu]} = \overline{\boldsymbol{\sigma}}_{k}(\widehat{\boldsymbol{L}}_{\text{EM}_{1}}^{[\nu-1]}),
\widehat{\boldsymbol{k}}_{i}^{[\nu]} = \underset{k}{\operatorname{arg\,max}} \frac{\mathcal{N}_{\boldsymbol{x}_{i}}(\widetilde{\boldsymbol{\mu}}_{k}^{[\nu]}, \mathbf{I}_{2})}{\exp((\widetilde{\boldsymbol{\sigma}}_{k}^{[\nu]})^{2})},$$
(54)

and:

$$\widehat{p}_{k,i}^{[\nu]} \propto \mathcal{N}_{\boldsymbol{x}_i}(\widehat{\boldsymbol{\mu}}_k^{[\nu]}, \mathbf{I}_2), \qquad (55)$$

$$\widehat{\boldsymbol{\mu}}_k^{[\nu]} = \overline{\boldsymbol{\mu}}_k(\widetilde{\boldsymbol{P}}^{[\nu-1]}) = \frac{\sum_{i=1}^N \widetilde{p}_{k,i}^{[\nu-1]} \boldsymbol{x}_i}{\sum_{i=1}^N \widetilde{p}_{k,i}^{[\nu-1]}}, \qquad (55)$$

 $\begin{array}{lll} \text{where} & \widehat{\mathbf{\Upsilon}}_{\mathrm{EM}_2}^{[\nu]} & \triangleq & [\widehat{\boldsymbol{\mu}}_1^{[\nu]}, \widehat{\boldsymbol{\mu}}_2^{[\nu]}, \dots, \widehat{\boldsymbol{\mu}}_K^{[\nu]}], \quad \widetilde{\mathbf{\Upsilon}}^{[\nu]} & \triangleq \\ [\widetilde{\boldsymbol{\mu}}_1^{[\nu]}, \widetilde{\boldsymbol{\mu}}_2^{[\nu]}, \dots, \widetilde{\boldsymbol{\mu}}_K^{[\nu]}], \quad \widehat{\boldsymbol{L}}_{\mathrm{EM}_1}^{[\nu]} & \triangleq & [\widehat{\boldsymbol{l}}_1^{[\nu]}, \widehat{\boldsymbol{l}}_2^{[\nu]}, \dots, \widehat{\boldsymbol{l}}_N^{[\nu]}] \quad \text{with} \\ \widehat{\boldsymbol{l}}_i^{[\nu]} & \triangleq & [\widehat{\boldsymbol{l}}_{1,i}, \widehat{\boldsymbol{l}}_{2,i}, \dots, \widehat{\boldsymbol{l}}_{K,i}]^T \quad \text{and} \quad \widehat{\boldsymbol{l}}_{k,i}^{[\nu]} & = \quad \delta[k - \widehat{\boldsymbol{k}}_i^{[\nu]}], \\ \widetilde{\boldsymbol{P}}^{[\nu]} & \triangleq & [\widetilde{\boldsymbol{p}}_1^{[\nu]}, \widetilde{\boldsymbol{p}}_2^{[\nu]}, \dots, \widetilde{\boldsymbol{p}}_N^{[\nu]}] \quad \text{with} \quad \sum_{k=1}^K \widehat{\boldsymbol{p}}_{k,i}^{[\nu]} & = \quad 1, \\ \forall i \in \{1, 2, \dots, N\}. \end{array}$

The forms $\overline{\mu}_k$ and $\overline{\sigma}_k$ are given in (46). By convention, we keep $\widetilde{\mu}_k^{[\nu]} = \widetilde{\mu}_k^{[\nu-1]}$ and $\widetilde{\sigma}_k^{[\nu]} = \widetilde{\sigma}_k^{[\nu-1]}$ unchanged if $\sum_{i=1}^N \widehat{l}_{k,i}^{[\nu-1]} = 0$ in (54).

Also, since $f(\Upsilon, L, X)$ is of CEF form (26), we can feasibly evaluate $\mathrm{KL}_{\widetilde{f}_{\mathrm{EM}}^{[\nu]}||f(X,\Upsilon,L)}$ directly for $\widetilde{f}_{\mathrm{EM}_1}^{[\nu]}$ and $\widetilde{f}_{\mathrm{EM}_2}^{[\nu]}$, as defined in (51). The convergence of $\mathrm{ELBO}_{\mathrm{EM}}^{[\nu]}$, as given in (33), is then computed as follows:

$$\begin{split} \text{ELBO}_{\text{EM}}^{[\nu]} &= -\text{KL}_{\widetilde{f}_{\text{EM}}^{[\nu]}||f(\boldsymbol{X},\boldsymbol{\Upsilon},\boldsymbol{L})} = \log \frac{\zeta_{\text{EM}}^{[\nu]}}{\zeta K^{N}}, \\ \zeta_{\text{EM}_{1}}^{[\nu]} &= \frac{f(\boldsymbol{\Upsilon} = \widetilde{\boldsymbol{\Upsilon}}^{[\nu]}, \boldsymbol{L} = \widehat{\boldsymbol{L}}_{\text{EM}}^{[\nu]}, \boldsymbol{X})}{\exp\left(\sum_{k=1}^{K} (\widetilde{\boldsymbol{\sigma}}_{k}^{[\nu]})^{2} \sum_{i=1}^{N} \widehat{l}_{k,i}^{[\nu]}\right)} \prod_{k=1}^{K} ((\widetilde{\boldsymbol{\sigma}}_{k}^{[\nu]})^{2} 2\pi e). \\ \zeta_{\text{EM}_{2}}^{[\nu]} &= \frac{f(\boldsymbol{\Upsilon} = \widehat{\boldsymbol{\Upsilon}}_{\text{EM}}^{[\nu]}, \boldsymbol{L} = \widetilde{\boldsymbol{P}}^{[\nu]}, \boldsymbol{X})}{\prod_{k=1}^{K} \prod_{i=1}^{N} \widetilde{\boldsymbol{p}}_{k,i}^{[\nu]} \widehat{\boldsymbol{p}}_{k,i}^{[\nu]}}. \end{split}$$

Remark 40. Comparing (54-55) with (50), we can see that the k-means algorithm only considers the mean (i.e. the first moment), while EM algorithm takes both mean and variance (i.e. the first and second moments) into account.

3) VB approximation: Let us now derive VB approximation $\widetilde{f}_{VB}(\Upsilon, L|X) = \widetilde{f}_{VB}(\Upsilon|X)\widetilde{f}_{VB}(L|X)$ in (29) for true posterior distribution $f(\Upsilon, L|X)$. Since $f(\Upsilon, L, X)$ is of CEF form (26), the VB form can be feasibly identified via (28), as follows:

$$\widetilde{f}_{VB}^{[\nu]}(\mathbf{\Upsilon}|\mathbf{X}) \propto \exp \mathbb{E}_{\widetilde{f}_{VB}^{[\nu-1]}(\mathbf{L}|\mathbf{X})} \log f(\mathbf{X}, \mathbf{\Upsilon}, \mathbf{L}) \qquad (56)$$

$$= \prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{\mu}_{k}} \left(\widetilde{\boldsymbol{\mu}}_{k}^{[\nu]}, \widetilde{\boldsymbol{\sigma}}_{k}^{[\nu]} \mathbf{I}_{2} \right),$$

$$\widetilde{f}_{VB}^{[\nu]}(\mathbf{L}|\mathbf{X}) \propto \exp \mathbb{E}_{\widetilde{f}_{VB}^{[\nu]}(\mathbf{\Upsilon}|\mathbf{X})} \log f(\mathbf{X}, \mathbf{\Upsilon}, \mathbf{L})$$

$$= \prod_{i=1}^{N} M u_{\boldsymbol{l}_{i}} \left(\widetilde{\boldsymbol{p}}_{i}^{[\nu]} \right),$$

Replacing $m{L}$ in (46) with $\widetilde{m{P}}^{[
u]} = \mathbb{E}_{\widetilde{f}_{\mathrm{vR}}^{[
u]}(m{L}|m{X})}(m{L})$, we then have:

$$\widetilde{\boldsymbol{\mu}}_{k}^{[\nu]} = \overline{\boldsymbol{\mu}}_{k}(\widetilde{\boldsymbol{P}}^{[\nu-1]}), \ \widetilde{\sigma}_{k}^{[\nu]} = \overline{\sigma}_{k}(\widetilde{\boldsymbol{P}}^{[\nu-1]}),
\widetilde{p}_{k,i}^{[\nu]} \propto \frac{\mathcal{N}_{\boldsymbol{x}_{i}}(\widetilde{\boldsymbol{\mu}}_{k}^{[\nu]}, \mathbf{I}_{2})}{\exp((\widetilde{\sigma}_{k}^{[\nu]})^{2})},$$
(57)

where $\widetilde{\boldsymbol{P}}^{[\nu]} \triangleq [\widetilde{\boldsymbol{p}}_1^{[\nu]}, \widetilde{\boldsymbol{p}}_2^{[\nu]}, \ldots, \widetilde{\boldsymbol{p}}_N^{[\nu]}]$ and $\sum_{k=1}^K \widetilde{\boldsymbol{p}}_{k,i}^{[\nu]} = 1, \ \forall i \in \{1, 2, \ldots, N\}$. The forms $\overline{\boldsymbol{\mu}}_k$ and $\overline{\sigma}_k$ are given in (46).

From (56), let $\zeta_{\mathrm{VB}}^{[\nu]}$ denote the normalizing constant of $\widetilde{f}_{\mathrm{VB}}^{[\nu]}(\boldsymbol{\Upsilon}|\boldsymbol{X}) = \frac{1}{\zeta_{\mathrm{VB}}^{[\nu]}} \exp \mathbb{E}_{\widetilde{f}_{\mathrm{VB}}^{[\nu-1]}(\boldsymbol{L}|\boldsymbol{X})} \log \frac{f(\boldsymbol{X},\boldsymbol{\Upsilon},\boldsymbol{L})}{\widetilde{f}_{\mathrm{VB}}^{[\nu-1]}(\boldsymbol{L}|\boldsymbol{X})}$, similarly

to (32). We then have:

$$\text{ELBO}_{VB}^{[\nu]} = \log \zeta_{VB}^{[\nu]} = \log \frac{1}{\zeta K^N} \prod_{k=1}^{K} \frac{\gamma_k(\widetilde{\boldsymbol{P}}^{[\nu]})}{\prod_{i=1}^{N} \widetilde{p}_{k,i}^{[\nu]} \widetilde{p}_{k,i}^{[\nu]}}, \quad (58)$$

where $\gamma_k(\widetilde{\boldsymbol{P}}^{[\nu]})$ is given in (46), with \boldsymbol{L} being replaced by $\widetilde{\boldsymbol{P}}^{[\nu]}$. The convergence of ELBO_{VB}, as mentioned in (33), can be used as a stopping rule.

Remark 41. From (56-57), we can see that the VB algorithm combines two EM algorithms (51-55) together and takes all moments of clustering data into account.

4) CVB approximation: Let us now derive CVB approximation $\widetilde{f}_{\text{CVB}}^{[1]} = \widetilde{f}_{2|1}^{[0]}\widetilde{f}_{1}^{[1]} = \widetilde{f}_{1|2}^{[1]}\widetilde{f}_{2}^{[1]}$ for true posterior distribution $f(\Theta|X) = f(\Upsilon, L|X)$, with $\Theta = [\Upsilon, L]$, via (25), (32). Firstly, let us note that, the denominator $f(\Upsilon, X)$ of $f(L|\Upsilon, X)$ in (48) is a mixture of K^N Gaussian components, which is not factorable over its marginals on μ_k , $k \in \{1, 2, \ldots, K\}$. Hence, a direct application of CVB algorithm (32) with $\theta_1 = L$ and $\theta_2 = \Upsilon$ would not yield a closed form for $\widetilde{f}_{CVB}(\Upsilon|X)$, when the total number K of clusters is not small.

• CVB's ternary partition:

For a tractable form of $\widetilde{f}_{CVB}(\Upsilon|X)$, let us now define two different binary partitions θ_1 and θ_2 for $\Theta = [\Upsilon, L]$ at each CVB's iteration, as explained in subsection IV-B2:

$$f(\boldsymbol{\Theta}|\boldsymbol{X}) = \underbrace{f(\boldsymbol{\Upsilon}|\boldsymbol{L},\boldsymbol{X})f(\boldsymbol{L}|\boldsymbol{X})}_{f_{2|1}} = \underbrace{f(\boldsymbol{L}_{\backslash j}|\boldsymbol{\Upsilon},\boldsymbol{X})f(\boldsymbol{\Upsilon},\boldsymbol{l}_{j}|\boldsymbol{X})}_{f_{1|2}},$$

$$\widetilde{f}(\boldsymbol{\Theta}|\boldsymbol{X}) = \underbrace{\widetilde{f}(\boldsymbol{\Upsilon}|\boldsymbol{l}_{j},\boldsymbol{X})\widetilde{f}(\boldsymbol{L}|\boldsymbol{X})}_{\widetilde{f}_{2|1}} = \underbrace{\widetilde{f}(\boldsymbol{L}_{\backslash j}|\boldsymbol{l}_{j},\boldsymbol{X})\widetilde{f}(\boldsymbol{\Upsilon},\boldsymbol{l}_{j}|\boldsymbol{X})}_{\widetilde{f}_{1|2}},$$

$$(59)$$

for any node $j \in \{1, 2, ..., N\}$. Note that, the true conditional $f_{1|2} \triangleq f(\boldsymbol{L}_{\backslash j}|\boldsymbol{\Upsilon}, \boldsymbol{l}_j, \boldsymbol{X}) = f(\boldsymbol{L}_{\backslash j}|\boldsymbol{\Upsilon}, \boldsymbol{X})$ in (59) does not depends on \boldsymbol{l}_j , since $f(\boldsymbol{L}|\boldsymbol{\Upsilon}, \boldsymbol{X})$ in (48) is conditionally independent, i.e. $f(\boldsymbol{L}|\boldsymbol{\Upsilon}, \boldsymbol{X}) = \prod_{i=1}^N f(\boldsymbol{l}_i|\boldsymbol{\Upsilon}, \boldsymbol{X})$, as illustrated in Fig. 12.

Hence, given a ternary partition $\boldsymbol{\Theta} = [\boldsymbol{L}_{\backslash j}, \boldsymbol{l}_j, \boldsymbol{\Upsilon}]$ for each node j in (59), we have set $\boldsymbol{\theta}_1 = \boldsymbol{L} = [\boldsymbol{L}_{\backslash j}, \boldsymbol{l}_j]$ and $\boldsymbol{\theta}_2 = \boldsymbol{\Upsilon}$ in the forward form, but $\boldsymbol{\theta}_1 = \boldsymbol{L}_{\backslash j}$ and $\boldsymbol{\theta}_2 = [\boldsymbol{l}_j, \boldsymbol{\Upsilon}]$ in reverse form in (59). The equality in CVB form $\widetilde{f}_{\text{CVB}}^{[1]} = \widetilde{f}_{2|1}^{[0]} \widetilde{f}_1^{[1]} = \widetilde{f}_{1|2}^{[1]} \widetilde{f}_2^{[1]}$ is still valid, since we still have the same joint parameters $\boldsymbol{\Theta} = [\boldsymbol{\Upsilon}, \boldsymbol{L}]$ on both sides.

• CVB's initialization:

Let us consider the left form in (59) first. For tractability, the initial CVB $\tilde{f}_{2|1}^{[0]} \triangleq \tilde{f}^{[0]}(\mathbf{\Upsilon}|\boldsymbol{l}_j, \boldsymbol{X})$ will be set as a restricted form of the true conditional $f_{2|1} \triangleq f(\mathbf{\Upsilon}|\boldsymbol{L}, \boldsymbol{X})$ in (45), as follows:

$$\widetilde{f}_{2|1}^{[0]} = \prod_{k=1}^{K} \widetilde{f}^{[0]}(\boldsymbol{\mu}_{k}|\boldsymbol{l}_{j}) = \prod_{k=1}^{K} \prod_{m=1}^{K} \mathcal{N}_{\boldsymbol{\mu}_{k}}^{l_{m,j}} \left(\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]}, \widetilde{\sigma}_{k,m,j}^{[0]} \mathbf{I}_{2} \right)$$
(60)

where $\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]} \in \mathbb{R}^2$ and $\widetilde{\sigma}_{k,m,j}^{[0]} > 0$ are initial means and variances of $\widetilde{f}^{[0]}(\boldsymbol{\mu}_k|\boldsymbol{l}_j) = \prod_{m=1}^K \widetilde{f}^{[0]}(\boldsymbol{\mu}_k|l_{m,j})$.

• CVB's iteration (forward step):

Let us now apply CVB algorithm (32) to (59) and approximate $f_1 \triangleq f(L|X)$ via $\tilde{f}_{2|1}^{[0]}$ in (60), as follows:

$$\widetilde{f}_{1}^{[1]} \triangleq \widetilde{f}^{[1]}(\boldsymbol{L}|\boldsymbol{X}) = \frac{1}{\zeta_{1}^{[1]}} \exp \mathbb{E}_{\widetilde{f}_{2|1}^{[0]}} \log \frac{f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L})}{\widetilde{f}_{2|1}^{[0]}} \qquad (61)$$

$$= \frac{1}{\zeta_{1}^{[1]} \zeta K^{N}} \prod_{k=1}^{K} \prod_{m=1}^{K} (\widetilde{\kappa}_{k,m,j}^{[1]} \prod_{i=1}^{N} (\widetilde{\gamma}_{k,m,i,j}^{[1]})^{l_{k,i}})^{l_{m,j}},$$

in which $f(X, \Upsilon, L)$ is given in (43-44) and, hence:

$$\widetilde{\kappa}_{k,m,j}^{[1]} = 2\pi e(\widetilde{\sigma}_{k,m,j}^{[0]})^2, \ \widetilde{\gamma}_{k,m,i,j}^{[1]} \triangleq \frac{\mathcal{N}_{\boldsymbol{x}_i}(\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]}, \mathbf{I}_2)}{\exp((\widetilde{\sigma}_{k,m,j}^{[0]})^2)}, \ (62)$$

since we have $\exp \mathbb{E}_{\mathcal{N}_{\boldsymbol{\mu}}(\widetilde{\boldsymbol{\mu}},\mathbf{I}_2)} \log \frac{\mathcal{N}_{w_i}(\boldsymbol{\mu},\mathbf{I}_2)}{\mathcal{N}_{\boldsymbol{\mu}}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\sigma}}\mathbf{I}_2)} = \frac{\mathcal{N}_{w_i}(\widetilde{\boldsymbol{\mu}},\mathbf{I}_2)\exp(-\widetilde{\boldsymbol{\sigma}}^2)}{1/(2\pi\widetilde{\boldsymbol{\sigma}}^2e)}$ in general, as shown in (28). Comparing the true updated weights $\gamma_k(\boldsymbol{L})$ of f_1 in (46) with the approximated weights $\widetilde{\gamma}_{k,m,i,j}^{[1]}$ of $\widetilde{f}_1^{[1]}$ in (62), we can see that CVB algorithm has approximated the intractable forms $\{\overline{\boldsymbol{\mu}}_k(\boldsymbol{L}), \overline{\sigma}_k^2(\boldsymbol{L})\}$ with total K^N elements by a factorized set of N tractable forms $\{\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]}, \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[0]}\}$ with only K^2N elements.

Comparing (59) with (61), we can identify the form of $\tilde{f}_{1|2}^{[1]}$ in (59), as follows:

$$\widetilde{f}_{1|2}^{[1]} \triangleq \widetilde{f}^{[1]}(\boldsymbol{L}_{\backslash j}|\boldsymbol{l}_{j}, \boldsymbol{X}) = \prod_{i \neq j} Mu_{\boldsymbol{l}_{i}} \left(\widetilde{\boldsymbol{W}}_{i,j}^{[1]} \boldsymbol{l}_{j}\right)$$
 (63)

where $\widetilde{\boldsymbol{W}}_{i,j}^{[1]}$ is a left stochastic matrix, whose $\{m,k\}$ -element is the updated transition probability from $l_{m,j}$ to $l_{k,i}$: $\widetilde{\boldsymbol{w}}_{k,m}^{[1]}(i,j) \triangleq \frac{\widetilde{\gamma}_{k,m,i,j}^{[1]}}{\sum_{k=1}^{K} \widetilde{\gamma}_{k,m,i,j}^{[1]}}$, $\forall i \neq j$. For later use, let us assign $\widetilde{\boldsymbol{W}}_{j,j}^{[\nu]} \triangleq \mathbf{I}_{K}$, with \mathbf{I}_{K} denoting $K \times K$ identity matrix, when i=j, at any iteration ν .

• CVB's iteration (reverse step):

Let us apply CVB (32) to (59) again and approximate $f_2 \triangleq f(\Upsilon, l_j | X)$ by $\tilde{f}_2^{[2]}$ in $\tilde{f}_{\text{CVB}}^{[2]} = \tilde{f}_{2|1}^{[2]} \tilde{f}_1^{[2]} = \tilde{f}_{1|2}^{[1]} \tilde{f}_2^{[2]}$, via $\tilde{f}_{1|2}^{[1]}$ in (63), as follows:

$$\widetilde{f}_{2}^{[2]} \triangleq \widetilde{f}^{[2]}(\boldsymbol{\Upsilon}, \boldsymbol{l}_{j} | \boldsymbol{X}) = \frac{1}{\zeta_{2}^{[2]}} \exp \mathbb{E}_{\widetilde{f}_{1|2}^{[1]}} \log \frac{f(\boldsymbol{X}, \boldsymbol{\Upsilon}, \boldsymbol{L})}{\widetilde{f}_{1|2}^{[1]}} \\
= \widetilde{f}^{[2]}(\boldsymbol{\Upsilon} | \boldsymbol{l}_{j}. \boldsymbol{X}) \widetilde{f}^{[2]}(\boldsymbol{l}_{j} | \boldsymbol{X}) \tag{64}$$

in which, similar to (45-46), we have:

$$\widetilde{f}_{2|1}^{[2]} \triangleq \widetilde{f}^{[2]}(\mathbf{\Upsilon}|\boldsymbol{l}_{j}.\boldsymbol{X}) = \prod_{k=1}^{K} \prod_{m=1}^{K} \mathcal{N}_{\boldsymbol{\mu}_{k}}^{l_{m,j}} \left(\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[1]}, \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[1]} \mathbf{I}_{2} \right)
\widetilde{f}^{[2]}(\boldsymbol{l}_{j}|\boldsymbol{X}) = \frac{1}{\zeta_{2}^{[2]} \zeta K^{N}} \prod_{k=1}^{K} \prod_{m=1}^{K} (\widetilde{\boldsymbol{\gamma}}_{k,m,j}^{[1]})^{l_{m,j}} \tag{65}$$

Note that, as shown in (28), we have replaced $l_{k,i}$ in (46) by $\widetilde{l}_{k,i}(\boldsymbol{l}_j) \triangleq \mathbb{E}_{\widetilde{f}_{1|2}^{[1]}}(l_{k,i}) = \sum_{m=1}^{M} \widetilde{w}_{k,m}(i,j)l_{m,j}$ in (65) and,

hence

$$\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[1]} \triangleq \frac{\sum_{i=1}^{N} \widetilde{w}_{k,m}^{[1]}(i,j) \boldsymbol{x}_{i}}{\sum_{i=1}^{N} \widetilde{w}_{k,m}^{[1]}(i,j)}, \ \widetilde{\sigma}_{k,m,j}^{[1]} \triangleq \frac{1}{\sqrt{\sum_{i=1}^{N} \widetilde{w}_{k,m}^{[1]}(i,j)}}$$

$$\widetilde{\gamma}_{k,m,j}^{[1]} = \frac{2\pi (\widetilde{\sigma}_{k,m,j}^{[1]})^2 \prod_{i=1}^{N} \mathcal{N}_{\boldsymbol{x}_i}^{\widetilde{w}_{k,m}^{[1]}(i,j)} (\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[1]}, \mathbf{I}_2)}{\prod_{i \neq j} \widetilde{w}_{k,m}^{[1]}(i,j)^{\widetilde{w}_{k,m}^{[1]}(i,j)}}, \quad (66)$$

in which, by convention, $\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[1]} = \widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]}, \ \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[1]} = \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[0]}$ are kept unchanged and $\widetilde{\gamma}_{k,m,j}^{[1]} = 1$ if $\sum_{i=1}^N \widetilde{w}_{k,m}^{[1]}(i,j) = 0$. It is feasible to recognize that $\widetilde{f}^{[2]}(\boldsymbol{l}_j|\boldsymbol{X})$ in (65) is actually a

It is feasible to recognize that $\widetilde{f}^{[2]}(\boldsymbol{l}_j|\boldsymbol{X})$ in (65) is actually a Multinomial distribution: $\widetilde{f}^{[2]}(\boldsymbol{l}_j|\boldsymbol{X}) = Mu_{\boldsymbol{l}_j}(\widetilde{\boldsymbol{p}}_j^{[2]})$, in which $\widetilde{\boldsymbol{p}}_j^{[2]} \triangleq [\widetilde{p}_{1,j}^{[2]}, \widetilde{p}_{2,j}^{[2]}, \dots, \widetilde{p}_{K,j}^{[2]}]^T$ and $\sum_{m=1}^K \widetilde{p}_{m,j}^{[2]} = 1$, as follows:

$$\widetilde{p}_{m,j}^{[2]} \triangleq \frac{\prod_{k=1}^{K} \widetilde{\gamma}_{k,m,j}^{[1]}}{\sum_{m=1}^{M} \prod_{k=1}^{K} \widetilde{\gamma}_{k,m,j}^{[1]}}, \ \zeta_{2}^{[2]} = \frac{\sum_{m=1}^{M} \prod_{k=1}^{K} \widetilde{\gamma}_{k,m,j}^{[1]}}{\zeta K^{N}}$$
(67)

• CVB's form at convergence:

From (60) and (65), we can see that $\tilde{f}_{2|1}^{[\nu]}$ can be updated iteratively from $\tilde{f}_{2|1}^{[\nu-2]}$, given that only one CVB marginal is updated per iteration ν . The iterative CVB then converges when the ELBO^[\nu] $\triangleq \log \zeta_2^{[\nu]}$, given in (67), converges at $\nu = \nu_c$, as shown in (33).

Then, for any chosen $j \in \{1,2,\ldots,M\}$, the marginals in converged CVB $\widetilde{f}_j^{[\nu_c]}$ can be derived from $\widetilde{f}^{[\nu_c]}(\boldsymbol{l}_j|\boldsymbol{X}) = Mu_{\boldsymbol{l}_j}(\widetilde{\boldsymbol{p}}_j^{[\nu_c]})$ in (67), as follows:

$$\widetilde{f}_{j}^{[\nu_{c}]}(\Upsilon|X) \triangleq \sum_{\boldsymbol{l}_{j}} \widetilde{f}^{[\nu_{c}]}(\Upsilon|\boldsymbol{l}_{j}.\boldsymbol{X}) \widetilde{f}^{[\nu_{c}]}(\boldsymbol{l}_{j}|\boldsymbol{X}), \qquad (68)$$

$$\widetilde{f}_{j}^{[\nu_{c}]}(\boldsymbol{L}|\boldsymbol{X}) \triangleq \prod_{i \neq j} \widetilde{f}^{[\nu_{c}]}(\boldsymbol{l}_{i}|\boldsymbol{l}_{j}.\boldsymbol{X}) \widetilde{f}^{[\nu_{c}]}(\boldsymbol{l}_{j}|\boldsymbol{X}),$$

in which $\widetilde{f}_j^{[\nu_c]}(\boldsymbol{\Upsilon}|\boldsymbol{X})$ is a mixture of M Gaussian components:

$$\begin{split} \widetilde{f}_{j}^{[\nu_{c}]}(\boldsymbol{\Upsilon}|\boldsymbol{X}) &= \sum_{m=1}^{K} \widetilde{p}_{m,j}^{[\nu_{c}]} \prod_{k=1}^{K} \mathcal{N}_{\boldsymbol{\mu}_{k}}(\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[\nu_{c}]}, \widetilde{\sigma}_{k,m,j}^{[\nu_{c}]} \mathbf{I}_{2}), \\ \widetilde{f}_{j}^{[\nu_{c}]}(\boldsymbol{l}_{i}|\boldsymbol{X}) &= \sum_{\boldsymbol{L} \setminus i} \widetilde{f}_{j}^{[\nu_{c}]}(\boldsymbol{L}|\boldsymbol{X}) = Mu_{\boldsymbol{l}_{i}}(\widetilde{\boldsymbol{q}}_{i}^{[\nu_{c}]}(j)), \end{split}$$

with $\widetilde{\boldsymbol{q}}_i^{[\nu_c]}(j) \triangleq [\widetilde{q}_{1,i}^{[\nu_c]}(j),\ldots,\widetilde{q}_{K,i}^{[\nu_c]}(j)]^T = \widetilde{\boldsymbol{W}}_{i,j}^{[\nu_c]}\widetilde{\boldsymbol{p}}_j^{[\nu_c]}, \ \forall i \in \{1,2,\ldots,N\}.$ The approximated posterior estimates for cluster's means and labels in this case are, respectively:

$$\widehat{\mathbf{\Upsilon}}(j) \triangleq \mathbb{E}_{\widetilde{f}_{j}^{[\nu_{c}]}(\mathbf{\Upsilon}|\mathbf{X})}(\mathbf{\Upsilon}) = \sum_{m=1}^{K} \widetilde{p}_{m,j}^{[\nu_{c}]} \widetilde{\mathbf{\Upsilon}}_{m,j}^{[\nu_{c}]},$$

$$\widehat{l}_{i}(j) \triangleq \arg \max_{\mathbf{l}_{i}} \widetilde{f}_{j}^{[\nu_{c}]}(\mathbf{l}_{i}|\mathbf{X}) = \epsilon_{\widehat{k}_{i}(j)},$$
(69)

 $\begin{array}{ll} \text{where} \quad \widetilde{\pmb{\Upsilon}}_{m,j}^{[\nu_c]} \triangleq \quad [\widetilde{\pmb{\mu}}_{1,m,j}^{[\nu_c]}, \dots, \widetilde{\pmb{\mu}}_{K,m,j}^{[\nu_c]}] \quad \text{and} \quad \widehat{k_i}(j) \quad \triangleq \\ \arg\max_k \widetilde{q}_{k,i}^{[\nu_c]}(j), \, \forall i \in \{1,2,\dots,N\}. \end{array}$

• Augmented CVB approximation:

As shown above, each value $j \in \{1, 2, ..., N\}$ yields a different network structure for CVB approximation, as mentioned in section V. Let us consider here three simple ways to make use of these N CVB's structures.

The first and heuristic way, namely CVB₁ scheme, is to choose $\widehat{l_j}(j)$ in (69), as the estimate for j-th label, because the CVB's ternary structure is more focused on l_j at each $j \in \{1,2,\ldots,N\}$, as shown in Fig. 12. Since every j-th structure is equally important in this way, we can pick the empirical average $\widehat{\mathbf{\Upsilon}} \triangleq \sum_{j=1}^N \frac{1}{N} \widehat{\mathbf{\Upsilon}}(j)$ as estimate for cluster's means.

The second way, namely CVB₂ scheme, is to pick j such that $\mathrm{KL}_{\widehat{f}_j^{[\nu_c]}||f}$ at convergence is minimized, as mentioned in (40) . From (33-34), we then have $\widehat{j} \triangleq \arg\min_j \mathrm{KL}_{\widehat{f}_j^{[\nu_c]}||f} = \arg\max_j \mathrm{ELBO}^{[\nu_c]}(j) = \arg\max_j \log \zeta_2^{[\nu_c]}(j)$, with $\zeta_2^{[\nu_c]}(j)$ given in (67). Then, $\widehat{l}_i(\widehat{j})$ and $\widehat{\Upsilon}(\widehat{j})$ in (69) will be used as estimates for categorical label l_i and cluster means, respectively, $\forall i \in \{1,2,\ldots,N\}$.

The third way, namely CVB₃ scheme, is to apply the augmented approach for CVB, given in (38). Then, from (41) and (69), the augmented CVB's estimates for cluster's means and labels in this case are $\hat{\Upsilon}^* \triangleq \sum_{j=1}^N q_j^* \hat{\Upsilon}(j)$ and $\hat{l}_i^* = \epsilon_{\hat{k}_i^*}$, respectively, with:

$$\widehat{k_i}^* \triangleq \arg\max_{k} \sum_{j=1}^{N} q_j^* \widetilde{q}_{k,i}^{[\nu_c]}(j),
q_j^* = \frac{\exp(-KL_{\widetilde{f}_j^{[\nu_c]}||f})}{\sum_{j=1}^{N} \exp(-KL_{\widetilde{f}_j^{[\nu_c]}||f})} = \frac{\zeta_2^{[\nu_c]}(j)}{\sum_{j=1}^{N} \zeta_2^{[\nu_c]}(j)}, \quad (70)$$

in which q_j^* is found via (38) and $\mathrm{KL}_{\widetilde{f}_j^{[\nu c]}||f} = -\mathrm{ELBO}^{[\nu_c]}(j) + \log f(\boldsymbol{X})$, as shown in (34) $\forall j \in \{1,2,\ldots,N\}$.

Although we can compute all moments of augmented CVB $\widetilde{f}_0^{[\nu_c]} = \sum_{j=1}^N q_j^* \widetilde{f}_j^{[\nu_c]} \text{via}$ (41) and (70), it is difficult to evaluate $\text{KL}_{\widetilde{f}_0^{[\nu_c]}||f}$ and its ELBO value directly, as mentioned in subsection V-B. Hence, for comparison with CVB $_2$ scheme in simulations, let us instead compute heuristic values $\sum_{j=1}^N \frac{1}{N} \text{ELBO}^{[\nu_c]}(j)$ and $\sum_{j=1}^N q_j^* \text{ELBO}^{[\nu_c]}(j)$ at convergence for CVB $_1$ and CVB $_3$ schemes, respectively, with ELBO $^{[\nu_c]}(j) \triangleq \log \zeta_2^{[\nu_c]}(j)$ given in (67).

Remark 42. Note that, the CVB $\widetilde{f}_j^{[\nu_c]}$ still belongs to a conditional structure class of node j at convergence, even if the initialization $\{\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]}, \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[0]}\}$ of CVB is exactly the same as that of VB. Indeed, in below simulations, even though initially we set $\widetilde{\boldsymbol{\mu}}_{k,m,j}^{[0]} = \widetilde{\boldsymbol{\mu}}_{k}^{[0]}, \ \widetilde{\boldsymbol{\sigma}}_{k,m,j}^{[0]} = \widetilde{\boldsymbol{\sigma}}_{k}^{[0]}, \ \forall m,j$ and, hence, $\widetilde{f}_{2|1}^{[0]} = \widetilde{f}^{[0]}(\boldsymbol{\Upsilon}|\boldsymbol{l}_j,\boldsymbol{X})$ in (60) independent of \boldsymbol{l}_j , the conditional $\widetilde{f}^{[\nu]}(\boldsymbol{\Upsilon}|\boldsymbol{l}_j,\boldsymbol{X})$ in (65-66) depends on \boldsymbol{l}_j again in subsequent iterations, as already explained in subsection IV-B2 for this case of ternary partition.

5) Simulation's results: Since k-means algorithm (50) works best for independent Normal clusters, let us illustrate the superior performance of CVB to mean-field approximations even in this case. For this purpose, a set of K=4 bivariate independent Normal clusters $\mathcal{N}(\boldsymbol{\mu}_k,\mathbf{I}_2)$ are generated randomly, with true means $\boldsymbol{\Upsilon}=\boldsymbol{\Upsilon}_0R+\begin{bmatrix}1\\1\end{bmatrix}$ and $\boldsymbol{\Upsilon}_0\triangleq\begin{bmatrix}-1&1&1&-1\\1&1&-1&-1\end{bmatrix}$. At each time i, a cluster is then

chosen with equal probabilities $p_k = \frac{1}{K}, k \in \{1, 2, ..., K\}$

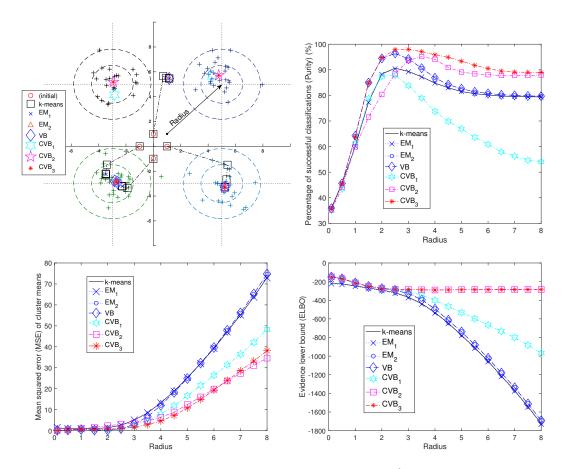


Figure 13. CVB and mean-field approximations for K=4 bivariate independent Normal clusters $\mathcal{N}(\boldsymbol{\mu},\mathbf{I}_2)$, with mean vectors $\boldsymbol{\mu}$ located diagonally and equally at radius R from the offset point $[1,1]^T$. The upper left panel shows the convergent results of approximated mean vectors for one Monte Carlo run in the case R=4, with true mean vectors located at intersections of four dotted lines. The dashed circles represent contours of true Normal distributions. The plus signs + are N=100 random data, generated with equal probability from each Normal cluster. The four smallest circles are the same initial guesses of true mean vectors for all algorithms. The dash-dot line illustrates the k-means algorithm, from initial to convergent points. The other panels show the Purity, MSE and ELBO values at convergence with varying radius. The higher Purity, the higher percentage of correct classification of data. The higher ELBO at each radius, the lower KL divergence and, hence, the better approximation for that case of radius, as shown in (33-34) and illustrated in Fig. 7, 9. The number of Monte Carlo runs for each radius is 10^4 .

in order to generate the data $x_i \in \mathbb{R}^2, i \in \{1,2,\ldots,N\}$, with N=100, as shown in Fig. 13. The varying radius R then controls the inter-distance between clusters. In order to quantify the algorithm's performance, let us compute the Purity and mean squared error (MSE) for estimates $\hat{\boldsymbol{l}}_i$, $\hat{\boldsymbol{\Upsilon}}$ of categorical labels \boldsymbol{l}_i and mean vectors $\boldsymbol{\Upsilon}$, respectively. The Purity, which is a common measure for percentage of successful label's classification [53], is calculated as follows: Purity $=\sum_{k=1}^K \frac{1}{N} \max_m \sum_{i=1}^N \delta[\hat{l}_{k,i} = l_{m,i}]$ in each Monte Carlo run. The higher Purity $\in [0,1]$, the better estimate for labels. The MSE in each Monte Carlo run is calculated as follows: MSE $=\frac{1}{K} \min_{\phi \in \Phi} ||\phi(\hat{\boldsymbol{\Upsilon}}) - \boldsymbol{\Upsilon}||^2$, where Φ is all K! possible permutations of K estimated cluster means in $\hat{\boldsymbol{\Upsilon}} \in \mathbb{R}^{2 \times K}$.

For comparison at convergence, the initialization $\widetilde{\Upsilon}^{[0]} = \Upsilon_0$ and $\widetilde{\sigma}^{[0]} = 1$ are the same for all algorithms. The k-means (50) and EM₁ algorithms (54) will converge at iteration ν_c if there is no update for categorical labels, i.e. $\widehat{L}^{[\nu_c]} = \widehat{L}^{[\nu_c-1]} \Leftrightarrow \text{ELBO}^{[\nu_c]} = \text{ELBO}^{[\nu_c-1]}$ in this case. The other algorithms are called converged at iteration ν_c if $0 \leq \text{ELBO}^{[\nu_c]} - \text{ELBO}^{[\nu_c-1]} \leq 0.01$. The averaged values of

 ν_c over all cases in Fig. 13 are $[16.4, 16.4, 27.2, 27.4, 27.8] \pm [5.0, 5.1, 10.4, 10.4, 7.8]$ for k-means, EM $_1$, EM $_2$,VB and CVB algorithms. Only one approximated marginal is updated per iteration.

We can see that both performance and number of iterations of k-means and EM_1 algorithms are almost identical to each other, since they use the same approach with point estimates for categorical labels. Although the EM_1 (54) takes one extra data-driven step, in comparison with k-means, by using the total number of classified labels in each cluster as an indicator for credibility, the EM_1 is virtually the same as k-means in estimate's accuracy. Likewise, since the point estimates of labels are data-driven and use hard decision approach, the k-means and EM_1 yield lower accuracy than other methods, which are model-driven and use soft decision approach.

The EM_2 (55) and VB (57) also have almost identical performance and number of iterations, even though EM_2 does not update the cluster mean's credibility via total number of classified labels like VB does. Hence, like the case of EM_1 versus k-means, this extra step of data-driven update seems insignificant in terms of estimate's accuracy. Nevertheless, since both EM_2 and VB use the model's probability of each

label as weighted credibility and make soft decision at each iteration, their performance is significantly better than k-means and EM_1 in the range of radius $R \in [2,4]$. Hence, the model-driven update step seems to exploit more information from the true model than the data-driven update step, when the clusters are close to each other.

For a large radius R > 4, there is not much difference between soft and hard decisions for these standard Normal clusters, since the tail of Normal distribution is very small in these cases. Hence, given the same initialization at origin, the performances of all mean-field approximations like k-means, EM_1 , EM_2 and VB are very close to each other when the interdistance between clusters is high. Also, since the computation of soft decision in VB and EM_2 requires almost double number of iterations, compared with hard decision approaches like k-means and EM_1 , the k-means is more advantageous in this case, owing to its low computational complexity.

The CVB algorithms are the slowest methods overall. Since the CVB in (70) requires nearly the same number of iterations as VB for each structure $j \in \{1, 2, \ldots, N\}$, as illustrated in Fig. 12, the CVB's complexity is at least N times slower than VB method, where N is the number of data. In practice, we may not have to update all N CVB's potential structures, since there might be some good candidates out of exponentially growing number of potential structures. In this paper, however, let us consider the case of N structures in order to illustrate the superior performance of augmented CVB form in CVB₃ (70), in comparison with VB, heuristic CVB₁ and hit-or-miss CVB₂ approaches.

The heuristic CVB_1 , which takes uniform average for mean vectors over all N potential structures, returns a lower MSE than mean-field approximations in all cases. This result seems reasonable, since cluster means are common parameters of all potential CVB structures in Fig. 12. In contrast, CVB_1 returns label's estimate \hat{l}_j via j-th structure only, without considering label's estimates from other CVB's structures. Hence, the label's Purity of CVB_1 is only on par with that of mean-field approximations for short radius $R \leq 2$ and deteriorates over longer radius R > 2. As illustrated in Fig. 11, CVB might be the worst approximation if the CVB's structure is too different from true posterior structure. In this case, a single j-th structure seems to be a bad CVB candidate for estimating label l_j at time $j \in \{1, 2, \ldots, N\}$.

The hit-or-miss CVB_2 , which picks the single best structure \widehat{j} in terms of KL divergence, yields the worst performance in the range $R \in [1, 2.5]$, while in other cases, it is the second-best method. The structure \widehat{j} , as illustrated in Fig. 12, concentrates on the \widehat{j} -th label. Hence, the classification's accuracy of CVB_2 depends on whether the hard decision on \widehat{j} -th label serves as a good reference for other labels, as illustrated in Fig. 11. For this reason, CVB_2 may be able to achieve globally optimal approximation, but it may also be worse than mean-field approximations. When R < 3, which is less than three standard deviation of a standard Normal cluster, the clusters data are likely overlapped with each other. Within this range, the hard decision of CVB_2 on \widehat{j} destroys the correlated information between clusters and, hence, becomes worse than other methods. For $R \geq 3$, the

CVB₂ becomes better, which indicates that the classification's accuracy now relies more on the most significantly correlated structure between labels.

Generalizing both schemes CVB_1 and CVB_2 , CVB_3 (70) can return the optimal weights for the mixture of N potential structures and achieve the minimum upper bound of KL divergence (37), as illustrated in Fig. 4. Hence, the CVB_3 yields the best performance in Fig. 13. When R < 3, the CVB_3 is on par with VB approximation, since the probabilities computed via Normal model are high enough for making soft decisions in VB. When R > 3, however, VB has to rely on hard decisions like k-means, since the standard Normal probabilities are too low. The CVB_3 , in contrast, automatically move the mixture's weights closer to hard decision on the best structures like CVB_2 .

Note that, although the computed ELBO values for CVB_2 in Fig. 13 are correct, the computed ELBO values for CVB_1 and CVB_3 are merely heuristic and not correct values, since their ELBO values are hard to compute in this case. Nonetheless, from their performance in Purity and MSE, we may speculate that the true ELBO values of CVB_1 and CVB_3 are lower and higher than those of CVB_2 , respectively. Equivalently, in terms of KL divergence, the CVB_3 seems to be the best posterior approximation for this independent Normal cluster model, followed by CVB_2 , CVB_1 and mean-field approximations, which yield almost identical ELBO values.

Intuitively, as shown in the case of R=4 in the upper left panel of Fig. 13, the mean-field approximations like VB, EM and k-means seems not to recognize the correlations between data of the same clusters, but focus more on the inter-distance between clusters as a whole. The CVB approximations, in contrast, exploit the correlations between each label l_j to all other labels, as shown in Fig. 12. Although the heuristic CVB₁ becomes worse when R increases, the CVB₂ and CVB₃ are still able to pick the best correlated structures to represent the data. When inter-distance of cluster is much higher than cluster's variance, these two CVB methods stabilize and accurately classify 90% of total data in average. The successful rate is only about 80% for all other state-of-the-art mean-field approximations.

VII. CONCLUSION

In this paper, the independent constraint of mean-field approximations like VB, EM and k-means algorithms has been shown to be a special case of a broader conditional constraint class, namely copula. By Sklar's theorem, which guarantees the existence of copula for any joint distribution, a copula Variational Bayes (CVB) algorithm is then designed in order to minimize the Kullback-Leibler (KL) divergence from the true joint distribution to an approximated copula class. The iterative CVB can converge to the true probability distribution when their copula structures are close to each other. From perspective of generalized Bregman divergence in information geometry, the CVB algorithm and its special cases in mean-field approximations have been shown to iteratively project the true probability distribution to a conditional constraint class until convergence at a local minimum KL divergence.

For a global approximation of a generic probabilistic network, the CVB is then further extended to the so-called augmented CVB form. This global CVB network can be seen as an optimally weighted hierarchical mixture of many local CVB approximations with simpler network structures. By this way, the locally optimal approximation in mean-field methods can be extended to be globally optimal in copula class for the first time. This global property was then illustrated via simulations of correlated bivariate Gaussian distribution and standard Normal clustering, in which the CVB's performance was shown to be far superior to VB, EM and k-means algorithms in terms of percentage of accurate classifications, mean squared error (MSE) and KL divergence. Despite being canonical, these popular Gaussian models illustrated the potential applications of CVB to machine learning and Bayesian network. The application of copula's design in statistics and a faster computational flow for augmented CVB network may be regarded as two out of many promising approaches for improving CVB approximation in future works.

Appendix A

BAYESIAN MINIMUM-RISK ESTIMATION

Let us briefly review the importance of posterior distributions in practice, via minimum-risk property of Bayesian estimation method. Without loss of generalization, let us assume that the unknown parameter θ in our model is continuous. In practice, the aim is often to return estimated value $\hat{\theta} \triangleq \hat{\theta}(\boldsymbol{x})$, as a function of noisy data \boldsymbol{x} , with least mean squared error $\text{MSE}(\hat{\theta}, \theta) \triangleq \mathbb{E}_{f(\boldsymbol{x}, \theta)} || \hat{\theta}(\boldsymbol{x}) - \theta ||^2$, where $|| \cdot ||$ is \mathcal{L}_2 -normed operator. Then, by basic chain rule of probability $f(\boldsymbol{x}, \theta) = f(\theta | \boldsymbol{x}) f(\boldsymbol{x})$, we have [1], [45]:

$$\hat{\theta} \triangleq \underset{\tilde{\theta}}{\operatorname{arg \, min}} \operatorname{MSE}(\tilde{\theta}, \theta)
= \underset{\tilde{\theta}}{\operatorname{arg \, min}} \mathbb{E}_{f(\theta|\boldsymbol{x})} ||\tilde{\theta}(\boldsymbol{x}) - \theta||^{2}
= \mathbb{E}_{f(\theta|\boldsymbol{x})}(\theta),$$
(71)

which shows that the posterior mean $\hat{\theta} = \mathbb{E}_{f(\theta|x)}(\theta)$ is the least MSE estimate. Note that, the result (71) is also a special case of Bregman variance theorem (7) when applied to Euclidean distance (8). In general, we may replace the \mathcal{L}_2 -norm in (71) by other normed functions. For example, it is well-known that the best estimators for the least total variation norm \mathcal{L}_1 and the zero-one loss \mathcal{L}_{∞} are the median and mode of the posterior $f(\theta|x)$, respectively [1], [45].

ACKNOWLEDGEMENT

I am always grateful to Dr. Anthony Quinn for his guidance on Bayesian methodology. He is the best Ph.D. supervisor that I could hope for.

REFERENCES

- V. H. Tran, "Variational Bayes inference in digital receivers," Ph.D. dissertation, Trinity College Dublin, 2014.
- [2] V. Smidl and A. Quinn, The Variational Bayes Method in Signal Processing. Springer, 2006.
- [3] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, Jun. 2003.

- [4] S. Watanabe and J.-T. Chien, Bayesian Speech and Language Processing. Cambridge University Press, 2015.
- [5] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.
- [6] S. M. Stigler, The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, 1986.
- [7] M. Karny and K. Warwick, Eds., Computer Intensive Methods in Control and Signal Processing - The Curse of Dimensionality. Birkhauser, Boston, MA, 1997.
- [8] A. Graves, "Practical variational inference for neural networks," Advances in Neural Information Processing Systems (NIPS), 2011.
- [9] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with Variational Bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, Mar. 2010.
- [10] S. Subedi and P. D. McNicholas, "Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions," *Advances in Data Analysis and Classification*, vol. 8, no. 2, pp. 167– 193, Jun. 2014.
- [11] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, Nov. 2008.
- [12] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.
- [13] A. Sklar, "Fonctions de répartition à N dimensions et leurs marges," Publications de l'Institut Statistique de l'Université de Paris, vol. 8, pp. 229–231, 1959.
- [14] F. Durante and C. Sempi, Principles of Copula Theory. Chapman and Hall/CRC, 2015.
- [15] A. Kolesarova, R. Mesiar, J. Mordelova, and C. Sempi, "Discrete copulas," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 5, pp. 698– 705, Oct. 2006.
- [16] A. Sklar, "Random variables, distribution functions, and copulas a personal look backward and forward," in *Distributions with fixed marginals and related topics*, ser. Lecture Notes-Monograph, L. Ruschendorf, B. Schweizer, and M. D. Taylor, Eds., vol. 28. Institute of Mathematical Statistics, Hayward, CA, 1996, pp. 1–14.
- [17] X. Zeng and T. Durrani, "Estimation of mutual information using copula density function," *IEEE Electronics Letters*, vol. 47, no. 8, pp. 493–494, Apr. 2011.
- [18] S. Grønneberg and N. L. Hjort, "The copula information criteria," Scandinavian Journal of Statistics, vol. 41, no. 2, pp. 436–459, 2014.
- [19] L. A. Jordanger and D. Tjøstheim, "Model selection of copulas AIC versus a cross validation copula information criterion," *Statistics and Probability Letters*, vol. 92, pp. 249–255, Jun. 2014.
- [20] S. ichi Amari, Information Geometry and Its Applications. Springer Japan, Feb. 2016.
- [21] P. Stoica and Y. Selen, "Cyclic minimizers, majorization techniques, and the Expectation-Maximization algorithm - a refresher," *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 112–114, Jan. 2004.
- [22] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, Feb. 2017.
- [23] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. B-48, pp. 259–302, 1986.
- [24] A. Dogandzic and B. Zhang, "Distributed estimation and detection for sensor networks using hidden Markov random field models," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 3200–3215, 2006.
- [25] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
- [26] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for k-means clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, Feb. 2015.
- [27] V. H. Tran, "Cost-constrained Viterbi algorithm for resource allocation in solar base stations," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4166–4180, Apr. 2017.
- [28] T. Jebara and A. Pentland, "On reversing Jensen's inequality," Advances in Neural Information Processing Systems 13 (NIPS), pp. 231–237, 2001.
- [29] P. Carbonetto and N. de Freitas, "Conditional mean field," Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS), pp. 201–208, Dec. 2006.

- [30] E. P. Xing, M. I. Jordan, and S. Russell, "A generalized mean field algorithm for variational inference in exponential families," *Proceedings* of the Nineteenth conference on Uncertainty in Artificial Intelligence (UAI), pp. 583–591, Aug. 2003.
- [31] D. Geiger, C. Meek, and C. Meek, "Structured variational inference procedures and their realizations," *Proceedings of Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.
- [32] D. Tran, D. M. Blei, and E. M. Airoldi, "Copula variational inference," 28th International Conference on Neural Information Processing Systems (NIPS), vol. 2, pp. 3564–3572, Dec. 2015.
- [33] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence and Bayesian estimation of distributions," *IEEE Transactions* on *Information Theory*, vol. 54, no. 11, pp. 5130–5139, Nov. 2008.
- [34] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman voronoi diagrams," Discrete & Computational Geometry (Springer), vol. 44, no. 2, pp. 281–307, Sep. 2010.
- [35] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, Oct. 2005.
- [36] S. ichi Amari, "Divergence, optimization and geometry," *International Conference on Neural Information Processing*, pp. 185–193, 2009.
- [37] M. Adamcík, "The information geometry of Bregman divergences and some applications in multi-expert reasoning," *Entropy*, vol. 16, no. 12, pp. 6338–6381, Dec. 2014.
- [38] M. A. Proschan and P. A. Shaw, Essentials of Probability Theory for Statisticians. Chapman and Hall/CRC, Apr. 2016.
- [39] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," IEEE Transactions on Information Theory, vol. 55, no. 6, pp. 2882– 2904, Jun. 2009.
- [40] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "An introduction to functional derivatives," Department of Electronic Engineering, University of Washington, Seattle, WA, Tech. Rep. 0001, 2008.
- [41] U. Cherubini, E. Luciano, and W. Vecchiato, Copula Methods in Finance. John Wiley & Sons, Oct. 2004.
- [42] J.-F. Mai and M. Scherer, Financial Engineering with Copulas Explained. Palgrave Macmilan, 2014.
- [43] A. Shemyakin and A. Kniazev, Introduction to Bayesian Estimation and Copula Models of Dependence. Wiley-Blackwell, May 2017.
- [44] S. Han, X. Liao, D. Dunson, and L. Carin, "Variational Gaussian copula inference (and supplementary materials)," *Proceedings of the* 19th International Conference on Artificial Intelligence and Statistics, vol. 51, pp. 829–838, May 2016.
- [45] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons Canada, 2006.
- [46] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference - a review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Feb. 2017.
- [47] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, Apr. 2005.
- [48] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, Nov. 2008.
- [49] M. Sugiyama, Introduction to Statistical Machine Learning. Elsevier, 2015.
- [50] D. J. MacKay, Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
- [51] R. Ranganath, D. Tran, and D. M. Blei, "Hierarchical variational models," 33rd International Conference on International Conference on Machine Learning (ICML), vol. 48, pp. 2568–2577, Jun. 2016.
- [52] D. Tran, R. Ranganath, and D. Blei, "Hierarchical implicit models and likelihood-free variational inference," Advances in Neural Information Processing Systems (NIPS), pp. 5529–5539, Dec. 2017.
- [53] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, Boston, 2006.



Viet Hung Tran received the B.Eng. degree from Hochiminh city University of Technology, Vietnam, in 2008, the master's degree from ENS Cachan, Paris, France, in 2009, and the Ph.D. degree from the Trinity College Dublin, Ireland, in 2014. From 2014 to 2016, he held a post-doctoral position with Telecom ParisTech. He is currently a Research Fellow at University of Surrey, London suburb, U.K. His research interest is optimal algorithms for Bayesian learning network and information theory. He was awarded the best mathematical paper prize at IEEE

Irish Signals and Systems Conference, 2011.