CONDITIONALLY FACTORIZED VARIATIONAL BAYES WITH IMPORTANCE SAMPLING

Runze Gan, Simon Godsill

Engineering Department, University of Cambridge, Cambridge, UK, CB2 1PZ

ABSTRACT

Coordinate ascent variational inference (CAVI) is a popular approximate inference method; however, it relies on a mean-field assumption that can lead to large estimation errors for highly correlated variables. In this paper, we propose a conditionally factorized variational family with an adjustable conditional structure and derive the corresponding coordinate ascent algorithm for optimization. The algorithm is termed Conditionally factorized Variational Bayes (CVB) and implemented with importance sampling. We show that by choosing a finer conditional structure, our algorithm can be guaranteed to achieve a better variational lower bound, thus providing a flexible trade-off between computational cost and inference accuracy. The validity of the method is demonstrated in a simple posterior computation task.

***Index Terms*—** variational Bayes, importance sampling, coordinate ascent variational inference, structured variational inference

1. INTRODUCTION

As one of the earliest formulations of variational inference (VI) [1, 2], CAVI (also known as mean-field VI [3, 4]) is a popular method for approximation of intractable posterior densities for Bayesian models. Compared to Markov chain Monte Carlo (MCMC), CAVI is faster, easier to monitor its convergence, and typically the posterior obtained by CAVI can be efficiently summarised with sufficient statistics. For these reasons, CAVI has been widely applied as an alternative strategy to MCMC. For example, in the signal processing field, CAVI has been employed to develop approximate Bayesian filters [5, 6], smoothers [3, 7], and to carry out approximate Bayesian parameter estimation for state space models (SSMs) [3, 8].

Despite the above-mentioned benefits, CAVI relies on the mean-field assumption: it assumes independent fully factorized variational posteriors. Such an assumption renders the approximation inaccurate for highly correlated variables and may introduce additional local optima [9, 2]. To mitigate this problem, in this paper, we present the CVB, a coordinate ascent optimizer that can theoretically yield a better approximation with an adjustable conditional variational distribution preserving the dependence between variables.

Problem formulation Consider a Bayesian model, where the set of all observed variables is denoted as Y, and let X denote the set of parameters and latent variables whose posterior $p(X|Y)$ is of interest but intractable. This paper focuses on approximating this posterior within the VI framework. To this end, we first propose a family of variational distributions q, and then find the member of this family which minimizes the Kullback-Leibler (KL) divergence to the exact posterior. Such a procedure is equivalent to finding

$$q^* = \arg\max_q \mathcal{F}(q), \tag{1}$$

subject to the restriction that q belongs to the predefined family. The evidence lower bound (ELBO) $\mathcal{F}(q)$ in (1) is defined as

$$\mathcal{F}(q) = \mathrm{E}_{q(X)} \log \frac{p(X,Y)}{q(X)}. \tag{2}$$

In this paper, we propose to use a conditionally factorized family for the variational distribution q – a generic and flexible family where the dependence between the desired variables can be constructed with user selected detail. It encompasses the standard mean-field family as a special case.

Background of CAVI Classical CAVI is a VI algorithm where the variational distribution q belongs to a mean-field family q_{mf}. This family assumes that the approximated distribution $q_{mf}(X)$ can be independently factorized as follows,

$$q_{mf}(X) = \prod_{i=1}^{\nu} q^i_{mf}(x_i), \tag{3}$$

where each $x_i (i = 1, 2, ..., \nu)$ is disjointly partitioned from X, i.e. $X = \{x_1, x_2, ..., x_\nu\}$, and each q^i_{mf} is a *free-form* distribution (a variational distribution q^i_{mf} is *free-form* if there is no predefined parametric form for it). With such an assumption in (3), the optimization problem in (1) can be solved by a coordinate ascent algorithm. Specifically, for $i = 1, 2, ..., \nu$, the algorithm iteratively updates $q^i_{mf}(x_i)$ by the optimization in (4) while keeping q^{i-}_{mf} fixed, where $q^{i-}_{mf} = \prod_{l \neq i} q^l_{mf}(x_l)$.

$$\arg\max_{q^i_{mf}} \mathcal{F}(q_{mf}) \propto \exp(\mathrm{E}_{q^{i-}_{mf}(x_{i-})} \log p(X,Y)) \tag{4}$$

This update is a fundamental result based on variational calculus, see [1, 4, 3]. As the name CAVI suggests, each update step is coordinate-wise and guarantees a non-negative increment of the ELBO, such that (2) will eventually reach a local optimum.

In some cases, the optimal distribution in (4) may not be analytically tractable. A remedy to this issue is to sample from intractable variational distributions, thereby using Monte Carlo methods to approximate the other optimal distributions in closed-form expressions. For example, to carry out VI for models with constrained prior, [10] uses MCMC to obtain samples for the intractable update in the standard CAVI procedure; and, [8] applies sequential Monte Carlo within CAVI procedure to estimate the parameters of a nonlinear SSM. Approximating intractable coordinate ascent updates with Monte Carlo methods is also a key concept utilised in the implementation of the proposed CVB algorithm.

Related work and contribution Restoring the dependence within the variational distributions has been studied since the early days of the VI method [11]. This is known as structured VI in the machine learning community, see [2, 1, 9]. Early research on structured VI is often model-specific and limited to the graphical model setting [11, 12]. Attempts to improve the mean-field approximation in a general probabilistic model, to our knowledge, began with [13], whose parameters are numerically optimized to increase the ELBO. More recently, the conditional variational distributions employed in [9, 14] are more closely related to our work; the conditional structure they implicitly assume (termed *conditional everywhere* in our paper)

can be regarded as a limiting case of the *setwise conditional* structure, which we propose and define in our paper for the conditionally factorized family. Moreover, whilst [9, 14] focus on deriving the update for stochastic VI , we consider deriving the coordinate-ascent update for *free-form* variational distributions.

Our main contribution is to present a conditionally factorized variational family that can flexibly account for the conditional structure between variables, encompassing the standard mean-field family as a special case. The resulting coordinate ascent updates are derived and approximated with importance sampling. The guaranteed performance improvement compared to the standard mean-field CAVI is discussed and demonstrated in a simple example.

2. CONDITIONALLY FACTORIZED VARIATIONAL BAYES

Consider the problem formulated in Section 1. Assume X includes at least two variables. We partition it into two disjoint parts, denoted by the global variable S and local variable Z (i.e. $X = \{S, Z\}$). If required, the local variable Z can be further partitioned into N_c ($N_c \geq 1$) variables, i.e. $Z = \{Z_1, Z_2, ..., Z_{N_c}\}$. Note that we do not impose any restriction on the type of the global or local variables: they can be continuous, discrete or mixed type.

Conditionally factorized variational family A member of the proposed conditionally factorized family should first satisfy the following factorization:

$$q(X) = q_g(S) \prod_{i=1}^{N_c} q_c^i(Z_i|S). \tag{5}$$

Note that our CVB method also applies to $N_c = 1$, whereas a factorization may improve the tractability of variational distributions.

Denote the set of all factorized variational distributions in (5) by $\{q\}_{cf} = \{q_g, q_c^1, ..., q_c^{N_c}\}$. No predefined relationship is assumed between distributions in $\{q\}_{cf}$ and thus coordinate-wise update for optimization in (1) may be carried out, i.e. one of the distributions in $\{q\}_{cf}$ is updated whilst the others are kept fixed. Furthermore, the global distribution q_g is a regular *free-form* variational distribution in a similar sense to the *free-form* distribution in (3). Each conditional variational distribution q_c^i is a function of both S and Z_i defined as follows. Denote the domain of S as U (i.e. $S \in U$).

Definition 2.1. $q_c^i(Z_i|S)$ is *setwise conditional* on a partition $\mathcal{P}_i$ if 1) $\mathcal{P}_i$ is a partition [15] of U; and 2) the distribution $q_c^i(Z_i|S)$ can be written as

$$q_c^i(Z_i|S) = \sum_{A \in \mathcal{P}_i} q_c^{i,A}(Z_i) I(S \in A), \tag{6}$$

where $I(S \in A)$ is the indicator function; $\{q_c^{i,A}(Z_i) : A \in \mathcal{P}_i\}$ consists of $|\mathcal{P}_i|$ *free-form* variational distributions of Z_i which have no predefined relationship with each other.

Remark. The ***conditional everywhere*** is a limiting case of *setwise conditional* where the partition $\mathcal{P}_i$ is the partition of singleton, i.e. $\mathcal{P}_i = \{\{u\} : u \in U\}$. In this case $|\mathcal{P}_i| = \infty$ if S includes a continuous variable. Another special case of *setwise conditional* is the partition $\mathcal{P}_i$ being the trivial partition, i.e. $\mathcal{P}_i = \{U\}$ when q_c^i no longer depends on S, and the proposed conditional mean-filed assumption in (5) degenerates to the standard mean-field assumption in (3) if $\mathcal{P}_i = \{U\}$ for all $i = 1, 2, ..., N_c$.

We now complete the definition of the conditionally factorized family. Each a realization of $\{q\}_{cf}$ that satisfies the above definition is a member of the conditionally factorized family. Given the factorized variables $\{S, Z_1, ..., Z_{N_c}\}$, the proposed conditionally factorized family is only parameterized by the partitions introduced in the Definition 2.1, i.e. $\{\mathcal{P}_i : i = 1, 2, ..., N_c\}$.

It is informative to consider how the conditionally factorized family is affected by the associated partitions. Specifically, it can be easily verified that for variables $\{S, Z_1, ..., Z_{N_c}\}$, one conditionally factorized family is *wider* than another (each member of the latter is a member of the former) if each associated partition $\mathcal{P}_i (i = 1, 2, ..., N_c)$ of the former family is *finer* than the corresponding partition of the latter (every set in the former partition is a subset of a set in the latter partition). Therefore, the conditionally factorized family with *conditional everywhere* q_c^i for all $i = 1, 2, ...N_c$ is the *widest* family of the proposed framework, since each q_c^i has the *finest* partition; also the proposed conditionally factorized family is always *wider* than the fully factorized mean-field family for variables $\{S, Z_1, ..., Z_{N_c}\}$, since the latter can be considered as a special case of the conditionally factorized family that assumes the trivial partitions for all q_c^i.

Coordinate ascent update The aim of VI is to find the member of the proposed conditionally factorized family that solves the optimization problem in (1). A *wider* family guarantees a better global optimum, but it may take more effort to search for it. Here a coordinate ascent algorithm is sought for the optimization. Specifically, the parametric form of each distribution from $\{q\}_{cf}$ is iteratively updated (whilst other variational distributions are kept fixed) by the optimal distribution given below. Denote the optimal $q_c^i(Z_i|S)$ by $\tilde{q}_c^i(Z_i|S) = \arg\max_{q_c^i} \mathcal{F}(\{q\}_{cf})$ for $i = 1, 2, ..., N_c$, and the optimal $q_g(S)$ by $\tilde{q}_g(S) = \arg\max_{q_g} \mathcal{F}(\{q\}_{cf})$, then

$$\tilde{q}_g(S) \propto \frac{\exp\left(\mathrm{E}_{\prod_{i=1}^{N_c} q_c^i(Z_i|S)} \log p(X, Y)\right)}{\prod_{i=1}^{N_c} \exp\left(\mathrm{E}_{q_c^i(Z_i|S)} \log q_c^i(Z_i|S)\right)}. \tag{7}$$

For each q_c^i which is *setwise conditional* on a partition $\mathcal{P}_i$, we have

$$\tilde{q}_c^i(Z_i|S) = \sum_{A \in \mathcal{P}_i} \tilde{q}_c^{i,A}(Z_i) I(S \in A),$$

$$\tilde{q}_c^{i,A}(Z_i) \propto \exp\left(\frac{\int_A q_g(S) \mathrm{E}_{q_c^{i-}(Z_{i-}|S)} \log p(X, Y) dS}{\int_A q_g(S) dS}\right), \tag{8}$$

and for each q_c^i which is *conditional everywhere*, we have

$$\tilde{q}_c^i(Z_i|S = u) \propto \exp\left(\mathrm{E}_{q_c^{i-}(Z_{i-}|S=u)} \log p(Z, Y, S = u)\right) \tag{9}$$

for any $u \in U$. The integral in (8) can be replaced by a summation for the discrete part of S. The $q_c^{i-}(Z_{i-}|S)$ in (8) and (9) denotes $\prod_{l \neq i} q_c^l(Z_l|S)$ for $N_c > 1$; and when $N_c = 1$, the operator $\mathrm{E}_{q_c^{i-}}$ should be neglected, i.e. $\mathrm{E}_{q_c^{i-}(Z_{i-}|S)} \log p(X, Y) = \log p(X, Y)$. The laborious derivation for (7)-(9) will be presented in full paper.

For any assumed factorized variables $\{S, Z_1, ..., Z_{N_c}\}$, the theoretical coordinate ascent algorithm, denoted as theoretical CVB, is summarized in Algorithm 1. As with the standard CAVI, the ELBO is guaranteed to monotonically increase over iterations of the theoretical CVB. This is due to: 1) each update in the Algorithm 1 is guaranteed to increase the ELBO since they are derived from the exact $\arg\max \mathcal{F}\{q\}_{cf}$ operator, and 2) the refinement step can keep the current ELBO since the resulting *wider* conditionally factorized family includes the latest updated parametric form of $\{q\}_{cf}$ as a member (as discussed above), and allows the ELBO to be optimized in a *wider* variational family. When all the partitions $\mathcal{P}_i, i = 1, 2, ..., N_c$ are fixed, the theoretical CVB will converge to a local optimum of the problem in (1).

We can now see why the theoretical CVB can be guaranteed to improve the performance of the standard CAVI. As the standard CAVI can be considered as the CVB with the trivial partitions, when such a CAVI has converged, we can always refine the partitions

Algorithm 1: Theoretical CVB

while *the ELBO* $\mathcal{F}(\{q\}_{cf})$ *not converged* **do**
  Update $q_g \leftarrow \tilde{q}_g$ according to (7).
  for $i = 1 : N_c$ **do**
    Refine the $\mathcal{P}_i$ if higher accuracy is required.
    Update $q_c^i \leftarrow \tilde{q}_c^i$ according to (8) or (9) with the $\mathcal{P}_i$.

within the proposed CVB framework and the ELBO is guaranteed to monotonically increase again. Therefore, the theoretical CVB with *conditionally everywhere* q_c^i which assumes the *finest* partition for all $i = 1, 2, ..., N_c$ is the most accurate setting of the proposed conditionally factorized family; however, it comes with the most intensive computations as there can be an infinite number of *free-form* distributions (i.e. $q_c^i(Z|S = u)$ for all $u \in U$) to be optimized. On the other hand, the *setwise conditional* q_c^i offers a flexible trade-off between accuracy and computational efficiency, i.e. a *finer* partition $\mathcal{P}_i$ can always yield a higher accuracy with more *free-form* distributions (i.e. $q_c^{i,A}$ for $A \in \mathcal{P}_i$) to be optimized.

3. IMPORTANCE SAMPLING BASED CVB

The implementation of Algorithm 1 is highly intractable for a general continuous variable S. Therefore, S is sampled from the optimal $\tilde{q}_g$ in (7), and these samples are then used to approximate the optimal update for q_c^i. The importance sampling is employed to carry out the required Monte Carlo approximation due to its relatively higher efficiency (compared to MCMC) and its ability to estimate the ELBO.

Algorithm We assume S is a low-dimensional variable such that a standard importance sampling can effectively carry out the sampling. Note that this framework can be extended to incorporate a high dimensional S in a sequential Monte Carlo scheme, and this case will be presented in future work. Suppose we have N_p particles sampled independently from the proposal $\lambda(S)$. We define the particle index set by $\mathcal{I} = \{1, 2, ..., N_p\}$, and denote each particle as $S^{(p)}$ where $p \in \mathcal{I}$. These particles can be uniquely partitioned according to the partition $\mathcal{P}_i$ introduced for a *setwise conditional* q_c^i in Definition 2.1. Specifically, for each $A \in \mathcal{P}_i$, the set of labels of the particles that lie in the region A can be denoted by $F(A) = \{p \in \mathcal{I} : S^{(p)} \in A\}$, and the partition of $\mathcal{I}$ which is resulted from $\mathcal{P}_i$ can be denoted by $\mathcal{S}_i = \{F(A) : A \in \mathcal{P}_i \wedge F(A) \neq \emptyset\}$. Note that by such a definition, the *conditional everywhere* q_c^i which assumes the *finest* $\mathcal{P}_i$ corresponds to the *finest* index partition $\mathcal{S}_i = \{\{p\} : p \in \mathcal{I}\}$.

The unnormalized weight $\tilde{\omega}$ between the optimal global distribution $\tilde{q}_g$ and the proposal λ, according to (7), can be defined as

$$\tilde{\omega}^{(p)} = \frac{\exp\left(\mathrm{E}_{\prod_{i=1}^{N_c} q_c^i(Z_i|S^{(p)})} \log p(S^{(p)}, Z, Y)\right)}{\lambda(S^{(p)}) \prod_{i=1}^{N_c} \exp\left(\mathrm{E}_{q_c^i(Z_i|S^{(p)})} \log q_c^i(Z_i|S^{(p)})\right)}. \quad (10)$$

By normalizing the weight $\omega^{*(p)} = \tilde{\omega}^{(p)} / \sum_p \tilde{\omega}^{(p)}$, we can approximate the $\tilde{q}_g(S)$ in (7) by an empirical distribution: $\tilde{q}_g(S) \approx \sum_{p=1}^{N_p} \omega^{*(p)} \delta(S^{(p)})$. The q_g update step in the theoretical CVB (Algorithm 1) can then be approximated by updating the $\tilde{\omega}^{(p)}$ according to (10) for all particles $S^{(p)}$.

We now consider the update for $q_c^i(Z_i|S)$. Instead of evaluating $q_c^i(Z_i|S)$ for all possible values of S, we only tackle the associated local distribution $q_c^i(Z_i|S = S^{(p)})$ for each particle $S^{(p)}$ since they are sufficient to calculate the weight in (10) to update q_g, and to produce the required posterior approximation, e.g. (12). Now for a particular particle $S^{(p)}$, suppose A is the set member of region partition $\mathcal{P}_i$ that includes $S^{(p)}$, i.e. $S^{(p)} \in A \in \mathcal{P}_i$, and B is the corresponding set member of index partition $\mathcal{S}_i$ that includes p , i.e. $p \in B \in \mathcal{S}_i$. Depending on the definition of q_c^i, the optimal distribution $\tilde{q}_c^i(Z_i|S = S^{(p)})$ is either (9) or equal to the $\tilde{q}_c^{i,A}(Z_i)$ in (8), where the latter is often analytically intractable. Here we approximate the optimal distribution $\tilde{q}_c^i(Z_i|S = S^{(p)})$ by $\hat{q}_c^{i,B}(Z_i)$ where

$$\hat{q}_c^{i,B}(Z_i) \propto \exp\left(\frac{\sum_{j \in B} \tilde{\omega}^{(j)} \mathrm{E}_{q_c^{i-}(Z_{i-}|S^{(j)})} \log p(S^{(j)}, Z, Y)}{\sum_{j \in B} \tilde{\omega}^{(j)}}\right). \quad (11)$$

Observe that $\hat{q}_c^{i,B}$ in (11) is the exact optimal $\tilde{q}_c^i(Z_i|S = S^{(p)})$ in (9) if q_c^i is *conditional everywhere* such that $B = \{p\}$; otherwise $\hat{q}_c^{i,B}(Z_i)$ is an approximation of $\tilde{q}_c^{i,A}(Z_i)$ in (8), where the exponent of (8) is approximated within the importance sampling framework.

Finally, we summarize the importance sampling based CVB (denoted as IS-CVB) in Algorithm 2, and the output can be extracted to produce the required approximated posterior, e.g. the joint distribution $p(Z_1, Z_2)$ can be approximated as the following mixture distributions where the dependence between Z_1 and Z_2 are retained.

$$p(Z_1, Z_2) \approx \sum_{p=1}^{N_p} \omega^{*(p)} q_c^1(Z_1|S^{(p)}) q_c^2(Z_2|S^{(p)}). \quad (12)$$

Algorithm 2: IS-CVB

Require: Particles $S^{(p)} \sim \lambda(S)$ for all $p \in \mathcal{I}$, maximum iteration limit M, tolerance threshold $\epsilon > 0$, initial index partitions $\mathcal{S}_i (i = 1 : N_c)$ and initial distributions $\{q\}_{cf}$.
Output: $\tilde{\omega}^{(p)}, q_c^i(Z_i|S^{(p)})$ for all $p \in \mathcal{I}$.
for $k = 1, 2, ..., M$ **do**
  Update $\tilde{\omega}^{(p)}$ according to (10) for all $p \in \mathcal{I}$.
  Evaluate the estimated ELBO $\hat{\mathcal{F}}_k$ according to (13).
  if $(\hat{\mathcal{F}}_k - \hat{\mathcal{F}}_{k-1}) < \epsilon \wedge k \geq 2$ **then**
    break
  for $i = 1, 2, ..., N_c$ **do**
    Refine the $\mathcal{S}_i$ if higher accuracy is required.
    for $B \in \mathcal{S}_i$ **do**
      Evaluate $\hat{q}_c^{i,B}$ according to (11).
      Set $q_c^i(Z_i|S^{(p)}) \leftarrow \hat{q}_c^{i,B}$ for all $p \in B$.

Estimated ELBO The IS-CVB (Algorithm 2) can offer an estimation of the exact ELBO as a by-product. Specifically, the exact ELBO can be biasedly estimated by $\hat{\mathcal{F}}$, which is defined as follows:

$$\hat{\mathcal{F}} = \log\left(\frac{1}{N_p} \sum_{p=1}^{N_p} \tilde{\omega}^{(p)}\right). \quad (13)$$

Theorem 3.1. *Suppose* $q_c^1, q_c^2, ..., q_c^{N_c}$ *are the local variational distributions which are employed in* (10) *to compute the* $\hat{\mathcal{F}}$ *in* (13)*, and suppose* $\tilde{q}_g(S)$ *is the exact optimal distribution computed according to* (7) *with the same variational distributions* $q_c^{1:N_c}$*, then we have*

$$\mathrm{E}_{\prod_p \lambda(S^{(p)})} \exp\left(\hat{\mathcal{F}}\right) = \exp\left(\mathcal{F}\left(\tilde{q}_g, q_c^1, q_c^2, ..., q_c^{N_c}\right)\right),$$
$$\mathrm{E}_{\prod_p \lambda(S^{(p)})} \hat{\mathcal{F}} \leq \mathcal{F}\left(\tilde{q}_g, q_c^1, q_c^2, ..., q_c^{N_c}\right),$$

where $\mathcal{F}$ *is the exact ELBO defined in* (2)*.*

The proof will be presented in a full paper. Theorem 3.1 suggests that the average of unnormalized weights is an unbiased estimate of the exponential of exact ELBO, and $\hat{\mathcal{F}}$ can be viewed as an unbiased estimate of exact ELBO's lower bound. Since now the exact ELBO $\mathcal{F}$ is intractable, we will instead monitor the $\hat{\mathcal{F}}$ in (13) to assess the convergence. It is well-known that the monotonically increasing property of the exact ELBO produced by the standard CAVI can be used to design the algorithm termination condition and to check the implementation of the algorithm. In fact, the $\hat{\mathcal{F}}$ produced by Algorithm 2 has similar monotonically increasing and convergent

properties, see Theorem 3.2. Therefore, a tolerance threshold of the increment of $\hat{\mathcal{F}}$ is set in Algorithm 2 to determine whether the convergence is reached. Moreover, it is useful to check whether the $\hat{\mathcal{F}}$ produced by Algorithm 2 always satisfies $\hat{\mathcal{F}}_k - \hat{\mathcal{F}}_{k-1} \geq 0$ for all $k \geq 2$; if not, then the implementation of Algorithm 2 is incorrect.

Theorem 3.2. *1) The estimated ELBO in Algorithm 2 is monotonically increasing across iterations, i.e.* $\hat{\mathcal{F}}_k - \hat{\mathcal{F}}_{k-1} \geq 0$ *for all* $k \geq 2$. *2) Assume the Algorithm 2 will not be terminated (e.g. set* $M = \infty, \epsilon < 0$*), the sequence* $(\hat{\mathcal{F}}_k)_{k\in\mathbb{N}}$ *is convergent if index partitions* $\mathcal{S}_i (i = 1, 2, ..., N_c)$ *are fixed after some iteration number.*

The elaborate proof will be presented in a full paper. Similar to the performance improvement achieved by the theoretical CVB (Algorithm 1), the IS-CVB (Algorithm 2) can also guarantee a higher $\hat{\mathcal{F}}$ compared to a more standard Monte Carlo-based mean-field CAVI. The latter can be considered as a special case of the Algorithm 2 where $\mathcal{S}_i = \{\mathcal{I}\}$ for all $i = 1, 2, ..., N_c$. When the produced $\hat{\mathcal{F}}$ in such a setting is converged (convergence is guaranteed by the Theorem 3.2), we can refine the $\mathcal{S}_i$ within the proposed CVB framework. This will increase the $\hat{\mathcal{F}}$ again according to the Theorem 3.2. The highest $\hat{\mathcal{F}}$ can be achieved by $\mathcal{S}_i$ being the *finest* partition, i.e. the *conditional everywhere* q_c^i. However, each *conditional everywhere* q_c^i requires computing N_p *free-form* variational distributions in the Algorithm 2. When the computational power is limited, the more flexible *setwise conditional* q_c^i is favored as it can achieve a competitive performance with fewer *free-form* variational distributions to evaluate by using a wisely chosen partition $\mathcal{S}_i$.

4. SIMULATION RESULT

In this section, we verify the proposed IS-CVB on a simple example. Consider the following Bayesian model:

$$p(\mu, \nu) = \mathcal{N}(\mu; 0, 2)\mathcal{G}(\nu; 2, 0.5),$$
$$p(x|\mu, \nu) = \mathcal{N}\left(x; \mu, \nu^{-1}\right), \quad p(y|x) = \mathcal{N}\left(y; 0.5x^2, 1 + e^{10x}\right),$$

where $\mathcal{G}(k, \theta)$ is the gamma distribution with shape parameter k and scale parameter θ. We are interested in the following two posteriors: 1) $p(x, \mu, \nu|y = 8)$, and 2) $p(x, \mu, \nu|y = 25)$. As the exact posteriors are analytically intractable, we approximate them within the proposed CVB framework. We consider the conditionally factorized families which satisfy the factorization $q(x, \mu, \nu) = q_g(x)q_c^\mu(\mu|x)q_c^\nu(\nu|x)$, and adopt the following five different partitions $\mathcal{P}_\mu, \mathcal{P}_\nu$ for q_c: **(i)** fully factorized: $\mathcal{P}_\mu = \{\mathbb{R}\}, \mathcal{P}_\nu = \{\mathbb{R}\}$; **(ii)** jointly factorized 1: *conditional everywhere* q_c^μ and $\mathcal{P}_\nu = \{\mathbb{R}\}$; **(iii)** jointly factorized 2: $\mathcal{P}_\mu = \{\mathbb{R}\}$ and *conditional everywhere* q_c^ν; **(iv)** setwise conditional: $\mathcal{P}_\mu = \mathcal{P}_\nu = \{\mathbb{R}_{\geq 0}, \mathbb{R}_{<0}\}$; **(v)** conditional everywhere: *conditional everywhere* q_c^μ and q_c^ν. Note that the conditionally factorized families with settings (i)-(iii) degenerate to the mean-field families. The above five settings are implemented in IS-CVB without refinement steps, i.e. the partitions are fixed all the time. We also consider another two settings: **(vi)**: initialized with (i), switches to (iii), then to (v) if converged; **(vii)**: initialized with (i), switches to (iv), then to (v) if converged, where both are initialized with the setting (i), and then switch to a setting with *finer* partitions twice if the algorithm has converged.

All settings (i)-(vii) are implemented by using IS-CVB with the same particle set ($N_p = 1000$) and the same initialized distributions. The proposal we use is $\lambda(x) = 0.3\mathcal{N}(x; m_1, c_1) + 0.7\mathcal{N}(x; 1, 1.5)$, where $m_1 = -\sqrt{2y}, c_1 = 1 + e^{10m_1}$. This proposal is designed to cover the possible range of x that can generate the given observation y. The implementation of IS-CVB with the settings (i)-(iii) are simply Monte Carlo based mean-field CAVI, and the standard CAVI is analytically intractable for this model.

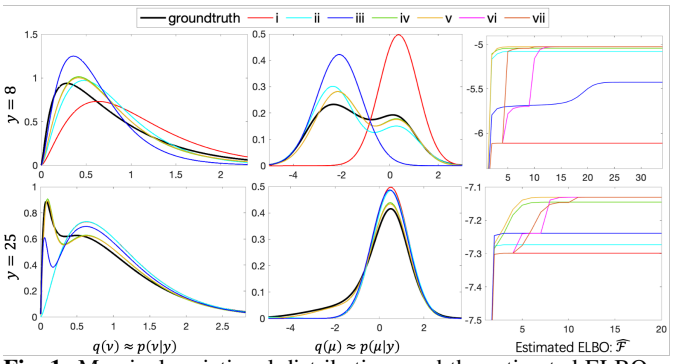

Fig. 1: Marginal variational distributions and the estimated ELBO. The groundtruth is obtained using a numerical grid-based method.

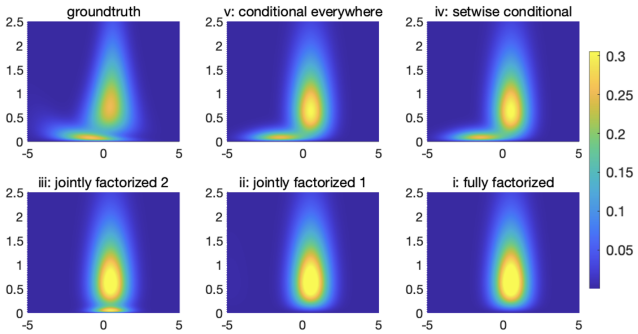

Fig. 2: Joint variational distributions $q(\mu, \nu) \approx p(\mu, \nu|y = 25)$. The groundtruth is obtained using a numerical grid-based method.

The marginal variational distribution $q(\mu)$ and $q(\nu)$ for all the settings are shown in the Figure 1, except (vi) and (vii) as they are highly overlapped with setting (v). We can see that the proposed conditionally factorized families (iv)-(v) effectively capture the multimodal behaviour, and produce better fitted marginal approximations than standard mean-field settings (i)-(iii) for all considered distributions. The advantage of our method is further supported by their higher estimated ELBO $\hat{\mathcal{F}}$ in Figure 1. The stepped curve of the $\hat{\mathcal{F}}$ produced by setting (vi) and (vii) satisfies Theorem 3.2, and guarantees a higher $\hat{\mathcal{F}}$ produced with the proposed CVB method compared to standard Monte Carlo mean-field CAVI. Furthermore, the joint variational distributions $q(\mu, \nu)$ which approximate the posterior with observation $y = 25$ (evaluated via (12)) are plotted in the Figure 2, where the dependence between μ and ν is clearly retained in the proposed conditionally factorized family, i.e. settings (iv) and (v), and in contrast lost in other mean-field settings, i.e. (i)-(iii). Finally from Figure 1 and Figure 2, the *setwise conditional* setting (iv) achieves a competitive performance as the most accurate conditional *conditional everywhere* setting. We emphasise that it only requires the evaluation of 4 *free-form* optimal variational distributions (two for each q_c) in the implementation of IS-CVB, which is even less than the jointly factorized mean-field cases (ii) and (iii) where N_p+1 *free-form* variational distributions need to be determined.

5. CONCLUSION

This paper presents a generic variational family that can account for the dependence between variables with user selected detail. The resulting CVB algorithm offers a flexible trade-off between inference accuracy and computational cost. Although the algorithm is only demonstrated here with a simple posterior computation task, we have also developed it into a full filtering and online parameter estimation strategy for time series in a conditionally Gaussian system (e.g. [16, 17, 18, 19]), results of which will be reported in a full paper.

6. REFERENCES

[1] 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, 2017.

[2] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[3] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] V. Smidl and A. Quinn, "Variational Bayesian filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 5020–5030, 2008.

[6] S. Sarkka and A. Nummenmaa, "Recursive noise adaptive Kalman filtering by variational Bayesian approximations," *IEEE Transactions on Automatic control*, vol. 54, no. 3, pp. 596–600, 2009.

[7] T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson, "Approximate Bayesian smoothing with unknown process and measurement noise covariances," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2450–2454, 2015.

[8] C. Li and S. J. Godsill, "Variational parameter learning in sequential state-space model via particle filtering," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5589–5593.

[9] M. D. Hoffman and D. M. Blei, "Structured stochastic variational inference," in *Artificial Intelligence and Statistics*, 2015, pp. 361–369.

[10] L. Ye, A. Beskos, M. De Iorio, and J. Hao, "Monte Carlo co-ordinate ascent variational inference," *Statistics and Computing*, pp. 1–19, 2020.

[11] L. K. Saul and M. I. Jordan, "Exploiting tractable substructures in intractable networks," *Advances in neural information processing systems*, pp. 486–492, 1996.

[12] P. Carbonetto and N. De Freitas, "Conditional mean field," in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*. MIT Press, 2007, vol. 19, p. 201.

[13] T. S. Jaakkola and M. I. Jordan, "Improving the mean field approximation via the use of mixture distributions," in *Learning in graphical models*, pp. 163–173. Springer, 1998.

[14] J. Y. Liu and X. Qiao, "Conditional variational inference with adaptive truncation for Bayesian nonparametric models," *arXiv preprint arXiv:2001.04508*, 2020.

[15] P. J. Cameron, *Naïve set theory*, pp. 1–36, Springer London, London, 1998.

[16] R. Gan and S. J. Godsill, "α-stable Lévy state-space models for manoeuvring object tracking," *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pp. 1–7, 2020.

[17] Q. Li, B. I. Ahmad, and S. J. Godsill, "Sequential dynamic leadership inference using Bayesian Monte Carlo methods," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 4, pp. 2039–2052, 2021.

[18] R. Gan, J. Liang, B. I. Ahmad, and S. J. Godsill, "Bayesian intent prediction for fast maneuvering objects using variable rate particle filters," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[19] R. Gan, B. I. Ahmad, and S. J. Godsill, "Lévy state-space models for tracking and intent prediction of highly maneuverable objects," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 4, 2021.