Computational Statistics (2021) 36:1901–1930
https://doi.org/10.1007/s00180-021-01063-w

ORIGINAL PAPER

Variational approximation for importance sampling

Xiao Su[1] · **Yuguo Chen**[2]

Received: 13 August 2019 / Accepted: 5 January 2021 / Published online: 13 January 2021

Abstract
We propose an importance sampling algorithm with proposal distribution obtained from variational approximation. This method combines the strength of both importance sampling and variational method. On one hand, this method avoids the bias from variational method. On the other hand, variational approximation provides a way to design the proposal distribution for the importance sampling algorithm. Theoretical justification of the proposed method is provided. Numerical results show that using variational approximation as the proposal can improve the performance of importance sampling and sequential importance sampling.

Keywords f-divergence · Monte Carlo · Proposal distribution · Variational inference

1 Introduction

Monte Carlo methods, such as importance sampling (IS) and Markov chain Monte Carlo (MCMC), are widely used in Bayesian inference when analytical computation based on the posterior distribution is difficult. The posterior distributions are sometimes hard to sample directly, especially for complex statistical models with both unknown parameters and latent variables. In that case, importance sampling draws samples from an easy-to-sample proposal distribution, and then corrects the bias by the importance weights. Choosing a good proposal distribution is essential to the efficiency of importance sampling algorithms. We often try to find a proposal distribution that is close to the target distribution to reduce the variance of the importance weight. For high

This work was supported in part by National Science Foundation Grant DMS-2015561.

✉ Yuguo Chen
yuguo@illinois.edu

Xiao Su
xaosu@amazon.com

[1] Amazon, Seattle, WA 98109, USA

[2] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

dimensional problems, sequential importance sampling (SIS) (Liu and Chen 1998; Doucet et al. 2000) gives a way to construct the proposal distribution sequentially.

Variational Bayes (VB) (Jordan et al. 1999) tackles the problem in a different way by deriving a tractable approximation to the posterior distribution. It minimizes the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) between the posterior and the variational approximation, and uses the variational approximation to make inference. In the optimization part, VB algorithm usually uses stochastic optimization (Robbins and Monro 1951) or coordinate optimization strategy. This method is also related to the EM algorithm (Dempster et al. 1977). VB, IS and MCMC can be used for general computational problems, but in this paper we focus on the application in Bayesian settings to make the discussion more concrete.

An advantage of VB is the variational approximation can be obtained quickly, and it usually runs faster than Monte Carlo sampling algorithms such as MCMC. The variational method has been applied in many fields, such as computational biology (Sanguinetti et al. 2006), network data analysis (Hofman and Wiggins 2008; Zreik et al. 2017; O'Hagan and White 2019), natural language processing (Blei et al. 2003), and statistical inference (Armagan and Dunson 2011; Depraetere and Vandebroek 2017). However, one issue with the variation method is that the gap between the variational approximation and the true posterior distribution may lead to biased inference based on variational approximation. In many problems, the estimate based on variation approximation may not be consistent. Also the uncertainty of the VB estimate is not available.

In this paper, we consider using the variational approximation as the proposal distribution for importance sampling, and then using the importance weight to correct the bias. Since the importance sampling estimate is consistent under mild conditions, the bias issue of VB is resolved. The uncertainty of the importance sampling estimate is also relatively easy to obtain. In the meantime, since the variational approximation is close to the true posterior distribution and is usually easy to sample, it is a good choice for the importance sampling proposal distribution. So this idea combines the strength of these two methods. We will provide theoretical justification of the proposed method using the f-divergence (Ali and Silvey 1966), and implement the proposed methods on several models to demonstrate its performance in practice.

The paper is organized as follows. We first review importance sampling and variational approximation in Sect. 2, and introduce the new method in Sect. 3. Then, we provide theoretical justification in Sect. 4, and give numerical results of the new method on several examples in Sect. 5. Section 6 concludes with a discussion.

2 Literature review

2.1 Importance sampling

Suppose $\mathbf{Z}$ is a random vector with probability density function $p(\mathbf{z})$, and we want to estimate the expectation of some function $h(\mathbf{Z})$:

$$\boldsymbol{\mu} = E_p(h(\mathbf{Z})) = \int h(\mathbf{z})p(\mathbf{z})d\mathbf{z}.$$

If $p(\mathbf{z})$ is hard to sample directly, we may consider importance sampling (IS) to generate samples from a proposal distribution $q(\mathbf{z})$. Then the expectation $\boldsymbol{\mu}$ can be estimated by the weighted average

$$\tilde{\boldsymbol{\mu}} = \frac{w(\mathbf{z}^{(1)})h(\mathbf{z}^{(1)}) + \cdots + w(\mathbf{z}^{(m)})h(\mathbf{z}^{(m)})}{w(\mathbf{z}^{(1)}) + \cdots + w(\mathbf{z}^{(m)})}, \tag{1}$$

where $w(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)})/q(\mathbf{z}^{(i)})$ are the importance weights. The estimate $\tilde{\boldsymbol{\mu}}$ is consistent, and it can also handle densities that are only known up to normalizing constants.

The standard error of $\tilde{\boldsymbol{\mu}}$ can be used to measure the efficiency of the IS algorithm. Another criterion is the effective sample size (ESS) (Kong et al. 1994; Kong 1992; Martino et al. 2017):

$$\text{ESS} = \frac{m}{1 + cv^2},$$

where the coefficient of variation (cv) is defined as:

$$cv^2 = \frac{Var_q[w(\mathbf{Z})]}{E_q^2[w(\mathbf{Z})]}.$$

The ESS roughly approximates the number of independent and identically distributed (i.i.d.) samples these m importance samples are equivalent to. Thus, a smaller cv^2 indicates that the IS algorithm is more effective in terms of the ESS. In addition, the cv^2 is also the χ^2 distance between the proposal distribution $q(\mathbf{z})$ and the target distribution $p(\mathbf{z})$, defined as

$$\chi^2(p||q) = \int \frac{(p-q)^2}{q} d\mathbf{z},$$

and this will be used later in our theoretical justification.

For high dimensional problems, it is often hard to find a good proposal for IS. To overcome this difficulty, Liu and Chen (1998) and Doucet et al. (2000) provided the general framework of sequential importance sampling (SIS) to build up the proposal $q(\mathbf{z})$ sequentially. For a d-dimensional vector $\mathbf{z} = (z_1, \ldots, z_d)$, the proposal

distribution can be decomposed as:

$$q(\mathbf{z}) = q_1(z_1)q_2(z_2|z_1)\cdots q_d(z_d|z_1,\ldots,z_{d-1}).$$

Each proposal distribution in the decomposition is for a low dimensional component, so it is relatively easier to design a good proposal. The target distribution $p(\mathbf{z})$ can be decomposed in a similar way by using auxiliary distributions to guide the choice of the proposal distribution (Liu and Chen 1998). The importance weight can also be computed recursively based on the decomposition. SIS has been successfully applied to many problems, including the filtering problem in hidden Markov models (or state space models).

Another variation of IS is adaptive importance sampling (AIS) (Cappé et al. 2004, 2008; Bugallo et al. 2017), which provides a scheme to find a good proposal distribution adaptively based on samples in previous steps. For multi-modal distributions, Owen (2013) suggested using mixture importance sampling as a way to carry out AIS. However, AIS does not work well for high dimensional distributions without incorporating an additional MCMC layer, and the computation time of AIS is usually much longer than importance sampling (Bugallo et al. 2017).

2.2 Variational approximation

Variational Bayesian method (Jordan et al. 1999) is a technique for approximating the intractable integrals in Bayesian inference. It is typically useful when the statistical models are relatively complex with a lot of parameters and latent variables. In Bayesian inference, suppose we have a set of n i.i.d. data $\mathbf{x}$, and all latent variables and parameters are denoted by $\mathbf{Z}$. We need to find an approximation to the posterior distribution $p(\mathbf{z}|\mathbf{x})$ that can minimize the KL divergence, i.e.,

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathcal{D}}{\operatorname{argmin}}\, \mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})),$$

where $\mathcal{D}$ is a restricted distribution family. Here $\mathcal{D}$ is usually a simpler family of distributions to make the optimization and inference tractable.

Xing et al. (2002) assumed the variational distribution $q(\mathbf{z})$ can be factorized over some partitions of the latent variables as follows:

$$q(\mathbf{z}) = \prod_{j=1}^{M} q_j(z_j),$$

where M is the number of parameters and latent variables. The best distribution q_j^* for each factor that solves the optimization problem can be expressed as:

$$q_j^*(z_j) = \frac{e^{E_{-j}[\log p(\mathbf{z},\mathbf{x})]}}{\int e^{E_{-j}[\log p(\mathbf{z},\mathbf{x})]}dz_j} \propto e^{E_{-j}[\log p(z_j,\mathbf{z}_{-j},\mathbf{x})]}. \tag{2}$$

Here $E_{-j}[\cdot]$ means the expectation with respect to all $q_i(z_i)$ with $i \neq j$ and $\mathbf{z}_{-j}$ means all the elements in the vector $\mathbf{z}$ except z_j. However, the optimal mean-field variational approximations $q_j^*(z_j)$ cannot be computed directly because $E_{-j}[z_i]$ $(i \neq j)$ are involved in the right hand side of (2). Thus, an iterative method is often used to obtain the best solution, and such mean-field variational algorithm can only guarantee to converge to a local minimum of $\mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$ (Blei et al. 2017).

Beal and Ghahramani (2003) proposed a variational Bayesian EM algorithm to estimate the marginal likelihood of probabilistic models with latent variables or incomplete data. They also compared the variational bound with a sampling-based method known as annealed importance sampling (Neal 2001). Dieng et al. (2017) proposed another variational algorithm by minimizing the χ-divergence between the variational approximation and the posterior distribution.

3 VB approximation for importance sampling

Although obtaining variational approximation is faster than some sampling based methods, such as MCMC, and it learns the approximate probability density functions through optimization, the inference based on the approximation is biased due to the gap between the variational approximation and the true posterior distribution. On the other hand, IS provides a consistent estimate, but the proposal distribution is hard to design. Here we combine VB with IS by using variational approximation $q(\mathbf{z})$ as the proposal distribution for IS. It avoids the bias from VB approximation and also provides a good way to construct the proposal distribution for IS.

Suppose we have a model with prior $p(\mathbf{z})$ and likelihood function $p(\mathbf{x}|\mathbf{z})$, where $\mathbf{z}$ contains both parameters and latent variables, then the posterior distribution is

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x})} \propto p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}, \mathbf{z}).$$

By the mean-field variational algorithm, we can obtain the variational approximation $q(\mathbf{z})$ to the posterior $p(\mathbf{z}|\mathbf{x})$. If the support of $q(\mathbf{z})$ includes the support of $p(\mathbf{z}|\mathbf{x})$, then the expectation of the function $h(\mathbf{Z})$ with respect to $p(\mathbf{z}|\mathbf{x})$ can be estimated by importance sampling as in (1), with $w(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})/q(\mathbf{z}^{(i)})$. The variational importance sampling algorithm is summarized in Algorithm 1.

Algorithm 1 Variational importance sampling

1. Obtain the analytical expression of $p(\mathbf{z}|\mathbf{x})$ (up to a normalizing constant)
2. Derive the variational approximation $q(\mathbf{z}) = \prod_{j=1}^{M} q_j(z_j)$ to $p(\mathbf{z}|\mathbf{x})$
3. For $i \in \{1, \ldots, m\}$
4. Draw $\mathbf{z}^{(i)}$ from the proposal distribution $q(\mathbf{z})$
5. Calculate importance weight $w(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})/q(\mathbf{z}^{(i)})$
6. Estimate the expectation of $h(\mathbf{Z})$ with respect to $p(\mathbf{z}|\mathbf{x})$ by (1).

Dowling et al. (2018) used the modes of the variational distributions to initialize the location parameters of the proposal distributions in adaptive importance sampling,

which is applicable when the variational approximation is in the location scale family. Our proposed method uses the variational approximation itself as the proposal distribution for importance sampling. It does not put restrictions on the proposal distribution, and it can be extended to sequential importance sampling as shown in the next section.

3.1 VB approximation for sequential importance sampling

If the dimension of the parameters and latent variables is high, or if the data arrive sequentially, SIS is often used. VB can be combined with SIS as well by constructing the proposal with VB sequentially.

Let $\mathbf{z}$ be all the hidden variables, and $\mathbf{z}_{1:t} = \{z_1, \ldots, z_t\}$ be the first t components. Let $\mathbf{x} = \{x_1, \ldots, x_T\}$ be the data which arrive sequentially. The posterior distribution of interest is $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$, $t = 1, \ldots, T$. In variational approximation, we assume the approximation $q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$ can be factorized in the following way:

$$q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t}) = \prod_{k=1}^{t} q(z_k|\mathbf{x}_{1:t}),\ t = 1, \ldots, T.$$

We consider two different approaches for constructing the proposal distribution sequentially.

VB-SIS1. In the first method, at each time $t = 1, \ldots, T$, we minimize the KL divergence between $q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$ and the true posterior distribution $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$, and obtain the variational distributions as follows:

$$\begin{aligned}
q(\mathbf{z}_{1:1}|\mathbf{x}_{1:1}) = q_1(\mathbf{z}_{1:1}) &= q_{11}(z_1),\\
q(\mathbf{z}_{1:2}|\mathbf{x}_{1:2}) = q_2(\mathbf{z}_{1:2}) &= q_{21}(z_1)\, q_{22}(z_2),\\
&\vdots\\
q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = q_T(\mathbf{z}_{1:T}) &= q_{T1}(z_1)\, q_{T2}(z_2) \cdots q_{TT}(z_T).
\end{aligned}$$

We will use $q_{tt}(z_t)$, $t = 1, 2, \ldots, T$, as the proposal distributions in SIS, and we call this method VB-SIS1 with general procedure given in Algorithm 2.

VB-SIS2. Another method is to obtain the proposal distribution in the current step t by reusing the proposals in previous steps. This procedure can be represented as follows:

$$\begin{aligned}
\tilde{q}(\mathbf{z}_{1:1}|\mathbf{x}_{1:1}) = \tilde{q}_1(\mathbf{z}_{1:1}) &= \tilde{q}_1(z_1),\\
\tilde{q}(\mathbf{z}_{1:2}|\mathbf{x}_{1:2}) = \tilde{q}_2(\mathbf{z}_{1:2}) &= \tilde{q}_1(z_1)\, \tilde{q}_2(z_2),\\
&\vdots\\
\tilde{q}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \tilde{q}_T(\mathbf{z}_{1:T}) &= \tilde{q}_1(z_1)\, \tilde{q}_2(z_2) \cdots \tilde{q}_T(z_T).
\end{aligned}$$

At time t, in order to obtain $\tilde{q}_t(\mathbf{z}_{1:t})$, we fix the proposals from previous steps $\tilde{q}_1(z_1), \ldots, \tilde{q}_{t-1}(z_{t-1})$, and obtain $\tilde{q}_t(z_t)$ by minimizing the KL divergence between

Algorithm 2 Variational sequential importance sampling 1 (VB-SIS1)

1. Set $w_0(\mathbf{z}_{1:0}^{(i)}) = 1, i = 1, \ldots, m$
2. For $t \in \{1, \ldots, T\}$
3. Obtain the analytical expression of $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$
4. Derive the variational approximation to $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$ using VB-SIS1:

$$q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t}) = q_{t1}(z_1)q_{t2}(z_2)\cdots q_{tt}(z_t)$$

5. For $i \in \{1, \ldots, m\}$
6. Draw $z_t^{(i)}$ from the proposal distribution $q_{tt}(z_t)$
7. Update importance weight $w_t(\mathbf{z}_{1:t}^{(i)}) = w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{p(\mathbf{z}_{1:t}^{(i)}|\mathbf{x}_{1:t})}{p(\mathbf{z}_{1:t-1}^{(i)}|\mathbf{x}_{1:t-1})q_{tt}(z_t^{(i)})}$
8. Using the sample $\mathbf{z}_{1:t}^{(i)}, i = 1, \ldots, m$, and importance weights $w_t(\mathbf{z}_{1:t}^{(i)})$ to estimate the expectation of $h(\mathbf{z}_{1:t})$ with respect to $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$

$\tilde{q}(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$ and the true posterior distribution $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$. Since we only need to determine the variational distribution for the last latent variable at each step, the running time will be shorter than VB-SIS1. We will use $\tilde{q}_t(z_t)$, $t = 1, \ldots, T$, as the proposal distribution, and we call this method VB-SIS2 with general procedure given in Algorithm 3.

Algorithm 3 Variational sequential importance sampling 2 (VB-SIS2)

1. Set $w_0(\mathbf{z}_{1:0}^{(i)}) = 1, i = 1, \ldots, m$
2. For $t \in \{1, \ldots, T\}$
3. Obtain the analytical expression of $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$
4. Derive the variational approximation to $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$ using VB-SIS2:

$$q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t}) = \tilde{q}_1(z_1)\tilde{q}_2(z_2)\cdots\tilde{q}_t(z_t)$$

5. For $i \in \{1, \ldots, m\}$
6. Draw $z_t^{(i)}$ from the proposal distribution $\tilde{q}_t(z_t)$
7. Update importance weight $w_t(\mathbf{z}_{1:t}^{(i)}) = w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{p(\mathbf{z}_{1:t}^{(i)}|\mathbf{x}_{1:t})}{p(\mathbf{z}_{1:t-1}^{(i)}|\mathbf{x}_{1:t-1})\tilde{q}_t(z_t^{(i)})}$
8. Using the sample $\mathbf{z}_{1:t}^{(i)}, i = 1, \ldots, m$, and importance weights $w_t(\mathbf{z}_{1:t}^{(i)})$ to estimate the expectation of $h(\mathbf{z}_{1:t})$ with respect to $p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})$

In some cases (such as the hidden Markov model example in Sect. 5.4), we use the following approximation to further simplify the variational approximation:

$$p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t}) \approx p(\mathbf{z}_{1:t}|\mathbf{x}_{\max(1,t-\Delta+1):t})$$

where Δ is a tuning parameter. This approximation assumes that the observations at time $k < t-\Delta$ almost provide no additional information to $\mathbf{z}_{1:t}$. Under this assumption, we can obtain the variational approximation at step t only based on the observations $\mathbf{x}_{\max(1,t-\Delta+1):t}$, that is,

$$q(\mathbf{z}_{1:t}|\mathbf{x}_{1:t}) = q(\mathbf{z}_{1:t}|\mathbf{x}_{\max(1,t-\Delta+1):t}).$$

Naesseth et al. (2018) considered approximating the posterior distribution for the state space model by introducing variational parameters and resampling procedures. The variational SIS algorithms we proposed are different because we obtain the proposal distribution at each step by deriving variational approximation sequentially. Our variational SIS can be used for general computation based on SIS, including state space models. Adding the resampling procedure can further improve the efficiency of SIS. We will not consider it here because we would like to compare the VB proposal with the standard proposal to evaluate the efficiency gain from VB proposal. Adding resampling steps will make it hard to distinguish where the efficiency gain is coming from. In practice, users can always combine resampling with variational SIS to make it more effective in high dimensional problems.

4 Theoretical justification

To simplify the notation, we will use p and q to denote the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ and the variational distribution $q(\mathbf{z})$ in this section. In variational inference, we minimize the KL divergence between q and p:

$$KL(q||p) = \int q \log \frac{q}{p} d\mathbf{z}.$$

In importance sampling, the cv^2 is the χ^2 distance between p and q:

$$\chi^2(p||q) = \int \frac{(p-q)^2}{q} d\mathbf{z},$$

and we hope to find a proposal distribution q with a relatively small cv^2.

In order to make connections between these two distances, we introduce a more general f-divergence (Ali and Silvey 1966) between p and q as:

$$D_f(p||q) = E_q\left[f\left(\frac{p}{q}\right)\right] = \int f\left(\frac{p}{q}\right) \cdot q \, d\mathbf{z},$$

where $f(\cdot)$ satisfies the following three conditions:

(i) $f(1) = 0$.
(ii) $f(x)$ is a convex function.
(iii) $f(x)$ is continuous at $x = 1$.

Let $u = p/q$, $f_1(u) = -\log u$ and $f_2(u) = (u-1)^2$, then we can see that the two distances can be written as:

$$KL(q||p) = D_{f_1}(p||q) \quad \text{and} \quad \chi^2(p||q) = D_{f_2}(p||q).$$

Sason and Verdú (2016) showed the following f-divergence inequality:

$$0 \le KL(p||q) \le \log(1 + \chi^2(p||q)),$$

and stated that there is no lower-bound on KL divergence in terms of χ^2 distance. That means the convergence in χ^2 distance implies the convergence in KL divergence, but the other way may not be true. We examine the relationship between KL divergence and χ^2 distance more closely below.

The Taylor expansion for $f_1(u)$ at $u = 1$ is :

$$f_1(u) = -\log u = -\log(1 + (u-1)) = -(u-1) + \frac{(u-1)^2}{2} - \frac{(u-1)^3}{3} + \cdots .$$

Taking expectation on both sides with respect to q and using the fact $E_q[u] = E_q[p/q] = 1$, we obtain the following equations

$$KL(q||p) = \frac{1}{2}\chi^2(p||q) + o((u-1)^2).$$

This indicates that when u is close to 1, these two distances are equivalent, i.e.,

$$KL(q||p) \asymp \frac{1}{2}\chi^2(p||q).$$

In order to quantify the value of u, we introduce two quantities β_1 and β_2 as follows (Sason and Verdú 2016):

$$\beta_1 = \operatorname{ess\,inf} \frac{q}{p} = \left(\operatorname{ess\,sup} \frac{p}{q}\right)^{-1}, \quad \beta_2 = \operatorname{ess\,inf} \frac{p}{q} = \left(\operatorname{ess\,sup} \frac{q}{p}\right)^{-1}. \tag{3}$$

The essential infimum and the essential supremum are defined as:

$$\operatorname{ess\,inf} \frac{p}{q} = \sup\{b \in \mathbb{R} : \mu(\{x : p(x)/q(x) < b\}) = 0\},$$
$$\operatorname{ess\,sup} \frac{p}{q} = \inf\{a \in \mathbb{R} : \mu(\{x : p(x)/q(x) > a\}) = 0\},$$

where $\mu(\cdot)$ denotes the Lebesgue measure.

Since $\int q(\mathbf{z})\, d\mathbf{z} = 1$ and $\int p(\mathbf{z}|\mathbf{x})\, d\mathbf{z} = 1$, we have $0 \le \beta_1, \beta_2 \le 1$, and $\beta_1 = 1 \Leftrightarrow \beta_2 = 1 \Leftrightarrow p = q$. Suppose $0 < \beta_1 < 1$ and $0 < \beta_2 < 1$. We say a sequence of probability measures with densities p_n converge to q if

$$\lim_{n\to\infty} \operatorname{ess\,inf} \frac{p_n}{q} = 1. \tag{4}$$

Lemma 1 *Suppose f is a function satisfying Conditions (i)–(iii), and a sequence of probability measures with densities p_n converge to q in the sense of* (4)*. Let*

$$\beta_{1,n}^{-1} = \operatorname{ess\,sup} \frac{p_n}{q}, \quad \beta_{2,n} = \operatorname{ess\,inf} \frac{p_n}{q}.$$

Then we have

$$\lim_{n\to\infty} \beta_{1,n} = \lim_{n\to\infty} \beta_{2,n} = 1,$$

and

$$\lim_{n\to\infty} D_f(p_n||q) = 0.$$

The proof of the lemma as well as the proof of the following theorem are in the Appendix. Define a function:

$$\kappa(t) = \frac{t\log t + (1-t)}{(t-1) - \log t}, \quad 0 < t < 1 \text{ and } t > 1, \tag{5}$$

which is increasing for $t \in (0,1) \cup (1,\infty)$. Then, from Sason and Verdú (2016), the following inequalities hold:

$$\kappa(\beta_{2,n}) \le \frac{KL(p_n||q)}{KL(q||p_n)} \le \kappa(\beta_{1,n}^{-1}), \tag{6}$$

$$\frac{1}{2}\beta_{1,n} \le \frac{KL(p_n||q)}{\chi^2(p_n||q)} \le \frac{1}{2}\beta_{2,n}^{-1}, \tag{7}$$

where p_n, $\beta_{1,n}$, and $\beta_{2,n}$ are defined in Lemma 1. Note that (6) and (7) do not require the convergence of p_n to q. The following theorem gives the limit of the ratios in (6) and (7).

Theorem 1 *Suppose a sequence of probability measures with densities p_n converge to q in the sense of (4). For KL divergence and χ^2 distance, we have*

$$\lim_{n\to\infty} \frac{KL(p_n||q)}{KL(q||p_n)} = 1, \quad \lim_{n\to\infty} \frac{KL(p_n||q)}{\chi^2(p_n||q)} = \frac{1}{2}.$$

From the above theorem, we immediately have the following corollary.

Corollary 1 *Suppose a sequence of probability measures with densities p_n converge to q in the sense of (4). For KL divergence and χ^2 distance, we have*

$$\lim_{n\to\infty} \frac{KL(q||p_n)}{\chi^2(p_n||q)} = \frac{1}{2}.$$

The theorem and corollary show that the KL divergence and χ^2 distance are equivalent (up to a constant) when the proposal distribution and the target distribution are getting close to each other. In practice, we cannot obtain a proposal that is arbitrarily close to the target distribution, but following the insight of the theorem and corollary,

we can use the KL divergence to bound the χ^2 distance . When we consider a proposal distribution q in importance sampling, we have from (6) and (7) that

$$2\beta_2\kappa(\beta_2) \leq \frac{\chi^2(p||q)}{KL(q||p)} \leq 2\beta_1^{-1}\kappa(\beta_1^{-1}). \tag{8}$$

Therefore the upper and lower bounds for χ^2 distance are

$$2\beta_2\kappa(\beta_2)KL(q||p) \leq \chi^2(p||q) \leq 2\beta_1^{-1}\kappa(\beta_1^{-1})KL(q||p). \tag{9}$$

Our goal is to find a proposal distribution q close to the target distribution p in terms of the χ^2 distance $\chi^2(p||q)$. The relation in (9) indicates that the distribution q that minimizes the KL divergence $KL(q||p)$ tends to give tighter bounds for the χ^2 distance $\chi^2(p||q)$, and under a smaller upper bound in (9), the χ^2 distance $\chi^2(p||q)$ tends to be small as well, which is what we hope to achieve in choosing the proposal distribution for importance sampling. Therefore, it is reasonable to use the distribution q that minimizes the KL divergence $KL(q||p)$ as the proposal distribution. This justifies the use of VB solution as the proposal distribution for importance sampling. The upper bound in (9) can give us an intuitive way to evaluate the choice of the proposal distribution since a smaller upper bound often indicates that the corresponding proposal distribution has better performance in importance sampling. This idea is illustrated in the example in Sect. 5.1 by computing β_1 and β_2 explicitly. However, the exact values of β_1 and β_2 are hard to calculate in some complex models.

5 Numerical results

All examples in this section were coded in R and run on a MacBook Pro with 2.3 GHz Intel Core i7 processor.

5.1 Univariate normal

This toy example is on Bayesian inference for a univariate normal distribution. Suppose our observed data $\mathbf{x} = \{x_1, \ldots, x_N\}$ is a random sample from a normal distribution with mean μ and precision τ. We use the normal-gamma conjugate prior for μ and τ as follows:

$$p(\mu|\tau) = \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}), \quad p(\tau) = \text{Gamma}(a_0, b_0).$$

We consider a factorized variational approximation to the posterior distribution $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$. The variational approximation algorithm gives $q_\mu(\mu) \sim \mathcal{N}(\nu, \lambda^{-1})$ with the mean and precision:

$$\nu = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \text{ and } \lambda = (\lambda_0 + N)E[\tau],$$

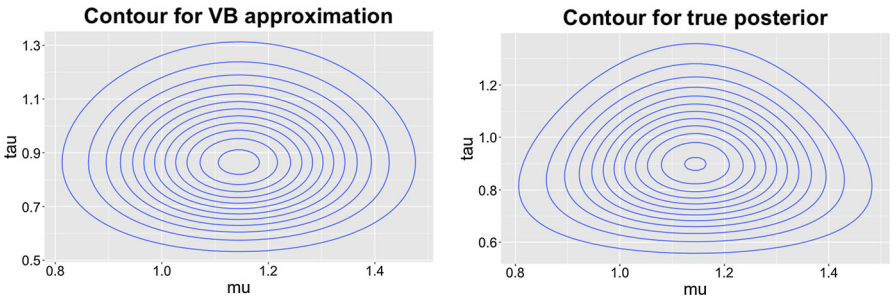

Fig. 1 Contour plots for the true posterior and the VB approximation

and $q_\tau(\tau) \sim \text{Gamma}(a, b)$ with two parameters:

$$a = a_0 + \frac{N}{2}, \quad b = b_0 + \frac{1}{2} E_\mu \left[\sum_{i=1}^{N} (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right].$$

If we follow the updating rules and compute the expectation with the parameter values from the previous step, we can obtain the variational distribution $q(\mu, \tau)$ as in Algorithm 4.

Algorithm 4 Variational algorithm for univariate normal

1. Initialize $b = 1, \lambda = 1$
2. Calculate $\nu = \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N}$ and $a = a_0 + \frac{N}{2}$
3. Repeat the following until convergence
4. $\quad \lambda = (\lambda_0 + N)\frac{a}{b}$
5. $\quad b = b_0 + \frac{1}{2}\left[(\sum_{i=1}^{N} x_n^2 + \lambda_0 \mu_0^2) - (2\sum_{i=1}^{N} x_n + 2\lambda_0 \mu_0)\nu + (\lambda_0 + N)(\nu^2 + \frac{1}{\lambda}) \right]$

We set the hyperparameters $\mu_0 = 1, \lambda_0 = 1, a_0 = 1, b_0 = 1$, and generated $N = 50$ data points from $N(1, 1)$. For this simple example, the true posterior distribution $p(\mu, \tau | \mathbf{x})$ can be derived as

$$p(\mu|\tau, \mathbf{x}) = \mathcal{N}\left(\frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N}, (\lambda_0 + N)^{-1} \right),$$

$$p(\tau|\mathbf{x}) = \text{Gamma}\left(a_0 + \frac{N}{2}, b_0 + \frac{1}{2}\left[\sum_{i=1}^{N} (x_i - \bar{x})^2 + \frac{\lambda_0 N (\bar{x} - \mu_0)^2}{\lambda_0 + N} \right] \right).$$

The contour plots in Fig. 1 show some resemblance between the true posterior distribution and the VB approximation.

We compared the performance of different methods in Table 1, including the variational Bayes method (denoted by "VB"), IS with variational distribution as the proposal (denoted by "VB as proposal"), IS with the prior as the proposal (denoted by "Prior as proposal"), and adaptive importance sampling (denoted by "AIS") (Bugallo et al.

Table 1 Simulation results for the univariate normal example

Parameter	VB	VB as proposal	Prior as proposal	AIS	True mean
μ	1.1445	1.1453 (0.0007)	1.1448 (0.0226)	1.1443 (0.0021)	1.1445
τ	0.8992	0.9170 (0.0006)	0.9192 (0.0183)	0.9181 (0.0015)	0.9169

Table 2 The values of β_1^{-1} and β_2 for the univariate normal example

	VB as proposal	Prior as proposal
β_1^{-1}	1.751	2.513
β_2	0.673	0.282

2017). The variational distributions are well-known standard distributions in this example, and the expectations are easy to compute. The three IS algorithms are based on $m = 100,000$ samples, and the numbers in parentheses are the standard errors. The true posterior mean is also provided (denoted by "True mean").

Table 1 shows that IS with variational distribution as proposal gives much smaller standard errors than IS with prior as the proposal and AIS. The computation time of AIS is much longer than IS with VB or prior as the proposal, since AIS needs to update the proposal distribution adaptively based on samples from previous steps, while the variational distribution and the prior are relatively easier to obtain. Using variational method directly gives a biased estimate for τ (the estimate for μ happens to be the same as the true mean), and the variability of the estimate is unknown.

Since the true posterior distribution is known in this example, we can calculate β_1 and β_2 defined in (3). The values of β_1 and β_2, which are presented in Table 2, are related to the ratio between the posterior distribution p and the proposal distribution q. They also appear in the upper and lower bounds of the χ^2 distance between p and q in (8) and (9). Since β_1^{-1} is smaller when VB is the proposal and $\kappa(t)$ is an increasing function for $0 < t < 1$, that implies the upper bounds $2\beta_1^{-1}\kappa(\beta_1^{-1})$ in (8) and $2\beta_1^{-1}\kappa(\beta_1^{-1})KL(q||p)$ in (9) are smaller when VB is the proposal (note that $KL(q||p)$ is minimized for VB proposal). Similarly, VB proposal has a larger β_2 which implies $2\beta_2\kappa(\beta_2)$ in the lower bound in (8) and (9) is larger for the VB proposal. All these suggest that using VB as the proposal may lead to a smaller χ^2 distance and better performance.

5.2 Gaussian mixture model

Suppose we have N i.i.d. observations $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ from a Gaussian mixture distribution, and each $\mathbf{x}_i$ is a D-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})^T$. Suppose there are K mixture components and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ denotes the mixture proportions. The labels that indicate the membership of the observations are denoted by the latent variables $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, where $\mathbf{z}_i \sim \text{Multinomial}(1, \boldsymbol{\pi})$. In other

words, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})^T$ is a K-dimensional vector with one element equal to 1, which specifies the label of $\mathbf{x}_i$, and all other elements equal to 0. If the k-th element of $\mathbf{z}_i$ is 1, we write

$$\mathbf{x}_i | z_{ik} = 1 \ \sim \ \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}),$$

where μ_i and $\boldsymbol{\Lambda}_i$ are the mean and precision matrix of each multivariate Gaussian component.

We use a symmetric Dirichlet distribution with hyperparameter α_0 as the prior distribution for $\boldsymbol{\pi}$:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_0}, \quad \text{where } \boldsymbol{\alpha}_0 = (\alpha_0, \alpha_0, \ldots, \alpha_0).$$

For the mean vector $\boldsymbol{\mu}_k$ and the precision matrix $\boldsymbol{\Lambda}_k$, we use a normal-Wishart prior distribution as the conjugate prior for these two parameters:

$$\boldsymbol{\Lambda}_k \sim \mathrm{Wishart}(\boldsymbol{W}_0, \nu_0) \ \ \Rightarrow \ \ p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, \nu_0),$$

$$\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) \ \ \Rightarrow \ \ p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right),$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \ldots, \boldsymbol{\Lambda}_K)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K)$. The likelihood function of the Gaussian mixture model is

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}},$$

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \right).$$

The posterior distribution is

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{x}) \propto p(\mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}).$$

For variational approximation, following Bishop (2006) we first factorize $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ into the following variational distribution:

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k).$$

After calculating the logarithm of the optimal distribution, we get:

$$\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1)\sum_{k=1}^{K}\ln\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}\ln\pi_k$$

$$\Rightarrow \quad q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \qquad \text{where } \alpha_k = \alpha_0 + N_k \text{ and } N_k = \sum_{i=1}^{N} r_{ik}.$$

Then we further decompose the variational distribution as $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^*(\boldsymbol{\Lambda}_k)$, and the variational joint posterior distribution of $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ is also normal-Wishart distribution with different parameters from the prior distributions:

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

If we follow the updating rules for each parameter, we can obtain the variation approximation for Gaussian mixture model as in Algorithm 5.

Algorithm 5 Variational algorithm for Gaussian mixture model

1. Initialize $\boldsymbol{\alpha}$, $\bar{\mathbf{x}}_k$, $\mathbf{W}_k$, $\mathbf{m}_k$, $\mathbf{S}_k$ and r_{ik}
2. Repeat the following steps until convergence
3. Calculate $N_k = \sum_{i=1}^{N} r_{ik}$ and update $\boldsymbol{\alpha}$ by $\alpha_k = \alpha_0 + N_k$
4. Update $\bar{\mathbf{x}}_k$ and $\mathbf{S}_k$ by

$$\bar{\mathbf{x}}_k = \frac{1}{N_k}\sum_{i=1}^{N} r_{ik}\mathbf{x}_i \text{ and } \mathbf{S}_k = \frac{1}{N_k}\sum_{i=1}^{N} r_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}}$$

5. Update $\mathbf{W}_k$ and ν_k by

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^{\mathrm{T}} \text{ and } \nu_k = \nu_0 + N_k$$

6. Update $\mathbf{m}_k$ and β_k by

$$\mathbf{m}_k = \frac{1}{\beta_k}(\beta_0\boldsymbol{\mu}_0 + N_k\bar{\mathbf{x}}_k) \text{ and } \beta_k = \beta_0 + N_k$$

7. Update r_{ik} by

$$\rho_{ik} = \exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\mathbf{x}_i - \mathbf{m}_k)^T\mathbf{W}_k(\mathbf{x}_i - \mathbf{m}_k)\right) \text{ and } r_{ik} = \frac{\rho_{ik}}{\sum_{j=1}^{K}\rho_{ik}}$$

8. Variational distribution is $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)$ and $q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$.

In the following simulation, we fix the hyperparameters $\alpha_0 = 1$, $\beta_0 = 5$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{W}_0 = \mathbf{I}_D$, and $\nu_0 = 5$. Tables 3, 4 and 5 show the results for different combinations

Table 3 Simulation results for Gaussian mixture model with $D = 1$, $K = 2$, and $\alpha_0 = 1$

Parameter	VB	VB as proposal	Prior as proposal	'True mean'	True parameter
ω_1	0.782	0.774 (0.004)	0.665 (0.125)	0.775	0.7
ω_2	0.218	0.227 (0.004)	0.335 (0.125)	0.235	0.3
μ_1	−2.671	−2.666 (0.007)	−1.149 (0.892)	−2.663	−3
μ_2	1.945	1.870 (0.032)	0.161 (1.193)	1.847	3
Λ_1	0.292	0.287 (0.006)	1.419 (0.726)	0.278	1
Λ_2	0.686	0.682 (0.005)	0.446 (0.316)	0.677	1

of the dimension of the data D and the number of mixture components K. The variational distributions are well-known standard distributions in this example, and the expectations of all parameters are easy to compute when applying VB directly. The two IS algorithms are based on $m = 10,000$ samples. The last column denotes the true parameters when we generated the observed data. The 'True mean' is an estimate of the true posterior mean based on $1,000,000$ samples from importance sampling with VB approximation as the proposal.

From Tables 3, 4 and 5, we can see that IS with variational distribution as proposal gives smaller standard errors than IS with prior as the proposal. In addition, using VB directly will introduce bias to the estimates.

5.3 Linear regression model

Let $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$ be the observed pairs of data, where $\mathbf{x}_i \in \mathbb{R}^p$. Consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\epsilon_i \sim N(0, \sigma^2)$. The likelihood function is

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$, and $\mathbf{I}$ is the identity matrix. Similar to You et al. (2014), we use inverse gamma and normal conjugate priors for $\boldsymbol{\beta}$ and σ^2 as follows:

$$\sigma^2 \sim \text{Inv-Gamma}(A, B), \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2\mathbf{I}),$$

where A, B, $\sigma_{\boldsymbol{\beta}}^2$ are hyperparameters.

Let $\mathbf{z}$ be all parameters of interests, i.e., $\mathbf{z} = [\boldsymbol{\beta}^T, \sigma^2]^T$. We consider a factorized variational approximation $q^*(\mathbf{z}) = q_{\boldsymbol{\beta}}^*(\boldsymbol{\beta}) q_{\sigma^2}^*(\sigma^2)$. Since we chose the conjugate priors for $\mathbf{z}$, the variational distributions can be written as:

$$q_{\boldsymbol{\beta}}^*(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}), \quad q_{\sigma^2}^*(\sigma^2) \sim \text{Inv-Gamma}\left(A + n/2, B_{q(\sigma^2)}\right).$$

Table 4 Simulation results for Gaussian mixture model with $D = 1$, $K = 3$, and $\alpha_0 = 1$

Parameter	VB	VB as proposal	Prior as proposal	'True mean'	True parameter
ω_1	0.499	0.482 (0.012)	0.449 (0.088)	0.484	0.5
ω_2	0.266	0.288 (0.010)	0.391 (0.125)	0.290	0.3
ω_3	0.235	0.230 (0.014)	0.159 (0.754)	0.226	0.2
μ_1	−3.670	−3.691 (0.102)	−1.020 (0.217)	−3.610	−5
μ_2	−0.136	−0.170 (0.021)	0.008 (0.157)	−0.158	0
μ_3	2.610	2.306 (0.102)	0.300 (0.136)	2.415	5
Λ_1	3.577	3.901 (0.153)	6.967 (1.833)	3.753	1
Λ_2	0.194	0.189 (0.006)	4.772 (2.148)	0.185	1
Λ_3	0.160	0.137 (0.005)	0.965 (0.813)	0.146	1

Table 5 Simulation results for Gaussian mixture model with $D = 2$, $K = 2$, and $\alpha_0 = 1$

Parameter	VB	VB as proposal	Prior as proposal	'True mean'	True parameter
ω_1	0.730	0.732 (0.002)	0.618 (0.125)	0.733	0.7
ω_2	0.270	0.268 (0.002)	0.382 (0.125)	0.367	0.3
μ_{11}	− 2.689	−2.689 (0.005)	−0.654 (1.412)	−2.688	−3
μ_{21}	− 2.671	−2.668 (0.006)	−0.135 (1.617)	−2.663	−3
μ_{12}	2.012	2.006 (0.012)	0.553 (0.925)	1.973	3
μ_{22}	1.592	1.578 (0.011)	−0.081 (0.825)	1.586	3
Λ_{111}	0.579	0.558 (0.006)	0.822 (0.183)	0.562	1
Λ_{121}	− 0.292	−0.282 (0.005)	−1.583 (0.902)	−0.276	0
Λ_{221}	0.700	0.687 (0.007)	1.937 (1.025)	0.688	1
Λ_{112}	1.901	1.918 (0.014)	5.772 (2.245)	1.920	2
Λ_{122}	− 1.655	−1.679 (0.014)	−2.791 (0.725)	−1.680	−1
Λ_{222}	2.234	2.262 (0.016)	3.129 (0.616)	2.271	3

By solving the optimization problem iteratively, we can obtain the updating rules of all the parameters, as well as the corresponding variational algorithm in Algorithm 6.

Algorithm 6 Variational algorithm for linear regression model

1. Initialize $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = I_p$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \mathbf{1}^T$, $B_{q(\sigma^2)} = 1$
2. Repeat the following until convergence
3. Update $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left[\left(\frac{A + n/2}{B_{q(\sigma^2)}} \right) \mathbf{X}^T \mathbf{X} + \sigma_{\boldsymbol{\beta}}^{-2} \mathbf{I} \right]^{-1}$$

4. Update $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$:

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \left(\frac{A + n/2}{\mathbf{B}_{q(\sigma^2)}} \right) \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \mathbf{X}^T \mathbf{y}$$

5. Update $B_{q(\sigma^2)}$:

$$B_{q(\sigma^2)} = B + \frac{1}{2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}||^2 + \frac{1}{2} \mathrm{tr} \left(\mathbf{X}^T \mathbf{X} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \right)$$

In the simulation, we generated $N = 50$ data pairs from the following true model

$$y = 3 + 0 \cdot x_1 - 3x_2 + 5x_3 + \epsilon \ , \ \ \epsilon \sim N(0, \sigma^2),$$

where x_1 has no influence on the response variable y. We fix the hyperparameters $\sigma_{\boldsymbol{\beta}} = 2$, $A = 2$, and $B = 5$. The variational distribution obtained from Algorithm 6 is used to estimate the parameters directly and also as the proposal for IS. The two IS algorithms with different proposals are both based on $m = 10,000$ samples. The 'True mean' is an estimate of the true posterior mean based on $1,000,000$ samples from IS with VB approximation as the proposal.

Table 6 shows that IS with variational distribution as proposal gives smaller standard errors than IS with prior as the proposal. Using variational method directly gives a biased estimate and variability of the estimate is unknown. For example, using VB directly gives an estimate of -0.096 for β_1 without quantification of the uncertainty of the estimate, so it is hard to tell whether the true value of β_1 is 0. On the other hand, the 95% confidence interval of the estimates based on both IS algorithms contain 0, which indicates that β_1 is not significant in the linear model.

5.4 Hidden Markov model

The hidden Markov model (HMM) consists of a Markov chain with hidden states $\mathbf{z} = \{z_0, z_1, z_2, \ldots, z_T\}$ and an observed sequence of data $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$,

Table 6 Simulation results for linear regression model

Parameter	VB	VB as proposal	Prior as proposal	'True mean'	True parameter
β_0	2.934	2.949 (0.034)	2.723 (0.174)	2.903	3
β_1	−0.096	−0.048 (0.054)	−0.314 (0.262)	−0.073	0
β_2	−2.745	−2.710 (0.050)	−1.989 (0.482)	−2.703	−3
β_3	4.442	4.350 (0.069)	3.259 (0.565)	4.371	5
σ^2	5.831	5.853 (0.138)	7.920 (1.401)	5.852	4

where z_0 is the initial state, and T is the length of the sequence. The hidden states evolve according to

$$Z_t|(Z_{t-1} = z_{t-1}) \sim f(z_t|z_{t-1}),$$

and the dependence between the observed data and hidden state can be represented as

$$X_t|(Z_t = z_t) \sim g(x_t|z_t).$$

Given the observed data, the posterior distribution of the hidden states can be written as:

$$p(\mathbf{z}_{0:T}|\mathbf{x}_{1:T}) = \frac{p(\mathbf{z}_{0:T}, \mathbf{x}_{1:T})}{p(\mathbf{x}_{1:T})} \propto p(\mathbf{z}_{0:T})p(\mathbf{x}_{1:T}|\mathbf{z}_{0:T}),$$

where

$$p(\mathbf{z}_{0:T}) = f(z_0)\prod_{t=1}^{T} f(z_t|z_{t-1}) \quad \text{and} \quad p(\mathbf{x}_{1:T}|\mathbf{z}_{0:T}) = \prod_{t=1}^{T} g(x_t|z_t).$$

We consider the filtering problem, which is to infer $\mathbf{z}_{1:t}$ from the observations $\mathbf{x}_{1:t}$, $t = 1, \ldots, T$. When applying SIS to the filtering problem, the naive choice of the proposal distribution is to sample z_t from $f(z_t|z_{t-1})$. However, this proposal is not very efficient because it does not take into account the information contained in the observations.

The two variational approximations in Sect. 3.1, VB-SIS1 and VB-SIS2, can be used to construct better proposals for SIS. The corresponding algorithm is the same as Algorithms 2 and 3, and the weight updating step for HMM can be written explicitly as

$$\begin{aligned} w_t(\mathbf{z}_{1:t}^{(i)}) &= w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})}{p(\mathbf{z}_{1:t-1}|\mathbf{x}_{1:t-1})q_{tt}(z_t^{(i)})} \\ &= w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{g(x_t|z_t^{(i)})f(z_t^{(i)}|z_{t-1}^{(i)})}{q_{tt}(z_t^{(i)})}, \end{aligned}$$

or

$$\begin{aligned} w_t(\mathbf{z}_{1:t}^{(i)}) &= w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_{1:t})}{p(\mathbf{z}_{1:t-1}|\mathbf{x}_{1:t-1})\tilde{q}_t(z_t^{(i)})} \\ &= w_{t-1}(\mathbf{z}_{1:t-1}^{(i)}) \frac{g(x_t|z_t^{(i)})f(z_t^{(i)}|z_{t-1}^{(i)})}{\tilde{q}_t(z_t^{(i)})}. \end{aligned}$$

We study two examples below, one is a discrete HMM and the other one is a continuous HMM.

Table 7 Simulation results for discrete HMM with $\Delta = 7$, $T = 50$, and varying sample size m

Proposal	m	cv^2	Time (s)
$f(z_t\|z_{t-1})$	1000	321.0979	0.8
VB-SIS1 $\Delta = 7$	1000	78.0338	235
VB-SIS2 $\Delta = 7$	1000	205.3263	52
$f(z_t\|z_{t-1})$	5000	342.0129	4.2
VB-SIS1 $\Delta = 7$	5000	75.1225	251
VB-SIS2 $\Delta = 7$	5000	202.2352	63
$f(z_t\|z_{t-1})$	30000	336.1599	20.6
VB-SIS1 $\Delta = 7$	30000	77.9406	306
VB-SIS2 $\Delta = 7$	30000	208.3262	75

Table 8 Simulation results for discrete HMM with $m = 5000$ and varying length of sequence T

Proposal	T	cv^2	Time (s)
$f(z_t\|z_{t-1})$	30	97.0153	3.1
VB-SIS1 $\Delta = 7$	30	18.0764	149
VB-SIS2 $\Delta = 7$	30	45.6237	34
$f(z_t\|z_{t-1})$	50	342.0129	4.2
VB-SIS1 $\Delta = 15$	50	75.1225	335
VB-SIS2 $\Delta = 15$	50	202.2352	63
$f(z_t\|z_{t-1})$	100	1252.2339	8.3
VB-SIS1 $\Delta = 32$	100	193.3824	703
VB-SIS2 $\Delta = 32$	100	527.2363	233

5.4.1 Discrete hidden Markov model

In the discrete HMM example, assume $z_t \in \{1, \ldots, K\}$ and $x_t \in \{1, \ldots, W\}$. Then the model can be specified by two matrices: transition matrix $\mathbf{A}_{K\times K}$ and emission matrix $\mathbf{B}_{K\times W}$, where A_{ij} denotes the probability of transitioning from state i to state j and B_{kw} denotes the probability of emitting observation w from state k. We propose the variational approximation similar to Wang and Blunsom (2013).

In the simulation study, we set $z_0 = 1$, $K = 3$ and $W = 4$, i.e., $z_t \in \{1, 2, 3\}$ and $x_t \in \{1, 2, 3, 4\}$. The transition and emission matrices are chosen to be:

$$A = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.4 & 0.2 & 0.4 \\ 0.6 & 0.2 & 0.2 \end{bmatrix}, \quad B = \begin{bmatrix} 0.3 & 0.3 & 0.3 & 0.1 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.1 & 0.6 & 0.2 & 0.1 \end{bmatrix}.$$

We considered different combinations of the length of the sequence T, the number of samples m, and the tuning parameter Δ. The results are presented in Tables 7 and 8 and Fig. 2.

From Table 7, we can see that if we fix Δ and the length of sequence T, the cv^2 for each method will not change much when we increase the number of samples m.

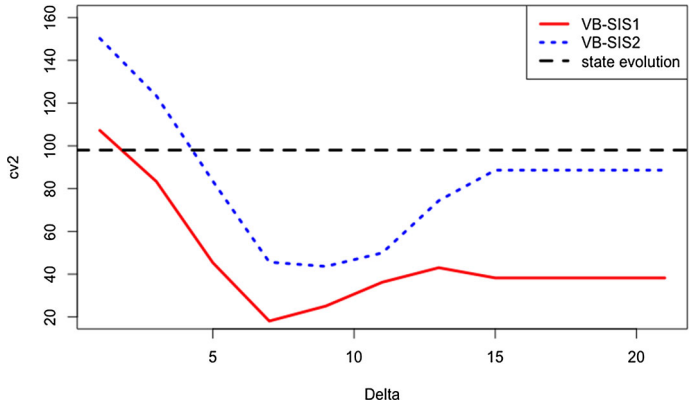

Fig. 2 cv^2 for variational SIS for discrete HMM with $m = 5000$, $T = 30$, and varying tuning parameter Δ

Table 8 shows that if we fix m, then T will influence both cv^2 and the computation time a lot. In general, using the state evolution $f(z_t|z_{t-1})$ takes less time, but the cv^2 is large. VB-SIS1 gives the smallest cv^2, but the computation time is the longest. The performance of VB-SIS2 is somewhere between the other two methods. Note that after the data are generated, we only need to compute the variational approximation once, so this time-consuming step will not be influenced by the sample size m. Figure 2 shows how the cv^2 of importance sampling changes with the value of Δ. The horizontal dashed line is the cv^2 when the state evolution $f(z_t|z_{t-1})$ is used as the proposal, and it can serve as a benchmark.

5.4.2 Stochastic volatility model

The stochastic volatility model consists of the following state equation and observation equation:

$$Z_t = \alpha Z_{t-1} + \sigma V_t, \quad X_t = \beta \exp(Z_t/2) W_t,$$

where $V_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$, $W_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$, and both the hidden state Z_t and the observation X_t are continuous real-valued random variables.

In the simulation study, the initial state $Z_0 \sim \mathcal{N}(0, \sigma^2/(1-\alpha^2))$, and we set $\alpha = 0.3$, $\sigma = 5$ and $\beta = 2$. In this case, the variational distributions $\{q_t(z_t)\}_{t=1}^T$ also follow the normal distribution. We considered different combinations of the length of the sequence T, the number of samples m, and the tuning parameter Δ. The results are in Tables 9 and 10.

From Table 9, we can see that if we fix Δ and the length of sequence T, the cv^2 for each method will not change much when we increase the number of samples m. Table 10 shows that if we increase the length of the observed sequence T, then the cv^2 increases for all proposal distributions we tested. Tables 9 and 10 indicate that using the state evolution $f(z_t|z_{t-1})$ as the proposal distribution takes less time, but the cv^2 is relatively large. VB-SIS1 gives the smallest cv^2, but the computation time is the

Table 9 Simulation results for stochastic volatility model with $\Delta = 7$, $T = 50$, and varying sample size m

Proposal	Estimate (s.e.)	m	cv^2	Time (s)
$f(z_t\|z_{t-1})$	15.32 (1.42)	1000	151.0883	0.7
VB-SIS1 $\Delta = 7$	13.90 (0.97)	1000	48.0338	15.3
VB-SIS2 $\Delta = 7$	13.63 (1.23)	1000	68.5262	3.6
$f(z_t\|z_{t-1})$	14.98 (0.42)	5000	134.9283	3.2
VB-SIS1 $\Delta = 7$	14.71 (0.25)	5000	45.1735	17.7
VB-SIS2 $\Delta = 7$	14.64 (0.36)	5000	62.2415	5.7
$f(z_t\|z_{t-1})$	14.53 (0.04)	30000	142.1737	17.5
VB-SIS1 $\Delta = 7$	14.48 (0.03)	30000	51.2624	24.2
VB-SIS2 $\Delta = 7$	14.44 (0.03)	30000	98.1525	19.7

Table 10 Simulation results for stochastic volatility model with $m = 5000$ and varying length of the sequence T

Proposal	Estimate (s.e.)	T	cv^2	Time (s)
$f(z_t\|z_{t-1})$	15.32 (1.42)	30	151.0883	0.7
VB-SIS1 $\Delta = 7$	13.90 (0.97)	30	48.0338	15.3
VB-SIS2 $\Delta = 7$	13.63 (1.23)	30	65.5262	3.6
$f(z_t\|z_{t-1})$	24.72 (2.42)	50	412.5422	2.2
VB-SIS1 $\Delta = 15$	26.37 (1.75)	50	73.2527	22.4
VB-SIS2 $\Delta = 15$	26.43 (1.98)	50	83.6236	8.4
$f(z_t\|z_{t-1})$	−24.52 (3.42)	100	1524.3532	15.3
VB-SIS1 $\Delta = 32$	−27.12 (2.52)	100	265.3262	32.5
VB-SIS2 $\Delta = 32$	−27.26 (2.97)	100	436.2363	20.3

longest. The performance of VB-SIS2 is somewhere between the other two methods. If we fix the running time, VB-SIS2 has a larger effective sample size than VB-SIS1.

5.5 Dirichlet process

The last example is a Dirichlet process (DP) mixture model widely used in Bayesian inference. Dirichlet Process can be written as $G \sim \mathrm{DP}(\alpha, G_0)$, where G_0 is the base distribution of this stochastic process, and α is a positive scalar parameter. In addition, G and G_0 should have the same support, but G is a discrete distribution with countably infinite number of point masses. Given the previous $n-1$ observations, we generate the next one as follows:

$$X_n|X_1,\ldots,X_{n-1} = \begin{cases} X_i & \text{w.p. } \frac{1}{n-1+\alpha} \quad (i=1,\ldots,n-1), \\ \text{a new draw from } G_0 & \text{w.p. } \frac{\alpha}{n-1+\alpha}, \end{cases}$$

where w.p. means "with probability". Let K be the unique values among $\{X_1, \ldots, X_{n-1}\}$, denoted by $\{X_k^*\}_{k=1}^K$, and we can rewrite the sampling procedure as

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_k^* & \text{w.p. } \frac{\text{num}_{n-1}(X_k^*)}{n-1+\alpha} \quad (k = 1, \ldots, K), \\ \text{a new draw from } G_0 & \text{w.p. } \frac{\alpha}{n-1+\alpha}, \end{cases}$$

where $\text{num}_{n-1}(X_k^*)$ is the number of X_k^* in the set $\{X_1, \ldots, X_{n-1}\}$. Then, the joint density function can be written as

$$\begin{aligned} P(X_1, \ldots, X_N) &= P(X_1)P(X_2|X_1)\cdots P(X_N|X_1, \ldots, X_{N-1}) \\ &= \frac{\alpha^k \prod_{k=1}^K (\text{num}_N(X_k^*) - 1)!}{\alpha(1+\alpha)\cdots(N-1+\alpha)} \prod_{k=1}^K G_0(X_K^*), \end{aligned}$$

which does not depend on the order of variables.

Dirichlet process can also be treated as a stick breaking process. We first draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$, then generate $X_1^*, X_2^*, \ldots \sim G_0$. A multinomial distribution can be derived as

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j).$$

The Dirichlet process G is a discrete distribution with $P(G = X_i^*) = \pi_i(\mathbf{v})$, and it can be written as $G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta_{X_i^*}$, where δ_x is the Dirac measure at point x. In Dirichlet process mixture model, data come from a mixture of an infinite number of distributions. If we have N observed data points $\{x_i\}_{i=1}^N$, they will be generated from at most N different components. The following is the generating procedure of DP mixture model.

- $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$
- $\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$
- $y_i \sim \text{Multinomial}(\pi)$
- $\eta_k \sim G_0$
- $X_i | y_i, \boldsymbol{\eta} \sim p(x_i | \eta_{y_i})$

Given the latent variable z_i, we assume the observation x_i follows a distribution from an exponential family with the likelihood function $p(x_i|\eta_{y_i})$.

Following Blei and Jordan (2006) and Hughes and Sudderth (2013), let $\mathbf{Z} = \{\mathbf{V}, \boldsymbol{\eta}, \mathbf{Y}\}$ be all latent variables and $\theta = \{\alpha\}$ be the hyper parameter. Since the number of different components is infinite, we introduce a truncated level T as an upper bound of the number of clusters, that is, mixture proportions $\pi_t(\mathbf{v}) = 0$ for $t > T$. Then we can factorize the posterior distribution and obtain the following variational

Table 11 Simulation results for Dirichlet process mixture models

T	α	cv^2 (naive proposal)	cv^2 (VB proposal)	s.e. ratio (naive/VB)
2	1	159.43	32.62	1.52
3	1	142.52	21.63	3.62
5	1	163.13	19.63	10.39
7	1	158.40	29.64	7.52
5	3	235.12	62.35	3.74
7	3	265.32	53.52	5.96
9	3	257.41	37.36	12.94
11	3	246.51	51.74	9.62

decomposition:

$$q(\mathbf{v}, \boldsymbol{\eta}, \mathbf{y}) = \prod_{t=1}^{T-1} q_{1,t}(v_t) \prod_{t=1}^{T} q_{2,t}(\eta_t) \prod_{n=1}^{N} q_{3,n}(y_n),$$

where $q_{1,t}(v_t)$ are beta distributions, $q_{2,t}(\eta_t)$ are exponential family distributions, and $q_{3,n}(y_n)$ are multinomial distributions. We can use the coordinate ascent algorithm to solve the optimization problem. A general rule to choose the truncated level T is to be close to the theoretical value of the expected number of clusters, given N observations:

$$E\,[\text{number of clusters}|x_1, \ldots, x_N] = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1} = \alpha(\psi(\alpha + N) - \psi(\alpha)),$$

where $\psi(\cdot)$ is the digamma function.

We generated $N = 50$ observed data from DP mixture model, and implemented IS with different proposal distributions based on $m = 1,000$ samples. We considered different combinations of the hyper parameters (α, T). Since the number of parameters is large, we only reported the cv^2 and the average of the ratios of the standard errors of the parameter estimates from different methods.

From the results in Table 11, we can see that IS with variational distribution as proposal gives smaller cv^2 than IS with prior as the proposal. The average of the ratios of the standard errors is greater than 1 in all settings, which means using VB as the proposal usually gives smaller standard errors than using the naive proposal. This average ratio becomes larger when the truncated level T is close to the theoretical expectation of the number of clusters (4.49 for $\alpha = 1$ and 9.11 for $\alpha = 3$).

6 Discussion

In this paper, we combine variational approximation and IS to improve the performance of both methods. Using variational approximation as the proposal distribution of IS

can avoid the biased estimate and the lack of uncertainty quantification of the VB estimate. It also provides a way to design a good proposal for IS. We provide theoretical justification of the proposed methods, and numerical results also show that using variational approximation as the proposal can enhance the performance of IS and SIS.

Using VB as proposal for IS tends to be computationally more expensive than some naive choice of the proposal. This is mainly due to the computational cost for finding the VB solution. Sometimes it might be worthwhile to stop the VB algorithm a little early to obtain a rough approximation and allow more time for IS to correct the bias. The tradeoff between VB-SIS1 and VB-SIS2 also illustrates this point.

There are several possible directions for future research in this area. One topic is to further improve the efficiency and performance of the algorithm, especially for models involving HMM and Dirichlet process. For example, the tuning parameter Δ plays an important role in the proposed VB-SIS algorithm for HMM. It is of interest to develop theory or find an analytic expression for the optimal Δ. Another direction is to consider other variational approximations beyond mean-field variational approximation, which may lead to good proposals for importance sampling. Applying the proposed method in more complex models or real data examples is valuable as well.

Appendices

A Proof of Lemma 1

Proof We have $\lim_{n\to\infty} \beta_{2,n} = 1$ immediately from the definition of convergence in (4).

Now we prove $\lim_{n\to\infty} \beta_{1,n} = 1$. For $\forall\ \epsilon > 0$ and $\delta > 0$, define $I_1^{(n)} = \{x : \frac{p_n(x)}{q(x)} < 1-\epsilon\}$, $I_2^{(n)} = \{x : 1-\epsilon \le \frac{p_n(x)}{q(x)} < 1+\delta\}$, and $I_3^{(n)} = \{x : \frac{p_n(x)}{q(x)} \ge 1+\delta\}$.

From (4), we have for any given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n > N$, we have

$$\operatorname{ess\,inf} \frac{p_n}{q} > 1-\epsilon.$$

By the definition of essential infimum, we have

$$\sup\{b \in \mathbb{R} : \mu(\{x : p_n(x)/q(x) < b\}) = 0\} > 1-\epsilon,$$

which implies

$$\mu(I_1^{(n)}) = \mu(\{x : p_n(x)/q(x) < 1-\epsilon\}) = 0.$$

Then we have

$$\int_{I_1^{(n)}} \frac{p_n}{q}\, q\, dx = \int_{I_1^{(n)}} p_n\, dx = 0 \quad \text{for } n > N.$$

So

$$1 = \int_{\mathbb{R}} p_n \, dx = \int_{I_1^{(n)}} \frac{p_n}{q} q \, dx + \int_{I_2^{(n)}} \frac{p_n}{q} q \, dx + \int_{I_3^{(n)}} \frac{p_n}{q} q \, dx$$
$$= \int_{I_2^{(n)}} \frac{p_n}{q} q \, dx + \int_{I_3^{(n)}} \frac{p_n}{q} q \, dx \quad \text{for } n > N.$$

From the definitions of $I_2^{(n)}$ and $I_3^{(n)}$, we have

$$1 = \int_{I_2^{(n)}} \frac{p_n}{q} q \, dx + \int_{I_3^{(n)}} \frac{p_n}{q} q \, dx$$
$$\geq (1-\epsilon) \int_{I_2^{(n)}} q \, dx + (1+\delta) \int_{I_3^{(n)}} q \, dx \quad \text{for } n > N. \tag{10}$$

Similarly, we also have

$$1 = \int_{\mathbb{R}} q \, dx = \int_{I_1^{(n)}} q \, dx + \int_{I_2^{(n)}} q \, dx + \int_{I_3^{(n)}} q \, dx$$
$$= \int_{I_2^{(n)}} q \, dx + \int_{I_3^{(n)}} q \, dx \quad \text{for } n > N. \tag{11}$$

From (10) and (11), the following inequality holds:

$$1 \geq (1-\epsilon) \int_{I_2^{(n)}} q \, dx + (1+\delta) \left(1 - \int_{I_2^{(n)}} q \, dx\right)$$
$$= (1+\delta) - (\epsilon + \delta) \int_{I_2^{(n)}} q \, dx \quad \text{for } n > N. \tag{12}$$

Suppose $\limsup_{n\to\infty} \int_{I_2^{(n)}} q \, dx = \theta(\epsilon, \delta) \in [0, 1]$, then $\liminf_{n\to\infty} \int_{I_3^{(n)}} q \, dx = 1 - \theta(\epsilon, \delta)$ based on (11). Since the definition of $I_3^{(n)}$ depends only on δ, not ϵ, we know that $\liminf_{n\to\infty} \int_{I_3^{(n)}} q \, dx$ also depends only on δ, not ϵ. Thus $\theta(\epsilon, \delta) = \theta(\delta)$ does not depend on ϵ.

Taking limit inferior on both sides of (12), we have

$$1 \geq (1+\delta) - \limsup_{n\to\infty} \left((\epsilon + \delta) \int_{I_2^{(n)}} q \, dx\right) = (1+\delta) - (\epsilon + \delta)\theta(\delta). \tag{13}$$

Therefore,

$$\theta(\delta) \geq \frac{\delta}{\delta + \epsilon}. \tag{14}$$

Note that (14) is true for any $\epsilon > 0$ and $\delta > 0$ selected at the beginning of the proof. Since the left hand side of (14) does not depend on ϵ, letting $\epsilon \to 0$ on the right hand

side of (14), we have $\theta(\delta) \geq 1$. On the other hand, $\limsup_{n\to\infty} \int_{I_2^{(n)}} q\,dx = \theta(\delta) \in [0, 1]$. Therefore we have $\theta(\delta) = 1$, which implies $\liminf_{n\to\infty} \int_{I_3^{(n)}} q\,dx = 1 - \theta(\delta) = 0$ for any $\delta > 0$. From the definition of $\beta_{1,n}$, we have $\lim_{n\to\infty} \beta_{1,n} = 1$.

Since $\mu(\{x : p_n(x)/q(x) < \beta_{2,n}\}) = \mu(\{x : p_n(x)/q(x) > \beta_{1,n}^{-1}\}) = 0$, we have

$$D_f(p_n||q) = \int_{\{\beta_{2,n} \leq \frac{p_n}{q} \leq \beta_{1,n}^{-1}\}} f\left(\frac{p_n}{q}\right) q\,dx \leq \sup_{\beta_{2,n} \leq \beta \leq \beta_{1,n}^{-1}} |f(\beta)|.$$

Letting $n \to \infty$, due to the continuity of f at 1, we have $\lim_{n\to\infty} D_f(p_n||q) \leq f(1) = 0$. □

B Proof of Theorem 1

Proof From Lemma 1, we have

$$\lim_{n\to\infty} \beta_{1,n} = \lim_{n\to\infty} \beta_{2,n} = 1.$$

By L'Hospital's rule, we have $\lim_{t\to 1} \kappa(t) = 1$, where $\kappa(t)$ is defined in (5). Therefore, take limit on the both sides of (6) and (7), we have

$$\lim_{n\to\infty} \frac{KL(p_n||q)}{KL(q||p_n)} = 1\,, \quad \lim_{n\to\infty} \frac{KL(p_n||q)}{\chi^2(p_n||q)} = \frac{1}{2}.$$

□

References

Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. J R Stat Soc Ser B 28:131–142

Armagan A, Dunson D (2011) Sparse variational analysis of linear mixed models for large data sets. Stat Probab Lett 81:1056–1062

Beal MJ, Ghahramani Z (2003) The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Stat 7:453–464

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. Bayesian Anal 1:121–143

Blei DM, Kucukelbir A, Mcauliffe JD (2017) Variational inference: a review for statisticians. J Am Stat Assoc 112:859–877

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Bugallo MF, Elvira V, Martino L, Luengo D, Miguez J, Djuric PM (2017) Adaptive importance sampling: the past, the present, and the future. IEEE Signal Process Mag 34:60–79

Cappé O, Douc R, Guillin A, Marin J-M, Robert CP (2008) Adaptive importance sampling in general mixture classes. Stat Comput 18:447–459

Cappé O, Guillin A, Marin J-M, Robert CP (2004) Population Monte Carlo. J Comput Graph Stat 13:907–929

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Depraetere N, Vandebroek M (2017) A comparison of variational approximations for fast inference in mixed logit models. Comput Stat 32:93–125
Dieng AB, Tran D, Ranganath R, Paisley J, Blei D (2017) Variational inference via χ upper bound minimization. Adv Neural Inf Process Syst 30:2732–2741
Doucet A, Godsill S, Andrieu C (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. Stat Comput 10:197–208
Dowling M, Nassar J, Djurić PM, Bugallo MF (2018) Improved adaptive importance sampling based on variational inference. In: Proceedings of the 26th European signal processing conference (EUSIPCO), IEEE, pp 1632–1636
Hofman JM, Wiggins CH (2008) Bayesian approach to network modularity. Phys Rev Lett 100:258701
Hughes MC, Sudderth E (2013) Memoized online variational inference for Dirichlet process mixture models. Adv Neural Inf Process Syst 1133–1141
Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233
Kong A (1992) A note on importance sampling using standardized weights. University of Chicago, Dept. of Statistics, Tech. Rep 348
Kong A, Liu JS, Wong WH (1994) Sequential imputations and Bayesian missing data problems. J Am Stat Assoc 89:278–288
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Liu JS, Chen R (1998) Sequential Monte Carlo methods for dynamic systems. J Am Stat Assoc 93:1032–1044
Martino L, Elvira V, Louzada F (2017) Effective sample size for importance sampling based on discrepancy measures. Signal Process 131:386–401
Naesseth C, Linderman S, Ranganath R, Blei D (2018) Variational sequential Monte Carlo. In: Proceedings of the twenty-first international conference on artificial intelligence and statistics, proceedings of machine learning research, pp 968–977
Neal RM (2001) Annealed importance sampling. Stat Comput 11:125–139
Owen AB (2013) Monte Carlo theory, methods and examples. http://statweb.stanford.edu/~owen/mc/
O'Hagan A, White A (2019) Improved model-based clustering performance using Bayesian initialization averaging. Comput Stat 34:201–231
Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407
Sanguinetti G, Lawrence ND, Rattray M (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. Bioinformatics 22:2775–2781
Sason I, Verdú S (2016) f-divergence inequalities. IEEE Trans Inf Theory 62:5973–6006
Wang P, Blunsom P (2013) Collapsed variational Bayesian inference for hidden Markov models. AISTATS 599–607
Xing EP, Jordan MI, Russell S (2002) A generalized mean field algorithm for variational inference in exponential families. In: Proceedings of the nineteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp 583–591
You C, Ormerod JT, Mueller S (2014) On variational Bayes estimation and variational information criteria for linear regression models. Aust N Z J Stat 56:73–87
Zreik R, Latouche P, Bouveyron C (2017) The dynamic random subgraph model for the clustering of evolving networks. Comput Stat 32:501–533