# 最大期望模拟退火的贝叶斯变分推理算法

刘浩然*①② 张力悦①② 苏昭玉①② 张 赟③ 张 磊③

①(燕山大学信息科学与工程学院 秦皇岛 066004)

②(河北省特种光纤与光纤传感重点实验室 秦皇岛 066004)

③(北京市机电研究院 北京 100027)

**摘 要**：针对贝叶斯变分推理收敛精度低和搜索过程中易陷入局部最优的问题，该文基于模拟退火理论(SA)和最大期望理论(EM)，考虑变分推理过程中初始先验对最终结果的影响和变分自由能的优化效率问题，构建了双重EM模型学习变分参数的初始先验，以降低初始先验的敏感性，同时构建逆温度参数改进变分自由能函数，使变分自由能在优化过程得到有效控制，并提出一种基于最大期望模拟退火的贝叶斯变分推理算法。该文使用收敛性准则理论分析算法的收敛性，利用所提算法对一个混合高斯分布实例进行实验仿真，实验结果表明该算法具有较优的收敛结果。

**关键词**：贝叶斯变分推理；模拟退火；最大期望；逆温度参数

**中图分类号**：TN911.7 **文献标识码**：A **文章编号**：1009-5896(2021)07-2046-09

**DOI**: 10.11999/JEIT200389

# Bayesian Variational Inference Algorithm Based on Expectation-Maximization and Simulated Annealing

LIU Haoran①② ZHANG Liyue①② SU Zhaoyu①② ZHANG Yun③ ZHANG Lei③

①(*School of Information Science and Engineering, Yanshan University, Qinhuangdao* 066004, *China*)

②(*The Key Laboratory for Special Fiber and Fiber Sensor of Hebei Province, Yanshan University, Qinhuangdao* 066004, *China*)

③(*Beijing Institute of Mechanical and Electrical Engineering, Beijing* 100027, *China*)

**Abstract**: For the problem that Bayesian variational inference with low convergence precision is easy to fall into local optimum during search process, a Bayesian variational inference algorithm based on Expectation-Maximization (EM) and Simulated Annealing (SA) is proposed. The influence of the initial prior on the final result and the optimization efficiency of the variational free energy in the process of variational inference can not be ignored. The double EM is introduced to construct the initial prior of the variational parameter to reduce the sensitivity of the initial prior. And the inverse temperature parameter is introducted to improve the free energy function, which makes the energy be effectively controlled in the optimization process. This paper uses convergence criterion theory to analyze the convergence of the algorithm. The proposed algorithm is used for experiments with an Gaussian mixture model and the experimental results show that the proposed algorithm has better convergence results.

**Key words**: Bayesian variational inference; Simulated Annealing(SA); Expectation-Maximization(EM); Inverse temperature parameter

## 1 引言

在机器学习中，贝叶斯推理已经成为求解不可观测变量后验概率的重要方法，它类似最大期望算法。当模型更为复杂时，贝叶斯精确推理求解具体的参数分布时间开销巨大，而许多时候对于复杂模型的参数分布精确求解不是必要的，而推理的真正目的是得到参数的期望或者近似分布，以了解数据遵循分布的规律[1,2]。精确的贝叶斯推理表现出计算量随变量增加呈指数增长以及计算时间也迅速增加的问题，近似贝叶斯推理针对于此类问题有较好的方法，其中基于蒙特卡罗抽样(Markov Chain

收稿日期：2020-05-15；改回日期：2021-03-19；网络出版：2021-04-15

*通信作者： 刘浩然 liu.haoran@ysu.edu.cn

基金项目：国家重点研发项目(2019YFB1707301)，河北省人才工程培养资助项目(A201903005)

Foundation Items: The National Key Research and Development Program of China (2019YFB1707301), Hebei Talent Engineering Training Support Project(A201903005)

Monte Carlo, MCMC)[3]的随机近似算法通过采样的方式计算参数后验概率，MCMC算法提供了从目标分布渐进地生成精确样本的精确率保证，然而MCMC方法在数据量较小时，无法保证采集过程的准确率，算法的结果较差，算法会随着数据量和采样量的增大，其结果越来越好，但与此同时计算消耗和时间消耗也越来越大，而另一种确定性近似算法即变分推理算法[4]将求解隐参数后验问题转化成变分优化问题，通过迭代找到较优的后验分布解，这种方法在应对更为复杂的模型下，表现出良好的计算效率[5]。

变分推理算法广泛应用于计算机科学、模式识别、图像处理、卡尔曼滤波、战场决策等领域[6–9]。近年来，许多学者深入研究变分推理算法并提出了有效的改进方案，以解决其存在的相关问题。Gianniotis等人[10]使用梯度下降优化变分参数实现模型似然下界重新构建，通过循环迭代得到模型参数后验分布，算法具有较强的泛化能力，可应用于多种场景，如贝叶斯线性回归、贝叶斯多目标分类、概率图像去噪等，但算法在引入梯度下降优化似然下界时加重了原算法的局部最优问题。Shekaramiz等人[11]提出使用贪婪策略筛选出支持度子集进行变分推理，该算法能有效地防止算法过拟合，但算法主要针对稀疏的贝叶斯学习，在数据样本属性较多时，该算法无法得到较好的模型参数后验结果。Katahira等人[12]将最大熵引入确定性退火算法控制退火过程，算法得到相对原变分算法更优的结果，但其随机初始先验对最终结果影响较大。Tabushi等人[13]使用非线性最大化非广延统计力学的Tsallis熵提出了广义确定性退火最大期望(Generalize Deterministic Annealing Expectation-Maximization, GDAEM)算法，通过控制参数得到全局较优解，由于控制参数设置的问题，收敛效率有待提升。Salimans等人[14]提出了基于马尔科夫链的变分近似，该算法将两种近似方式有效结合，实现了速度和精度的相对平衡，但算法针对复杂的模型时，依然表现出采样时间消耗大的问题。

本文针对部分算法对求解模型参数后验分布时间消耗长、收敛精度低的问题，提出一种基于最大期望模拟退火的贝叶斯变分推理算法(Expectation-maximization and Simulated annealing for Variational Bayesian Inference, ES-VBI)，将双重最大期望(Expectation-Maximization, EM)策略应用于初始先验的生成，将模拟退火策略的逆温度参数用于似然下界的优化，最终返回算法迭代最优解，利用收敛性准则理论分析算法的收敛性，将该算法应用于混合高斯分布实例的实验仿真，说明本算法的优势。

## 2　问题描述

变分推理经常被用于计算模型的参数后验分布问题，通过变分法将求解模型参数后验问题转换为变分自由能最大化迭代寻优问题，最终返回全局较优解，即为最合理的近似模型参数后验分布。本文提出的算法针对初始先验和变分自由能在寻优过程中进行优化：(1)在保证算法速率的情况下，满足模型参数的后验分布的近似度最高；(2) 尽量降低算法对初始先验的敏感性，保证算法具有对全局有效的初始先验，防止算法过早地陷入局部最优。针对变分推理过程中的上述特点，给出如下推理模型。

### 2.1　参数描述

本文给出3个定义描述推理模型。

(1)观测变量和不可观测变量。贝叶斯网络中，变量类型分为观测变量和不可观测变量，其中观测变量是直接可观察或可采集的变量；不可观测变量是不可直接观察或不可直接采集的变量，不可观测变量包括潜变量和模型参数，潜变量用于解释观测变量，可以看作其对应观测变量的抽象和概括。模型参数是描述模型数据自身存在规律的参数。

(2)库尔贝克距离(Kullback-Leibler, KL)。衡量假设的近似分布$q$与真实分布$p$之间差异的量称为KL散度，也称KL距离[15]。其表达式如式(1)所示

$$\mathrm{KL}\left(q\,\|p\right)=\int q\left(W\right)\lg\frac{q\left(W\right)}{p\left(W\,|X\right)}\mathrm{d}W \tag{1}$$

其中，$\mathrm{KL}\left(q\,\|p\right)$表示近似分布$q$与真实分布$p$的KL距离；$W$表示模型参数集合(隐变量集合)；$X$表示观测变量数据。

在变分推理中，通过随机赋值或者其他赋值方法给出一个初始先验分布$q$，假设$q$是精确条件的候选假设分布，算法将$\mathrm{KL}\left(q\,\|p\right)$不断优化达到最小，达到假设分布与真实分布的差异无限接近于0，得到对应的$q$为

$$q^{*}=\arg\min\mathrm{KL}\left(q\,\|p\right) \tag{2}$$

(3)证据下界(Evidence Lower BOund, ELBO)。观测变量数据似然对数$\lg p\left(X\right)$与KL距离的差值称为证据下界(ELBO)[16]。在贝叶斯变分推理过程中，初始计算参数后验分布时，根据贝叶斯公式，将数据分布表示为似然对数的形式$\lg p\left(X\right)$，经过化简得到ELBO与$\lg p\left(X\right)$, KL距离的关系表达式

$$L(q)=\lg p\left(X\right)-\mathrm{KL}\left(q\,\|p\right) \tag{3}$$

其中，$L(q)$称为证据下界，即ELBO。此时$\lg p\left(X\right)$是常数并保持不变，要使KL距离最小，则使ELBO

达到最大，KL距离与ELBO的关系如图1所示。由于在变分推理过程中，最小化KL距离无法通过式(1)直接计算，故通过最大化ELBO(即最大化$L(q)$)，最终得到最优的近似分布$q^*=\arg\max L(q)$。

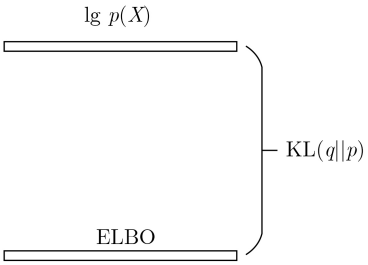


图 1 KL距离与ELBO的关系

### 2.2 模型后验分布描述

设推理模型的观测变量集合为$\boldsymbol{X}=\{x_1, x_2,\cdots,x_n\}$，其中$x_i,(i=1,2,\cdots,n)$表示第$i$个观测变量，$n$表示观测变量的总数；设不可观测变量集合为模型参数$\omega$和潜变量集合$\boldsymbol{z}=\{z_1,z_2,\cdots,z_m\}$，其中模型参数$\omega$表示描述模型数据在统计模型下的自身存在的规律，理论上最优的模型参数与数据无限接近吻合(拟合度无限接近100%)，$z_j,(j=1,2,\cdots,m)$表示第$j$个潜变量，$m$表示潜变量的总数。通过对联合分布$p(z,\omega,X)$积分得到观测数据的边缘密度分布(边际似然) $p(X)$，如式(4)所示

$$p(X)=\iint p(z,\omega,X)\,\mathrm{d}z\mathrm{d}\omega \tag{4}$$

根据贝叶斯定理，$z$和$\omega$在$X$上的后验概率表示为

$$p(z,\omega\,|X)=\frac{p(z,\omega,X)}{p(X)}=\frac{p(X\,|z,\omega)\,p(z,\omega)}{p(X)} \tag{5}$$

将式(4)代入式(5)，式(5)分母部分表示成边缘密度分布形式

$$p(z,\omega\,|X)=\frac{p(X\,|z,\omega)\,p(z,\omega)}{\iint p(z,\omega,X)\,\mathrm{d}z\mathrm{d}\omega} \tag{6}$$

对于许多复杂的模型，式(6)分母部分的积分很难在多项式时间内完成，使得要求解的后验分布$p(z,\omega\,|X)$无法通过贝叶斯定理直接求解，所以本文使用变分推理间接完成对$p(z,\omega\,|X)$的求解。

设$p(z,\omega\,|X)$的近似分布为$q(z,\omega)$，通过优化计算使它不断逼近真实分布$p(z,\omega\,|X)$，依据平均场理论[17,18]，$q(z,\omega)$表示为

$$q(z,\omega)=q_z(z)\,q_\omega(\omega) \tag{7}$$

将边际似然等式(式(4))的两边进行对数运算，式(4)进一步分解为

$$\begin{aligned}\ln p(X)&=\ln\iint p(z,\omega,X)\,\mathrm{d}z\mathrm{d}\omega\\&=\ln\iint q(z,\omega)\cdot\frac{p(z,\omega,X)}{q(z,\omega)}\mathrm{d}z\mathrm{d}\omega\end{aligned} \tag{8}$$

根据Jeason不等式的性质[11]，式(8)等式右边部分存在如式(9)的不等关系

$$\begin{aligned}&\ln\iint q(z,\omega)\frac{p(z,\omega,X)}{q(z,\omega)}\mathrm{d}z\mathrm{d}\omega\\&\geq\iint q(z,\omega)\ln\frac{p(z,\omega,X)}{q(z,\omega)}\mathrm{d}z\mathrm{d}\omega\end{aligned} \tag{9}$$

将式(9)中不等式右边部分进一步分解为

$$\begin{aligned}&\iint q(z,\omega)\ln\frac{p(z,\omega,X)}{q(z,\omega)}\mathrm{d}z\mathrm{d}\omega\\&=\iint q(z,\omega)\ln p(z,\omega,X)\,\mathrm{d}z\mathrm{d}\omega\\&\quad-\iint q(z,\omega)\ln q(z,\omega)\mathrm{d}z\mathrm{d}\omega\\&=\mathrm{E}_{q(z,\omega)}\left[\ln p(z,\omega,X)\right]\\&\quad+H_{q(z,\omega)}(z,\omega)\leq\ln p(X)\end{aligned} \tag{10}$$

其中，$H_q$表示$q$的联合熵，即$-\iint q(z,\omega)\ln q(z,\omega)\,\mathrm{d}z\mathrm{d}\omega=H_{q(z,\omega)}(z,\omega)$；$\mathrm{E}_q(*)$表示在分布$q$上作$*$的数学期望，即$\iint q(z,\omega)\ln p(z,\omega,X)\,\mathrm{d}z\mathrm{d}\omega=\mathrm{E}_{q(z,\omega)}[\ln p(z,\omega,X)]$。

由式(8)、式(9)、式(10)可得$\ln p(X)$的下界为$\mathrm{E}_{q(z,\omega)}[\ln p(z,\omega,X)]+H_{q(z,\omega)}(z,\omega)$，即证据下界ELBO，变分推理就是要最大化$\ln p(X)$的下界，即最大化似然下界ELBO，故将其设为目标函数$L(q)$

$$L(q)=\mathrm{E}_{q(z,\omega)}\left[\ln p(z,\omega,X)\right]+H_{q(z,\omega)}(z,\omega) \tag{11}$$

算法通过不断更新$q(z,\omega)$，使目标函数$L(q)$达到最大，最终求得$q(z,\omega)$的最优近似。

### 2.3 双重EM模型

当前多数的推理算法模型参数初始化都是在一定范围内进行随机赋值操作，但较差的初始化结果，导致算法存在无法收敛到较优值的情况，因此模型参数初始化对初始先验敏感，在增加模型参数的数量时，这个问题显得更加突出。针对先验初始化问题，本算法采用双重最大期望(EM)[19,20]方法，算法在第1次EM算法结果的基础上运行第2次EM算法。在每次循环迭代，在第1阶段，对一组随机初始化数据使用EM算法，得到了第1阶段的模型参数。在第2阶段，针对第1阶段结果再进行第2阶段EM操作，完成ES-VBI的初始化。

求解含观测变量、模型参数的条件概率最大时的模型参数$\omega$, $z$，以求解$\omega$为例，如式(12)所示

$$\omega = \arg\max \sum_{z\in Z} p\left(X, z \left|\omega\right.\right) \tag{12}$$

根据对数函数的单调性可得式(12)的等价形式为

$$\omega = \arg\max \lg \sum_{z\in Z} p\left(X, z \left|\omega\right.\right) \tag{13}$$

算法根据上式进行迭代找到最优解：首先给定一个$\omega$的随机估计值$\omega^k$，然后计算$p\left(X\left|\omega^k\right.\right)$，通过计算更新$\omega^k$作为$\omega^{k+1}$，同时保证

$$\lg p\left(X\left|\omega^{k+1}\right.\right) > \lg p\left(X\left|\omega^k\right.\right) \tag{14}$$

$$\begin{aligned}
&\lg p\left(X\left|\omega\right.\right)\\
=&\lg \sum_{z\in Z}\left[p\left(z\left|\omega\right.\right) p\left(X\left|z,\omega\right.\right)\right]\\
=&\lg \sum_{z\in Z}\left[p\left(z\left|X,\omega^k\right.\right)\right.\\
&\left.\cdot\frac{p\left(z\left|\omega\right.\right) p\left(X\left|z,\omega\right.\right)}{p\left(z\left|X,\omega^k\right.\right)}\right]\\
\geq&\sum_{z\in Z}\left\{p\left(z\left|X,\omega^k\right.\right)\right.\\
&\left.\cdot\lg\left[\frac{p\left(z\left|\omega\right.\right) p\left(X\left|z,\omega\right.\right)}{p\left(z\left|X,\omega^k\right.\right)}\right]\right\}\\
=&U\left(\omega^k,\omega\right)
\end{aligned} \tag{15}$$

式(15)中等号成立的条件是$\omega=\omega^k$，此时有

$$\lg p\left(X\left|\omega\right.\right) = U\left(\omega^k,\omega\right) \tag{16}$$

EM算法的递推过程通过不断更新$\omega^k$直至算法收敛。算法通过迭代至终止条件得到第1阶段的模型分布。

算法将最大似然估计准则应用到第1阶段所提供的模型分布。在执行该步骤过程中，选取来源于第1阶段模型分布的数据样本作为第2阶段的初始先验，该阶段的参数分布后验概率为

$$p\left(\omega\left|X'\right.\right) = \frac{p\left(X'\left|\omega_1\right.\right) p\left(\omega_1\right)}{\iint p\left(z,\omega_1,X'\right)\mathrm{d}z\mathrm{d}\omega_1} \tag{17}$$

其中，$\omega_1$为第1阶段所得到的参数估计；$X'$为第2阶段模型参数估计的数据样本。

在第2阶段的计算模型分布对应的最大似然估计作为最终的初始分布$\omega^*$，如式(18)所示。同理得到$z$的运行结果$z^*$

$$\omega^* = \arg\max l\left(\omega\left|X\right.\right) = \lg\prod_{l=1}^{m} p\left(X_l\left|\omega\right.\right) \tag{18}$$

### 2.4 模拟退火模型

统计物理学的基本公式有$L=E-TS$,($L$是变分自由能，$E$是总能量，$T$是温度，$S$是熵)[21]，根据此公式原理，本文在原目标函数(式(11))的基础上引入逆温度参数$\phi=1/T$构建新的目标函数，则此时变分目标函数表示为

$$\begin{aligned}
L_\phi(q) =&\mathrm{E}_{q(z,\omega)}\left[\ln p\left(z,\omega,X\right)\right]\\
&+\frac{1}{\phi}H_{q(z,\omega)}\left(z,\omega\right)
\end{aligned} \tag{19}$$

逆温度参数有两种特殊情况：如果$\phi\to 0$，则第2项成为整个式子主导，使整个式子最大化相当于使第2个式子最大化，因均匀分布的熵最大，于是整个式子趋于一个均匀分布。如果$\phi=1$，整个式子与式(11)相同，可知得到原始的后验分布，模拟退火对算法的控制逐渐降低。

独立随机变量的联合熵等于各变量熵的和，因此式(19)最右边一项变为式(20)最右边两项，式(20)为

$$\begin{aligned}
L_\phi(q) =&\mathrm{E}_{q_z(z)q_\omega(\omega)}\left[\ln p\left(z,\omega,X\right)\right]\\
&+\frac{1}{\phi}H_{q_z(z)}\left(z\right)+\frac{1}{\phi}H_{q_\omega(\omega)}\left(\omega\right)
\end{aligned} \tag{20}$$

通过迭代计算$q_z\left(z\right)$和$q_\omega\left(\omega\right)$以实现最大化变分自由能$L_\phi(q)$，其中第$k$次的最大化变分自由能${L_\phi}^k(q)$

$${L_\phi}^k(q) = \arg\max L_\phi({q_\omega}^k\left(\omega\right),{q_z}^k\left(\omega\right)) \tag{21}$$

其中，${q_\omega}^k\left(\omega\right)$和${q_z}^k\left(\omega\right)$表示第$k$次迭代的$q_\omega\left(\omega\right)$和$q_z\left(z\right)$。

算法在模拟退火过程中，逐渐向最优解靠拢，相比于未加入逆温度参数的推理过程，算法能控制算法优化的速率，提升算法的效率。变分贝叶斯(Variational Bayes, VB)算法其实是广义的梯度下降算法或广义的EM算法，通过计算模型的分布$q_z\left(z\right)$,$q_\omega\left(\omega\right)$优化最大变分自由能$L_\phi(q)$，实现变分推理。ES-VBI类似于确定性退火变分贝叶斯(Deterministic Annealing Variational Bayesian, DAVB)算法[12]的确定性退火变体。DAVB算法的关键区别在于，温度参数$\phi$修改后置隐藏变量熵对目标函数的权重，而在ES-VBI改变后验参数和隐藏变量熵的权值。

## 3 算法实现

### 3.1 算法执行过程

贝叶斯变分(VB)推理的核心是通过构造一个简单分布$q$去近似待求解的复杂分布$p$，通过不断缩小它们之间的差异即KL距离，不断增大变分自由能(ELBO)，使分布$q$与$p$无限接近，最终求得分布$q$即为最终的参数分布。

在迭代过程中引入拉格朗日乘数[22]$\lambda_1$和$\lambda_2$构造

拉格朗日乘数式$F_1(z,\omega,X,\lambda_1)$和$F_2(z,\omega,X,\lambda_2)$，其中以$L_\phi(q)$为目标函数，对变分边缘分布$q(z,\omega)$进行优化，根据式(7) $q(z,\omega)$的两项$q_z(z)$, $q_\omega(\omega)$可以独立地改变，在优化$q_z(z)$时，$q_\omega(\omega)$当作常数来处理，因此$L_\phi(q)$仅为$q_z(z)$的函数，此时构造的拉格朗日函数仅以$\int q_z(z)\mathrm{d}z-1$为约束条件，故构造了式(22)的拉格朗日函数；在优化$q_\omega(\omega)$时，$q_z(z)$当作常数来处理，因此$L_\phi(q)$仅为$q_\omega(\omega)$的函数，此时构造的拉格朗日函数仅以$\int q_\omega(\omega)\mathrm{d}\omega-1$为约束条件，故构造了式(23)的拉格朗日函数

$$F_1(z,\omega,X,\lambda_1)=L_\phi(q)+\lambda_1\left(\int q_z(z)\mathrm{d}z-1\right) \tag{22}$$

$$F_2(z,\omega,X,\lambda_2)=L_\phi(q)+\lambda_2\left(\int q_\omega(\omega)\mathrm{d}\omega-1\right) \tag{23}$$

对以式(22)和式(23)构造的拉格朗日式求导并令其等于0(其中$F_1(z,\omega,X,\lambda_1)$对$q_z(z)$求导，$F_2(z,\omega,X,\lambda_2)$对$q_\omega(\omega)$求导)，得

$$\mathrm{E}_{q_\omega(\omega)}[\ln p(z,X|\omega)]-\frac{1}{\phi}(\ln q_z(z)+1)+\lambda_1=0 \tag{24}$$

$$\mathrm{E}_{q_z(z)}[\ln p(z,X|\omega)]-\frac{1}{\phi}(\ln q_\omega(\omega)+1)+\lambda_2=0 \tag{25}$$

将式(24)和式(25)化简求解得ES-VBI的$z$和$\omega$的迭代公式

$$q^{t+1}(z)=\frac{1}{N_z}q(z)^\phi*\exp\mathrm{E}_{q(\omega)}[\phi\ln p(X,z|\omega)] \tag{26}$$

$$q^{t+1}(\omega)=\frac{1}{N_\omega}q(\omega)^\phi*\exp\mathrm{E}_{q(z)}[\phi\ln p(X,z|\omega)] \tag{27}$$

其中，$N_z$和$N_\omega$均为归一化常数，$q^{t+1}(z)$和$q^{t+1}(\omega)$是第$t+1$代的后验值估计。

在原变分推理算法流程中加入双重EM和逆温度参数构建退火循环，可以推导出如表1所示的ES-VBI算法。

**表 1 ES-VBI算法**

(1)根据式(12)—式(18)方式构建基于最大似然估计的双重EM模型计算出初始先验$\omega^*$，$z^*$；
(2)设置模拟退火的初始逆温度参数$\phi$,$0<\phi<1$, $t=0$，构建基于逆温参数变分自由能的目标函数；
(3)根据拉格朗日算法子求得$z$和$\omega$的迭代公式；
(4)执行以下迭代步骤直至收敛：
　执行迭代式(26)更新$q(z)$；
　执行迭代式(27)更新$q(\omega)$；
　执行$t=t+1$；
(5)$\phi\leftarrow\phi\times$ const；
(6)如果$\phi<1$，则跳转第(4)步，否则终止算法。

其中，关于模拟退火的迭代参数const设置，退火过程是根据控制该参数的取值使得迭代的逆温度参数$\phi$逐渐从一个小正数(正向趋于0小于1的数)增加至1，以达到最大终止迭代条件，而且逆温度参数初值也是正向趋于0小于1，所以const的初值应设为大于1，这样$\phi\leftarrow\phi\times$ const才能逐渐从0趋于1。本算法作为DA-VB和SA-VB算法的变体，对于该参数的设置结合了退火算法的整体性能以及大体参照DA-VB的参数const设置 1.1和1.2[12]，通过实验发现两种设置情况下均能达到最终的退火性能，因此本文选择使用const=1.1。

### 3.2 算法收敛性分析

**定理1** ES-VBI算法中每次迭代增加$L(q)$，即$L(q^{t+1})\geq L(q^t)$，当且仅当$\phi(q^{t+1}|q^t)=\phi(q^t)$时，等号成立。

**证明** 在给定初始先验$q^0$时，$L(q^t)$收敛到局部最大值，加入退火逆温度参数之后，式(19)中的第2项会随$\phi$的增加而增大，因此保证了$L(q^{t+1})\geq L(q^t)$成立。

当两代后验相同时，即$\phi(q^{t+1}|q^t)=\phi(q^t)$，在两代中，目标函数(式(19))中的第2项相同，即两代的逆温度参数在此时没有改变，所以此时有$L(q^{t+1})=L(q^t)$，综上目标函数$L(q)$是增函数。证毕

**定理2** 如果给定$q^t$，计算$q^{t+1}$作为$q^t$，使E达到最小，此时有$L(q^{t+1})\leq L(q^t)$。

**证明** 计算变分自由能差

$$\begin{aligned}\Delta L=&L(q^{t+1})\leq L(q^t)\\=&\mathrm{E}^{t+1|t}_{q(z,\omega)}[\ln p(z,\omega,X)]-\mathrm{E}^{t}_{q(z,\omega)}[\ln p(z,\omega,X)]\\&+\frac{1}{\phi}\left(H^{t+1|t}_{q(z,\omega)}(z,\omega)-H^{t}_{q(z,\omega)}(z,\omega)\right)\end{aligned} \tag{28}$$

当$\phi>0$时，$\Delta L$的第3项为0或负数。因此如果设置$q^{t+1}(z,\omega)=\arg\min E^{t+1}_{q(z,\omega)}[\ln p(z,\omega,X)]$，则有$\mathrm{E}^{t+1}_{q(z,\omega)}[\ln p(z,\omega,X)]\leq \mathrm{E}^{t}_{q(z,\omega)}[\ln p(z,\omega,X)]$，$\Delta L\leq 0$。

自由能差的第3项可以分解为

$$\begin{aligned}
&H^{t}_{q(z,\omega)}(z,\omega)-H^{t+1|t}_{q(z,\omega)}(z,\omega)\\
&\quad=-\iint q^{t}(z,\omega)\ln q^{t}(z,\omega)\mathrm{d}z\mathrm{d}\omega\\
&\qquad+\iint q^{t}(z,\omega)\ln q^{t+1}(z,\omega)\mathrm{d}z\mathrm{d}\omega\\
&\quad=\iint q^{t}(z,\omega)\ln\frac{q^{t+1}(z,\omega)}{q^{t}(z,\omega)}\mathrm{d}z\mathrm{d}\omega\\
&\quad=E_{q^t}\left\{\ln\frac{q^{t+1}(z,\omega)}{q^{t}(z,\omega)}\right\}\\
&\quad\leq\ln E_{q^t}\left\{\ln\frac{q^{t+1}(z,\omega)}{q^{t}(z,\omega)}\right\}\\
&\quad=\ln\iint\ln q^{t+1}(z,\omega)\mathrm{d}z\mathrm{d}\omega\\
&\quad=\ln 1=0
\end{aligned}\tag{29}$$

因此$H^{t}_{q(z,\omega)}(z,\omega)\leq H^{t+1|t}_{q(z,\omega)}(z,\omega)$成立。　证毕

当$\phi=1$时，变分自由能与改进前的自由能一致。因此，使ELBO最大化的$q(z,\omega)$与使不完全数据似然对数最大化的ML估计值完全相等。换句话说，可以解释为ML估计，即将对数似然函数最大化问题重新表述为自由能最大化问题。由于自由能依赖于$\phi$，通过在原有的EM步骤中加入另一个循环作为退火过程，得到了ES-VBI算法。根据收敛性准则[23]，定理1和定理2在理论上保证了ES-VBI算法的收敛到全局最优解空间内，即算法收敛于全局较优解。

## 4　实验仿真

### 4.1　高斯混合实例和实验设置

高斯模型混合模型是比较经典的包含隐参数的数据模型，我们利用提出的算法从混合高斯模型的数据近似拟合出混合高斯模型中的相关参数的估计值，而且对于在高斯混合模型可以较为准确地衡量算法对模型的鲁棒性和准确性，具有较强的说服力，文献[13]使用了高斯混合模型，文献[14]双变量高斯模型，而且提到可以推广到多变量的混合模型。除此之外在关于变分推理算法对比文献[10–12]中还使用了隐马尔可夫模型、多目标分类、线性回归、逻辑回归等模型进行实验仿真，对比算法的性能，提升算法的可信度。本算法与对比算法属于同类型的推理算法可以针对上述模型进行实验对比，本文由于文章内容限制，仅在高斯混合模型下进行仿真实验，具体如下：本文给出一个高斯混合模型对提出的ES-VBI算法性能进行实验验证，其中该混合模型包含$K$个高斯分布，以及$n$个可观测变量$x_i,i=1,2,\cdots,n$，潜在变量包含$K$个高斯分布的均值$\mu_k,k=1,2,\cdots,K$和$n$个类别参数$c_i,i=1,2,\cdots,n$，该模型表示为

$$\left.\begin{aligned}
&\mu_k\sim N\left(0,\sigma^2\right),k=1,2,\cdots,K\\
&c_i\sim\text{categorical}\left(1/K,\cdots,1/K\right),i=1,2,\cdots,n\\
&x_i\left|c_i\right.,\mu\sim N\left({c_i}^{\mathrm{T}}\mu,1\right),i=1,2,\cdots,n
\end{aligned}\right\}\tag{30}$$

其中，$c_i$表示样本$x_i$对应哪个高斯分布，其服从多项式分布。

因此需要通过算法估计隐变量$\mu_k$和$c_i$，潜变量与观测变量的联合密度如式(31)所示

$$p(\mu,c,x)=p(\mu)\prod_{i=1}^{n}p(x_i)*p(x_i\left|c_i\right.,\mu)\tag{31}$$

假设变分参数为：$m=(m_1,m_2,\cdots,m_k)$,$s^2=\left({s_1}^2,{s_2}^2,\cdots,{s_K}^2\right)$,$\sigma=(\sigma_1,\sigma_2,\cdots,\sigma_n)$，其中第1个和第2个是均值$\mu_k$对应的两个变量，第3个参数为类别$c_i$对应的变量。这3个参数决定了变分分布$q$。通过把联合分布与平均场结合起来，形成高斯混合的变分自由能ELBO，它是一个关于$m$,$s^2$,$\sigma$的函数, 如式(32)所示

$$\begin{aligned}
&\text{ELBO}\left(m,s^2,\sigma\right)\\
&\quad=\sum_{k=1}^{K}\mathrm{E}\left[\lg p\left(\mu_k\right);m_k,{s_k}^2\right]\\
&\qquad+\sum_{i=1}^{n}\left(\mathrm{E}\left[\lg p\left(c_i\right);\varphi_i\right]\right.\\
&\qquad\left.+\mathrm{E}\left[\lg p\left(x_i|c_i,\mu\right);\varphi_i,m,s^2\right]\right)\\
&\qquad-\sum_{i=1}^{n}\mathrm{E}\left[\lg q\left(c_i;\varphi_i\right)\right]\\
&\qquad-\sum_{k=1}^{K}\mathrm{E}\left[\lg q\left(\mu_k;m_k,{s_k}^2\right)\right]
\end{aligned}\tag{32}$$

基于指数族的变分性质，计算得到变分参数$m$,$s^2$,$\sigma$的更新公式为

$$\left.\begin{aligned}
&\sigma_{ik}\propto\exp\left(m_kx_i-\frac{{m_k}^2+{s_k}^2}{2}\right)\\
&m_k=\frac{\sum{}_i\varphi_{ik}x_i}{1/\sigma^2+\sum{}_i\varphi_{ik}}\\
&{s_k}^2=\frac{1}{1/\sigma^2+\sum{}_i\varphi_{ik}}
\end{aligned}\right\}\tag{33}$$

为检验ES-VBI算法的迭代效率、收敛精度，将本文提出的ES-VBI算法与A-VI算法(原文中未给出算法具体名称，本文记作A-VI(Advanced-Variational Inference)算法)[10]、OSBL-VB(Ordinary Sparse Bayesiam Learning-Variational Bayesian)算法[11]、DA-VB(Deterministic Annealing-Variational Bayesian)算法[12]、GDAEM(Gen-

eralized Deterministic Annealing Expectation-Max-imization)算法[13]、MCVI(Markov Chain Variational Inference)算法[14]在混合模型中进行仿真对比，另外本文使用了一些常用的变分推理的评价量：ELOB和时间$t$(s)。每个算法在相同数据集下独立运行100次求取平均值为最后的统计结果。实验环境：处理器Intel(R) Core(TM),CPU i7-7700HQ，主频2.80 GHz，内存为16 GB，Windows10 64 bit操作系统，python3.8版本。

### 4.2 各个算法针对高斯混合模型的实验对比

在混合高斯模型优化中生成器数据为1000时各个算法的ELOB(表示变分自由能，越大越好)和时间$t$(s)的对比结果如表2所示，它们的迭代过程如图2所示。

由表2数据可得，在6种算法中，ES-VBI算法的变分自由能ELOB最大，且时间消耗仅比GDAEM算法多一些，但ES-VBI算法的收敛精度高于GDAEM算法，这是由于ES-VBI算法使用了双重EM相对于GDAEM算法的单层EM时间消耗有所增加，但ES-VBI通过双重EM的时间增加代价换来了算法的收敛精度提升。按照ELOB的收敛结果可以将6种算法的收敛精度从大到小排序为ES-VBI算法、OSBL-VB算法、A-VI算法、GDAEM算法、MCVI算法、DA-VB算法。而且可以看出ES-VBI算法对比于确定性退火算法DA-VB收敛精度提升较大，说明基于逆温度参数对算法的收敛进行有效的控制，使得算法在迭代寻优过程的目标函数值ELOB更大，在每次迭代中更能优化出较好的模型参数。

**表 2 各算法ELOB和时间对比**

| 算法名称 | ELOB | $t$(s) |
|---|---|---|
| ES-VBI | –664428.51 | 9.28 |
| A-VI | –721994.83 | 15.44 |
| OSBL-VB | -707239.90 | 27.97 |
| DA-VB | –922489.46 | 12.83 |
| GDAEM | –790262.55 | 5.62 |
| MCVI | –894487.22 | 30.25 |

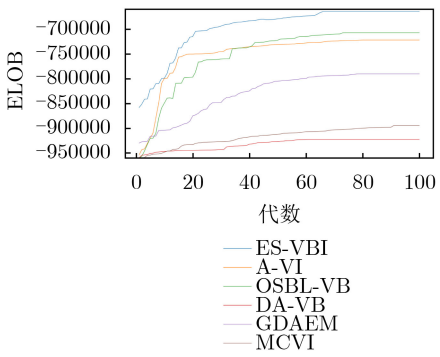


图 2 各算法迭代过程对比图

图2显示各个算法在1至100代的目标函数优化图，可以看出ES-VBI算法的初始化过程的结果显然优于其他的随机初始过程结果，通过双重EM模型初始化模型参数使得算法最终的收敛结果较优。其中A-VI算法和OSBL-VB算法由于分别使用了梯度优化和贪婪策略使得其相对另外3种算法最终的结果较好，通过图2中A-VI的迭代线可以看出其在35代以后很难摆脱梯度优化对局部最优的加强影响，使得后期的迭代线较平，算法后期的优化效率较低，且结果精度有待提升。

将高斯混合模型的$K$设置为5，通过数据生成器生成5000数据样本，对比各个算法对高斯混合模型各个分量的拟合程度，其对比图如图3所示，图3中每种颜色代表一个高斯分量。

由各个算法对应的高斯混合模型拟合图，可以看出本文提出的ES-VBI算法拟合程度最好，很且明显优于DA-VB算法和MCVI算法，DA-VB算法通过最大熵控制退火过程，其并不能很好地适应算法迭代的过程，且随机初始先验也影响了最终的收敛结果。而本算法通过模拟退火的逆温度参数有效地控制迭代的过程，相比于DA-VB算法的确定性退火原理在前期和后期都有对目标函数相对较优的寻优效率，使得算法最终的收敛精度较高，拟合混合高斯分布的各个分量较优。

虽然本算法设计了一个改进的变分推理算法，算法对初始化的敏感性降低和迭代过程的有效控制，但所提出的方案仍然没有得到全局最优，只是在模型参数求解方面提供了一个优化的方法，且上述实验验证了提出算法有效性和较优性能。

## 5 结束语

针对贝叶斯变分推理的初始先验和变分自由能优化两个问题，本文基于模拟退火理论和最大期望理论提出了一种针对降低初始先验影响和变分自由能优化的变分推理方法。通过双重EM模型的构建，使得算法保证近似精度的情况下，进一步提升算法在前期的参数分布质量，通过模拟退火的逆温度参数控制迭代优化过程，进一步提升求解后验分布的精度。通过收敛性理论证明了该推理模型收敛于全局较优解。经过实验仿真证明本文的双重EM模型和模拟退火的逆温度参数的有效性，针对高斯混合模型本文提出算法表现出更好的性能，拟合模型参数的效果较优，同时该算法能够为卡尔曼

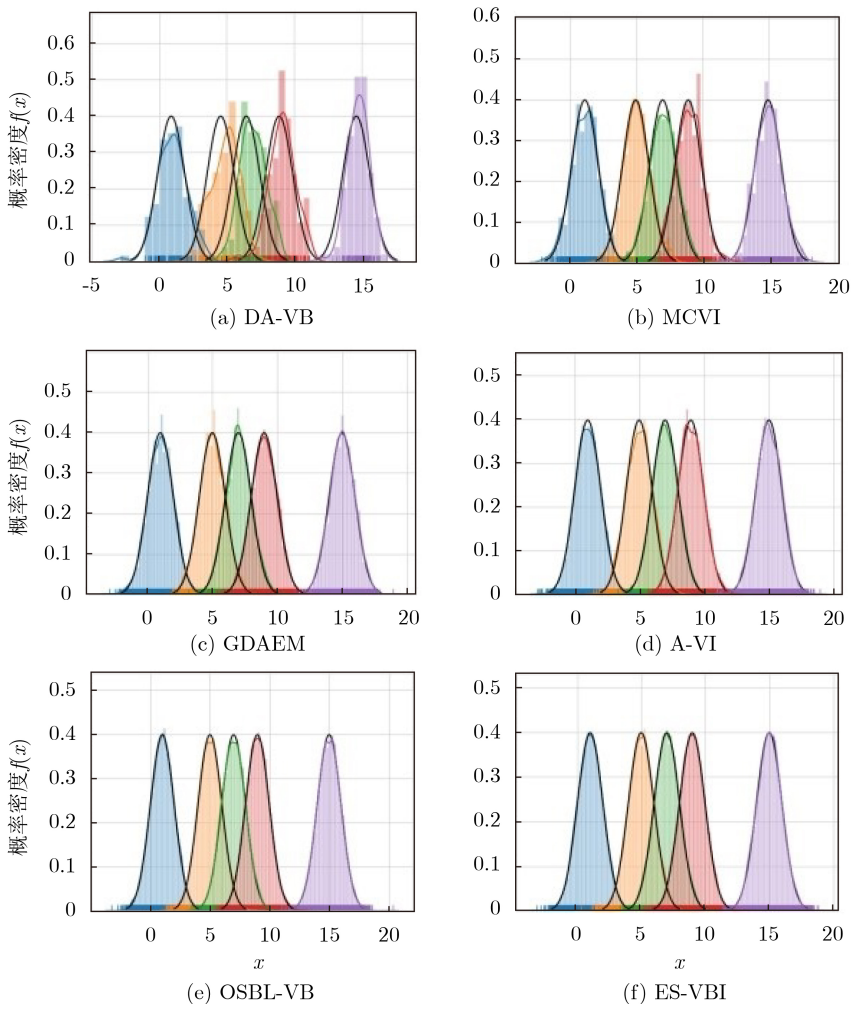


图 3 各算法针对高斯混合模型拟合图

滤波、贝叶斯神经网络参数求解、图像去噪等提供理论支持。

## 参 考 文 献


[1] SEEGER M W and WIPF D P. Variational Bayesian inference techniques[J]. *IEEE Signal Processing Magazine*, 2010, 27(6): 81–91. doi: 10.1109/msp.2010.938082.

[2] MA Yanjun, ZHAO Shunyi, and HUANG Biao. Multiple-model state estimation based on variational Bayesian inference[J]. *IEEE Transactions on Automatic Control*, 2019, 64(4): 1679–1685. doi: 10.1109/TAC.2018.2854897.

[3] JORDAN M I, GHAHRAMANI Z, JAAKKOLA T S, *et al*. An introduction to variational methods for graphical models[J]. *Machine Learning*, 1999, 37(2): 183–233. doi: 10.1023/a: 1007665907178.

[4] LATOUCHE P and ROBIN S. Variational Bayes model averaging for graphon functions and motif frequencies inference in *W*-graph models[J]. *Statistics and Computing*, 2016, 26(6): 1173–1185. doi: 10.1007/s11222-015-9607-0.

[5] WALTER J C and BARKEMA G T. An introduction to Monte Carlo methods[J]. *Physica A*: *Statistical Mechanics and its Applications*, 2015, 418: 78–87. doi: 10.1016/j.physa.2014.06.014.

[6] 孙海英, 李锋, 商慧亮. 改进的变分自适应中值滤波算法[J]. 电子与信息学报, 2011, 33(7): 1743–1747. doi: 10.3724/SP.J.1146.2010.01295.

SUN Haiying, LI Feng, and SHANG Huiliang. Salt-and-pepper noise removal by variational method based on improved adaptive median filter[J]. *Journal of Electronics & Information Technology*, 2011, 33(7): 1743–1747. doi: 10.3724/SP.J.1146.2010.01295.

[7] GUINDANI M and JOHNSON W O. More nonparametric Bayesian inference in applications[J]. *Statistical Methods & Applications*, 2018, 27(2): 239–251. doi: 10.1007/s10260-017-0399-6.

[8] 王瑞, 芮国胜, 张洋. 基于变分贝叶斯推断的半盲信道估计[J]. 哈尔滨工业大学学报, 2018, 50(5): 192–198. doi: 10.11918/j.issn.0367-6234.201708062.

WANG Rui, RUI Guosheng, and ZHANG Yang. Semi-blind channel estimation based on variational Bayesian inference[J]. *Journal of Harbin Institute of Technology*, 2018, 50(5): 192–198. doi: 10.11918/j.issn.0367-6234.

201708062.

[9] DE CASTRO M and VIDAL I. Bayesian inference in measurement error models from objective priors for the bivariate normal distribution[J]. *Statistical Papers*, 2019, 60(4): 1059–1078. doi: 10.1007/s00362-016-0863-7.

[10] GIANNIOTIS N, SCHNÖRR C, MOLKENTHIN C, *et al*. Approximate variational inference based on a finite sample of Gaussian latent variables[J]. *Pattern Analysis and Applications*, 2016, 19(2): 475–485. doi: 10.1007/s10044-015-0496-9.

[11] SHEKARAMIZ M, MOON T K, and GUNTHER J H. Sparse Bayesian learning using variational Bayes inference based on a greedy criterion[C]. 2017 51st Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, USA, 2017: 858–862. doi: 10.1109/ACSSC.2017.8335470.

[12] KATAHIRA K, WATANABE K, and OKADA M. Deterministic annealing variant of variational Bayes method[J]. *Journal of Physics*: *Conference Series*, 2008, 95(1): 012015. doi: 10.1088/1742-6596/95/1/012015.

[13] TABUSHI K and INOUE J. Improvement of EM algorithm by means of non-extensive statistical mechanics[C]. Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop, North Falmouth, USA, 2001: 133–142. doi: 10.1109/NNSP.2001.943118.

[14] SALIMANS T, KINGMA D P, and WELLING M. Markov Chain Monte Carlo and variational inference: Bridging the gap[C]. The 32nd International Conference on International Conference on Machine Learning, Lille, France, 2015: 1218–1226. doi: arxiv.org/pdf/1410.6460.

[15] GHODHBANI E, KAANICHE M, and BENAZZA-BENYAHIA A. Close approximation of kullback–leibler divergence for sparse source retrieval[J]. *IEEE Signal Processing Letters*, 2019, 26(5): 745–749. doi: 10.1109/LSP.2019.2907374.

[16] HE Xingyu, TONG Ningning, and HU Xiaowei. Superresolution radar imaging based on fast inverse-free sparse Bayesian learning for multiple measurement vectors[J]. *Journal of Applied Remote Sensing*, 2018, 12(1): 015013. doi: 10.1117/1.JRS.12.015013.

[17] LALAZISSIS G A, KÖNIG J, and RING P. New parametrization for the Lagrangian density of relativistic mean field theory[J]. *Physical Review C*, 1997, 55(1): 540–543. doi: 10.1103/PhysRevC.55.540.

[18] FORTUNATO S. Community detection in graphs[J]. *Physics Reports*, 2010, 486(3/5): 75–174. doi: 10.1016/j.physrep.2009.11.002.

[19] HUI Zhenyang, LI Dajun, JIN Shuanggen, *et al*. Automatic DTM extraction from airborne LiDAR based on expectation-maximization[J]. *Optics & Laser Technology*, 2019, 112: 43–55. doi: 10.1016/j.optlastec.2018.10.051.

[20] 胡磊, 周剑雄, 石志广, 等. 利用期望-最大化算法实现基于动态词典的压缩感知[J]. 电子与信息学报, 2012, 34(11): 2554–2560. doi: 10.3724/SP.J.1146.2012.00347.

HU Lei, ZHOU Jianxiong, SHI Zhiguang, *et al*. An EM-based approach for compressed sensing using dynamic dictionaries[J]. *Journal of Electronics & Information Technology*, 2012, 34(11): 2554–2560. doi: 10.3724/SP.J.1146.2012.00347.

[21] LALAOUI M, EL AFIA A, and CHIHEB R. A self-tuned simulated annealing algorithm using hidden Markov model[J]. *International Journal of Electrical and Computer Engineering*, 2018, 8(1): 291–298. doi: 10.11591/ijece.v8i1.pp291-298.

[22] HUANG Longbo and NEELY M J. Delay reduction via Lagrange multipliers in stochastic network optimization[J]. *IEEE Transactions on Automatic Control*, 2011, 56(4): 842–857. doi: 10.1109/TAC.2010.2067371.

[23] 陈志敏, 田梦楚, 吴盘龙, 等. 基于蝙蝠算法的粒子滤波法研究[J]. 物理学报, 2017, 66(5): 050502. doi: 10.7498/aps.66.050502.

CHEN Zhimin, TIAN Mengchu, WU Panlong, *et al*. Intelligent particle filter based on bat algorithm[J]. *Acta Physica Sinica*, 2017, 66(5): 050502. doi: 10.7498/aps.66.050502.

刘浩然：男，1980年生，教授，研究方向为贝叶斯算法、工业故障诊断及预测.

张力悦：男，1994年生，博士生，研究方向为贝叶斯算法、工业故障诊断及预测.

苏昭玉：女，1994年生，硕士生，研究方向为贝叶斯算法、工业故障诊断及预测.

张 赟：女，1979年生，博士，研究方向为机械设计及原理、系统建模.

张 磊：男，1991年生，学士，研究方向为数控机床在线测量及系统建模.

责任编辑：余 蓉