Neural Comput & Applic (2013) 23:2039–2050
DOI 10.1007/s00521-013-1445-4

INVITED REVIEW

A review of deterministic approximate inference techniques for Bayesian machine learning

Shiliang Sun

Received: 16 June 2013 / Accepted: 18 June 2013 / Published online: 3 July 2013

Abstract A central task of Bayesian machine learning is to infer the posterior distribution of hidden random variables given observations and calculate expectations with respect to this distribution. However, this is often computationally intractable so that people have to seek approximation schemes. Deterministic approximate inference techniques are an alternative of the stochastic approximate inference methods based on numerical sampling, namely Monte Carlo techniques, and during the last 15 years, many advancements in this field have been made. This paper reviews typical deterministic approximate inference techniques, some of which are very recent and need further explorations. With an aim to promote research in deterministic approximate inference, we also attempt to identify open problems that may be helpful for future investigations in this field.

Keywords Uncertainty · Probabilistic models · Bayesian machine learning · Posterior distribution · Deterministic approximate inference

1 Introduction

Uncertainty is one of the key concepts in modern artificial intelligence and human decision making, which naturally arises in situations where insufficient information is provided or some determining factors are not observed [5, 35, 40]. Probabilistic models, which represent a probability distribution over random variables, provide a principled and solid framework to resolve problems involving uncertainty.

A probabilistic model usually consists of three components: deterministic parameters, hidden variables including latent variables and stochastic parameters, and observable variables, which jointly specify the probability distribution. The hidden and observable variables are both random variables, though the latter are usually clamped to their observed values. The distinction between latent variables and stochastic parameters lies in the fact that the number of latent variables grows with the size of the observed data set, while the number of stochastic parameters is fixed independently of that size [6]. The existence of hidden variables may correspond to missing data or may be imaginary to allow complicated and powerful distributions to be formed. Note that for easy visualization and investigation of properties, a probabilistic model is often represented as a graphical model.

Determining a sole value or a distribution of values for parameters and latent variables in a probabilistic model from experience (i.e., data) is one of the core missions of machine learning. The determined value or distribution can then be used for decision making such as classification and regression. For this purpose, people have to resort to some measure of model appropriateness for the data. For example, one common principle for learning deterministic parameters is maximum likelihood estimation, which returns a parameter setting that maximizes the probability distribution of observed data.

However, maximum likelihood estimation is not appropriate for determining posterior distributions of hidden variables given observed data in which case the principle of Bayesian machine learning should be used. Here, we explicitly distinguish the meanings of estimation and inference. The term estimation refers to determining an

S. Sun (✉)
Department of Computer Science and Technology,
East China Normal University, 500 Dongchuan Road,
Shanghai 200241, China
e-mail: slsun@cs.ecnu.edu.cn; shiliangsun@gmail.com

approximate value for a deterministic parameter, and in contrast inference refers to the process to infer the probability distribution of a random variable. Given observed data D, Bayesian machine learning obtains the posterior distribution over all hidden variables denoted by H through the use of the prior distribution $p(H)$, the likelihood $p(D|H)$, and the model evidence $p(D)$ by Bayes' theorem:

$$p(H|D) = \frac{p(H,D)}{p(D)} = \frac{p(H)p(D|H)}{\int_H p(H,D)\mathrm{d}H}. \tag{1}$$

This process is called Bayesian inference [40, 45, 76]. If we are only interested in some of the hidden variables, a further marginalization of the above posterior over the other hidden variables should be performed. Note that our treatment applies to both discrete and continuous variables, where probability density functions and integrations are used for continuous variables and probability mass functions and summations are used for discrete variables. Since Bayesian machine learning employs a probability distribution rather than a single parameter setting to represent hidden variables, an appropriate mathematical expectation with respect to this distribution is usually necessary at the decision-making stage.

However, for many probabilistic models, an exact evaluation of the needed posterior distribution or the computation of expectations with respect to this distribution is intractable. Therefore, approximate inference is needed. Deterministic approximate inference is an important branch of approximate inference methodologies, and it has been actively studied, especially during the past 15 years. The goal of this paper is to review key advancements and typical techniques in the field of deterministic approximate inference some of which are quite latest, and give suggestions for further research by providing open problems. This review can be helpful for successful applications of deterministic approximate inference techniques to complicated probabilistic models and for the development of novel deterministic approximate inference methods.

The remainder of this paper proceeds as follows. In Sect. 2, we summarize major places where inference is needed and thus also deliver motivations for approximate inference. A concise comparison of stochastic and deterministic approximation inference is also provided. Section 3 surveys representative methods for deterministic approximate inference. Section 4 lists some open problems which may be helpful for promoting research on deterministic approximate inference. Finally, Section 5 concludes this paper.

2 Motivations of approximate inference

In this section, we first summarize three types of computations that are often encountered in Bayesian machine learning and need effective Bayesian inference. This leads naturally to the motivations of approximate inference for complicated probabilistic models. We also briefly compare two different categories of approximation schemes.

2.1 Model selection

For model selection or learning deterministic parameters from the data, one often needs to calculate the data likelihood function and then maximize it. However, for probabilistic models involving hidden variables, these hidden variables should be marginalized out by integration or summation. For many probabilistic models, the integration may not return analytically tractable formulations and the summation may involve exponentially many operations which are also intractable. This makes the exact computation of the likelihood function and thus direct maximum likelihood estimation infeasible.

The expectation maximization (EM) algorithm is an elegant substitution for parameter estimation in this case, which iteratively maximizes the expectation of the complete-data log likelihood [18]. In the E-step, inference is performed, i.e., the posterior distribution of the hidden variables is computed given a current estimate of the parameters. Actually, here the posterior can be known with respect to a multiplicative constant, i.e., we may use the joint distribution of the data and hidden variables as a surrogate without any influence on the final estimated parameters of the EM algorithm. But if we would like to estimate the value or a bound of the likelihood, the multiplicative constant cannot be omitted. In the M-step, the expectation of the complete-data log likelihood is evaluated with respect to this posterior and then maximized. However, the same problem of computational intractability can still exist for the evaluation of the expectation. By resorting to approximate inference which provides a convenient surrogate for the posterior distribution, these problems can be resolved.

Just a note that, for parameter estimation, alternative objective functions (e.g., the pseudolikelihood objective) rather than the current likelihood function can be adopted, whose optimization will not require much inference [35].

2.2 Hidden structure discovery

Sometimes, we are interested in the posterior distribution of hidden variables itself and the statistical information it provides. For example, suppose we have one million scanned books and would like to organize them by their hidden subjects for user-friendly navigation [29]. If we estimate the mode of the posterior distribution for this purpose (i.e., maximum a posteriori estimation), it can be computationally infeasible for complicated models, though the posterior distribution can just be known with respect to a multiplicative constant. In other cases, we may be interested in computing the posterior mean and variance of some hidden variables, which requires their exact posterior distributions. However, the posterior and the involved expectation with integration or summation can be intractable to compute.

All these difficulties can be eliminated if approximate inference is adopted. For instance, an appropriate surrogate distribution which decouples hidden variables or has an analytically convenient form is used to replace the true posterior, or Monte Carlo techniques are used to approximate the true posterior with random samples.

2.3 Bayesian model averaging

Decision making in Bayesian machine learning often requires Bayesian model averaging. The intuition is that we have a set of possible values of related hidden variables including stochastic parameters, so that the final decision should be the weighted average of the hidden variables using their posterior distribution, that is, an evaluation of expectation is needed.

Bayesian model averaging is a process involving integration or summation and thus can be intractable and needs approximate inference. Of course, for simplicity, point estimation of the hidden variables (even on the approximate posterior), e.g., the posterior mean values, may be used to provide a single setup of the involved random variables [70, 89]. In addition, if the integrand in Bayesian model averaging includes multiple functions, any one of them is appropriate to be approximated to make the computation feasible, e.g., the method used in Bayesian logistic regression [6].

2.4 Approximate inference

Now, it is clear that we need to resort to approximate inference when it is intractable to infer posterior distributions or calculate expectations with respect to these distributions [6]. There are two broad categories of approximation schemes: stochastic and deterministic approximate inference techniques.

Stochastic approximation, also known as Monte Carlo techniques, is based on numerical sampling methods. Although they are guaranteed to give exact results with enough samples, Monte Carlo techniques, especially Markov chain Monte Carlo, have two drawbacks: (1) the sampling process can be computationally demanding and thus impractical for large-scale problems; (2) it is hard to assess convergence, namely deciding the burn-in stage and when to stop sampling to get satisfying estimates [6, 22, 45, 78]. This paper will not address such methods. Interested readers are referred to [1, 5, 24, 51].

The strengths and weaknesses of deterministic approximate inference are complementary to those of Monte Carlo techniques. Deterministic approximation uses analytical approximations to the posterior distributions, e.g., the approximate distribution is factorized or has a convenient formulation such as Gaussian, and thus almost never leads to exact results [6]. Some deterministic approximate inference techniques are applicable to large-scale problems.

3 Methods for deterministic approximation inference

In this section, we survey representative methods for deterministic approximate inference. They can be divided into five large categories, with both classical and latest methods included.

3.1 Laplace approximation

The Laplace approximation first finds a mode of the posterior distribution and then construct an approximation with a Gaussian distribution by the second-order Taylor expansion about the mode [6, 35, 40]. A benefit of this approximation is its relative simplicity compared to other approximation techniques [5].

Suppose $\mathbf{h}$ denotes a set of continuous variables, and its posterior distribution $p(\mathbf{h})$ is given by

$$p(\mathbf{h}) = \frac{f(\mathbf{h})}{Z_0}, \tag{2}$$

where Z_0 is a normalization coefficient whose value is probably unknown. At a typical mode $\mathbf{h}_0$ of $f(\mathbf{h})$, which is also a mode of $\ln f(\mathbf{h})$, the gradient $\nabla \ln f(\mathbf{h})$ will vanish and the Hessian matrix (i.e., second-order derivative matrix) is negative definite. Using a second-order Taylor expansion of $\ln f(\mathbf{h})$ centered on $\mathbf{h}_0$, we have

$$\ln f(\mathbf{h}) \approx \ln f(\mathbf{h}_0) - \frac{1}{2}(\mathbf{h}-\mathbf{h}_0)^\top A(\mathbf{h}-\mathbf{h}_0), \tag{3}$$

where A is the negative of the Hessian matrix at $\mathbf{h}_0$ and thus positive definite. Now, we have

$$f(\mathbf{h}) \approx f(\mathbf{h}_0)\exp\left\{-\frac{1}{2}(\mathbf{h}-\mathbf{h}_0)^\top A(\mathbf{h}-\mathbf{h}_0)\right\}. \tag{4}$$

Hence, the approximate distribution $q(\mathbf{h})$, which is a multivariate Gaussian, is given by

$$\begin{aligned} q(\mathbf{h}) &= \frac{|A|^{1/2}}{(2\pi)^{d/2}}\exp\left\{-\frac{1}{2}(\mathbf{h}-\mathbf{h}_0)^\top A(\mathbf{h}-\mathbf{h}_0)\right\} \\ &= \mathcal{N}(\mathbf{h}|\mathbf{h}_0, A^{-1}), \end{aligned} \tag{5}$$

where d is the dimensionality of $\mathbf{h}$ and A is called the precision matrix of the Gaussian distribution [5, 6].

Note that optimization algorithms are usually needed to find the mode $\mathbf{h}_0$ and for multimodal distributions, there will be different choices for the mode which lead to different approximations [6]. In addition, since the Laplace approximation only considers the properties of the true distribution in the locality of a mode, it can fail to represent the global properties.

Recently, Rue et al. [63] proposed an integrated nested Laplace approximation to approximate posterior marginals in latent Gaussian models. The main tool involves applying the Laplace approximation more than once and using numerical integration with respect to low-dimensional random variables.

3.2 Variational inference

Variational inference itself constitutes a large family of methods for deterministic approximate inference. Its idea is to approximate the posterior distribution with a simpler family of probability distributions and then seek the distribution from this family that is closest to the true posterior [22, 31, 77]. The common measure for matching two probability distributions is the Kullback–Leibler (KL) divergence. The optimization of the KL divergence is often transformed to optimize some related bounds on the log data likelihood.

3.2.1 Mean field approximation

Suppose we are approximating the posterior distribution $p(H|D)$ in (1). One way to restrict the family of approximate distributions is to use factorized distributions, i.e.,

$$q(H) = \prod_{i=1}^{M} q_i(H_i), \tag{6}$$

where the elements of H are partitioned into M disjoint groups $\{H_i\}_{i=1}^{M}$ and each factor $q_i(H_i)$ is a probability distribution with a free functional form. This leads to the mean field approximation (a.k.a. variational Bayes approximation) framework [6, 55]. Concretely, the naive mean field approximation refers to the case that all hidden variables of interest are forced to be independent, namely a fully factorized form. The structured mean field approximation corresponds to posterior model structures more complex than the fully factorized form [66].

A useful decomposition for the data likelihood is

$$\ln p(D) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \tag{7}$$

where

$$\mathcal{L}(q) = \int q(H)\ln\left\{\frac{p(H,D)}{q(H)}\right\}\mathrm{d}H,$$
$$\mathrm{KL}(q\|p) = \int q(H)\ln\left\{\frac{q(H)}{p(H|D)}\right\}\mathrm{d}H. \tag{8}$$

The decomposition holds for any probability distribution $q(H)$. As the KL divergence $\mathrm{KL}(q\|p)$ is nonnegative, it is clear that $\mathcal{L}(q)$ is the lower bound of the log data likelihood. Note that, in the variational inference literature, $-\ln p(D) + \mathrm{KL}(q\|p)$ is often called the variational free energy and $-\ln p(D)$ is termed the free energy [82].

The mean field approximation maximizes the lower bound $\mathcal{L}(q)$, which is equivalent to minimizing the KL divergence $\mathrm{KL}(q\|p)$, to find an approximate distribution obeying the factorization requirement given in (6). An iterative algorithm is usually used, which optimizes the objective with respect to each q_i in turn, holding other factors fixed. Suppose we are solving q_j. The lower bound can be written as

$$\begin{aligned}\mathcal{L}(q) &= \int\left\{\prod_i q_i\right\}\left\{\ln p(H,D) - \sum_i \ln q_i\right\}\mathrm{d}H\\ &= \int q_j\left\{\int \ln p(H,D)\prod_{\{i\neq j\}} q_i \mathrm{d}H_i\right\}\mathrm{d}H_j\\ &\quad - \int q_j \ln q_j \mathrm{d}H_j + \mathrm{const}\\ &= \int q_j \ln \tilde{p}(H_j,D)\mathrm{d}H_j - \int q_j \ln q_j \mathrm{d}H_j + \mathrm{const},\end{aligned} \tag{9}$$

where const represents constants, and $\tilde{p}(H_j,D)$ [6] is a new defined probability distribution

$$\ln \tilde{p}(H_j,D) = \mathbb{E}_{\{i\neq j\}}[\ln p(H,D)] + \mathrm{const},$$
$$\mathbb{E}_{\{i\neq j\}}[\ln p(H,D)] = \int \ln p(H,D)\prod_{\{i\neq j\}} q_i \mathrm{d}H_i. \tag{10}$$

The last line of (9) includes a negative KL divergence, which indicates that the optimal q_j is equal to $\tilde{p}(H_j,D)$. Formally, the solution $q_j(H_j)$ is given by

$$\ln q_j(H_j) = \mathbb{E}_{\{i\neq j\}}[\ln p(H,D)] + \mathrm{const}, \tag{11}$$

from which the normalization constant for the distribution is easy to be obtained if necessary.

The above iterative procedure to seek $q(H)$ is guaranteed to converge to a local minimum, since the bound is convex with respect to each factor [6, 10]. There are some potential problems [63] for the mean field approximation: (1) the dependence between some hidden variables is not captured; (2) the posterior variance can be underestimated; (3) the integration computation involved may be intractable for nonconjugate models. For the last point, one can use parametric representations for the approximate distribution or some of its factors, which may permit tractable optimization algorithms to determine parameters.

The mean field approximation has been applied successfully to various areas, e.g., infinite mixtures of Gaussian processes [70], Gaussian process regression networks [84], the stick-breaking construction of beta processes [54], multiple kernel learning [25], and probabilistic matrix factorization [49, 50, 67]. Zhang and Schneider [90] proposed to minimize the composite divergence instead of the KL divergence to find factorized distributions in the context of multi-label classification.

A recent research topic is online variational inference which attempts to provide scalable algorithms that are applicable to large and streaming data. The main technique for reducing the computation time is to avoid an entire pass through all the data at each iteration using stochastic optimization which proceeds by iteratively subsampling the data and adjusting variational parameters based only on the obtained subset. Online mean field approximation algorithms were introduced for latent Dirichlet allocation and the hierarchical Dirichlet process topic model [29, 81]. Based on the mean field approximation, Bryant and Sudderth [12] further developed a split-merge online variational algorithm for hierarchical Dirichlet processes, which allows the truncation level to dynamically vary during learning.

The mean field approximation is not readily applicable to some models (e.g., nonconjugate models [13]) for which the integration computation does not return closed-form functions. For a certain class of nonconjugate models, Wang and Blei [80] developed two methods for variational mean field approximation: Laplace variational inference and delta method variational inference [11]. Laplace variational inference uses Laplace approximations within the coordinate ascent updates, while delta method variational inference applies Taylor expansions to approximate the lower bound $\mathcal{L}(q)$ and then derives variational updates. As a general algorithm implementation of the mean field approximation, variational message passing [85] is mainly applicable to conjugate-exponential models. Knowles and Minka [34] proposed a variational message passing algorithm for some nonconjugate models by deriving lower bounds to approximate the required expectations. Paisley et al. [53] proposed a method to learn variational parameters using a stochastic approximation of the gradient of the variational lower bound with respect to the parameters. The stochastic approximation is given by the Monte Carlo integration, and a variance reduction method based on control variates is further used to reduce the number of samples required to construct the stochastic search direction.

3.2.2 *Parametric distributions*

The family of approximate distributions can also be restricted by parametric distributions, that is,

$$q(H) = q(H|\boldsymbol{\omega}), \tag{12}$$

where $\boldsymbol{\omega}$ denotes the parameters of the distribution. Then, the variational lower bound $\mathcal{L}(q)$ can be optimized as a function of $\boldsymbol{\omega}$ to determine the optimal parameter setting. This kind of variational inference has the potential to capture the dependence between hidden variables.

The variational Gaussian approximation adopts a Gaussian distribution parameterized by the mean and covariance as the approximate posterior, and then finds these parameters through optimizing the variational lower bound. Opper and Archambeau [52] showed that for models with Gaussian priors and factorizing likelihoods, the number of variational parameters in the variational Gaussian approximation is actually very economical. Different from this type of approximation, Archambeau et al. [2, 3] proposed the variational Gaussian process approximation for models with non-Gaussian stochastic process priors and Gaussian likelihoods, where the Gaussian and non-Gaussian processes are both represented by stochastic differential equations.

Ding et al. [19] generalized the idea of variational inference by using approximate distributions from the *t*-exponential family [75] to improve the model robustness over noise. They defined and adopted a new divergence measure called the *t*-divergence, which is the Bregman divergence based on the *t*-entropy and plays the same role as the common KL divergence for variational inference. Challis and Barber [14] proposed affine independent variational inference which optimizes the KL divergence over a class of approximate distributions formed from an affine transformation of independently distributed hidden variables. The resultant approximate distributions can have skewness or other non-Gaussian properties.

3.2.3 *Refined lower bounds*

It is clear from (7) that $\mathcal{L}(q)$ is a lower bound of the log data likelihood. However, the bound can be tightened for specific models. To improve convergence and performance, some refined lower bounds have been presented.

King and Lawrence [33] proposed to optimize the KL-corrected bound, which is a lower bound of the log data likelihood and an upper bound on the standard variational lower bound $\mathcal{L}(q)$, to find deterministic parameters in a Gaussian process model and improved the speed of convergence. This bound is obtained by lower bounding the noise model involved in the data likelihood. They used the mean field approximation for posterior inference and also discussed the possibility of using the KL-corrected bound for posterior updates. This method was also used in Lázaro-Gredilla et al. [44].

Lázaro-Gredilla and Titsias [43] proposed a marginalized variational bound for posterior inference in the heteroscedastic Gaussian process model based on the mean field approximation and the Gaussian parametric approximation. This bound is also a lower bound of the log data likelihood, but tighter than the standard variational lower bound. It holds for models whose approximate posterior distributions are a product of two independent distributions one of which can be optimally represented by the other.

3.2.4 Collapsed variational Bayesian inference

Teh et al. [72] proposed a collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, which combines the mean field approximation and collapsed Gibbs sampling [26]. They made reasonable assumptions, namely the stochastic parameters depend on the latent variables in an exact fashion and the latent variables are assumed to be mutually independent. This algorithm is equivalent to first marginalizing out the stochastic parameters and then approximating the posterior over the latent variables with the mean field approximation. A Gaussian approximation and second-order Taylor expansion are further applied to compute the expectation terms involved in the posterior for computational efficiency. To evaluate test set probabilities, the stochastic parameters are fixed to their mean values with respect to the posterior of the latent variables.

Kurihara et al. [38] applied the collapsed variational Bayesian inference to Dirichlet process mixture models where only partial stochastic parameters are marginalized out. Hensman et al. [27] discussed the difference between the collapsed variational inference and the KL-corrected bound approach where the order of the marginalization and variational approximation is the key. Using the α-divergence, Sato and Nakagawa [65] proposed an interpretation of the collapsed variational Bayesian inference with a zero-order Taylor expansion for latent Dirichlet allocation. Wang and Blei [79] presented a locally collapsed variational inference algorithm, which enables truncation-free variational inference for Bayesian nonparametric models. They used a collapsed Gibbs sampler as a subroutine, which can operate in an unbounded space and thus the resultant algorithm is truncation-free.

3.2.5 Auxiliary-variable methods

By auxiliary-variable methods, we refer to the techniques using auxiliary hidden variables which are not explicitly included in the original models for variational inference.

Gaussian processes have a cubic time complexity with respect to the size of the training set, which makes them intractable for large data sets. To overcome this disadvantage, models for sparse Gaussian processes (e.g., [30, 91]) and mixtures of Gaussian processes (e.g., [68, 70]) have been proposed. Titsias [73] introduced a variational method for sparse Gaussian process regression with additive Gaussian noise that jointly learns the inducing inputs (a.k.a. support inputs) and hyperparameters by maximizing the variational lower bound of the log marginal likelihood. The inducing inputs can be selected from the training data or considered as auxiliary pseudo inputs and determined by applying continuous optimization. Unlike previous sparse Gaussian process methods, here the inducing inputs are defined to be variational parameters.

The auxiliary hidden variables are $\mathbf{f}_m$ evaluated at the inducing inputs X_m [73]. They are function values drawn from the same Gaussian process prior as the training function values $\mathbf{f}$ whose corresponding observations are $\mathbf{y}$. $\mathbf{f}_m$ is assumed to be a sufficient statistic in the sense that $\mathbf{z}$ and $\mathbf{f}$ are independent given $\mathbf{f}_m$ where $\mathbf{z}$ denotes any finite set of function values. To determine the involved quantities, the KL divergence between the augmented variational posterior $q(\mathbf{f}, \mathbf{f}_m)$ and the augmented true posterior $p(\mathbf{f}, \mathbf{f}_m| \mathbf{y})$ is minimized, where $q(\mathbf{f}, \mathbf{f}_m) = p(\mathbf{f}|\mathbf{f}_m)\ \phi(\mathbf{f}_m)$ and $\phi(\mathbf{f}_m)$ is a variational distribution [73].

This method was extended by Titsias and Lawrence [74] to the Gaussian process latent variable model (GP-LVM). The GP-LVM is an application of Gaussian process models to nonlinear dimensionality reduction (e.g., [69]) and can be regarded as a multivariate Gaussian process regression model where the inputs are treated as latent variables. They computed a closed-form variational lower bound of the GP-LVM log marginal likelihood, which depends on a lower bound on the log marginal likelihood of a Gaussian process regression model where the auxiliary hidden variables appear to eliminate a cumbersome term. The full variational distribution that results in the final lower bound is given by

$$q\left(\{\mathbf{f}_d, \mathbf{u}_d\}_{d=1}^{\widetilde{D}}, X\right) = \left(\prod_{d=1}^{\widetilde{D}} p(\mathbf{f}_d|\mathbf{u}_d, X)\phi(\mathbf{u}_d)\right) q(X), \quad (13)$$

where $\widetilde{D}$ is the dimensionality of the observed data vector, $\mathbf{f}_d$ is the Gaussian process latent function values evaluated at latent inputs X, $\mathbf{u}_d$ is the auxiliary hidden variables, $\phi(\mathbf{u}_d)$ is an arbitrary variational distribution over $\mathbf{u}_d$, and $q(X)$ is a variational distribution which has a factorized Gaussian form [74].

The GP-LVM framework was further extended to variational Gaussian process dynamical systems for modeling time series data [16, 56]. The variational approximation approach in Titsias [73] was extended to the multiple-output case by Álvarez et al. [4].

3.2.6 *Mixtures of distributions*

To enhance the representation capability of the family of approximate distributions and capture multimodality in the true posterior distribution, variational inference with mixtures of distributions has been presented. For example, mixtures of factorized distributions and mixtures of Gaussian distributions were used as variational distributions [7, 9, 40].

Gershman et al. [23] developed a variational inference method that can capture multiple modes of the posterior and is applicable to many nonconjugate models with continuous hidden variables. The family of approximate distributions is a uniform mixture of Gaussians whose means and variances are variational parameters. They termed their method nonparametric variational inference (NPV). To approximate the variational lower bound, NPV employs the Taylor series approximation of the log joint distribution and a bound on the entropy term.

3.2.7 *Convex relaxation*

Variational inference in probabilistic models can be represented as a constrained optimization problem of a certain functional. This motivates a class of deterministic approximate inference methods known as convex relaxation [28, 77]. The essence of convex relaxation is to construct an appropriate convex optimization problem that can be conveniently handled by optimization tools.

There are two key ingredients in the convex relaxation algorithms, a convex set as the constraint set and a convex surrogate of the functional to be optimized [58]. Examples of convex relaxation techniques include linear programming relaxations (e.g., for maximum a posterior estimation) [36, 37, 62, 71] and the more expressive conic programming relaxations [58, 77].

3.3 Assumed-density filtering

Assumed-density filtering is a fast sequential method for deterministic approximate inference, which minimizes the KL divergence between the true posterior and the approximate posterior (the reverse form of the KL divergence used in the mean field approximation). It has been independently developed in the control, statistics, and artificial intelligence literature [45], and is often encountered in conjunction with other terms such as moment matching and online Bayesian learning.

Suppose now we are minimizing $\mathrm{KL}(p\|q)$ with respect to an approximate distribution $q(H)$, where $p(H)$ is a fixed distribution and $q(H)$ is from the exponential family with the following parametric form

$$q(H) = h(H)g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^\top \mathbf{u}(H)\}, \tag{14}$$

where $\boldsymbol{\eta}$ represents the natural parameters of the distribution, and $\mathbf{u}(H)$ is the sufficient statistic function [6, 35]. The specific type of the exponential family, e.g., a Gaussian distribution or Dirichlet distribution, is usually determined from the context. We only need to seek the natural parameters $\boldsymbol{\eta}$ in order to determine $q(H)$. The KL divergence can be written as

$$\mathrm{KL}(p\|q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top \mathbb{E}_{p(H)}[\mathbf{u}(H)] + \mathrm{const}, \tag{15}$$

where const indicates terms independent of $\boldsymbol{\eta}$. Setting the gradient with respect to $\boldsymbol{\eta}$ to zero results in

$$-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(H)}[\mathbf{u}(H)]. \tag{16}$$

Since $-\nabla_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}) = \mathbb{E}_{q(H)}[\mathbf{u}(H)]$ for distributions from the exponential family [6], we have

$$\mathbb{E}_{q(H)}[\mathbf{u}(H)] = \mathbb{E}_{p(H)}[\mathbf{u}(H)]. \tag{17}$$

Therefore, the optimal parameters should match the expected sufficient statistics. The optimization process is actually moment matching [6].

Let the joint distribution over observed data D and hidden variables H be $p(H, D)$. Now, we show how to use assumed-density filtering to approximate the posterior $p(H|D)$ by $q(H)$ and estimate the model evidence $p(D)$. For specific illustrative examples, see [45].

First, decompose $p(H, D)$ into a product of simple factors

$$p(H,D) = \prod_{i=1}^{L} t_i(H). \tag{18}$$

Second, choose the proper parametric distribution for $q(H)$ from the exponential family.

Finally, incorporate the factors $t_i(H)$ sequentially into the approximate posterior [45]. Initialize with $q(H) = 1$. When incorporating the factor $t_i(H)$, calculate the exact posterior

$$p_i(H) = \frac{q(H)t_i(H)}{Z_i}, \tag{19}$$

where $Z_i = \int_H q(H)t_i(H)\mathrm{d}H$. By minimizing the KL divergence $\mathrm{KL}(p_i(H)\|q(H))$ through (17) where $p(H)$ is replaced with $p_i(H)$, we can update $q(H)$.

It is clear that the finally obtained $q(H)$ is an approximation of $p_L(H)$ in the sense of minimizing the KL divergence and is also used as the approximate distribution for $p(H|D)$. Using the former relationship recursively, we get

$$p(H|D) \approx \frac{\prod_{i=1}^{L} t_i(H)}{\prod_{i=1}^{L} Z_i} = \frac{p(H,D)}{\prod_{i=1}^{L} Z_i}. \tag{20}$$

Therefore, the model evidence $p(D)$ can be estimated by accumulating the normalization factors Z_i generated by each update, i.e., $p(D) \approx \prod_{i=1}^{L} Z_i$.

Note that assumed-density filtering performs worse than some off-line deterministic approximate inference methods due to its sequential nature [45]. Factors discarded early can be useful later to return a better approximate posterior.

However, the online nature of assumed-density filtering is indeed an appealing characteristic for some learning scenarios. For example, assumed-density filtering was successfully combined with an entropy-reduction based point selection criterion to provide sparse Gaussian processes [30, 41, 42, 91].

3.4 Expectation propagation

Expectation propagation [46] extends assumed-density filtering to batch situations, by incorporating iterative refinements of the approximate posterior. For some probabilistic models, its performance is significantly superior to assumed-density filtering and several other approximation methods [39, 45]. Recent applications of expectation propagation include approximate inference for sparse Gaussian processes [59] and Gaussian process dynamical systems [17], and marginal approximations in latent Gaussian models [15].

For the joint distribution $p(H, D)$ given in (18), we now show how to use expectation propagation [6, 45, 46] to get the approximate posterior $q(H)$ and the estimate of the model evidence $p(D)$. Expectation propagation assumes that the approximate posterior is a member of the exponential family and has the following factorized form

$$q(H) = \frac{1}{Z}\prod_{i=1}^{L} \tilde{t}_i(H), \tag{21}$$

where each factor $\tilde{t}_i(H)$ is approximation to $t_i(H)$, and Z is the normalization constant to make $q(H)$ be a probability distribution.

In expectation propagation, each factor is optimized in turn with the remaining factors fixed. First, initialize the factors $\tilde{t}_i(H)$ properly, and accordingly $q(H)$ is initialized by

$$q(H) = \frac{\prod_{i=1}^{L} \tilde{t}_i(H)}{\int_H \prod_{i=1}^{L} \tilde{t}_i(H)\mathrm{d}H}. \tag{22}$$

Then, cycle through the factors updating only one of them at a time until all factors converge. Suppose we are refining $\tilde{t}_j(H)$ from the current $q(H)$. Define a function $q^{\setminus j}(H)$ which is an unnormalized distribution as

$$q^{\setminus j}(H) = \frac{q(H)}{\tilde{t}_j(H)}, \tag{23}$$

and combine it with the true term $t_j\ (H)$ to induce a distribution

$$\frac{1}{Z_j} q^{\setminus j}(H) t_j(H), \tag{24}$$

where $Z_j = \int_H q^{\setminus j}(H) t_j(H)\mathrm{d}H$. By minimizing the KL divergence

$$\mathrm{KL}\left(\frac{1}{Z_j} q^{\setminus j}(H) t_j(H) \| q^{\mathrm{new}}(H)\right) \tag{25}$$

with moment matching, we can obtain the distribution $q^{\mathrm{new}}(H)$. Therefore, we have

$$\tilde{t}_j^{\mathrm{new}}(H) = K \frac{q^{\mathrm{new}}(H)}{q^{\setminus j}(H)}, \tag{26}$$

which is determined up to a scale. Because $\tilde{t}_j^{\mathrm{new}}(H)$ is an approximation to the true factor $t_j\ (H)$, to fix K, we can require

$$\int q^{\setminus j}(H) \tilde{t}_j^{\mathrm{new}}(H)\mathrm{d}H = \int q^{\setminus j}(H) t_j(H)\mathrm{d}H. \tag{27}$$

It follows that $K = Z_j$. Hence, the refinement of $\tilde{t}_j(H)$ is given by

$$\tilde{t}_j(H) = Z_j \frac{q^{\mathrm{new}}(H)}{q^{\setminus j}(H)}. \tag{28}$$

Of course, the $q(H)$ should also be updated to $q^{\mathrm{new}}(H)$.

The model evidence

$$p(D) = \int \prod_{i=1}^{L} t_i(H)\mathrm{d}H \tag{29}$$

can be approximated by replacing the factors $t_i(H)$ with their approximations $\tilde{t}_i(H)$, that is, $p(D) \approx \int \prod_{i=1}^{L} \tilde{t}_i(H)\mathrm{d}H$, where $\int \prod_{i=1}^{L} \tilde{t}_i(H)\mathrm{d}H$ is also the normalization constant of the final $q(H)$ as indicated by (21).

One disadvantage of expectation propagation is that in general, it is not guaranteed to converge. Moreover, since moment matching requires the evaluation of expectations, it is limited to the class of models for which this operation is possible [85]. In addition, expectation propagation is likely to find weak solutions when applied to multimodal distributions such as mixtures of certain distributions [6, 85].

Recently, Riihimäki et al. [61] proposed a nested expectation propagation algorithm for Gaussian process multiclass classification with the multinomial probit likelihood. It applies inner expectation propagation approximations for each likelihood term within the outer expectation propagation iterations.

3.5 Loopy belief propagation

Belief propagation [57] provides an efficient framework for exact inference of marginal posterior distributions in tree-structured probabilistic graphical models. It has different algorithmic formulations, and the most modern treatment is the sum-product algorithm on the factor graph representation [5, 6]. The use of the distributive law makes message passing operations efficient.

Since the message passing rules in belief propagation are regardless of the global structure of the graphs and thus purely local, one can apply belief propagation to graphs with loops though there is no guarantee that good results will be obtained. This method is known as loopy belief propagation [6, 48, 88]. Because messages can propagate many times around the graphs, loopy belief propagation can fail to converge. However, when it converges, the approximations to the correct marginals can be surprisingly accurate [5, 48].

To understand the success of loopy belief propagation, people have provided some theoretical results. Yedidia et al. [87] showed that the fixed points of loopy belief propagation correspond to stationary points of a simple approximation to the free energy, known as the Bethe free energy in statistical physics. This result makes connections with variational inference approaches. As a generalization of the Bethe free energy, the Kikuchi free energy [32] can give better approximations to the free energy. Inspired by this, Yedidia et al. [87] proposed generalized belief propagation whose fixed points can be shown to be equivalent to the stationary points of the Kikuchi free energy [86]. By establishing a connection between the Hessian of the Bethe free energy and the edge zeta function, Watanabe and Fukumizu [83] recently gave a new theoretical analysis of loopy belief propagation.

A disadvantage of loopy and generalized belief propagation is that they do not always converge to a fixed point. Thereby, alternatively one can explicitly minimize the Bethe or Kikuchi free energy to perform approximate inference [28]. Note that, generally, the Bethe free energy is not an upper or lower bound on the true free energy. Ruozzi [64] showed that for graphical models with binary variables and log-supermodular potential functions, the Bethe partition function always lower bounds the true partition function. In addition, people have proposed another class of algorithms, called loop corrections [47, 60], for approximate inference in loopy graphical models, which are based on the concept of cavity distributions.

4 Open problems

Now, we proceed to give open problems which can be important for further developments of the field of deterministic approximate inference for Bayesian machine learning.

4.1 Nonconjugate models and complex approximate distributions

To better explain the data, some highly flexible probabilistic models have to be adopted, which can be nonconjugate. This necessitates deterministic approximate inference methods for nonconjugate models. As nonconjugate models can differ largely from one to another, we conjecture that it is hard to give a generic deterministic approximate inference method which is good to all nonconjugate models. However, for specific nonconjugate models, it is quite possible to give proper deterministic approximate inference methods. Therefore, providing a categorization of nonconjugate models and identifying corresponding deterministic approximate inference methods can be interesting research topics. Moreover, for a specific nonconjugate model or a specific class of nonconjugate models, determining the performance limit of deterministic approximate inference is also of interest.

In addition, to enlarge the family of approximate distributions and capture some desirable posterior properties, people have considered mixtures of distributions and complex parametric distributions. Exploring more complex approximate distributions to extend the scope of feasible approximate distributions is worth studying.

4.2 Deterministic approximation inference for new models

Stochastic and deterministic approximation inference are two different kinds of approximation schemes. People may adopt either of them for approximate inference, given their personal expertise and like. Some recently proposed probabilistic models such as HDP–HMM [21], BP–HMM [20], distance-dependent Chinese restaurant processes [8], and infinite mixtures of multiple-output Gaussian processes [68] adopt stochastic approximation inference methods. Given the richness and advantages of deterministic approximate inference, it is therefore interesting to apply deterministic approximation inference techniques to these models and improve the efficiency and scalability. Furthermore, to analyze massive and streaming data, one can consider to use fast or online deterministic approximate inference techniques.

4.3 Different measures for matching two distributions

For deterministic approximate inference, most methods use the KL divergence as a measure to match two distributions. The KL divergence appears naturally when we derive the lower bound of the log data likelihood. But it should not be the only choice. We have indeed mentioned a t-divergence based variational inference method. Therefore, here we present two natural questions for deterministic approximate inference, which are how many appropriate measures we can choose and when we prefer one to another.

4.4 Prediction accuracy driven deterministic approximate inference

Current deterministic approximate inference methods usually find an approximate distribution by optimizing a likelihood or KL divergence related functional. However, this kind of objective is not directly related to the final prediction accuracies of the used probabilistic models.

Can we characterize the prediction accuracy of the true probabilistic models and find approximate distributions by optimizing some bound on the prediction accuracy? Can we directly represent the prediction accuracy using the approximate distribution and then determine a specific distribution by maximizing the accuracy or some bound of the accuracy? It would also be interesting to evaluate how much approximation is induced in terms of the loss of the prediction accuracies when the approximate posterior is substituted for the true posterior.

5 Conclusion

In this paper, we have summarized motivations for approximate inference, reviewed the major classes of deterministic approximate inference techniques, and presented open problems which are probably useful to the advancement of the research of deterministic approximate inference.

This paper can reflect the whole picture of the current deterministic approximate inference methodologies in Bayesian machine learning, although it is not possible and necessary to enumerate every deterministic approximate inference technique that has been used so far. We hope this review could provide readers fundamental techniques to implement approximate inference in their concrete probabilistic models and even inspire them to propose new deterministic approximate inference methods.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Project 61075005, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

1. Andrieu C, Freitas N, Doucet A, Jordan M (2003) An introduction to MCMC for machine learning. Mach Learn 50:5–43
2. Archambeau C, Cornford D, Opper M, Shawe-Taylor J (2007) Gaussian process approximation of stochastic differential equations. JMLR Workshop Conf Proc 1:1–16
3. Archambeau C, Opper M, Shen Y, Cornford D, Shawe-Taylor J (2008) Variational inference for diffusion processes. Adv Neural Inf Process Syst 20:17–24
4. Álvarez M, Luengo D, Titsias M, Lawrence N (2010) Efficient multioutput Gaussian processes through variational inducing kernels. In: Proceedings of the 13th International conference on artificial intelligence and statistics, pp 25–32
5. Barber D (2012) Bayesian reasoning and machine learning. Cambridge University Press, Cambridge
6. Bishop C (2006) Pattern recognition and machine learning. Springer, New York
7. Bishop C, Lawrence N, Jaakkola T, Jordan M (1998) Approximating posterior distributions in belief networks using mixtures. Adv Neural Inf Process Syst 10:416–422
8. Blei D, Frazier P (2011) Distance dependent Chinese restaurant processes. J Mach Learn Res 12:2461–2488
9. Bouchard G, Zoeter O (2009) Split variational inference. In: Proceedings of the 26th international conference on machine learning, pp 57–64
10. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
11. Braun M, McAuliffe J (2010) Varitional inference for large-scale models of discrete choice. J Am Stat Assoc 105:324–335
12. Bryant M, Sudderth E (2012) Truly nonparametric online variational inference for hierarchical Dirichlet processes. Adv Neural Inf Process Syst 25:2708–2716
13. Carlin B, Polson N (1991) Inference for nonconjugate Bayesian models using the Gibbs sampler. Can J Stat 19:399–405
14. Challis E, Barber D (2012) Affine independent variational inference. Adv Neural Inf Process Syst 25:2195–2203
15. Cseke B, Heskes T (2011) Approximate marginals in latent Gaussian models. J Mach Learn Res 12:417–454
16. Damianou A, Titsias M, Lawrence N (2011) Variational Gaussian process dynamical systems. Adv Neural Inf Process Syst 24:2510–2518
17. Deisenroth M, Mohamed S (2012) Expectation propagation in Gaussian process dynamical systems. Adv Neural Inf Process Syst 25:2618–2626
18. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incompute data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39:1–38
19. Ding N, Vishwanathan S, Qi Y (2011) *t*-divergence based approximate inference. Adv Neural Inf Process Syst 24:1494–1502
20. Gao Q, Sun S (2013) Human activity recognition with beta process hidden Markov models. In: Proceedings of the international conference on machine learning and cybernetics, pp 1–6
21. Gao Q, Sun S (2013) Trajectory-based human activity recognition with hierarchical Dirichlet process hidden Markov models. In: Proceedings of the 1st IEEE China summit and international conference on signal and information processing, pp 1–5
22. Gershman S, Blei D (2012) A tutorial on Bayesian nonparametric models. J Math Psychol 56:1–12
23. Gershman S, Hoffman M, Blei D (2012) Nonparametric variational inference. In: Proceedings of the 29th International Conference on machine learning, pp 1–8
24. Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London
25. Gönen M (2012) Bayesian efficient multiple kernel learning. In: Proceedings of the 29th international conference on machine learning, pp 1–8
26. Griffiths T, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101:5228–5235
27. Hensman J, Rattray M, Lawrence N (2012) Fast variational inference in the conjugate exponential family. Adv Neural Inf Process Syst 25:2897–2905

28. Heskes T (2006) Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. J Artif Intell Res 26:153–190
29. Hoffman M, Blei D, Wang C, Paisley J (2013) Stochastic variational inference. J Mach Learn Res 14 (in press)
30. Huang R, Sun S (2012) Sequential training of semi-supervised classification based on sparse Gaussian process regression. In: Proceedings of the international conference on machine learning and cybernetics, pp 702–707
31. Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233
32. Kikuchi R (1951) The theory of cooperative phenomena. Phys Rev 81:988–1003
33. King N, Lawrence N (2006) Fast variational inference for Gaussian process models through KL-correction. Lect Notes Aritif Intell 4212:270–281
34. Knowles D, Minka T (2011) Non-conjugate variational message passing for multinomial and binary regression. Adv Neural Inf Process Syst 25:1701–1709
35. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge
36. Kolter J, Jaakkola T (2012) Approximate inference in additive factorial HMMs with application to energy disaggregation. In: Proceedings of the 15th international conference on artificial intelligence and statistics, pp 1472–1482
37. Kulesza A, Pereira F (2008) Structured learning with approximate inference. Adv Neural Inf Process Syst 20:785–792
38. Kurihara K, Welling M, Teh Y (2007) Collapsed variational Dirichlet process mixture models. In: Proceedings of the 20th international joint conference on artificial intelligence, pp 2796–2801
39. Kuss M, Rasmussen C (2006) Assessing approximations for Gaussian process classification. Adv Neural Inf Process Syst 18:699–706
40. Lawrence N (2000) Variational inference in probabilistic models. PhD thesis, Computer Laboratory, University of Cambridge, Cambridge, UK
41. Lawrence N, Seeger M, Herbrich R (2003) Fast sparse Gaussian process methods: the informative vector machine. Adv Neural Inf Process Syst 15:625–632
42. Lawrence N, Seeger M, Herbrich R (2005) The informative vector machine: a practical probabilistic alternative to the support vector machine. Technical Report CS-04-07, Department of Computer Science, University of Sheffield, Sheffield, UK
43. Lázaro-Gredilla M, Titsias M (2011) Variational heteroscedastic Gaussian process regression. In: Proceedings of the 28th international conference on machine learning, pp 841–848
44. Lázaro-Gredilla M, Vaerenbergh S, Lawrence N (2012) Overlapping mixutres of Gaussian processes for the data association problem. Pattern Recognit 45:1386–1395
45. Minka T (2001) A family of algorithms for approximate Bayesian inference. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA
46. Minka T (2001) Expectation propagation for approximate Bayesian inference. In: Proceedings of the 17th conference on uncertainty in artificial intelligence, pp 362–369
47. Mooij J, Kappen H (2007) Loop corrections for approximate inference on factor graphs. J Mach Learn Res 8:1113–1143
48. Murphy K, Weiss Y, Jordan M (1999) Loopy belief propagation for approximate inference: an empirical study. In: Proceedings of the 15th conference on uncertainty in artificial intelligence, pp 467–475
49. Nakajima S, Sugiyama M, Babacan S, Tomioka R (2013) Global analytic solution of fully-observed variational Bayesian matrix factorization. J Mach Learn Res 14:1–37
50. Nakajima S, Tomioka R, Sugiyama M, Babacan S (2012) Perfect dimensionality recovery by variational Bayesian PCA. Adv Neural Inf Process Syst 25:980–988
51. Neal R (1993) Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, Toronto, Canada
52. Opper M, Archambeau C (2009) The variational Gaussian approximation revisited. Neural Comput 21:786–792
53. Paisley J, Blei D, Jordan M (2012) Variational Bayesian inference with stochastic search. In: Proceedings of the 29th international conference on machine learning, pp 1–8
54. Paisley J, Carin L, Blei D (2011) Variational inference for stick-breaking beta process priors. In: Proceedings of the 28th international conference on machine learning, pp 889–896
55. Parisi G (1988) Statistical field theory. Addison-Wesley, New York
56. Park H, Yun S, Park S, Kim J, Yoo C (2012) Phoneme classification using constrained variational Gaussian process dynamical system. Adv Neural Inf Process Syst 25:2015–2023
57. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco
58. Peng J, Hazan T, Srebro N, Xu J (2012) Approximate inference by intersecting semidefinite bound and local polytope. In: Proceedings of the 15th international conference on artificial intelligence and statistics, pp 868–876
59. Qi Y, Abdel-Gawad A, Minka T (2010) Sparse-posterior Gaussian processes for general likelihoods. In: Proceedings of the 26th conference on uncertainty in artificial intelligence, pp 450–457
60. Ravanbakhsh S, Yu C, Greiner R (2012) A generalized loop correction method for approximate inference in graphical models. In: Proceedings of the 29th international conference on machine learning, pp 1–8
61. Riihimäki J, Jylänki P, Vehtari A (2013) Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. J Mach Learn Res 14:75–109
62. Roth D, Yih W (2004) A linear programming formulation for global inference in natural language tasks. In: Proceedings of the conference on computational natural language learning, pp 1–8
63. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximation. J R Stat Soc Ser B (Stat Methodol) 71:319–392
64. Ruozzi N (2012) The Bethe partition function of log-supermodular graphical models. Adv Neural Inf Process Syst 25:117–125
65. Sato I, Nakagawa H (2012) Rethinking collapsed varitional Bayes inference for LDA. In: Proceedings of the 29th international conference on machine learning, pp 999–1006
66. Saul L, Jordan M (1996) Exploiting tractable substrutures in intractable networks. Adv Neural Inf Process Syst 8:486–492
67. Seeger M, Bouchard G (2012) Fast variational Bayesian inference for non-conjugate matrix factorization models. In: Proceedings of the 15th international conference on artificial intelligence and statistics, pp 1012–1018
68. Sun S (2013) Infinite mixtures of multivariate Gaussian processes. In: Proceedings of the international conference on machine learning and cybernetics, pp 1–6
69. Sun S (2013) Tangent space intrinsic manifold regularization for data representation. In: Proceedings of the 1st IEEE China summit and international conference on signal and information processing, pp 1–5
70. Sun S, Xu X (2011) Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow forecasting. IEEE Trans Intell Transp Syst 12:466–475
71. Taskar B, Chatalbashev V, Koller D (2004) Learning associate markov networks. In: Proceedings of the 21st international conference on machine learning, pp 102–109

72. Teh Y, Newman D, Welling M (2007) A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Adv Neural Inf Process Syst 19:1353–1360
73. Titsias M (2009) Variational learning of inducing variables in sparse Gaussian processes. In: Proceedings of the 12th international conference on artificial intelligence and statistics, pp 567–574
74. Titsias M, Lawrence N (2010) Bayesian Gaussian process latent variable model. In: Proceedings of the 13th international conference on artificial intelligence and statistics, pp 844–851
75. Tsallis C (1988) Possible generaliation of Boltzmann–Gibbs statistics. J Stat Phys 52:479–487
76. Tzikas D, Likas A, Galatsanos N (2008) The variational approximation for Bayesian inference. IEEE Signal Process Mag25:131–146
77. Wainwright M, Jordan M (2008) Graphical models, exponential families, and variational inference. Found Trends Machine Learn1:1–305
78. Wang B, Titterington D (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Anal 1:625–650
79. Wang C, Blei D (2012) Truncation-free stochastic variational inference for Bayesian nonparametric models. Adv Neural Inf Process Syst 25:422–430
80. Wang C, Blei D (2013) Variational inference in nonconjugate models. J Mach Learn Res 14 (in press)
81. Wang C, Paisley J, Blei D (2011) Online variational inference for the hierarchical Dirichlet process. In: Proceedings of the 14th international conference on artificial intelligence and statistics, pp 752–760
82. Watanabe K (2012) An alternative view of variational Bayes and asymptotic approximations of free energy. Mach Learn 86:273–293
83. Watanabe Y, Fukumizu K (2009) Graph zeta function in the Bethe free energy and loopy belief propagation. Adv Neural Inf Process Syst 22:2017–2025
84. Wilson A, Knowles D, Ghahramani Z (2012) Gaussian process regression networks. In: Proceedings of the 29th international conference on machine learning, pp 1–8
85. Winn J, Bishop C (2005) Variational message passing. J Mach Learn Res 6:661–694
86. Yedidia J, Freeman W, Weiss Y (2005) Constructing free energy approximations and generalized belief propagation algorithms. IEEE Trans Inf Theory 51:2282–2312
87. Yedidia J, Freeman W, Weiss Y (2001) Generalized belief propagation. Adv Neural Inf Process Syst 13:689–695
88. Yedidia J, Freeman W, Weiss Y (2003) Understanding belief propagation and its generalizations. Explor Artif Intell New Millenn 8:239–269
89. Yuan C, Neubauer C (2009) Variational mixture of Gaussian process experts. Adv Neural Inf Process Syst 21:1897–1904
90. Zhang Y, Schneider J (2012) A composite likelihood view for multi-label classification. In: Proceedings of the 15th international conference on artificial intelligence and statistics, pp 1407–1415
91. Zhu J, Sun S (2013) Single-task and multitask sparse Gaussian processes. In: Proceedings of the international conference on machine learning and cybernetics, pp 1–6