Second-Order Latent Space Variational Bayes for Approximate Bayesian Inference

Jaemo Sung, Zoubin Ghahramani, *Member, IEEE*, and Sung-Yang Bang, *Member, IEEE*

***Abstract*—In this letter, we consider a variational approximate Bayesian inference framework, *latent-space variational Bayes (LSVB)*, in the general context of *conjugate-exponential* family models with latent variables. In the LSVB approach, we integrate out model parameters in an exact way and then perform the variational inference over only the latent variables. It can be shown that LSVB can achieve better estimates of the model evidence as well as the distribution over the latent variables than the popular variational Bayesian expectation-maximization (VBEM). However, the distribution over the latent variables in LSVB has to be approximated in practice. As an approximate implementation of LSVB, we propose a *second-order LSVB (SoLSVB)* method. In particular, VBEM can be derived as a special case of a first-order approximation in LSVB (Sung *et al.* [1]). SoLSVB can capture higher order statistics neglected in VBEM and can therefore achieve a better approximation. Examples of Gaussian mixture models are used to illustrate the comparison between our method and VBEM, demonstrating the improvement.**

***Index Terms*—Bayesian inference, conjugate-exponential family, latent variable, mixture of Gaussians, model selection, variational method.**

I. Introduction

In the Bayesian approach [2], we give a prior $P(\boldsymbol{\theta}|\mathcal{M})$ over model parameters given a model $\mathcal{M}$. From this, we can obtain a posterior over latent variables X and model parameters $\boldsymbol{\theta}$ given data set Y

$$P(X,\boldsymbol{\theta}|Y,\mathcal{M}) = \frac{P(X,Y|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(Y|\mathcal{M})} \tag{1}$$

$$P(Y|\mathcal{M}) = \int dX d\boldsymbol{\theta} P(X,Y|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M}). \tag{2}$$

The posterior distribution $P(X,\boldsymbol{\theta}|Y,\mathcal{M})$ is useful for cluster analysis, dimensionality reduction, classification, and prediction tasks. In particular, the probability $P(Y|\mathcal{M})$, called *marginal likelihood* or *model evidence*, is an important quantity for model comparison [3] since it penalizes overcomplex models by automatically encoding Occam's Razor [4].

Unfortunately, true Bayesian inferences are generally intractable due to difficult integrals associated with the model evidence in (2). Therefore, it has to be approximated in practice. Based on factorization between latent variables and model parameters, variational Bayesian expectation-maximization (VBEM) [3], [5], a standard variational approximate Bayesian inference method, alternatively maximizes a lower bound $\mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}}$ of the log model evidence

$$\log P(Y) \geq \int dX d\boldsymbol{\theta} Q_X(X) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(Y,X|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q_X(X)Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \equiv \mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}} \tag{3}$$

with respect to approximating distributions Q_X in the VBE-step and $Q_{\boldsymbol{\theta}}$ in the VBM-step. At the maximum, the lower bound gives an approximate log model evidence and the approximating distribution provides an approximate posterior. The quality of the approximations are evaluated by the tightness of the lower bound. Compared with Markov chain Monte Carlo (MCMC) methods, VBEM can require much less computation and come with an easy to evaluate convergence criterion. However, VBEM can result in a significantly loose lower bound and often fail to find a correct model from a given data set since it ignores nontrivial correlations between latent variables and model parameters due to its inherent independence assumption. To simplify notation, we have here and will henceforth assume a given particular model, $\mathcal{M}$, even when this is not explicitly stated in the notation as in (3).

In this letter, we consider a more general variational Bayesian approximate inference method, which we named *latent-space variational Bayes (LSVB)*. In this approach, we first integrate out the model parameters in an exact way, leaving only the latent variables. Assuming weak dependencies among the latent variables over samples, we next attempt to maximize the lower bound $\mathcal{F}_{Q_X}$ in the form of

$$\log P(Y) \geq \int dX Q_X(X) \log \frac{P(Y,X)}{Q_X(X)} \equiv \mathcal{F}_{Q_X} \tag{4}$$

with respect to a factorized approximating distribution Q_X over samples, where $P(Y,X) \equiv \int d\boldsymbol{\theta} P(Y,X|\boldsymbol{\theta})P(\boldsymbol{\theta})$ denotes complete-data marginal likelihood. Fundamentally, LSVB can give a better solution than VBEM since its lower bound is always tighter than the lower bound of VBEM, that is, $\mathcal{F}_{Q_X} \geq \max_{Q_{\boldsymbol{\theta}}} \mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}}$. Next, we will focus LSVB on a general class of latent variable models called *conjugate-exponential* family and introduce a *second-order LSVB (SoLSVB)* method as a tractable implementation of LSVB.

II. Conjugate-Exponential Family

Consider a data set $Y = \{\boldsymbol{y}_i\}_{i=1}^n$ and a latent variable set $X = \{\boldsymbol{x}_i\}_{i=1}^n$, where the index i runs over samples. We assume both $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ to be multidimensional. Each of them is independently drawn from an *exponential family* [6] distribution $P(\boldsymbol{y}_i,\boldsymbol{x}_i|\boldsymbol{\theta})$ parameterized by the model parameters $\boldsymbol{\theta}$

$$P(\boldsymbol{y}_i,\boldsymbol{x}_i|\boldsymbol{\theta}) = f(\boldsymbol{y}_i,\boldsymbol{x}_i)g(\boldsymbol{\theta})\exp\left\{\boldsymbol{\Phi}(\boldsymbol{\theta})^{\mathsf{T}}\mathbf{u}_i(\boldsymbol{y}_i,\boldsymbol{x}_i)\right\} \tag{5}$$

Manuscript received January 27, 2008; revised April 14, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Champagne.

J. Sung and S.-Y. Bang are with the Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang, Korea (e-mail: emtidi; sybang@postech.ac.kr).

Z. Ghahramani is with the Information Engineering Department of Engineering, University of Cambridge, Cambridge, U.K. (e-mail: zoubin@eng.cam.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2008.2001557

where $\mathbf{u}_i(\boldsymbol{y}_i, \boldsymbol{x}_i)$ is a function of complete-data and $\boldsymbol{\phi}$ is called a natural parameter. The function g is a constant with respect to $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ ensuring that the distribution normalizes to one. A conjugate prior over the model parameters to the complete-data likelihood, $P(X, Y|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{x}_i, \boldsymbol{y}_i|\boldsymbol{\theta})$, also has the same form of exponential family

$$P(\boldsymbol{\theta}|\eta^\circ, \boldsymbol{\nu}^\circ) = h(\eta^\circ, \boldsymbol{\nu}^\circ)^{-1} g(\boldsymbol{\theta})^{\eta^\circ} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}^\circ\right\} \tag{6}$$

where η° and $\boldsymbol{\nu}^\circ$ denote prior hyper-parameters and $h(\eta^\circ, \boldsymbol{\nu}^\circ) \equiv \int d\boldsymbol{\theta} g(\boldsymbol{\theta})^{\eta^\circ} \exp\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}^\circ\}$ denotes the normalizing function. In particular, $\log h$ is a convex function of hyper-parameters since its Hessian matrix is always positive semi-definite as given by a covariance matrix of the natural parameters [6].

A class of models represented by the exponential family distribution in (5) with the conjugate prior in (6) is called the *conjugate-exponential* family [3] which has the posterior over the model parameters in the same form as the prior, that is, $P(\boldsymbol{\theta}|Y, X, \eta^\circ, \boldsymbol{\nu}^\circ) = P(\boldsymbol{\theta}|\eta, \boldsymbol{\nu})$ with posterior hyper-parameters $\eta \equiv n + \eta^\circ$ and $\boldsymbol{\nu} \equiv \sum_{i=1}^{n} \mathbf{u}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) + \boldsymbol{\nu}^\circ$. The conjugate-exponential family is broad and includes many interesting latent variable models such as mixtures of Gaussians, mixtures of factor analyzers, state-space models, hidden Markov models, linear dynamical systems, and some kinds of graphical models.

III. Latent-Space Variational Bayes

The conjugate-exponential family gives the complete-data marginal likelihood which comprises analytically known functions in the form of

$$P(Y, X|\eta^\circ, \boldsymbol{\nu}^\circ) = \frac{h(\eta, \boldsymbol{\nu})}{h(\eta^\circ, \boldsymbol{\nu}^\circ)} \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{x}_i). \tag{7}$$

This means that no optimization is needed to compute this complete-data marginal likelihood. From this, the lower bound $\mathcal{F}_{Q_X}$ in (4) with respect to a factorized approximating distribution $Q_X = \prod_{i=1}^{n} Q_{\boldsymbol{x}_i}$ over samples can be formulated by

$$\mathcal{F}_{Q_X} = R_{Q_X} + \left\langle \log h(\eta, \boldsymbol{\nu}) \right\rangle_{Q_X} \tag{8}$$

where $R_{Q_X} \equiv \sum_{i=1}^{n} \left\langle \log f(\boldsymbol{y}_i, \boldsymbol{x}_i) \right\rangle_{Q_{\boldsymbol{x}_i}} + \sum_{i=1}^{n} \mathcal{H}(Q_{\boldsymbol{x}_i}) - \log h(\eta^\circ, \boldsymbol{\nu}^\circ)$. We use the notation $\langle\cdot\rangle_Q$ for the expectation under a distribution Q and $\mathcal{H}$ for the entropy defined by $\mathcal{H}(Q) \equiv -\int dt Q(t) \log Q(t)$.

Since the lower bound $\mathcal{F}_{Q_X}$ is a concave functional over Q_X, if we set the functional derivative of $\mathcal{F}_{Q_X}$ with respect to $Q_{\boldsymbol{x}_i}$ to zero, we can find the optimal $Q_{\boldsymbol{x}_i}$ at the maximum of $\mathcal{F}_{Q_X}$ in the form of

$$Q_{\boldsymbol{x}_i}(\boldsymbol{x}_i) \propto f(\boldsymbol{y}_i, \boldsymbol{x}_i) \exp\left\{ \left\langle \log h(\eta, \boldsymbol{\nu}) \right\rangle_{Q_{X_{\neg i}}} \right\} \tag{9}$$

where $Q_{X_{\neg i}} = \prod_{j=1, \neq i}^{n} Q_{\boldsymbol{x}_j}$. The notation $\neg i$ denotes the exclusion of the ith sample. Generally, an analytical solution of Q_X does not exist due to couplings among $Q_{\boldsymbol{x}_i}$. However, we can locally maximize $\mathcal{F}_{Q_X}$ by iteratively updating $Q_{\boldsymbol{x}_i}$ at one time by fixing the others in somewhat round-robin (or random) order, starting from an initial guess. We call this iterative maximization procedure *LSVB* algorithm, which never decreases the lower bound $\mathcal{F}_{Q_X}$ and guarantees finding a local maximum of $\mathcal{F}_{Q_X}$.

It can be shown that LSVB is a more general and theoretically better approximate inference approach than VBEM. Theorem 1 shows the relationship between LSVB and VBEM.

Theorem 1: For the conjugate-exponential family, the lower bound of LSVB is tighter than the lower bound of VBEM: for all Q_X

$$\mathcal{F}_{Q_X} \geq \mathcal{F}_{Q_X}^{\mathrm{1st}} \equiv R_{Q_X} + \log h(\eta, \langle\boldsymbol{\nu}\rangle_{Q_X}) = \max_{Q_{\boldsymbol{\theta}}} \mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}} \tag{10}$$

where the equality is satisfied when $\log h$ is a linear function of $\boldsymbol{\nu}$. (Proof is given in Appendix A)

Since the lower bound $\mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}}$ of VBEM becomes tight after the VBM step, it essentially reduces to the same form of the first-order (linear) lower bound $\mathcal{F}_{Q_X}^{\mathrm{1st}}$ of $\mathcal{F}_{Q_X}$. In other words, VBEM for the conjugate-exponential family actually maximizes the first-order lower bound of LSVB and can be therefore viewed as a special case of first-order approximate LSVB [1]. We note that the first-order approximation of a nonlinear function is normally unreliable because some of the important high-order information is ignored. This means that VBEM can give a poor approximation of LSVB. An advantage of our LSVB approach is that it provides a theoretical framework to incorporate higher order information ignored by VBEM. In Section IV, we show an example by the second-order approximation method.

IV. Second-Order LSVB

The exact LSVB algorithm is generally hard to be done in practice due to difficult expectations of the nonlinear function $\log h$. A standard and practical way to approximate the expectation of a nonlinear function is to approximate the nonlinear function by a simpler function. We consider here the second-order (Gaussian) approximation, $\log \tilde{h}$, of $\log h$ around $\tilde{\boldsymbol{\nu}}$

$$\log \tilde{h}(\eta, \boldsymbol{\nu}; \tilde{\boldsymbol{\nu}}) = \log h(\eta, \tilde{\boldsymbol{\nu}}) + (\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}})^\top \nabla(\tilde{\boldsymbol{\nu}}) + \frac{1}{2} \mathrm{tr}[\nabla^2(\tilde{\boldsymbol{\nu}})(\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}})(\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}})^\top]$$

where $\nabla(\tilde{\boldsymbol{\nu}})$ and $\nabla^2(\tilde{\boldsymbol{\nu}})$ are gradient vector and Hessian matrix of $\log h$ evaluated at $\tilde{\boldsymbol{\nu}}$, respectively. Generally, the expectation of nonlinear function $\log h(\eta, \boldsymbol{\nu})$ under a distribution Q_X can be approximated by substituting $\log h(\eta, \boldsymbol{\nu})$ by $\log \tilde{h}(\eta, \boldsymbol{\nu}; \tilde{\boldsymbol{\nu}})$ around $\tilde{\boldsymbol{\nu}} = \langle\boldsymbol{\nu}\rangle_{Q_X}$

$$\langle \log h(\eta, \boldsymbol{\nu}) \rangle_{Q_X} \approx \log h(\eta, \langle\boldsymbol{\nu}\rangle_{Q_X}) + \frac{1}{2} \mathrm{tr}[\mathbf{C}_{\boldsymbol{\phi}|\langle\boldsymbol{\nu}\rangle_{Q_X}} \mathbf{C}_{\boldsymbol{\nu}}] \tag{11}$$

where $\mathbf{C}_{\boldsymbol{\phi}|\langle\boldsymbol{\nu}\rangle_{Q_X}} \equiv \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \boldsymbol{\phi}(\boldsymbol{\theta})^\top \rangle - \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle \langle \boldsymbol{\phi}(\boldsymbol{\theta})^\top \rangle$ under $P(\boldsymbol{\theta}|\eta, \langle\boldsymbol{\nu}\rangle_{Q_X})$ and $\mathbf{C}_{\boldsymbol{\nu}} \equiv \langle \boldsymbol{\nu}\boldsymbol{\nu}^\top \rangle - \langle\boldsymbol{\nu}\rangle\langle\boldsymbol{\nu}^\top\rangle$ under Q_X denote covariance matrices of natural parameters and posterior hyper-parameters, respectively. The notation $\mathrm{tr}[\cdot]$ denotes the standard matrix trace operation. For the exponential family distribution, the covariance matrix of natural parameters is particularly given by the Hessian of $\log h$, that is, $\mathbf{C}_{\boldsymbol{\phi}|\langle\boldsymbol{\nu}\rangle_{Q_X}} = \nabla^2(\langle\boldsymbol{\nu}\rangle_{Q_X})$. This approximation technique for the difficult expectation can be directly incorporated in $Q_{\boldsymbol{x}_i}$ and $\mathcal{F}_{Q_X}$ in LSVB.

Incorporating

$$\langle \log h(\eta, \boldsymbol{\nu}) \rangle_{Q_{X_{\neg i}}} \approx \langle \log \tilde{h}(\eta, \boldsymbol{\nu}; \langle\boldsymbol{\nu}\rangle_{Q_{X_{\neg i}}}) \rangle_{Q_{X_{\neg i}}}$$

in (9) gives the second-order approximating $Q_{\boldsymbol{x}_i}$ in the form of

$$Q_{\boldsymbol{x}_i}(\boldsymbol{x}_i) \propto f(\boldsymbol{y}_i, \boldsymbol{x}_i) h(\eta, \mathbf{u}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) + \langle\boldsymbol{\nu}_{\neg i}\rangle_{Q_{X_{\neg i}}}) \times \exp\left\{ \frac{1}{2} \mathrm{tr}\left[\mathbf{C}_{\boldsymbol{\phi}|\mathbf{u}_i(\boldsymbol{y}_i, \boldsymbol{x}_i) + \langle\boldsymbol{\nu}_{\neg i}\rangle_{Q_{X_{\neg i}}}} \mathbf{C}_{\boldsymbol{\nu}_{\neg i}} \right] \right\} \tag{12}$$

where $\boldsymbol{\nu}_{\neg i} = \boldsymbol{\nu}^{\circ} + \sum_{j=1,\neq i}^{n} \mathbf{u}_j(\boldsymbol{y}_j, \boldsymbol{x}_j)$. We call an iterative algorithm to update this second-order approximating $Q_{\boldsymbol{x}_i}$ instead of the exact one *SoLSVB*. The SoLSVB algorithm incorporates higher-order information and therefore captures some correlations neglected by the VBEM algorithm.

The SoLSVB algorithm gives an estimate of the optimal Q_X of LSVB. From this, we can also estimate the optimal $\mathcal{F}_{Q_X}$ of LSVB. Incorporating $\langle \log h(\eta, \boldsymbol{\nu}) \rangle_{Q_X} \approx \langle \log \tilde{h}(\eta, \boldsymbol{\nu}; \langle \boldsymbol{\nu} \rangle_{Q_X}) \rangle_{Q_X}$ with Q_X estimated by the SoLSVB algorithm provides the second-order approximate $\mathcal{F}_{Q_X}^{\text{2st}}$ of $\mathcal{F}_{Q_X}$

$$\mathcal{F}_{Q_X}^{\text{2st}} \equiv \mathcal{F}_{Q_X}^{\text{1st}} + \frac{1}{2} \operatorname{tr}[\mathbf{C}_{\boldsymbol{\phi}|\langle \boldsymbol{\nu} \rangle_{Q_X}} \mathbf{C}_{\boldsymbol{\nu}}]. \tag{13}$$

The second-order $\mathcal{F}_{Q_X}^{\text{2st}}$ in SoLSVB compensates for rough $\mathcal{F}_{Q_X}^{\text{1st}}$ in VBEM by taking into account uncertainties about the natural parameters and the posterior hyper-parameters. Therefore, SoLSVB can provide a much more accurate approximation of $\mathcal{F}_{Q_X}$ than VBEM.

In contrast to VBEM, both LSVB and SoLSVB no longer explicitly give an estimate of the posterior over the model parameters as they integrate out the model parameters. However, from the estimated distribution over the latent variables, one can later estimate a posterior over the model parameters. A simple way used here is to take a single VBM step with Q_X estimated by the SoLSVB algorithm, which reduces to

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})^{\eta} \exp\left\{ \boldsymbol{\phi}(\boldsymbol{\theta})^{\mathsf{T}} \langle \boldsymbol{\nu} \rangle_{Q_X} \right\}. \tag{14}$$

V. Example of Mixture of Gaussians

In order to demonstrate our method, we consider the mixture of Gaussians (MoG), a standard latent variable model for cluster analysis and density estimation of an unknown distribution. The number of components, K, specifies the model $\mathcal{M}$ for MoGs. Suppose a d-dimensional continuous data $\boldsymbol{y}_i$ and a discrete latent variable $x_i \in \{1, 2, \ldots, K\}$. For MoG, the joint distribution $P(\boldsymbol{y}_i, x_i | \boldsymbol{\theta})$ given the model parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k\}_{k=1}^{K}$ can be written in the form of

$$P(\boldsymbol{y}_i, x_i | \boldsymbol{\theta}) = \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)]^{\delta_k(x_i)} \tag{15}$$

where the indicator function $\delta_k(x_i)$ equals one if $x_i = k$ and zero otherwise. The mixing coefficient π_k satisfies $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The standard normal density $\mathcal{N}(\cdot | \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$ with mean vector $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\lambda}_k$ represents the kth mixture component. A standard conjugate prior over the model parameters of MoG consists of Dirichlet ($\mathcal{D}$) distribution on $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and Normal ($\mathcal{N}$)-Wishart ($\mathcal{W}$) distribution on $(\boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$ for all $k = 1, \ldots, K$

$$\mathcal{D}(\boldsymbol{\pi} | \alpha_1^{\circ}, \ldots, \alpha_K^{\circ}) \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{\xi}_k^{\circ}, \tau_k^{\circ} \boldsymbol{\lambda}_k) \mathcal{W}(\boldsymbol{\lambda}_k | r_k^{\circ}, \boldsymbol{B}_k^{\circ}) \tag{16}$$

where all standard distributions are given in Table I.

After converting (15) and (16) into the standard conjugate-exponential form in (5) and (6), we can find the posterior hyper-parameters $\boldsymbol{\nu} = \{\bar{\alpha}_k, \bar{\tau}_k, \bar{r}_k, \bar{\boldsymbol{\xi}}_k, \bar{\boldsymbol{B}}_k\}_{k=1}^{K}$ given by

$$\begin{aligned} \bar{\alpha}_k &= \alpha_k^{\circ} - 1 + n_k, \quad \bar{\tau}_k = \tau_k^{\circ} + n_k, \quad \bar{r}_k = 2r_k^{\circ} - d + n_k \\ \bar{\boldsymbol{\xi}}_k &= \tau_k^{\circ} \boldsymbol{\xi}_k^{\circ} + \boldsymbol{\rho}_k, \quad \bar{\boldsymbol{B}}_k = 2\boldsymbol{B}_k^{\circ} + \tau_k^{\circ} \boldsymbol{\xi}_k^{\circ} \boldsymbol{\xi}_k^{\circ\mathsf{T}} + \boldsymbol{W}_k \end{aligned} \tag{17}$$

TABLE I
Standard Distributions

$$\mathcal{D}(\pi_1, \ldots, \pi_K | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} {\pi_k}^{\alpha_k - 1}$$

$$\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\xi}, \boldsymbol{\lambda}) = (2\pi)^{-d/2} |\boldsymbol{\lambda}|^{1/2} \exp\left\{ -\tfrac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\xi})^{\mathsf{T}} \boldsymbol{\lambda} (\boldsymbol{\mu} - \boldsymbol{\xi}) \right\}$$

$$\mathcal{W}(\boldsymbol{\lambda} | r, \boldsymbol{B}) = \frac{\pi^{-d(d-1)/4} |\boldsymbol{B}|^{r}}{\prod_{l=1}^{d} \Gamma(\frac{2r+1-l}{2})} |\boldsymbol{\lambda}|^{r-(d+1)/2} \exp\left\{ -\operatorname{tr}[\boldsymbol{B}^{\mathsf{T}} \boldsymbol{\lambda}] \right\}$$

where n_k, $\boldsymbol{\rho}_k$, and $\boldsymbol{W}_k$ are sufficient statistics given by $n_k = \sum_{i=1}^{n} \delta_k(x_i)$, $\boldsymbol{\rho}_k = \sum_{i=1}^{n} \delta_k(x_i) \boldsymbol{y}_i$, and $\boldsymbol{W}_k = \sum_{i=1}^{n} \delta_k(x_i) \boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}}$. The hyper-parameter η does not play any role for MoG and can be safely dropped. In addition, the normalizing function h has the form of

$$h(\boldsymbol{\nu}) = \frac{(4\pi)^{(d(d+1)K/4)}}{\Gamma\left(\sum\limits_{k=1}^{K} \bar{\alpha}_k + K\right)} \times \prod_{k=1}^{K} \frac{2^{(d/2)\bar{r}_k} \Gamma(\bar{\alpha}_k + 1) \prod_{l=1}^{d} \Gamma(\frac{\bar{r}_k + d + 1 - l}{2})}{\bar{\tau}^{(d/2)} \left| \bar{\boldsymbol{B}}_k - \frac{1}{\bar{\tau}} \bar{\boldsymbol{\xi}}_k \bar{\boldsymbol{\xi}}_k^{\mathsf{T}} \right|^{(\bar{r}_k + d/2)}}$$

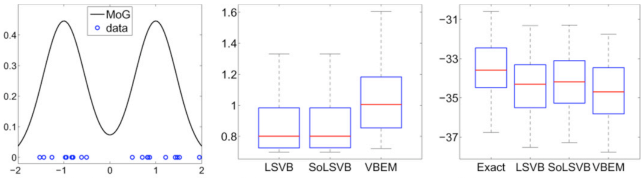
where $\Gamma(\cdot)$ denotes the standard gamma function. From the standard conjugate-exponential form for MoG above, both LSVB and SoLSVB algorithms are straightforward as given in the previous Sections III and IV. Due to the lack of space, we leave more detailed derivations to readers.

A. Numerical Results

We used default prior hyper-parameters with $\alpha_k^{\circ} = 1$, $r_k^{\circ} = 1 + 0.5d$, and $\boldsymbol{\xi}_k^{\circ} = (1/n) \sum_{i=1}^{n} \boldsymbol{y}_i$ (sample mean) for all components. In particular, $\boldsymbol{B}_k^{\circ}$ was set for $\langle \boldsymbol{\lambda}_k \rangle$ under the prior to be $(0.3\sigma_{\max})^{-2} \boldsymbol{I}_d$ and then τ_k° was set for the precision of $\boldsymbol{\mu}_k$ to be $(10\sigma_{\max})^{-2} \boldsymbol{I}_d$, where $\sigma_{\max}$ denotes the maximum standard deviation of data set among dimensions. Also, we initialized the distribution over the latent variables such as $Q_{x_i}(x_i = k) \propto \mathcal{N}\left(\boldsymbol{y}_i | \boldsymbol{c}_k, (0.3\sigma_{\max})^{-2} \boldsymbol{I}_d\right)$ with the center $\boldsymbol{c}_k$ of the kth cluster found by the standard k-means algorithm. The algorithms were considered to be converged when the successive changes in Q_X were very small such as $(1/nK) \sum_{i=1}^{n} \sum_{k=1}^{K} \left| Q_{x_i}(x_i = k) - Q_{x_i}^{old}(x_i = k) \right| < 10^{-6}$.

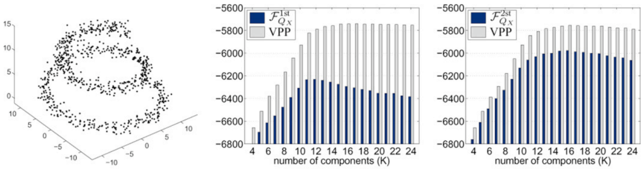
To see basic properties of the algorithms, we first used one-dimensional toy data sets of 20 data samples shown in Fig. 1(a). They were generated from the mixture of two Gaussians with $\pi_1 = \pi_2 = 0.5$, $\mu_1 = 1$, $\mu_2 = -1$, and $\lambda_1 = \lambda_2 = 5$. For these small data sets, we can perform the exact inference and the LSVB algorithm. Fig. 1 shows the results performed on the model with $K = 2$. We can see that VBEM results in the poorest distribution over the latent variables, showing the largest KL divergence to the true distribution (Fig. 1(b)). In contrast, directly approximating the distribution over the latent variables in LSVB, SoLSVB finds a good distribution over the latent variables, which is very close to LSVB (Fig. 1(b)). In addition, as a result of correcting the loose first-order lower bound in VBEM, SoLSVB gives a better estimate of the model evidence than VBEM, showing a slightly more accurate approximation to LSVB (Fig. 1(c)).

To examine differences of the algorithms in the task of model selection for density estimation, we next used a three-dimensional noisy shrinking spiral data set (see Fig. 2(a)), which has been often used to demonstrate inference algorithms in the machine learning literature [5]. For this data set, neither the exact inference nor the LSVB algorithm is allowed. In



(a) data set (b) KL divergence (c) model evidence

Fig. 1. Results based on 30 trials with different data sets of 20 data samples, randomly drawn from the mixture of two Gaussians. All inferences were performed on the model with $K = 2$. (a) Single case of random data set, annotated by sampling distribution. (b) KL divergence of the true posterior distribution over the latent variables from the approximate distribution estimated by the algorithms. (c) Estimated log model evidences: from left to right, $\log P(Y)$ for Exact, $\mathcal{F}_{Q_X}$ for LSVB, $\mathcal{F}^{2\text{st}}_{Q_X}$ for SoLSVB, and $\mathcal{F}^{1\text{st}}_{Q_X}$ for VBEM.



(a) data (b) VBEM (c) SoLSVB

Fig. 2. Example of spiral data set. (a) Data set of 800 data samples. (b) and (c): Estimated log model evidences, $\mathcal{F}^{1\text{st}}_{Q_X}$ for VBEM (b) and $\mathcal{F}^{2\text{st}}_{Q_X}$ for SoLSVB (c), compared with log variational predictive probabilities of the validation data set, denoted by VPP, based on $Q_{\boldsymbol{\theta}}$ estimated by VBEM (b) and SoLSVB (c). The results are averaged over 30 trials with different initial Q_X.

order to validate the estimated model evidence based on training data set $Y = \{\boldsymbol{y}_i\}_{i=1}^{800}$, we prepared an independent validation data set $\hat{Y} = \{\hat{\boldsymbol{y}}_i\}_{i=1}^{800}$ on Y, both of which were drawn from the same distribution. For each model, we then evaluated the log variational predictive probability[1] of $\hat{Y}$, given by VPP $= \log \int d\boldsymbol{\theta} P(\hat{Y}|\boldsymbol{\theta}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, where the variational approximate posterior $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \approx P(\boldsymbol{\theta}|Y)$ was estimated as given in (14) based on the training data set Y. VPP measures how well a learned model based on the training data set represents the underlying true distribution in terms of generalization performances. We note that VPP can be also used as a model criterion but the additional validation data set required by VPP is not always allowed to obtain in practice. In contrast, the model evidence allows us to compare models based solely on the training data set Y. For MoG, VPP has the form of the mixture of student distributions and its detailed form can be found in [7]. Fig. 2 shows the results over 30 trials with different initial Q_X. Both VPPs based on $Q_{\boldsymbol{\theta}}$ estimated by VBEM (Fig. 2(b)) and SoLSVB (Fig. 2(c)) are very similar and commonly find the best model nearby $K = 16$. Also, we can see that in all cases, VBEM underestimates the model evidence compared with SoLSVB, finding a simpler model with $K = 12$ as the best (Fig. 2(b)). Moreover, VBEM gives a significantly different shape from VPP, showing a poor generalization performance. Correcting the loose first-order lower bound of VBEM, SoLSVB gives a more reliable estimate of the model evidence and shows a similar tendency with VPP (Fig. 2(c)). It finds the best model with $K = 16$ like VPP but, in contrast to VPP, does not require an additional validation data set to do so.

VI. Conclusion

In this letter, we introduced the LSVB approach for variational approximate Bayesian inference and proposed the SoLSVB method as its approximate implementation in the general context of conjugate-exponential family. We successfully illustrated our method using examples of Gaussian mixture models, compared it with the popular VBEM method. It was shown in numerical results that SoLSVB can be a more reliable approximate inference method than VBEM since it captures higher-order statistics ignored by VBEM. SoLSVB is generally more expensive than VBEM since it requires an additional computation of Hessian, cost of which depends on models. However, it is still much more efficient than the MCMC methods. A similar idea, integrating out model parameters first, has been independently developed in [8], but they focused on a specific latent variable model called latent dirichlet allocation (LDA).[2] Our approach to the conjugate-exponential family is more general including LDA as its special case. We believe that our proposed method will be promising to other interesting latent variable models in the conjugate-exponential family.

Appendix

Proof of Theorem 1: Incorporating (5) and (6) into (3), the lower bound $\mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}}$ of VBEM for the conjugate-exponential model is reduced to $\mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}} = R_{Q_X} + \eta \langle \log g(\boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}} + \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle^{\mathsf{T}}_{Q_{\boldsymbol{\theta}}} \langle \boldsymbol{\nu} \rangle_{Q_X} + \mathcal{H}(Q_{\boldsymbol{\theta}})$. By the definition of convex function $\log h$, we get the following inequality $\langle \log h(\eta, \boldsymbol{\nu}) \rangle_{Q_X} \geq \log h(\eta, \langle \boldsymbol{\nu} \rangle_{Q_X})$. Furthermore, we can obtain a lower bound of $\log h(\eta, \langle \boldsymbol{\nu} \rangle_{Q_X})$ with respect to an arbitrary $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ by using Jensen's inequality [7]: $\log h(\eta, \langle \boldsymbol{\nu} \rangle_{Q_X}) \geq \int d\boldsymbol{\theta} Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \left(g(\boldsymbol{\theta})^{\eta} \exp \left\{ \boldsymbol{\phi}(\boldsymbol{\theta})^{\mathsf{T}} \langle \boldsymbol{\nu} \rangle_{Q_X} \right\} / Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \right)$. Plugging both inequalities above into $\mathcal{F}_{Q_X} = R_{Q_X} + \langle \log h(\eta, \boldsymbol{\nu}) \rangle_{Q_X}$, it is trivial to see the relation $\mathcal{F}_{Q_X} \geq R_{Q_X} + \log h(\eta, \langle \boldsymbol{\nu} \rangle_{Q_X}) = \max_{Q_{\boldsymbol{\theta}}} \mathcal{F}_{Q_X Q_{\boldsymbol{\theta}}}$, where the equality is satisfied when $\log h$ is a linear function of $\boldsymbol{\nu}$.

References

[1] J. Sung, Z. Ghahramani, and S.-Y. Bang, *Latent-space variational Bayes*, submitted for publication ***[Submitted where?]***.
[2] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 2000.
[3] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*. Oxford, U.K.: Oxford Univ. Press, 2003.
[4] W. H. Jefferys and J. O. Berger, "Occam's razor and Bayesian analysis," *Amer. Scientist*, vol. 80, pp. 64–72, 1992.
[5] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. Uncertainty in Artificial Intelligence*, 1999.
[6] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. Math. Statist., 1986.
[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[8] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2007, vol. 19.

[1] The variational predictive probability is a standard approximation for intractable true predictive probability, $P(\hat{Y}|Y) = \int d\boldsymbol{\theta} P(\hat{Y}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|Y)$.

[2] Our work was known to those authors in the form of a technical report.