Integrated Non-Factorized Variational Inference

Shaobo Han, Xuejun Liao, Lawrence Carin

Duke University, Durham, NC, USA

PROBLEM OF INTEREST

Problem: Consider a general Bayesian hierarchical model with observations: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$, latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$, and hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, the exact joint posterior $p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})/p(\boldsymbol{y})$ can be difficult to evaluate, since usually the normalization $p(\boldsymbol{y}) = \iint p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}d\boldsymbol{\theta}$ is intractable.

Solution:

- Markov chain Monte Carlo (MCMC) Methods
- Deterministic alternatives: Laplace approximation (LA), variational methods, expectation propagation (EP), and integrated nested Laplace approximation (INLA)

INLA IN A NUTSHELL

Main Idea: Discretizing the low-dimensional space $\boldsymbol{\theta}$ using a grid $\mathcal{G}$

(a) **Local Laplace approximation:**

$$q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x};\boldsymbol{x}^*(\boldsymbol{\theta}_k),\boldsymbol{H}(\boldsymbol{x}^*(\boldsymbol{\theta}_k))^{-1}), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G},$$

where $\boldsymbol{x}^*(\boldsymbol{\theta}_k) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ is the mode, and $\boldsymbol{H}(\boldsymbol{x}^*(\boldsymbol{\theta}_k))$ is the Hessian

(b) **Laplace's method of integration:**

$$q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) = \left.\frac{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta}_k)} \tag{1}$$

(c) **Numerical integration:** $q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)q(\boldsymbol{\theta}_k|\boldsymbol{y})\boldsymbol{\Delta}_k$

Benefits:

- Provides full posterior inference (i.e. joint density $q(\boldsymbol{x},\boldsymbol{\theta}|y)$)
- Computationally efficient (MCMC: hours or days, INLA: seconds or minutes)

Limitations:

- Applies only to latent Gaussian models (LGMs)
- No quantization for the accuracy of approximation $q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})q_{LA}(\boldsymbol{\theta}|\boldsymbol{y})$
- The dimension of parameter $\boldsymbol{\theta}$ should be no more than 5 or 6

NON-FACTORIZED VARIATIONAL INFERENCE

Evidence Lower Bound (ELBO): Applying Jensen's inequality to the log-marginal data likelihood, one obtains

$$\ln p(\boldsymbol{y}) = \ln \iint q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\frac{p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})}{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}d\boldsymbol{x}d\boldsymbol{\theta} \geq \iint q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})}{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}d\boldsymbol{x}d\boldsymbol{\theta} := \mathcal{L},$$

which holds for any proposed approximating distributions $q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$.

- The Jensen's gap is the Kullback-Leibler (KL) divergence
- The variational distribution $q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ is commonly restricted to tractable families

Mean-Field Variational Bayes (VB):

- Assumes **factorized** form: $q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = q(\boldsymbol{x})q(\boldsymbol{\theta})$, ignores posterior dependencies
- Provides an analytical lower bound to the model evidence
- Easily derived and in close form for conjugate models
- Challenging for non-conjugate models (especially when posterior is multimodal)

Proposed Method (INF-VB): We consider a **non-factorized** approximation $q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{y})$ and propose a hybrid continuous-discrete variational approximation $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})q_d(\boldsymbol{\theta}|\boldsymbol{y})$, where $q_d(\boldsymbol{\theta}|\boldsymbol{y})$ is a finite mixture of Dirac-delta distributions, $q_d(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_k \omega_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ with $\omega_k = q_d(\boldsymbol{\theta}_k|\boldsymbol{y})$ and $\sum_k \omega_k = 1$. The optimal hybrid variational distribution is

$$\begin{aligned} q^\star(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) &= \arg\min_{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})} \mathrm{KL}\left(q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\right) \\ &= \arg\min_{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}),\ q(\boldsymbol{\theta}|\boldsymbol{y})} \int q(\boldsymbol{\theta}|\boldsymbol{y})\left[\int q(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}d\boldsymbol{x} + \ln q(\boldsymbol{\theta}|\boldsymbol{y})\right]d\boldsymbol{\theta} \\ &= \arg\min_{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}),\ q_d(\boldsymbol{\theta}|\boldsymbol{y})} \sum_k q_d(\boldsymbol{\theta}_k|\boldsymbol{y})\left[\int q(\boldsymbol{x}|\boldsymbol{\theta}_k,\boldsymbol{y}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}d\boldsymbol{x} + \ln q_d(\boldsymbol{\theta}_k|\boldsymbol{y})\right]. \end{aligned}$$

This leads to the approximate marginal posterior $q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)q_d(\boldsymbol{\theta}_k|\boldsymbol{y})$.

Remarks:

- Preserves **full posterior dependencies** and could be therefore more accurate than the mean-field variational Bayes (VB) method
- The continuous approximation is flexible enough for various kinds of latent fields
- The discretized approximation of $p(\boldsymbol{\theta}|\boldsymbol{y})$ overcomes the potential non-conjugacy and multimodal posterior problems in $\boldsymbol{\theta}$

Variational Optimization Algorithm

- **Step 1 (Model Fitting):** For each $\boldsymbol{\theta}_k \in \mathcal{G}$, independently solving,

$$\begin{aligned} q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) &= \arg\min_{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)} \sum_k q(\boldsymbol{\theta}_k|\boldsymbol{y})\left[\int q(\boldsymbol{x}|\boldsymbol{\theta}_k,\boldsymbol{y}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}d\boldsymbol{x} + \ln q(\boldsymbol{\theta}_k|\boldsymbol{y})\right] \\ &= \arg\min_{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)) \end{aligned}$$

- **Step 2 (Model Comparison):** Given $\{q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) : \boldsymbol{\theta}_k \in \mathcal{G}\}$, one solves

$$\{q^\star(\boldsymbol{\theta}_k|\boldsymbol{y})\} = \arg\min_{\{q(\boldsymbol{\theta}_k|\boldsymbol{y})\}} \sum_k \underbrace{q_d(\boldsymbol{\theta}_k|\boldsymbol{y})\left[\int q^\star(\boldsymbol{x}|\boldsymbol{\theta}_k,\boldsymbol{y}) \ln \frac{q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}d\boldsymbol{x} + \ln q_d(\boldsymbol{\theta}_k|\boldsymbol{y})\right]}_{l(q_d(\boldsymbol{\theta}_k|\boldsymbol{y}))=l(\omega_k)}$$

Setting $\partial l(\omega_k)/\partial\omega_k = 0$ (also $\partial^2 l(\omega_k)/\partial\omega_k^2 > 0$), which is solved to give

$$q_d^\star(\boldsymbol{\theta}_k|\boldsymbol{y}) \propto \exp\left(\int q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \ln \frac{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}{q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}d\boldsymbol{x}\right) \tag{2}$$

Remarks:

- The optimal variational distribution $q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ is the exact posterior $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$. In case it is not available, we may further constrain $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ to a parametric form (e.g. multivariate Gaussian, or skew-normal densities)
- The computations on the grid are independent. INF-VB is **highly parallelizable**, with dominant computational load $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ distributed on each grid point

LINKS BETWEEN INF-VB AND INLA

Variational Unification of INLA: We compare the three computation steps of INLA with INF-VB in below:

(a) **Variational Gaussian (VG) approximation:** INF-VB with the Gaussian parametric constraint on $q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ provides a global Gaussian approximation

(b) **Hyperparameter learning:** The optimality justifications of $q_d^\star(\boldsymbol{\theta}|\boldsymbol{y})$ by INF-VB (2) explain the often observed empirical successes of hyperparameter selection based on the ELBO of $\ln p(\boldsymbol{y}|\boldsymbol{\theta})$, when only $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ with fixed θ is of interest

$$\ln p(\boldsymbol{\theta}|\boldsymbol{y}) = \ln\left[\int \frac{p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{x}\right] \geq \int \ln\left[\frac{p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\right]d\boldsymbol{x} = \ln q_d^\star(\boldsymbol{\theta}|\boldsymbol{y})$$

(c) **Dirac-delta mixture interpretation:** Let $\omega_k = q(\boldsymbol{\theta}_k|\boldsymbol{y})\boldsymbol{\Delta}_k$, we obtain

$$q(\boldsymbol{x}|\boldsymbol{y}) = \int q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})q_d(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} = \sum_k q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)q_d(\boldsymbol{\theta}_k|\boldsymbol{y}) = \sum_k q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)q(\boldsymbol{\theta}_k|\boldsymbol{y})\boldsymbol{\Delta}_k$$

Remarks:

- Applicable in more general scenarios than typically assumed by INLA, such as Bayesian Lasso, which is characterized by the **strong non-Guassianity** of $p(\boldsymbol{x}|\boldsymbol{\theta})$ and **non-differentiability** of the ℓ_1 norm
- INF-VB provides **optimal** variational $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ and $q_d(\boldsymbol{\theta}|\boldsymbol{y})$, within the proposed hybrid distribution family
- The Dirac-delta mixture interpretation of numerical integration enables us to quantitize the accuracy of INLA approximation $q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})q_{LA}(\boldsymbol{\theta}|\boldsymbol{y})$ using the KL divergence to $p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ under the variational framework

APPLICATION TO BAYESIAN LASSO

Model: Consider the Bayesian Lasso regression model, $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{\Phi} \in \mathbb{R}^{n\times p}$ is the design matrix containing predictors, $\boldsymbol{y} \in \mathbb{R}^n$ are responses, and $\boldsymbol{e} \in \mathbb{R}^n$ contain independent zero-mean Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{e};\boldsymbol{0},\sigma^2\boldsymbol{I}_n)$. We assume

$$x_j|\sigma^2,\lambda^2, \sim \frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}\|x_j\|_1\right), \quad \sigma^2 \sim \mathrm{InvGa}(\sigma^2;a,b), \quad \lambda^2 \sim \mathrm{Ga}(\lambda^2;r,s).$$

Inference:

(a) Concave variational Gaussian (VG) approximation: $q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{D})$, which minimizes $g(\boldsymbol{\mu},\boldsymbol{D}) = \mathrm{KL}(q(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{D})||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}))$

(b) Hyperparameter posterior: $q^\star(\boldsymbol{\theta}|y)$ can be evaluated analytically using either (1) or (2); both will yield a finite mixture of Gaussian for the marginal posterior $q(\boldsymbol{x}|\boldsymbol{y})$

(c) Negative ELBO: KL divergence up to a constant, can also be evaluated analytically

ACKNOWLEDGEMENTS

The work reported here was supported in part by grants from ARO, DARPA, DOE, NGA and ONR.

THEORETICAL ANALYSIS

Connection between Lasso and Bayesian Lasso: Let $(\boldsymbol{\mu}^*,\boldsymbol{D}^*)$ be the minimizer of the KL divergence $g(\boldsymbol{\mu},\boldsymbol{D})$, and $g_1(\boldsymbol{\mu}|\boldsymbol{D})$ collect the terms of $g(\boldsymbol{\mu},\boldsymbol{D})$ that are related to $\boldsymbol{\mu}$. Then the Bayesian posterior mean obtained via VG, i.e.,

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} g_1(\boldsymbol{\mu}|\boldsymbol{D}^*) = \arg\min_{\boldsymbol{\mu}} \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}\left(\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2\lambda\sigma\|\boldsymbol{x}\|_1\right),$$

is a counterpart of the deterministic Lasso, $\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}}\left(\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\mu}\|_2^2 + 2\lambda\sigma\|\boldsymbol{\mu}\|_1\right)$.

Remarks:

- Note that the Lasso solution **cannot** be found by gradient methods due to non-differentiability. By taking the expectation, the objective function is smoothed around 0 and thus differentiable
- This connection indicates that in VG for Bayesian Lasso, the conditions of deterministic Lasso **hold on average**, with respect to the variational distribution $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, in the parameter space of $\boldsymbol{\mu}$

EXPERIMENTAL RESULTS

Synthetic Dataset:

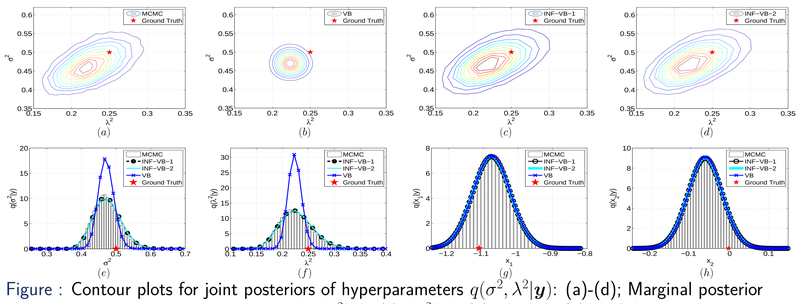

Figure : Contour plots for joint posteriors of hyperparameters $q(\sigma^2,\lambda^2|\boldsymbol{y})$: (a)-(d); Marginal posterior of hyperparameters and coefficients: (e) $q(\sigma^2|\boldsymbol{y})$, (f)$q(\lambda^2|\boldsymbol{y})$; (g) $q(x_1|\boldsymbol{y})$, (h)$q(x_2|\boldsymbol{y})$

Diabetes Dataset:

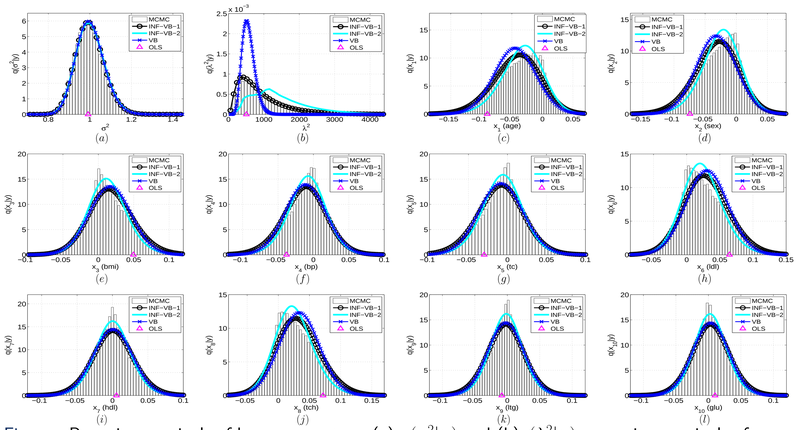

Figure : Posterior marginals of hyperparameters: (a) $q(\sigma^2|\boldsymbol{y})$ and (b)$q(\lambda^2|\boldsymbol{y})$; posterior marginals of coefficients: (c)-(l) $q(x_j|\boldsymbol{y})$ $(j = 1,\ldots,10)$

Comparison (Accuracy and Speed):

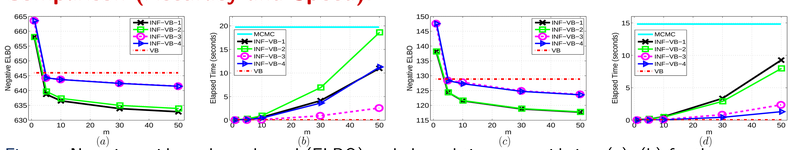

Figure : Negative evidence lower bound (ELBO) and elapsed time v.s. grid size; (a), (b) for the Diabetes dataset ($n = 442$, $p = 10$). (c), (d) for the Prostate cancer dataset ($n = 97$, $p = 8$)

Notes:

- Both MCMC and INF-VB preserve the strong posterior dependence among hyper parameters, while mean-field VB not
- While mean-field VB approximates the posterior mode well, the posterior variance can be (sometimes severely) underestimated
- The accuracy of INF-VB method with a 1×1 grid is worse than mean-field VB, which corresponds to the partial Bayesian learning of $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ with a fixed $\boldsymbol{\theta}$. As the grid size increases, the accuracies of INF-VB also increase