Integrated Non-Factorized Variational Inference

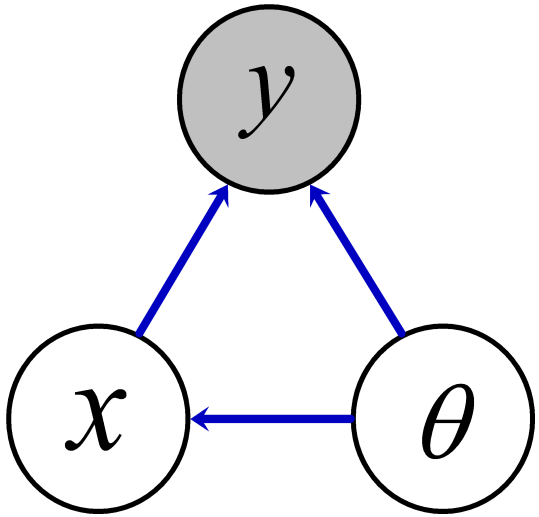
Shaobo Han, Xuejun Liao and Lawrence Carin

Duke University

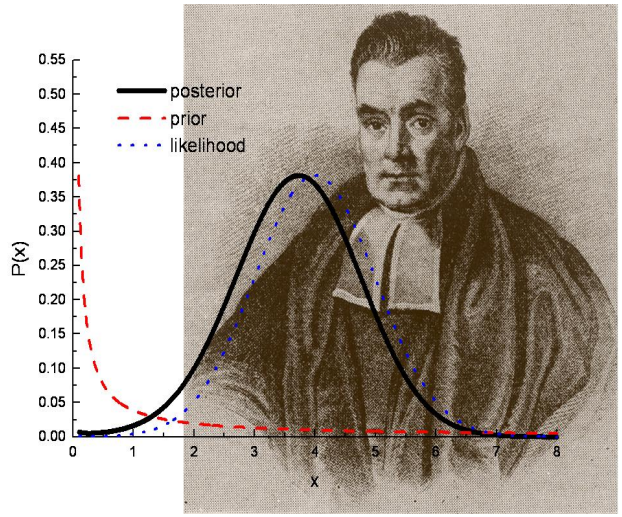
February 27, 2014

Overview

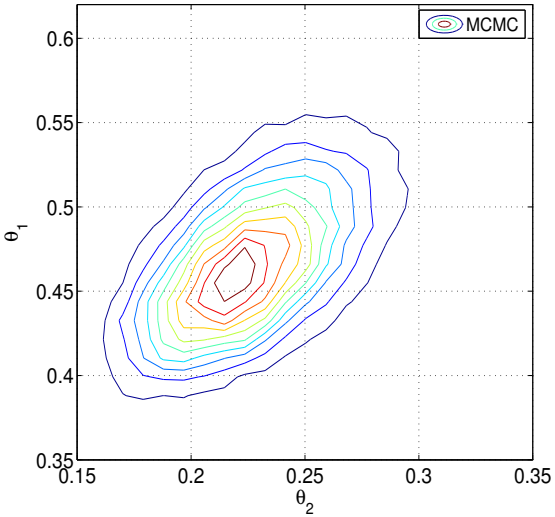
World



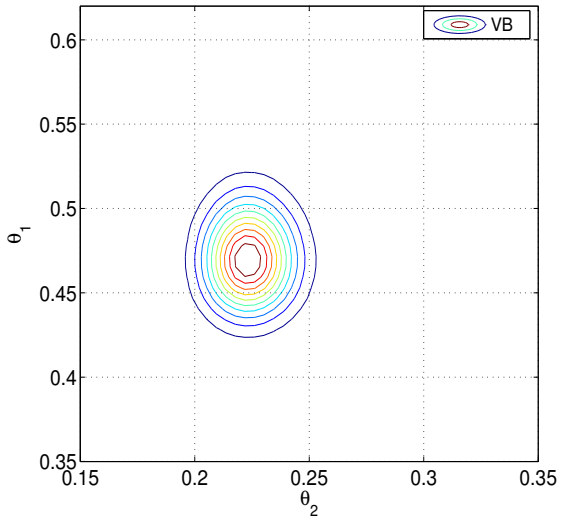
Graphical Models



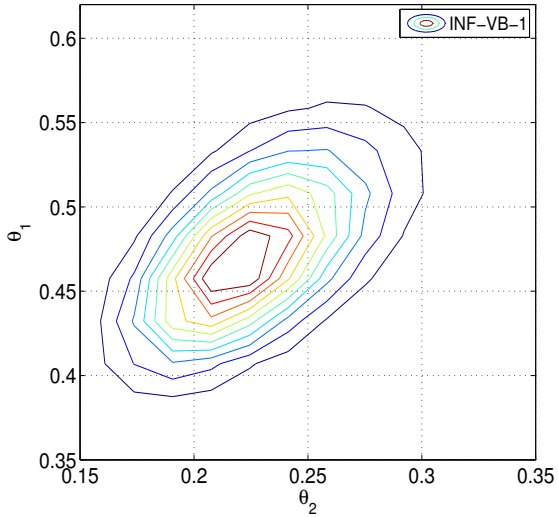
Posterior Inference



MCMC



VB



Our method

For full posterior inference, our method is

- A fast deterministic alternative to MCMC
- More accurate than mean-field variational Bayes (VB)

Outline

- Introduction
- Integrated Nested Laplace Approximation (INLA)
- **Integrated Non-Factorized Variational Bayes (INF-VB)**
- Applications
- Summary

Problem of Interest

Consider a general Bayesian hierarchical model

- Observation model: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
- Latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$
- Hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Problem of Interest

Consider a general Bayesian hierarchical model

- Observation model: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
- Latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$
- Hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Posterior inference:

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$$

Problem of Interest

Consider a general Bayesian hierarchical model

- Observation model: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
- Latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$
- Hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Posterior inference:

$$\begin{array}{ccc} p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) & \longrightarrow & p(\boldsymbol{x}|\boldsymbol{y}) \\ \downarrow & \searrow & \\ p(\boldsymbol{\theta}|\boldsymbol{y}) & & p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \end{array}$$

Problem of Interest

Consider a general Bayesian hierarchical model

- Observation model: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
- Latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$
- Hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Posterior inference:

$$\begin{array}{ccc} p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) & \longrightarrow & p(\boldsymbol{x}|\boldsymbol{y}) \\ \downarrow\uparrow & \searrow\nwarrow & \\ p(\boldsymbol{\theta}|\boldsymbol{y}) & + & p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \end{array}$$

Problem of Interest

Consider a general Bayesian hierarchical model

- Observation model: $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
- Latent variables: $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta})$
- Hyperparameters: $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Posterior inference:

$$\begin{array}{ccc} p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) & \longrightarrow & p(\boldsymbol{x}|\boldsymbol{y}) \\ \downarrow\uparrow & \searrow\nwarrow & \\ p(\boldsymbol{\theta}|\boldsymbol{y}) & + & p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) \end{array}$$

The exact joint posterior

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int\int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{x}d\boldsymbol{\theta}}$$

can be difficult to evaluate.

Approximate Posterior Inference

Sampling based methods:

- Markov chain Monte Carlo (MCMC)

Deterministic alternatives:

- Laplace approximation (LA)
- Variational inference
- Expectation propagation (EP)
- **Integrated nested Laplace approximation (INLA)**[1]

[1] Rue et al., 2009

Outline

- Introduction
- **Integrated Nested Laplace Approximation (INLA)**
- Integrated Non-Factorized Variational Bayes (INF-VB)
- Applications
- Summary

INLA in a Nutshell (1/3)

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

Demo:

$$\boldsymbol{\theta}_k \in \mathcal{G}$$

[2] Kass & Steffey, 1989

INLA in a Nutshell (1/3)

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

Demo:

$$q_G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_k) \quad \longleftarrow \quad \boldsymbol{\theta}_k \in \mathcal{G}$$

[2] Kass & Steffey, 1989

INLA in a Nutshell (1/3)

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

Demo:

$$q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \quad \longleftarrow \quad \boldsymbol{\theta}_k \in \mathcal{G}$$

1. Laplace approximation[2] :

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

where $\boldsymbol{x}^\star(\boldsymbol{\theta}_k) = \text{argmax}_{\boldsymbol{x}}\, p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$ is the posterior mode, and $\boldsymbol{H}(\boldsymbol{x}^\star(\boldsymbol{\theta}_k))$ is the Hessian matrix of the log posterior evaluated at the mode.

[2] Kass & Steffey, 1989

INLA in a Nutshell (2/3)

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

Demo:

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \longleftarrow q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \longleftarrow \boldsymbol{\theta}_k \in \mathcal{G}$$

[3] Tierney & Kadane, 1986

INLA in a Nutshell (2/3)

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

Demo:

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \longleftarrow q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \longleftarrow \boldsymbol{\theta}_k \in \mathcal{G}$$

According to the Bayes rule,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}, \quad \forall \boldsymbol{x} \tag{1}$$

[3] Tierney & Kadane, 1986

INLA in a Nutshell (2/3)

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

Demo:

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \longleftarrow q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \longleftarrow \boldsymbol{\theta}_k \in \mathcal{G}$$

According to the Bayes rule,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}, \quad \forall \boldsymbol{x} \tag{1}$$

2. Laplace's method of integration[3] :

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) = \left.\frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{p(\boldsymbol{y})q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}\right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})} \tag{2}$$

[3] Tierney & Kadane, 1986

INLA in a Nutshell (3/3)

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

Demo:

$$
\begin{array}{ccccc}
 & & q(\boldsymbol{x}|\boldsymbol{y}) & & \\
 & \nearrow & \uparrow & & \\
q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) & + & q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) & \longleftarrow & \boldsymbol{\theta}_k \in \mathcal{G}
\end{array}
$$

INLA in a Nutshell (3/3)

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

Demo:

$$\begin{array}{ccccc} & & q(\boldsymbol{x}|\boldsymbol{y}) & & \\ & \nearrow & \uparrow & & \\ q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) & + & q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) & \longleftarrow & \boldsymbol{\theta}_k \in \mathcal{G} \end{array}$$

3. Numerical integration:

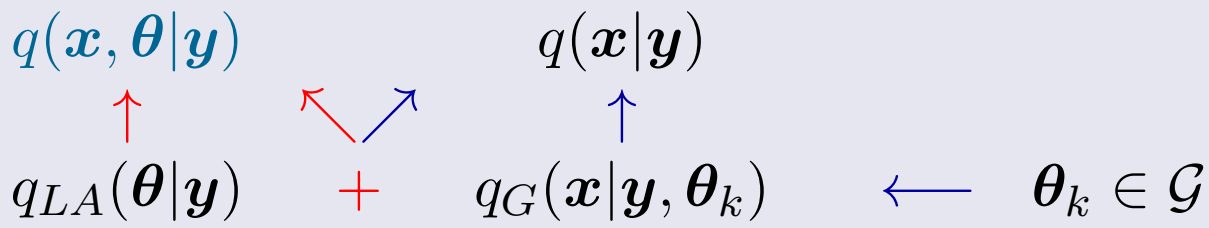
$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

with area weights $\boldsymbol{\Delta}_k$.

INLA in a Nutshell (3/3)

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

Demo:



3. Numerical integration:

$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

with area weights $\boldsymbol{\Delta}_k$.

INLA in a Nutshell (3/3)

Benefits:

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

Limitations:

1. Applies only to latent Gaussian models (LGMs)
2. No quantization for the accuracy of approximation $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$
3. The dimension of $\boldsymbol{\theta}$ has to be no more than 5 or 6

INLA in a Nutshell (3/3)

Benefits:

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

Limitations:

1. Applies only to latent Gaussian models (LGMs)
2. No quantization for the accuracy of approximation $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$
3. The dimension of $\boldsymbol{\theta}$ has to be no more than 5 or 6

Our method addresses the first two limitations with INLA.

Outline

- Introduction
- Integrated Nested Laplace Approximation (INLA)
- **Integrated Non-Factorized Variational Bayes (INF-VB)**
- Applications
- Future Research

Variational Inference

Variational inference turns Bayesian inference into optimization.

$$\min_{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})} \mathrm{KL}[q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})] \quad s.t. \quad q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \in \mathcal{Q} \tag{3}$$

Evidence lower bound (ELBO):

Applying Jensen's inequality,

$$\begin{aligned} \ln p(\boldsymbol{y}) &= \ln \int\int q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \frac{p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})}{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta} \\ &\geq \int\int q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta})}{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta} := \mathcal{L} \end{aligned} \tag{4}$$

- The Jensen's gap: $\ln p(\boldsymbol{y}) - \mathcal{L} = \mathrm{KL}(q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}))$
- The variational distribution $q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ is commonly restricted to tractable families $\mathcal{Q}$

Mean-Field Variational Bayes (VB)

Assumes **factorized** form: $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = q(\boldsymbol{x})q(\boldsymbol{\theta})$, then

$$\begin{aligned} q^\star(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &= \underset{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\operatorname{argmin}} \mathrm{KL}(q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})) \\ &= \underset{q(\boldsymbol{x}),q(\boldsymbol{\theta})}{\operatorname{argmin}} \int\int q(\boldsymbol{x})q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{x})q(\boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta} \end{aligned}$$

Remarks:

- Easily derived and in close form for conjugate models
- Challenging for non-conjugate models
- Ignores posterior dependencies and impairs the accuracy
- A poor approximation for a multi-modal distribution

Mean-Field Variational Bayes (VB)

Assumes **factorized** form: $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = q(\boldsymbol{x})q(\boldsymbol{\theta})$, then

$$q^{\star}(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \underset{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\operatorname{argmin}} \mathrm{KL}(q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}))$$

$$= \underset{q(\boldsymbol{x}),q(\boldsymbol{\theta})}{\operatorname{argmin}} \int\int q(\boldsymbol{x})q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{x})q(\boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta}$$

Remarks:

- Easily derived and in close form for conjugate models
- Challenging for non-conjugate models
- Ignores posterior dependencies and impairs the accuracy
- A poor approximation for a multi-modal distribution

Our non-factorized variational method addresses these issues with mean-field VB.

Hybrid Continuous-Discrete Family

Consider **non-factorized** form:

$$q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\boldsymbol{y}) \tag{5}$$

- $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are still coupling

1. The continuous approximation $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ is very flexible
   - Gaussian
   - Mean-Field
2. The discretized approximation $q_d(\boldsymbol{\theta}|\boldsymbol{y})$ is a finite mixture of Dirac-delta distributions,

$$q_d(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_k \omega_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}), \quad \omega_k = q_d(\boldsymbol{\theta}_k|\boldsymbol{y}), \quad \sum_k \omega_k = 1 \tag{6}$$

Proposed Method

Within the proposed hybrid family, the optimal variational distribution is

$$
\begin{aligned}
q^\star(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) &= \underset{q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\operatorname{argmin}} \mathrm{KL}(q(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})) \\
&= \underset{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}),q_d(\boldsymbol{\theta}|\boldsymbol{y})}{\operatorname{argmin}} \int\int q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta} \\
&= \underset{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k),q_d(\boldsymbol{\theta}_k|\boldsymbol{y})}{\operatorname{argmin}} \sum_k \int q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_d(\boldsymbol{\theta}_k|\boldsymbol{y}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_d(\boldsymbol{\theta}_k|\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})} d\boldsymbol{x}
\end{aligned}
$$

We give the name integrated non-factorized variational Bayes (INF-VB) to this method.

Variational Optimization Algorithm

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

$$q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \underset{q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)}{\operatorname{argmin}} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)) \tag{7}$$

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

$$q_d^\star(\boldsymbol{\theta}_k|\boldsymbol{y}) \propto \exp\left(\int q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) \ln \frac{p(\boldsymbol{x},\boldsymbol{\theta}_k|\boldsymbol{y})}{q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)} d\boldsymbol{x}\right) \tag{8}$$

- INF-VB is **parallelizable**, with dominant computational load distributed on each grid point
- INF-VB requires no iteration between Step 1 and Step 2

Our approach unifies INLA under the variational framework.

INLA v.s. INF-VB

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

1. Gaussian approximation

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

2. Hyperparameter learning

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}\right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})}$$

3. Marginal posterior of $\boldsymbol{x}$

$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

with area weights $\boldsymbol{\Delta}_k$.

INLA v.s. INF-VB

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

1. Gaussian approximation

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

(INF-VB) Step 1: Variational Gaussian approximation

$$q^\star_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \operatorname{argmin} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G}$$

2. Hyperparameter learning

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}\right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})}$$

3. Marginal posterior of $\boldsymbol{x}$

$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

INLA v.s. INF-VB

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

(INF-VB) Step 1: Variational Gaussian approximation

$$q^{\star}_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \operatorname{argmin} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G}$$

2. Hyperparameter learning

$$q_{LA}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})}{q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} \right|_{\boldsymbol{x}=\boldsymbol{x}^{\star}(\boldsymbol{\theta})}$$

(INF-VB) Step 2:

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

3. Marginal posterior of $\boldsymbol{x}$

$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

INLA v.s. INF-VB

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

(INF-VB) Step 1: Variational Gaussian approximation

$$q^{\star}_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \operatorname{argmin} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G}$$

(INF-VB) Step 2: Hyperparameter learning

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

3. Marginal posterior of $\boldsymbol{x}$

$$q(\boldsymbol{x}|\boldsymbol{y}) = \sum_k q_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q_{LA}(\boldsymbol{\theta}_k|\boldsymbol{y}) \boldsymbol{\Delta}_k$$

(INF-VB) Step 3:

$$q(\boldsymbol{x}|\boldsymbol{y}) = \int q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \sum_k q^{\star}_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q^{\star}_d(\boldsymbol{\theta}_k|\boldsymbol{y})$$

INLA v.s. INF-VB

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

(INF-VB) Step 1: Variational Gaussian approximation

$$q^{\star}_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) = \operatorname{argmin} \mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)), \quad \forall \boldsymbol{\theta}_k \in \mathcal{G}$$

(INF-VB) Step 2: Hyperparameter learning

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

(INF-VB) Step 3: Marginal posterior of $\boldsymbol{x}$

$$q(\boldsymbol{x}|\boldsymbol{y}) = \int q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) q_d(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \sum_k q^{\star}_{VG}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k) q^{\star}_{d}(\boldsymbol{\theta}_k|\boldsymbol{y})$$

Remarks

Benefits:

- Applicable to more general scenarios
- Optimal variational distributions $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}_k)$ and $q_d(\boldsymbol{\theta}|\boldsymbol{y})$
- Negative ELBO provides quantitization of the accuracy

Limitations:

- The dimension of $\boldsymbol{\theta}$ has to be no more than 5 or 6

Application to Bayesian Lasso

1. **Non-differentiability** of the ℓ_1 norm
2. The Laplace approximation of INLA cannot be applied

Bayesian Lasso Regression[4] (1/3)

Model:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e}, \quad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{\Phi} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{e} \in \mathbb{R}^n$. We assume

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

Problem: Given $\boldsymbol{y}$ and $\boldsymbol{\Phi}$, find posterior distributions for $\boldsymbol{x}$ and $\boldsymbol{\theta} = \{\lambda^2, \sigma^2\}$

[4] Park & Casella, 2008

Bayesian Lasso Regression (2/3)

Inference:

1. Data augmentation Gibbs sampler
2. Mean-Field VB
3. **INF-VB**

INF-VB for Bayesian Lasso

(1) $q^\star(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}_k)$: constrain $q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{C}\boldsymbol{C}^T)$, then

$$\mathrm{KL}(q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})||p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})) := g(\boldsymbol{\mu},\boldsymbol{C}) \tag{9}$$

is concave in $(\boldsymbol{\mu},\boldsymbol{C})$[a], $\boldsymbol{D} = \boldsymbol{C}\boldsymbol{C}^T$.

(2) $q^\star(\boldsymbol{\theta}|\boldsymbol{y})$: can be evaluated analytically

(3) $q^\star(\boldsymbol{x}|\boldsymbol{y})$: finite mixture of Gaussians

[a] Challis & Barber, 2011

Bayesian Lasso Regression (3/3)

Denote $(\boldsymbol{\mu}^\star, \boldsymbol{D}^\star) = \operatorname{argmin}_{\boldsymbol{\mu},\boldsymbol{D}} g(\boldsymbol{\mu}, \boldsymbol{D})$, the variational Bayesian Lasso,

$$\boldsymbol{\mu}^\star = \operatorname*{argmin}_{\boldsymbol{\mu}} g(\boldsymbol{\mu}), \quad g(\boldsymbol{\mu}) := \mathbb{E}_{\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{D}^\star)}\left(\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2\lambda\sigma\|\boldsymbol{x}\|_1\right) \quad (10)$$

is a counterpart of Lasso[5],

$$\hat{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{x}} f(\boldsymbol{x}), \quad f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + 2\lambda\sigma\|\boldsymbol{x}\|_1 \quad (11)$$

Remarks:

- The conditions of Lasso **hold on average**
- Smoothing around origin and thus **differentiable**
- Optimize a non-differential function by operating on a sequence of differentiable functions

[5] Tibshirani, 1996

Results (1/4): Diabetes Dataset[6]

This benchmark dataset contains

- Measurements on $n = 442$ diabetes patients
- $p = 10$ clinical covariates (age, sex, body mass index, average blood pressure, and six blood serum measurements)
- Response variable, a quantitative measure of disease progression

Goal:

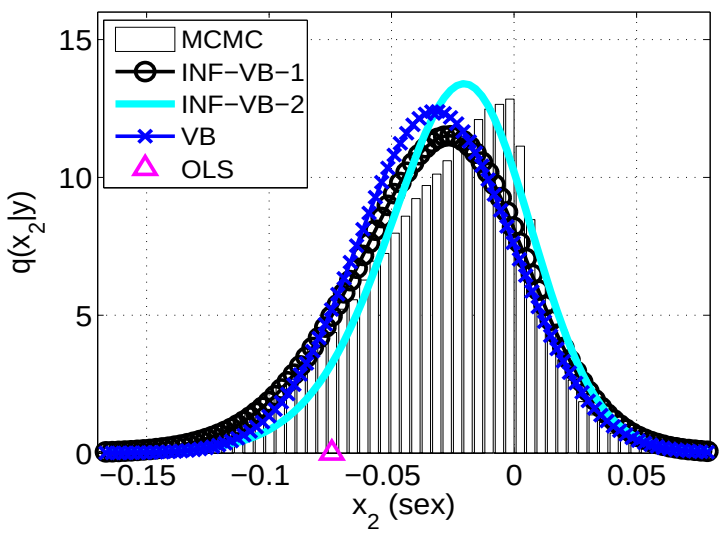
- Identify which covariates are important factors

Methods:

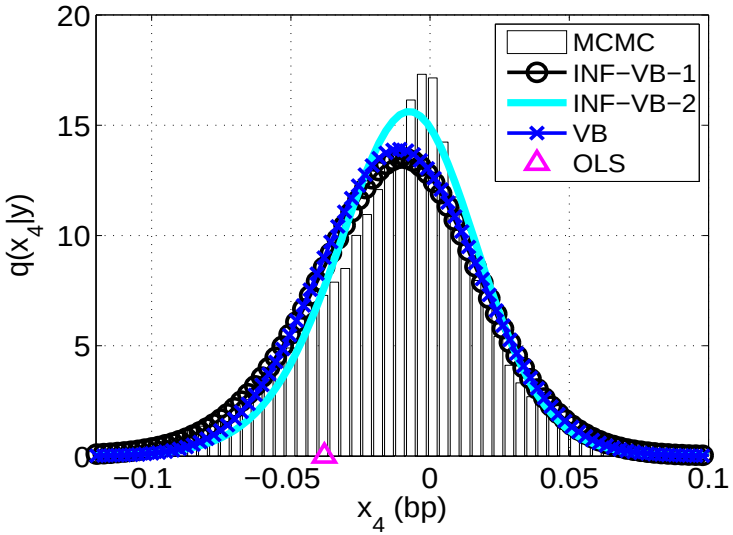
- **Intensive MCMC runs (ground truth)**
- Mean-Field VB
- **INF-VB-1**
- INF-VB-2 (INLA, replace LA with VG)
- Ordinary least square (OLS)

[6] Efron et al., 2004

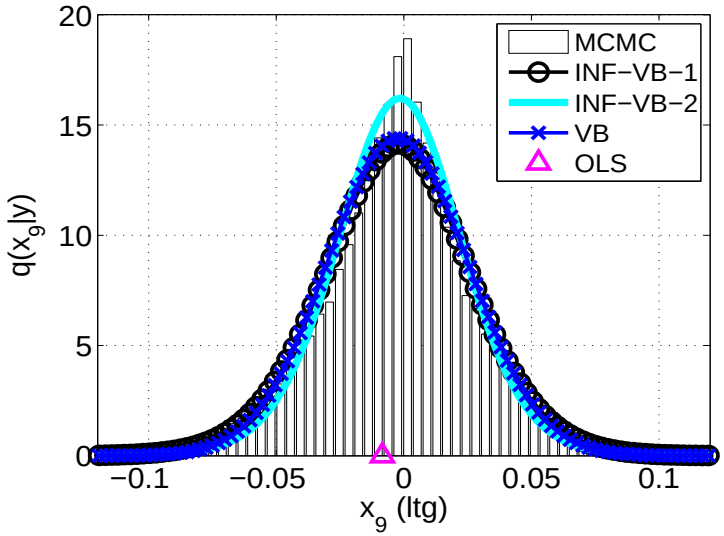
Results (2/4): Marginal Posteriors $q(x_j|\boldsymbol{y})$



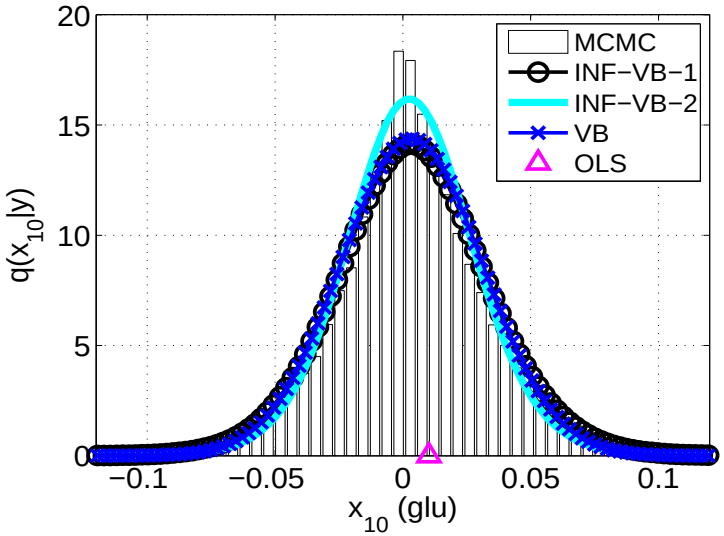
(a)



(b)

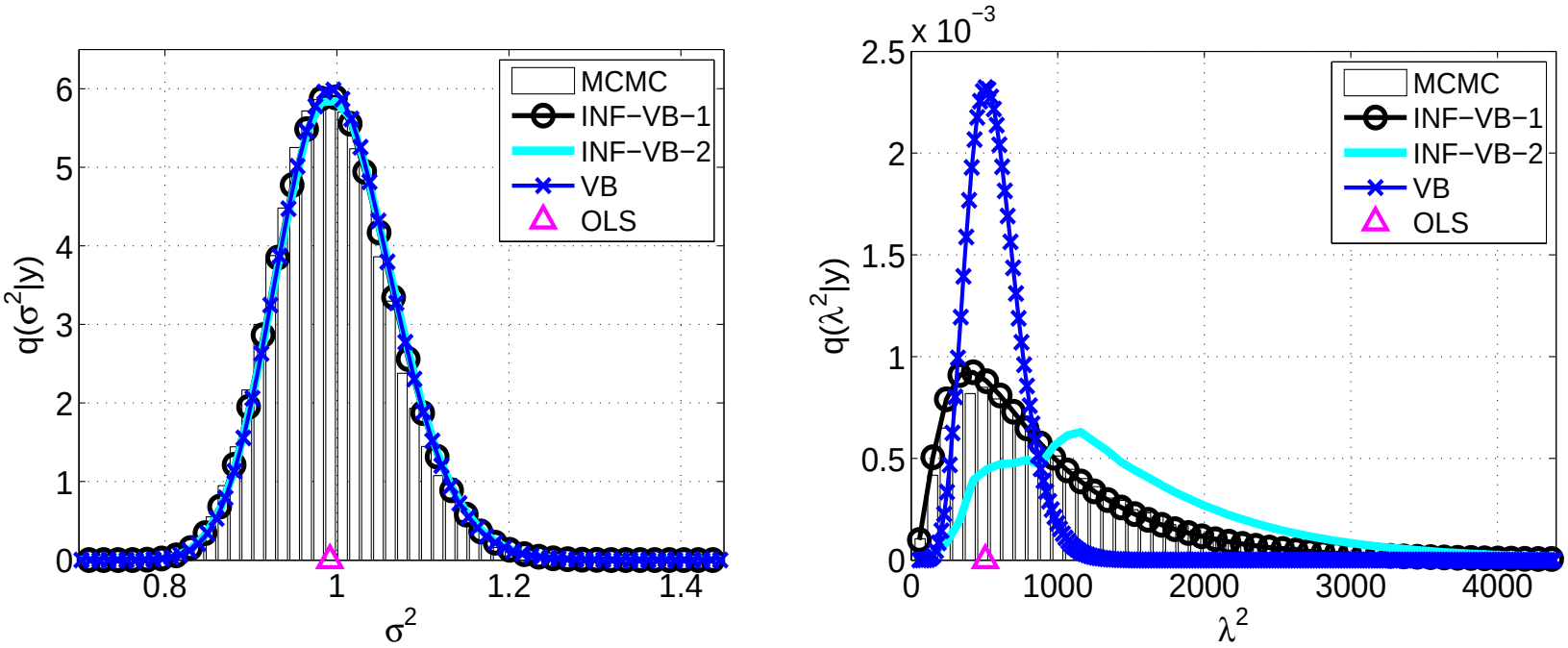


(c)



(d)

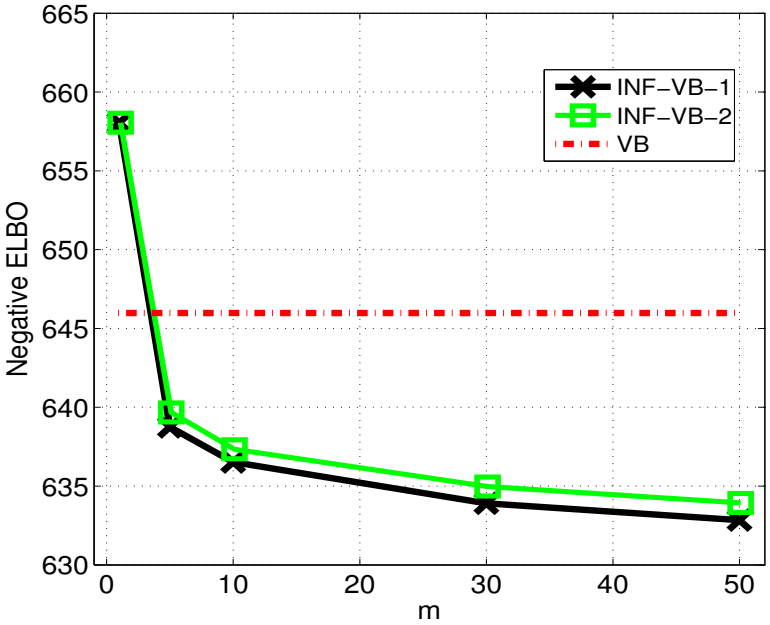
Results (3/4): Marginal Posteriors $q(\sigma^2|\boldsymbol{y})$ and $q(\lambda^2|\boldsymbol{y})$



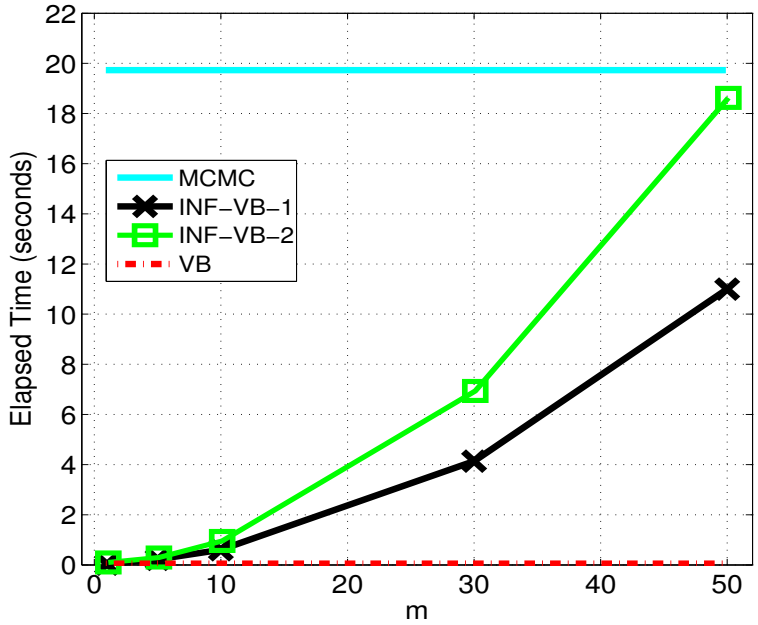
Posterior marginals of hyperparameters: (a) $q(\sigma^2|\boldsymbol{y})$ and (b) $q(\lambda^2|\boldsymbol{y})$

- Mean-Field VB could severely underestimate the posterior variance
- INF-VB-2 offers suboptimal solution

Results (4/4): Accuracy and Speed



(a) Accuracy



(b) Time

Grid size $m \times m$ and $m = 1, 5, 10, 30, 50$.

- INF-VB with a 1×1 grid: partial Bayesian learning of $q(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ with a fixed $\boldsymbol{\theta}$

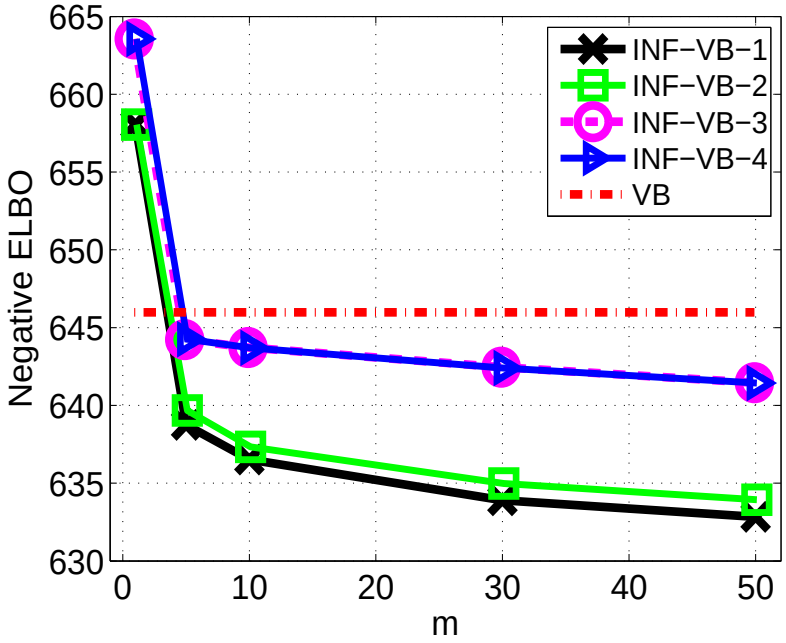
Summary

Our method:

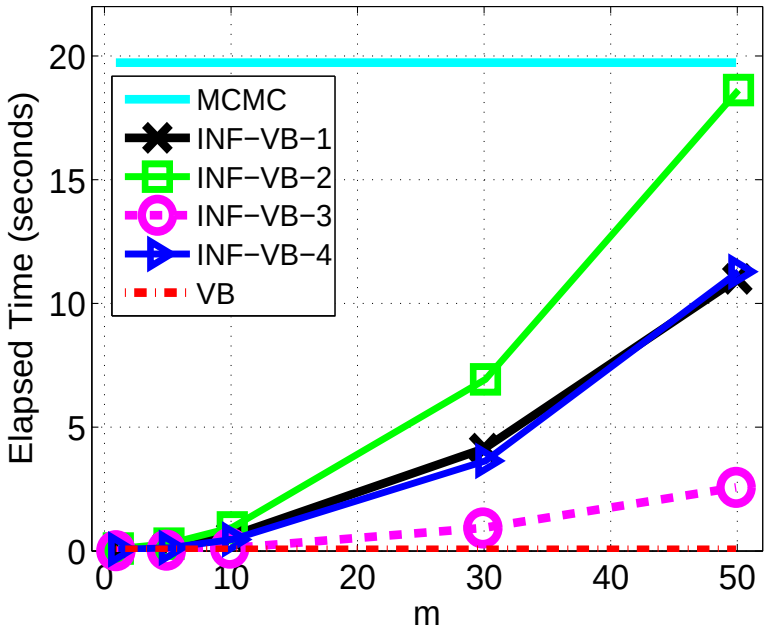
1. Tractable family $\mathcal{Q}$: **non-factorized**
2. Conditional conjugacy: **not required**
3. Multimodal posterior: **could handle**
4. Parallelizable: **yes**

More could be done...

Q&A: Accuracy and Speed



(a)



(b)

In INF-VB-3 and INF-VB-4 (INLA, replace LA with VG), we obtain a fast VG solution by minimizing a KL divergence upper bound

Grid size $m \times m$ and $m = 1, 5, 10, 30, 50$.