Variational Bayesian Methods Beyond Parametric and Continuous Assumptions

Mingyuan Zhou[⋆§]
Joint work with Mingzhang Yin[§]

[⋆]IROM Department, McCombs School of Business
[§]Department of Statistics and Data Sciences
The University of Texas at Austin

Workshop on Bayesian Nonparametrics for Signal and Image Processing
Bordeaux, July 3, 2018

Joint work with

Mingzhang Yin

PhD student (since Fall 2015)
in Statistics and Data Sciences

Bayesian Inference

- Bayes' rule:

$$P(\boldsymbol{z} \mid X) = \frac{P(X \mid \boldsymbol{z})P(\boldsymbol{z})}{P(X)} = \frac{P(X \mid \boldsymbol{z})P(\boldsymbol{z})}{\int P(X \mid \boldsymbol{z})P(\boldsymbol{z})d\boldsymbol{z}}$$

$$\text{Posterior of } \boldsymbol{z} \text{ given } X = \frac{\text{Conditional Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}}$$

- Two main ways for approximate Bayesian inference:
  - Draw $\boldsymbol{z} \sim P(\boldsymbol{z} \mid X)$ using Markov chain Monte Carlo (MCMC) based methods such as Gibbs sampling: iteratively sample $P(z_k \mid X, \boldsymbol{z} \backslash z_k)$
  - Approximate the posterior $P(\boldsymbol{z} \mid X)$ with $Q(\boldsymbol{z})$, which is straightforward to sample from, using an optimization method such as Laplace approximation and variational inference

Inference via Gibbs sampling

- Gibbs sampling:
  - One of the simplest MCMC algorithm for multivariate distributions
  - Widely used for statistical inference
- For a multivariate distribution $P(z_1, \ldots, z_K)$ that is difficult to sample from, if it is simpler to sample each of its variables conditioning on all the others, then we may use Gibbs sampling to obtain samples from this distribution as
  - Initialize $(z_1, \ldots, z_K) = (z_1^0, \ldots, z_K^0)$ at some values
  - For $s = 1 : S$
    For $k = 1 : K$
    Sample z_k^s conditioning on the others from
    $P(z_k^s \mid z_1^s, \ldots, z_{k-1}^s, z_{k+1}^{s-1}, \ldots, z_K^{s-1})$
    End
    End
- Restriction of Gibbs sampling: conjugacy is often required

Variational inference

- With variational distribution $Q(\boldsymbol{z})$, we have

$$\begin{aligned}\ln P(X) &= \int Q(\boldsymbol{z}) \ln \frac{P(X, \boldsymbol{z})}{Q(\boldsymbol{z})} d\boldsymbol{z} + \int Q(\boldsymbol{z}) \ln \frac{Q(\boldsymbol{z})}{P(\boldsymbol{z} \mid X)} d\boldsymbol{z} \\ &= \mathcal{L}(Q) + \mathrm{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z} \mid X)).\end{aligned}$$

- Since $\mathrm{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z} \mid X)) \geq 0$, minimizing the Kullback-Leibler (KL) divergence from $P(\boldsymbol{z} \mid X)$ to $Q(\boldsymbol{z})$ is the same as maximizing the evidence lower bound:

$$\min_Q \mathrm{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z} \mid X)) \Leftrightarrow \max_Q \mathrm{ELBO}$$

$$\begin{aligned}\mathrm{ELBO} = \mathcal{L}(Q) &= \mathbb{E}_Q[\ln P(X, \boldsymbol{z})] - \mathbb{E}_Q[\ln Q(\boldsymbol{z})] \\ &= \mathbb{E}_Q[\ln P(X \mid \boldsymbol{z})] - \mathrm{KL}(Q(\boldsymbol{z}) || P(\boldsymbol{z}))\end{aligned}$$

- Variational inference converts the problem of posterior inference into an optimization problem

Mean-field variational inference

- Mean-field variational inference (VI) factorizes the Q distribution of $\boldsymbol{z} = (z_1, \ldots, z_K)^T$ as

$$Q(\boldsymbol{z}) = \prod_{i=1}^{K} q_{\phi_i}(z_i)$$

- The factorized assumption allows for closed-form coordinate ascent updates:

$$q^*(z_k) = \frac{\exp\left\{\mathbb{E}_{q(\boldsymbol{z}_{-k})}[\log p(X, z_k, \boldsymbol{z}_{-k})]\right\}}{\int \exp\left\{\mathbb{E}_{q(\boldsymbol{z}_{-k})}[\log p(X, z_k, \boldsymbol{z}_{-k})]\right\} dz_k}, \quad k = 1, \cdots, K$$

  where $\boldsymbol{z}_{-k} = \{z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_K\}$.

- However, mean-field VI often clearly underestimates the variance of the posterior, due to the use of KL divergence and two restrictive constraints:
  - $q(z_k)$ are often restricted to the exponential family
  - The dependencies between z_k cannot be captured

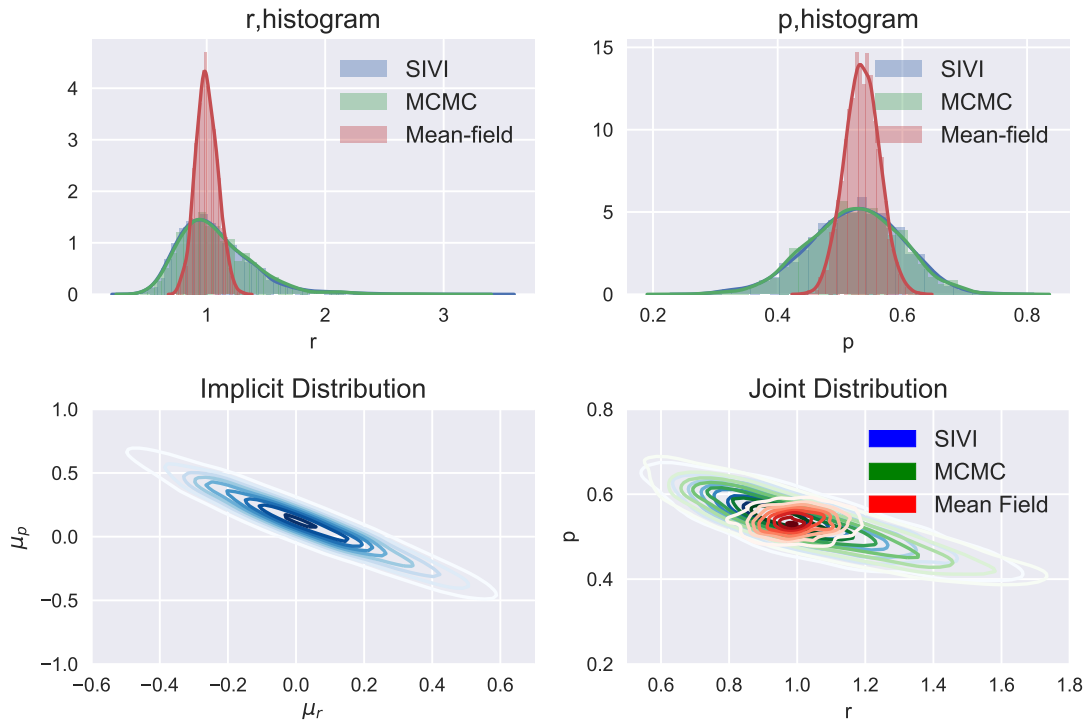
Model:

$$x_i \overset{i.i.d.}{\sim} \text{NB}(r, p), \;\; r \sim \text{Gamma}(a, 1/b), \;\; p \sim \text{Beta}(\alpha, \beta),$$

Mean-filed VI:

$$Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}),$$

Mean-filed VI underestimates variance (mainly due to the factorized assumption):

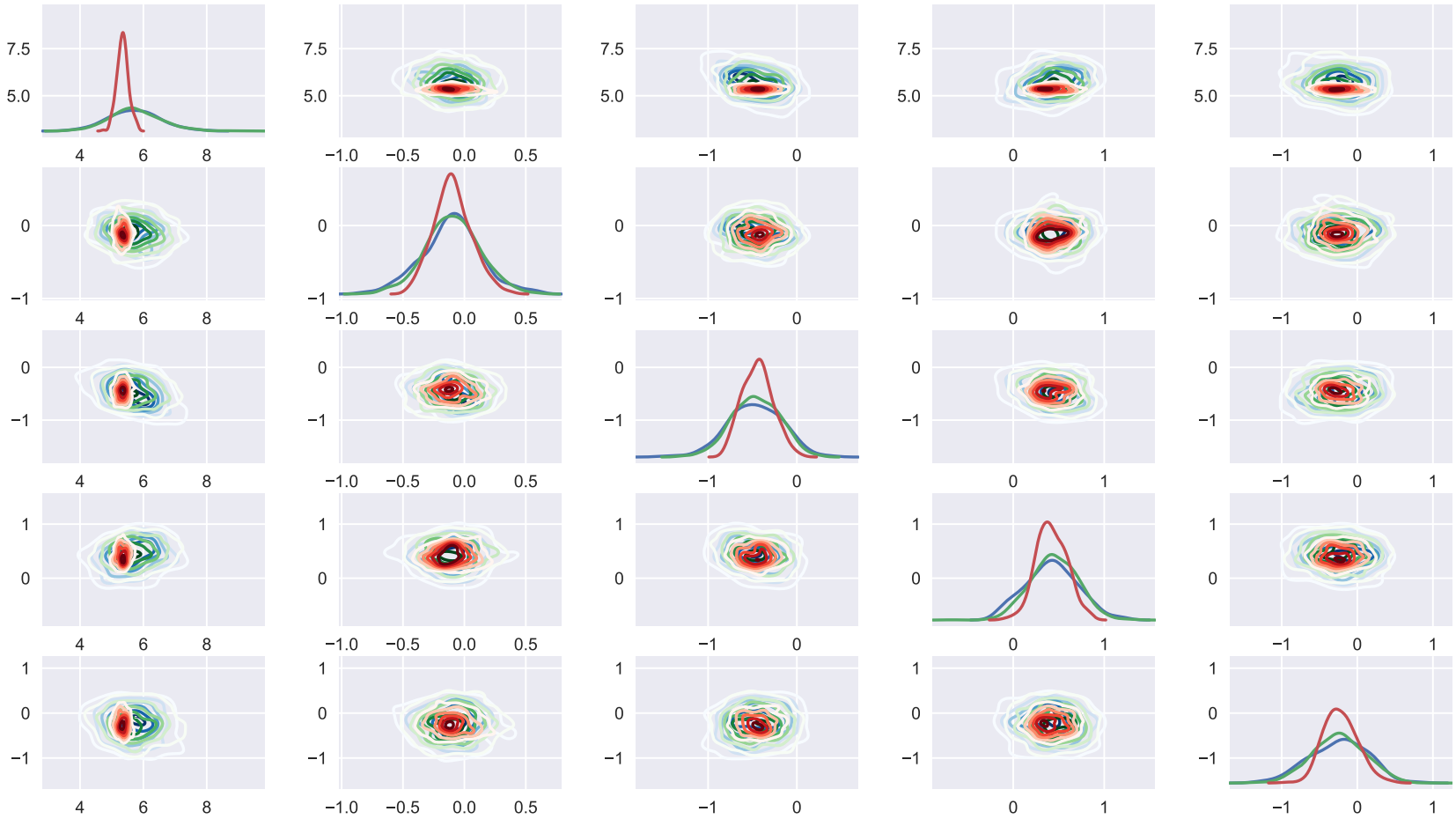


Bayesian logistic regression:

$$y_i \sim \text{Bernoulli}[(1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}})^{-1}], \quad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{V+1})$$

VI: $Q(\beta) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which underestimates variance (mainly due to the mismatch between $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the true posterior)

Blue: MCMC, Red: VI:



"Modern" variational inference

- Choose a more flexible $Q_\phi(\boldsymbol{z})$ and infer the variational parameter ϕ to maximize the ELBO via (stochastic) gradient ascent

$$\nabla_\phi \mathcal{L}(Q_\phi(\boldsymbol{z})) = \nabla_\phi \mathbb{E}_{\boldsymbol{z}\sim Q_\phi(\boldsymbol{z})}\left[\ln \frac{P(X,\boldsymbol{z})}{Q_\phi(\boldsymbol{z})}\right]$$

- Compute the gradient of the ELBO with respect to ϕ:
  - Score function gradient (a.k.a. REINFORCE, often suffering from high Monte Carlo estimation variance):

$$\nabla_\phi \mathcal{L}(Q_\phi(\boldsymbol{z})) = \mathbb{E}_{\boldsymbol{z}\sim Q_\phi(\boldsymbol{z})}\left[\ln \frac{P(X,\boldsymbol{z})}{Q_\phi(\boldsymbol{z})} \nabla_\phi \ln Q_\phi(\boldsymbol{z})\right]$$

  - If $\boldsymbol{z}\sim Q_\phi(\boldsymbol{z})$ is reparameterizable such that $\boldsymbol{z} = T_\phi(\boldsymbol{\epsilon}),\ \boldsymbol{\epsilon}\sim q(\boldsymbol{\epsilon})$, then one can use the reparameterization trick:

$$\nabla_\phi \mathcal{L}(Q_\phi(\boldsymbol{z})) = \mathbb{E}_{\boldsymbol{\epsilon}\sim q(\boldsymbol{\epsilon})}\left[\nabla_\phi \ln \frac{P(X,T_\phi(\boldsymbol{\epsilon}))}{Q_\phi(T_\phi(\boldsymbol{\epsilon}))}\right]$$

Challenges remain for "modern" variational inference

Parametric and continuous assumptions are commonly made, since

- There is a conflict between the ease of evaluating the log density ratio $\ln \frac{P(X,z)}{Q_\phi(z)}$ and the richness of $Q_\phi(z)$:
  - $\ln \frac{P(X,z)}{Q_\phi(z)}$ is straightforward to compute if $Q_\phi(z)$ is restricted to be analytic and point-wise evaluable
  - $\ln \frac{P(X,z)}{Q_\phi(z)}$ becomes difficult to compute if expanding the richness of the variational distribution family, e.g., allowing $Q_\phi(z)$ to be implicit
- One needs to control the Monte Carlo estimation variance:
  - The REINFORCE estimator often has large variance and needs to introduce appropriate control variates for variance reduction
  - The reparameterization trick often leads to low variance, but in general, it is not applicable to
    - Discrete distributions such as Bernoulli, categorical, and Poisson
    - Some commonly used continuous distributions such as gamma, beta, and Dirichlet

Relax parametric assumption with implicit distribution

Implicit distribution consists of a source of randomness $q(\boldsymbol{\epsilon})$ and a deterministic transform $T_{\phi} : \mathbb{R}^p \to \mathbb{R}^d$

$$\boldsymbol{z} = T_{\phi}(\boldsymbol{\epsilon}),\ \boldsymbol{\epsilon} \sim q(\boldsymbol{\epsilon})$$

- When T_{ϕ} is invertible and the dimension is low, the density

$$q_{\phi}(\boldsymbol{z}) = \frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_d} \int_{T_{\phi}(\boldsymbol{\epsilon}) \leq \boldsymbol{z}} q(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$$

  can be calculated using change of variables
- But in general $\{T_{\phi}(\boldsymbol{\epsilon}) \leq \boldsymbol{z}\}$ cannot be calculated and hence the high dimension integral is intractable, making $q_{\phi}(\boldsymbol{z})$ become implicit
- Even if $q_{\phi}(\boldsymbol{z})$ is implicit and hence difficult to evaluate, sampling $\boldsymbol{z} \sim q_{\phi}(\boldsymbol{z})$ is straightforward
- T_{ϕ} often corresponds to a Deep Neural Network

Why Deep Learning?

- TensorFlow, PyTorch, CNTK, Theano, . . . :
  - Automatic differentiation
  - Neural network libraries
  - Off-the-shelf optimization packages
- Exciting opportunities to be combined with Bayesian methods
  - Incorporating deep neural networks into hierarchical Bayesian models to define deep generative models
  - Empower variational inference with deep neural networks
  - Move beyond parametric and continuous assumptions
  - Define implicit distributions with MCMC
  - Build deep neural networks with many stochastic hidden layers
  - Design MCMC transition kernels with deep neural networks
  - Gradient backpropagation for discrete latent variables
  - . . .

Hierarchical variational family

The key to accurately capture the uncertainty is to capture the latent variable dependencies and expand the richness of the variational family

- One way is to add a hierarchical structure that assumes z_k to be conditional independent but marginally dependent, using

$$q(\boldsymbol{z} \mid \boldsymbol{\psi}) = \prod_{k=1}^{K} q(z_k \mid \psi_k), \quad \boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$$

- Marginalizing $\boldsymbol{\psi}$ out, we can view $\boldsymbol{z}$ as a variable drawn from the distribution family $\mathcal{H}$

$$\mathcal{H} = \left\{ h_{\boldsymbol{\phi}}(\boldsymbol{z}) : h_{\boldsymbol{\phi}}(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}[q(\boldsymbol{z} \mid \boldsymbol{\psi})] = \int_{\boldsymbol{\psi}} \left[\prod_{k=1}^{K} q(z_k \mid \psi_k) \right] q_{\boldsymbol{\phi}}(\boldsymbol{\psi}) d\boldsymbol{\psi} \right\}$$

- It is evident that $q(\boldsymbol{z} \mid \boldsymbol{\psi}) \in \mathcal{Q} \subseteq \mathcal{H}$, i.e., $\mathcal{H}$ expands the original variational distribution family

Semi-implicit variational inference (SIVI)

- SIVI chooses $h_\phi(\boldsymbol{z}) = \mathbb{E}_{q_\phi(\psi)} q(\boldsymbol{z} \mid \psi)$ as its variational distribution
- Optimize ELBO $= \mathbb{E}_{h_\phi(\boldsymbol{z})}[\ln p(\boldsymbol{x}, \boldsymbol{z}) - \ln h_\phi(\boldsymbol{z})]$ for SIVI is generally intractable if $h_\phi(\boldsymbol{z}) = \mathbb{E}_{q_\phi(\psi)} q(\boldsymbol{z} \mid \psi)$ is not analytic
- KL convexity and Jensen's inequality lead to an ELBO lower bound:

$$\begin{aligned}
&\underline{\mathcal{L}}(q(\boldsymbol{z} \mid \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \psi)} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \psi)} \\
&= -\mathbb{E}_{\psi \sim q_\phi(\psi)} \mathrm{KL}(q(\boldsymbol{z} \mid \psi) || p(\boldsymbol{z} | \boldsymbol{x})) + \log p(\boldsymbol{x}) \\
&\leq -\mathrm{KL}(\mathbb{E}_{\psi \sim q_\phi(\psi)} q(\boldsymbol{z} \mid \psi) || p(\boldsymbol{z} | \boldsymbol{x})) + \log p(\boldsymbol{x}) \\
&= \mathcal{L} = \mathbb{E}_{\boldsymbol{z} \sim h_\phi(\boldsymbol{z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{h_\phi(\boldsymbol{z})}
\end{aligned}$$

- Using the concavity of the logarithmic function, we have $\log h_\phi(\boldsymbol{z}) \geq \mathbb{E}_{\psi \sim q_\phi(\psi)} \log q(\boldsymbol{z} \mid \psi)$ and hence an ELBO upper bound:

$$\bar{\mathcal{L}}(q(\boldsymbol{z} \mid \psi), q_\phi(\psi)) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\boldsymbol{z} \sim h_\phi(\boldsymbol{z})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \psi)} \geq \mathcal{L}$$

- Note there is a subtle but critical difference between $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$

To compute $\nabla_\phi \underline{\mathcal{L}}$, we require

- $q(\boldsymbol{z} \mid \boldsymbol{\psi})$ is reparameterizable and has an explicit probability density function
- $q_\phi(\boldsymbol{\psi})$ is reparameterizable and easy to sample from (does not have to be explicit), e.g., $q_\phi(\boldsymbol{\psi})$ can be constructed by transforming random noise $\boldsymbol{\epsilon}$ via a deep neural network parameterized by $\boldsymbol{\phi}$

Maximizing the surrogate lower bound $\underline{\mathcal{L}}$ may lead to degeneracy that $q_\phi(\boldsymbol{\psi})$ converges to a point mass density:

Proposition (Degeneracy)

Let us denote $\boldsymbol{\psi}^* = \arg\max_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \boldsymbol{\psi})} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{\psi})}$, *then*

$$\underline{\mathcal{L}}(q(\boldsymbol{z} \mid \boldsymbol{\psi}), q_\phi(\boldsymbol{\psi})) \leq \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z} \mid \boldsymbol{\psi}^*)} \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z} \mid \boldsymbol{\psi}^*)},$$

where the equality is true if and only if $q_\phi(\boldsymbol{\psi}) = \delta_{\boldsymbol{\psi}^*}(\boldsymbol{\psi})$.

Asymptotically exact ELBO (regularizing the lower bound)

- Avoid degeneracy by adding regularization $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$

$$B_K = \mathbb{E}_{\boldsymbol{\psi},\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\mathrm{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{\psi})||\tilde{h}_K(\boldsymbol{z})), \tag{1}$$

  where $\tilde{h}_K(\boldsymbol{z}) = \frac{1}{K+1}[q(\boldsymbol{z}\,|\,\boldsymbol{\psi}) + \sum_{k=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})]$, $B_K \geq 0$, with $B_K = 0$ if and only if $K = 0$ or $q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$ degenerates to a point mass density

- The regularized surrogate ELBO can also be expressed as

$$\underline{\mathcal{L}}_K = \mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{\frac{1}{K+1}\left[q(\boldsymbol{z}\,|\,\boldsymbol{\psi}) + \sum_{k=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})\right]}$$

Proposition

The regularized lower bound $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$ is an asymptotically exact ELBO that satisfies $\underline{\mathcal{L}}_0 = \underline{\mathcal{L}}$ and $\lim_{K\to\infty} \underline{\mathcal{L}}_K = \mathcal{L}$

Asymptotically exact ELBO (correcting the upper bound)

- Avoid divergence by adding correction $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$

$$A_K = \mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim h_{\boldsymbol{\phi}}(\boldsymbol{z})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\Big[\log\Big(\frac{1}{K}\sum_{k=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})\Big) - \log q(\boldsymbol{z}\,|\,\boldsymbol{\psi})\Big].$$

- The corrected upper bound can be expressed as

$$\bar{\mathcal{L}}_K = \mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}\,|\,\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{\psi}^{(1)},\ldots,\boldsymbol{\psi}^{(K)}\sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})} \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{\frac{1}{K}\sum_{k=1}^{K} q(\boldsymbol{z}\,|\,\boldsymbol{\psi}^{(k)})}$$

Proposition

The corrected upper bound $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$ monotonically converges from the above towards the ELBO, satisfying $\bar{\mathcal{L}}_1 = \bar{\mathcal{L}}$, $\bar{\mathcal{L}}_{K+1} \leq \bar{\mathcal{L}}_K$, and $\lim_{K\to\infty} \bar{\mathcal{L}}_K = \mathcal{L}$.

Algorithm for SIVI

Algorithm 1 Semi-Implicit Variational Inference (SIVI)

input : Data $\{x_i\}_{1:N}$, joint likelihood $p(\boldsymbol{x}, \boldsymbol{z})$, explicit variational distribution $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid \boldsymbol{\psi})$ with reparameterization $\boldsymbol{z} = f(\boldsymbol{\epsilon}, \boldsymbol{\xi}, \boldsymbol{\psi})$, $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$, implicit layer neural network $T_{\boldsymbol{\phi}}(\epsilon)$ and source of randomness $q(\boldsymbol{\epsilon})$

output : Variational parameter $\boldsymbol{\xi}$ for the conditional distribution $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid \boldsymbol{\psi})$, variational parameter $\boldsymbol{\phi}$ for the mixing distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$

Initialize $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ randomly

while *not converged* **do**

  Set $\underline{L}_{K_t} = 0$, ρ_t and η_t as step sizes, and $K_t \geq 0$ as a non-decreasing integer; Sample $\boldsymbol{\psi}^{(k)} = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(k)})$, $\boldsymbol{\epsilon}^{(k)} \sim q(\boldsymbol{\epsilon})$ for $k = 1, \ldots, K_t$; take subsample $\mathbf{x} = \{x_i\}_{i_1:i_M}$

  for $j = 1$ **to** J **do**

    Sample $\boldsymbol{\psi}_j = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j)$, $\boldsymbol{\epsilon}_j \sim q(\boldsymbol{\epsilon})$

    Sample $\boldsymbol{z}_j = f(\tilde{\boldsymbol{\epsilon}}_j, \boldsymbol{\xi}, \boldsymbol{\psi}_j)$, $\tilde{\boldsymbol{\epsilon}}_j \sim p(\boldsymbol{\varepsilon})$

    $\underline{L}_{K_t} = \underline{L}_{K_t} + \frac{1}{J}\{ -\log \frac{1}{K_t+1}\big[\sum_{k=1}^{K_t} q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid \boldsymbol{\psi}^{(k)}) + q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid \boldsymbol{\psi}_j)\big] + \frac{N}{M} \log p(\mathbf{x} \mid \boldsymbol{z}_j) + \log p(\boldsymbol{z}_j)\}$

  end

  $t = t + 1$

  $\boldsymbol{\xi} = \boldsymbol{\xi} + \rho_t \nabla_{\boldsymbol{\xi}} \underline{L}_{K_t}\big(\{\boldsymbol{\psi}^{(k)}\}_{1,K_t}, \{\boldsymbol{\psi}_j\}_{1,J}, \{\boldsymbol{z}_j\}_{1,J}\big)$

  $\boldsymbol{\phi} = \boldsymbol{\phi} + \eta_t \nabla_{\boldsymbol{\phi}} \underline{L}_{K_t}\big(\{\boldsymbol{\psi}^{(k)}\}_{1,K_t}, \{\boldsymbol{\psi}_j\}_{1,J}, \{\boldsymbol{z}_j\}_{1,J}\big)$

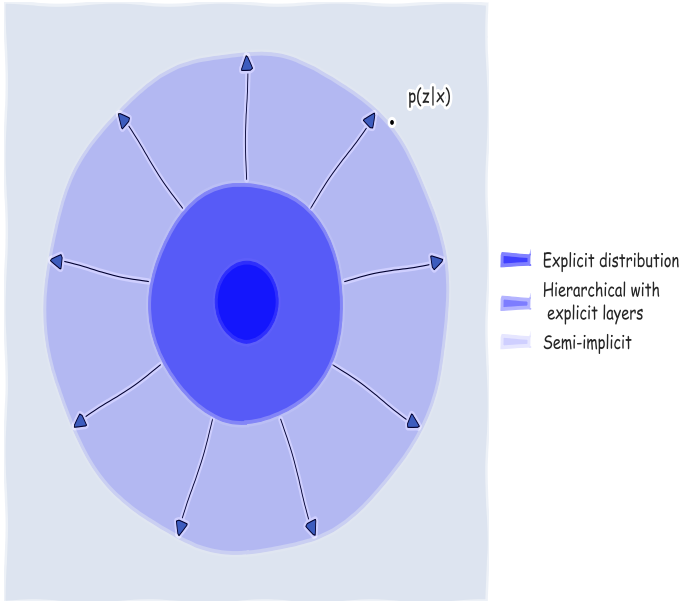
end

Methods to expand variational distribution family

Expand variational family via stochastic hierarchy and/or deterministic nonlinear transform

- Hierarchy with explicit layers:
  Negative Binomial⇔Poisson-Gamma hierarchy

- Normalizing Flow: transfer simple distribution with a chain of simple invertible mapping $\boldsymbol{z}_t = f_t \circ \cdots \circ f_0(\boldsymbol{z}_0)$
- Modeling the dependencies between univariate marginals with copula
- Implicit distribution $\boldsymbol{z} = f(\boldsymbol{\epsilon})$, where f is not invertible
- Our approach: hierarchy with explicit conditional layer, implicit mixing layers (semi-implicit)



Related work

Hierarchical models in VI:

- Hierarchical variational models (Ranganath et al., 2016)
- Auxiliary deep generative models (Maaløe et al., 2016)
- Hierarchical implicit models and likelihood-free variational inference (Tran et al., 2017)

Implicit models in VI:

- Learning in implicit generative models (Mohamed and Lakshminarayanan, 2016)
- Variational inference using implicit distributions (Huszár, 2017)
- Implicit variational inference with kernel density ratio fitting (Shi et al., 2017)

Auxiliary deep generative models:

- Generative model (decoder): $\boldsymbol{x} \mid \boldsymbol{z} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}),\ \boldsymbol{z} \sim p(\boldsymbol{z})$
- Auxiliary model in the decoder: $\boldsymbol{\psi} \mid \boldsymbol{z}, \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{\psi} \mid \boldsymbol{z}, \boldsymbol{x})$
- Inference model (encoder): $q(\boldsymbol{z}, \boldsymbol{\psi} \mid \boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x})$
- $\text{ELBO}_{\text{Auxiliary}} = \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x})} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x})} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\psi} \mid \boldsymbol{z}, \boldsymbol{x}) p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x})}$
- This lower bound is also used by the hierarchical variational models of Ranganath et al. (2016)

Semi-implicit variational inference (SIVI):

- Generative model (decoder): $\boldsymbol{x} \mid \boldsymbol{z} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}),\ \boldsymbol{z} \sim p(\boldsymbol{z})$
- Auxiliary model in the decoder: N/A
- Inference model (encoder): $h_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x}) d\boldsymbol{\psi} = \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x})} [q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x})]$
- $\text{ELBO}_{\text{SIVI}} = \mathbb{E}_{\boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x})} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x})} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z})}{\int q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{\psi}, \boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{\psi} \mid \boldsymbol{x}) d\boldsymbol{\psi}}$

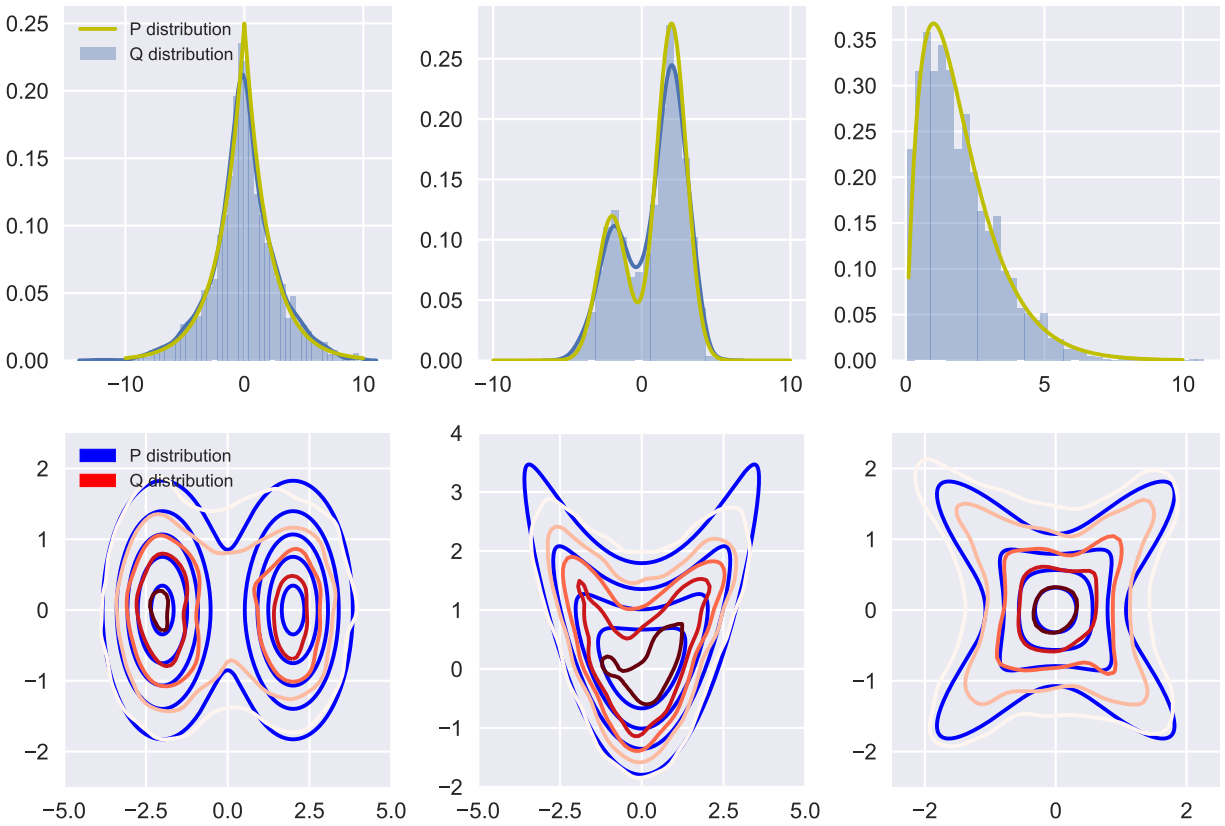
- $\text{ELBO}_{\text{Auxiliary}} = \mathbb{E}_{\psi\sim q_\phi(\psi\,|\,x)}\mathbb{E}_{z\sim q_\phi(z\,|\,\psi,x)} \log \frac{p_\theta(\psi\,|\,z,x)p_\theta(x\,|\,z)p(z)}{q_\phi(z\,|\,\psi,x)q_\phi(\psi\,|\,x)}$
- $\text{ELBO}_{\text{SIVI}} = \mathbb{E}_{\psi\sim q_\phi(\psi\,|\,x)}\mathbb{E}_{z\sim q_\phi(z\,|\,\psi,x)} \log \frac{p_\theta(x\,|\,z)p(z)}{\int q_\phi(z\,|\,\psi,x)q_\phi(\psi\,|\,x)d\psi}$
- Key differences:
  - SIVI has a tighter ELBO: $\log p(x) \geq \text{ELBO}_{\text{SIVI}} \geq \text{ELBO}_{\text{Auxiliary}}$

$$
\begin{aligned}
&\text{ELBO}_{\text{SIVI}} - \text{ELBO}_{\text{Auxiliary}} \\
&= \mathbb{E}_{\psi\sim q_\phi(\psi\,|\,x)}\mathbb{E}_{z\sim q_\phi(z\,|\,\psi,x)} \log \frac{\frac{q_\phi(z\,|\,\psi,x)q_\phi(\psi\,|\,x)}{\int q_\phi(z\,|\,\psi,x)q_\phi(\psi\,|\,x)d\psi}}{p_\theta(\psi\,|\,z,x)} \\
&= \mathbb{E}_{z\sim h_\phi(z\,|\,x)}\mathbb{E}_{\psi\sim q_\phi(\psi\,|\,z,x)} \log \frac{q_\phi(\psi\,|\,z,x)}{p_\theta(\psi\,|\,z,x)} \\
&= \mathbb{E}_{z\sim h_\phi(z\,|\,x)}\text{KL}(q_\phi(\psi\,|\,z,x)||p_\theta(\psi\,|\,z,x)) \\
&\geq 0
\end{aligned}
$$

  - SIVI allows $q_\phi(\psi\,|\,x)$ to be implicit
  - SIVI sandwiches $\text{ELBO}_{\text{SIVI}}$ between a lower bound and an upper bound, and uses an asymptotically exact surrogate ELBO for optimization

Expressiveness of SIVI

$h(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\psi} \sim q(\boldsymbol{\psi})} q(\boldsymbol{z} \mid \boldsymbol{\psi})$	$p(\boldsymbol{z})$
$z \sim \mathcal{N}(\psi, 0.1)$, $\psi \sim q(\psi)$	$\text{Laplace}(z; \mu = 0, b = 2)$
	$0.3\mathcal{N}(z; -2, 1) + 0.7\mathcal{N}(z; 2, 1)$
$z \sim \text{Log-Normal}(\psi, 0.1)$, $\psi \sim q(\psi)$	$\text{Gamma}(z; 2, 1)$
$\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{\psi}, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\right)$, $\boldsymbol{\psi} \sim q(\boldsymbol{\psi})$	$0.5\mathcal{N}(\boldsymbol{z}; -2, I) + 0.5\mathcal{N}(\boldsymbol{z}; 2, I)$
	$\mathcal{N}(z_1; z_2^2/4, 1)\mathcal{N}(z_2; 0, 4)$
	$0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & 1.8 \\ 1.8 & 2 \end{bmatrix}\right) + 0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & -1.8 \\ -1.8 & 2 \end{bmatrix}\right)$



Expressiveness of SIVI

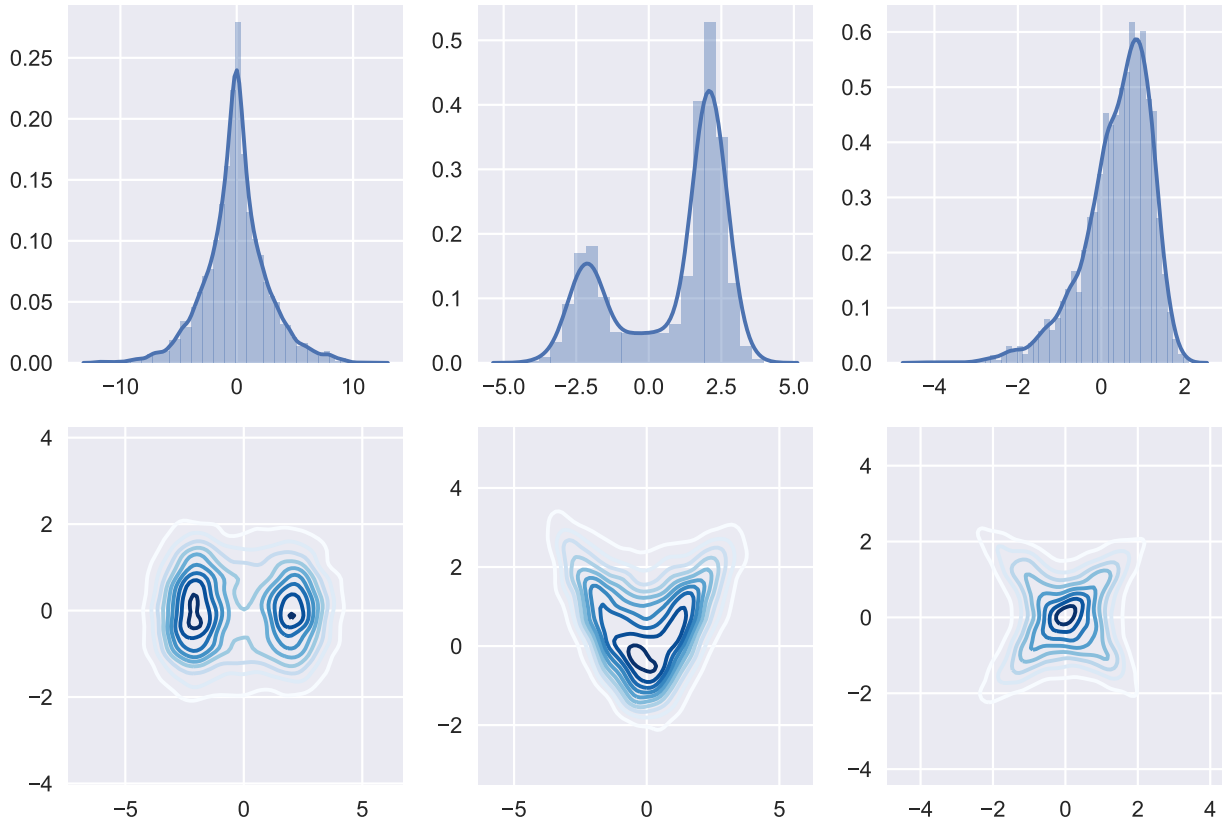


Figure: Visualization of the MLP based implicit distributions $\psi \sim q(\psi)$, which are mixed with isotropic Gaussian (or Log-Normal) distributions to approximate the target distributions.

Model:

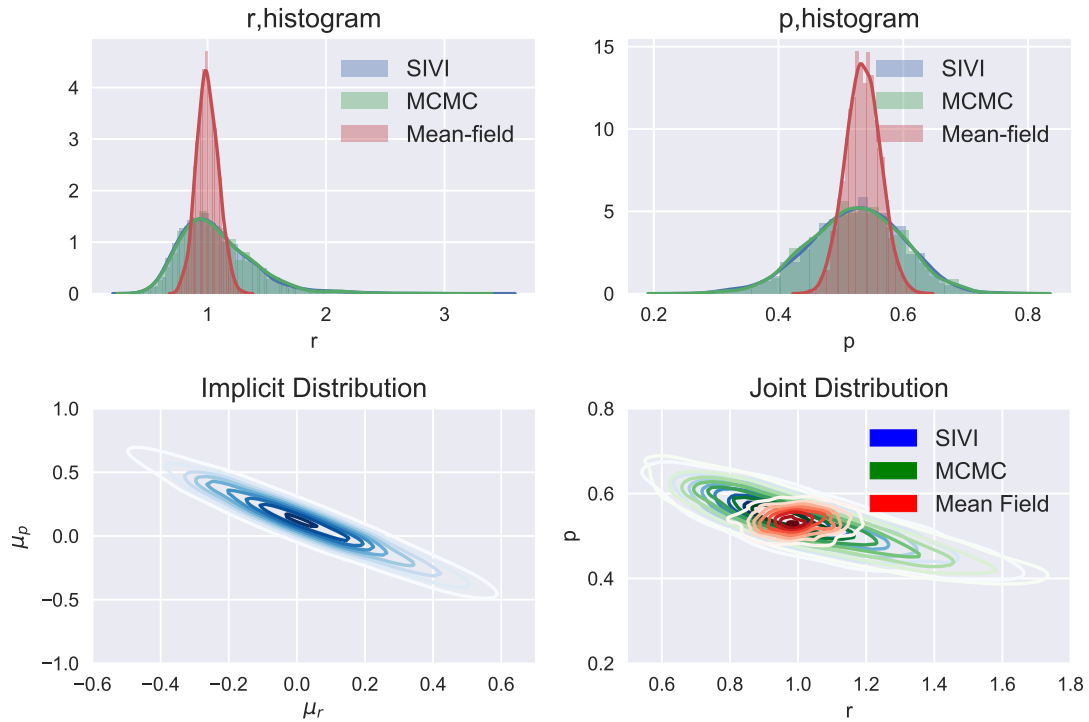
$$x_i \overset{i.i.d.}{\sim} \text{NB}(r, p), \;\; r \sim \text{Gamma}(a, 1/b), \;\; p \sim \text{Beta}(\alpha, \beta),$$

Mean-filed VI:

$$Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}),$$

SIVI (both the conditional and mixing q distributions are reparameterizable) :

$$q(r, p \,|\, \psi) = \text{Log-Normal}(r; \mu_r, \sigma_0^2)\text{Logit-Normal}(p; \mu_p, \sigma_0^2),$$
$$\psi = (\mu_r, \mu_p) \sim q(\psi),$$



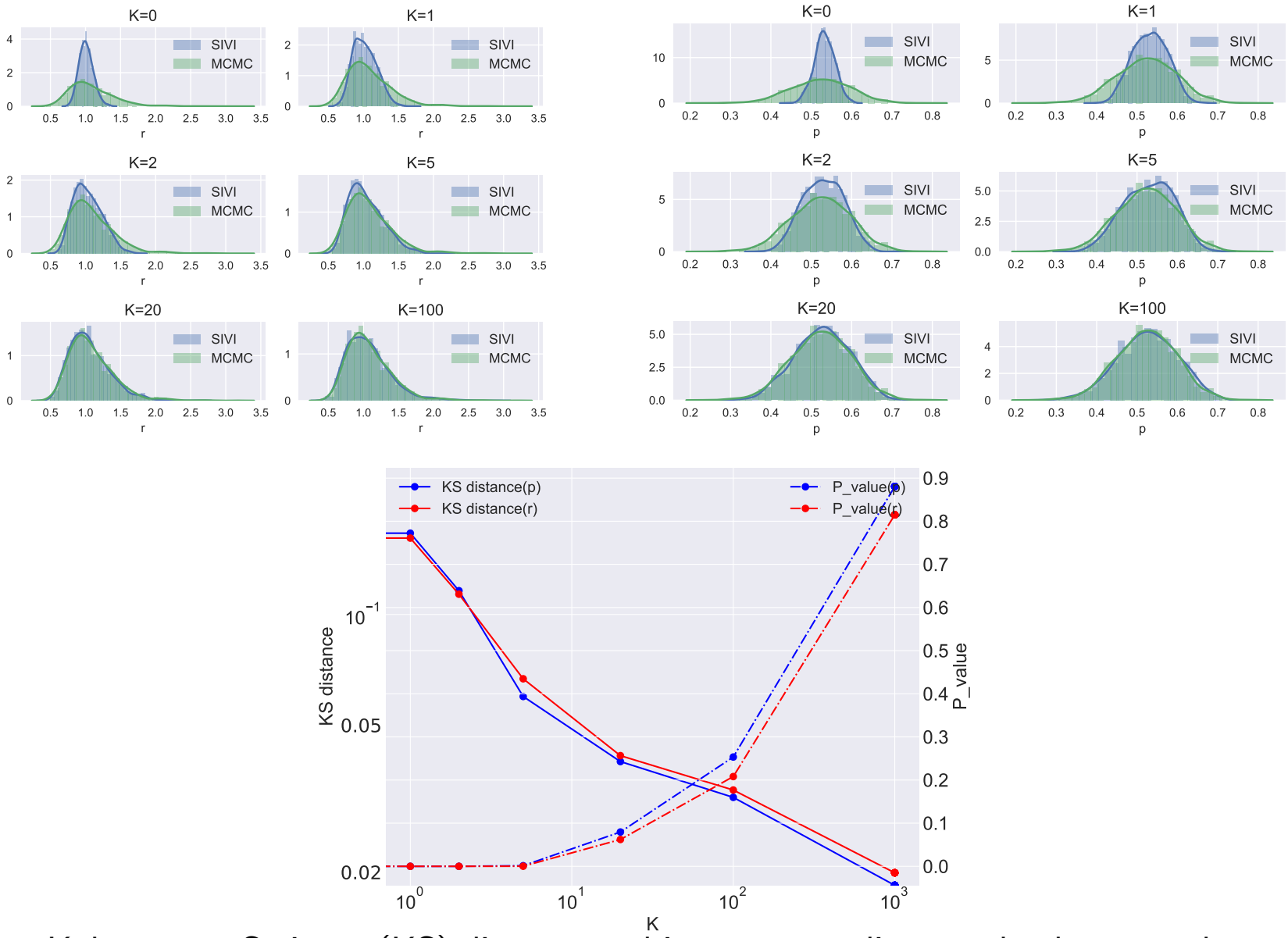


Figure: Kolmogorov-Smirnov (KS) distance and its corresponding p-value between the marginal posteriors of r and p inferred by SIVI and MCMC. SIVI rapidly improves as K increases.

Score function gradient for conjugate model

If $q(\boldsymbol{z} \mid \boldsymbol{\psi})$ is not reparameterizable, then we introduce a density ratio as

$$r_{\boldsymbol{\xi},\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}^{(1:K)}) = \frac{q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})))}{\frac{1}{K+1}[q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})) + \sum_{k=1}^{K} q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(k)}))]}$$

and approximate the gradient of $\underline{\mathcal{L}}_K$ with respect to $\boldsymbol{\phi}$ as

$$\begin{aligned}\nabla_{\boldsymbol{\phi}} \underline{\mathcal{L}}_K \approx \frac{1}{J} \sum_{j=1}^{J} \Big\{ & - \nabla_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{z} \sim q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j))} \log \frac{q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j))}{p(\boldsymbol{x}, \boldsymbol{z})} \\ & + \nabla_{\boldsymbol{\phi}} \log r_{\boldsymbol{\xi},\boldsymbol{\phi}}(\boldsymbol{z}_j, \boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}^{(1:K)}) \\ & + [\nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\xi}}(\boldsymbol{z}_j \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j))] \log r_{\boldsymbol{\xi},\boldsymbol{\phi}}(\boldsymbol{z}_j, \boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}^{(1:K)}) \Big\},\end{aligned}$$

- The first summation term is equivalent to the gradient of MFVI's ELBO
- Both the second and third terms correct the restrictions of $q_{\boldsymbol{\xi}}(\boldsymbol{z} \mid T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_j))$
- $\log r_{\boldsymbol{\xi},\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{\epsilon}, \boldsymbol{\epsilon}^{(1:K)})$ in the third term is expected to be small regardless of convergence, effectively mitigating the variance of score function gradient estimation that is usually high in basic black-box VI

Model:

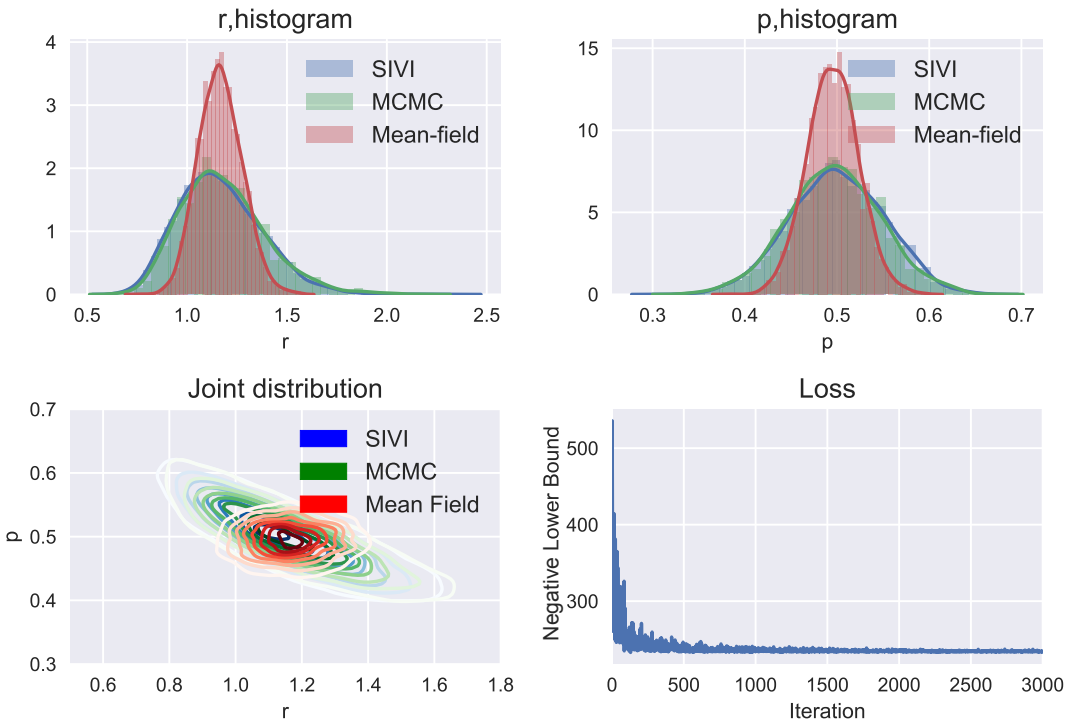
$$p(n_i, l_i \mid r, p) = r^{l_i} p^{n_i} (1-p)^r / Z_i, \;\; r \sim \text{Gamma}(a, 1/b), \;\; p \sim \text{Beta}(\alpha, \beta)$$

Mean-filed VI:

$$Q(r, p) = q(r)q(p) = \text{Gamma}(r; \tilde{a}, \tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta}),$$

SIVI (non-reparameterizable conditional q distribution but conjugate model):

$$q(r, p \mid \boldsymbol{\psi}) = \text{Gamma}(r; \psi_1, \psi_2)\text{Beta}(p; \psi_3, \psi_4), \;\; \boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3, \psi_4) \sim q(\boldsymbol{\psi})$$

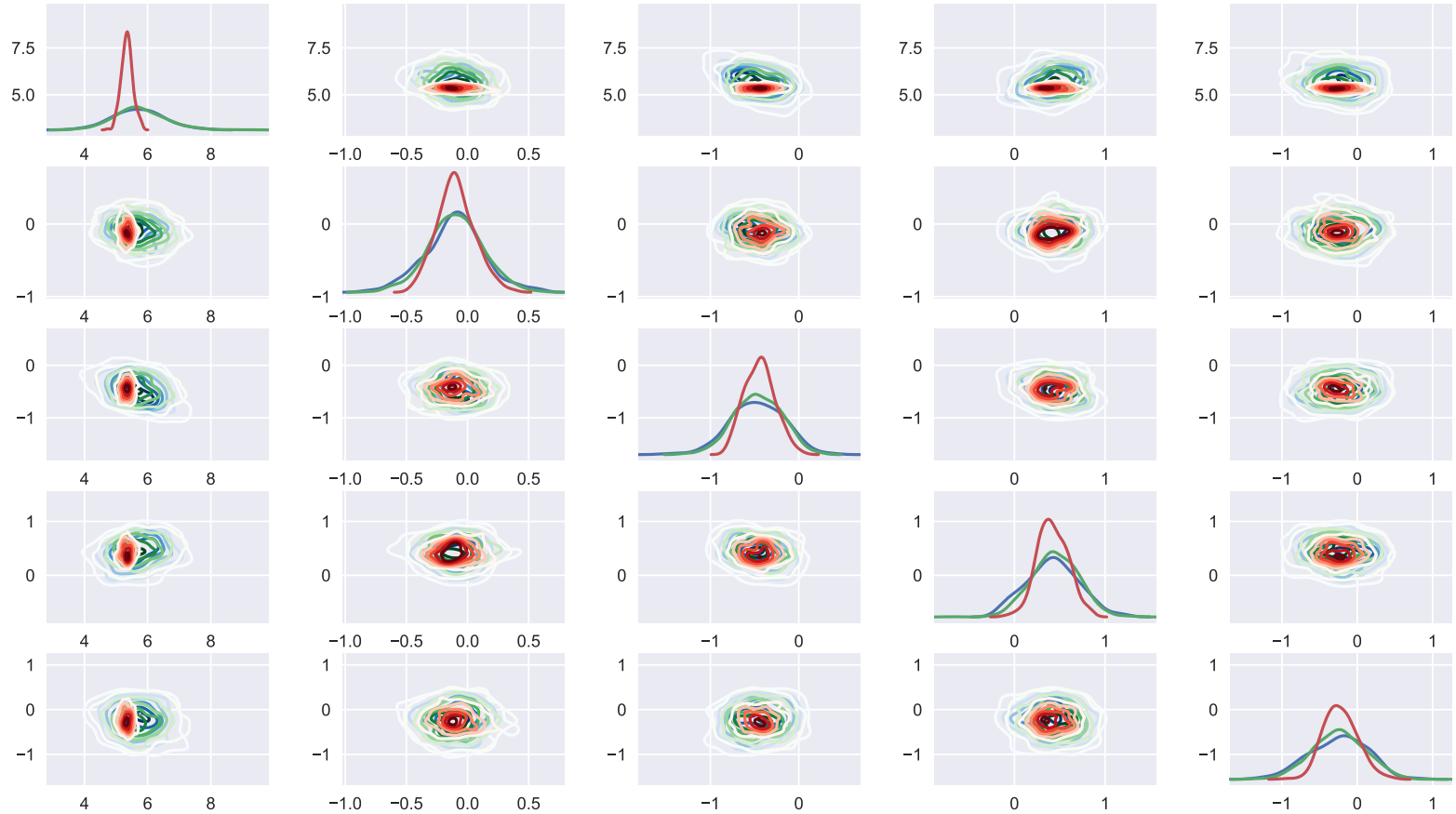


Bayesian logistic regression (pairwise joint distributions)

$$y_i \sim \text{Bernoulli}[(1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}})^{-1}], \quad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_{V+1})$$

SIVI: $q(\boldsymbol{\beta}\,|\,\boldsymbol{\psi}) = \mathcal{N}(\boldsymbol{\psi}, \boldsymbol{\Sigma}),\ \boldsymbol{\psi} \sim q_{\boldsymbol{\phi}}(\boldsymbol{\psi})$

(Blue: MCMC, Red: VI, Green: SIVI):



Bayesian logistic regression (univariate marginals)

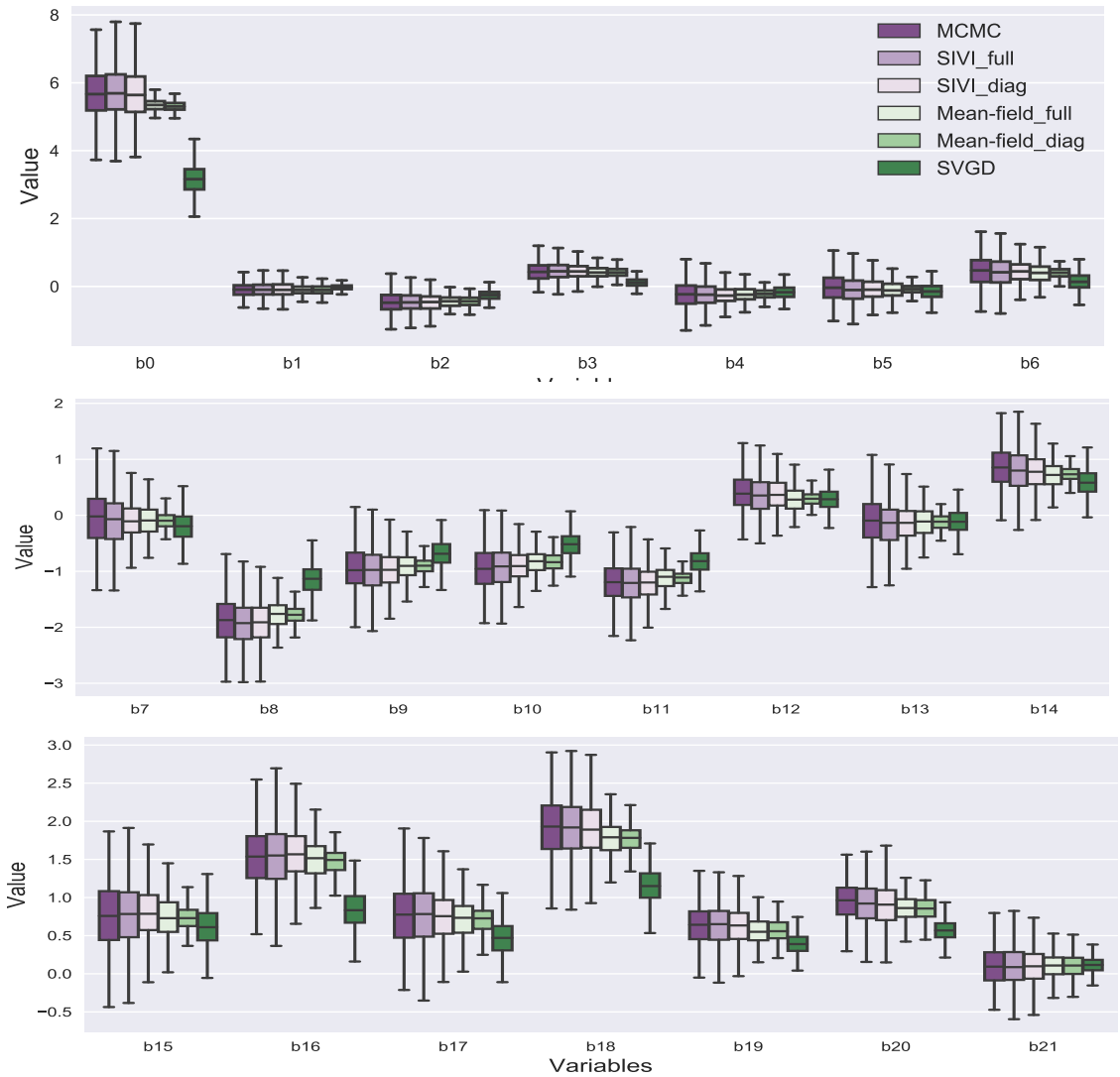


Figure: Comparison of all marginal posteriors of β_v inferred by various methods for Bayesian logistic regression on *waveform*.

Bayesian logistic regression (correlation coefficients)

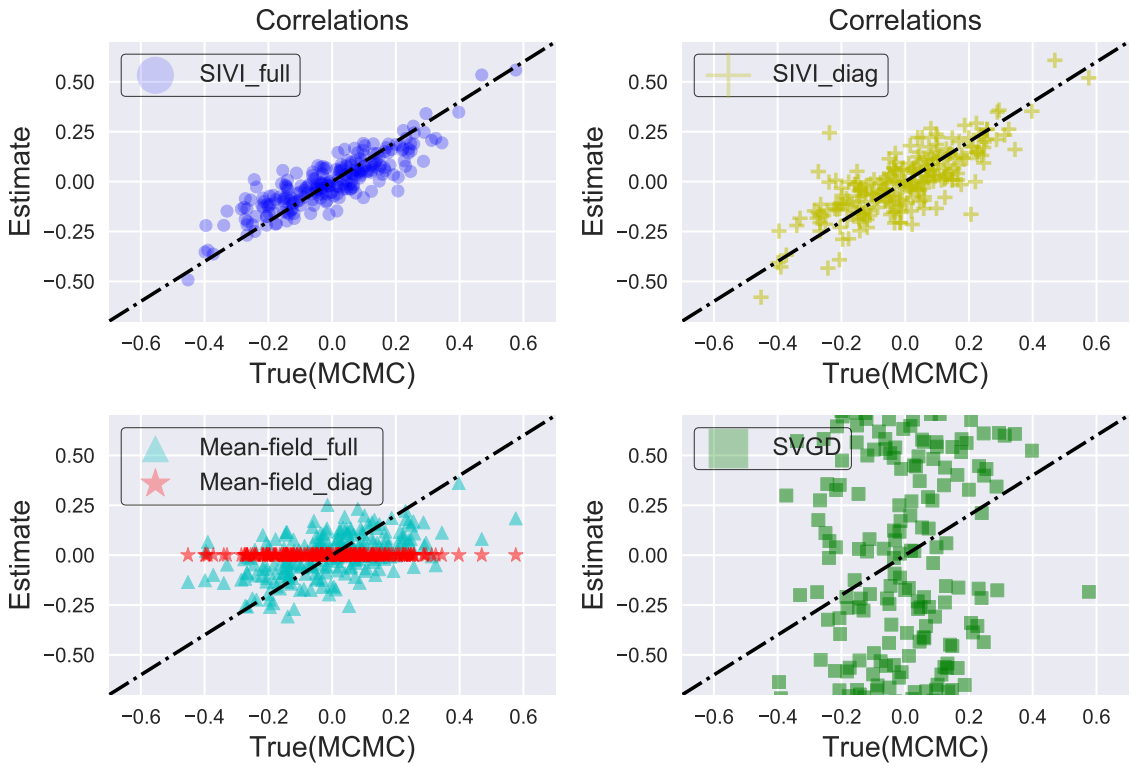


Figure: Against the correlation coefficients of $\boldsymbol{\beta}$ estimated from the posterior samples $\{\boldsymbol{\beta}_i\}_{i=1:1000}$ of MCMC on *waveform*, top left/right plots the correlation coefficients of SIVI with a full/diagonal covariance matrix, bottom left plots these of MFVI with a full/diagonal covariance matrix, and bottom right plots these of SVGD. The closer to the dashed line the better.

Bayesian logistic regression (predictive uncertainty)

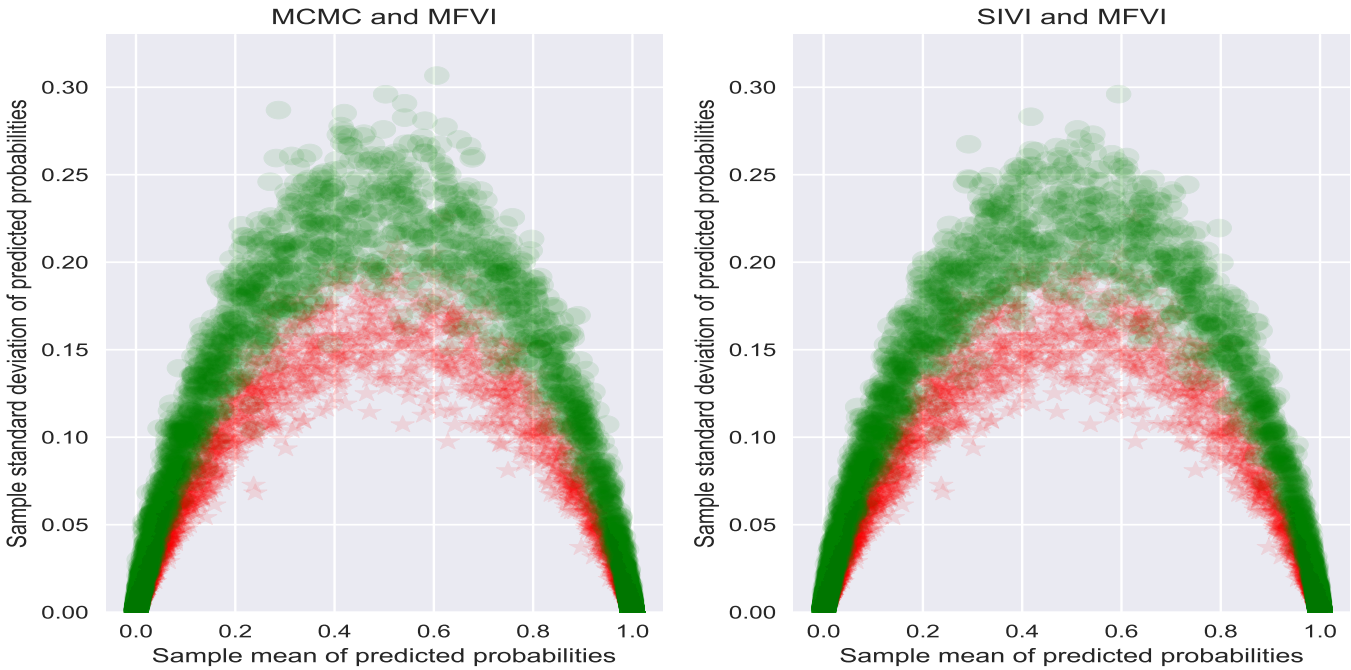


Figure: Comparison of MFVI (red) with a full covariance matrix, MCMC (green on left), and SIVI (green on right) with a full covariance matrix on quantifying predictive uncertainty for Bayesian logistic regression on *waveform*

Variational autoencoder

Variational Autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) is a popular generative model based approach for unsupervised feature learning and amortized inference.

- VAE iteratively infers the encoder parameter ϕ and decoder parameter θ to maximize the ELBO as

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{z \sim q_\phi(z \mid x)}[\log(p_\theta(x \mid z))] - \mathrm{KL}(q_\phi(z \mid x) || p(z)).$$

- The encoder distribution $q_\phi(z \mid x)$ is required to be reparameterizable and simple to compute its PDF, which usually restricts it to a small family of exponential distributions. A canonical form of the encoder is

$$q_\phi(z \mid x) = \mathcal{N}(z \mid \mu(x; \phi), \Sigma(x; \phi)),$$

  where the Gaussian distribution parameters are deterministically transformed from the observed data x, via non-probabilistic deep neural networks parameterized by ϕ.

- Thus, given observation x_i, its corresponding code z_i is forced to follow a Gaussian distribution, no matter how powerful the deep neural networks are. The Gaussian assumption, however, is often too restrictive to model skewed, heavy-tailed, and/or multi-modal distributions.

Semi-implicit variational autoencoder

We construct semi-implicit VAE (SIVAE) by using a hierarchical encoder that injects random noise at M different stochastic layers as

$$\begin{aligned}
&\boldsymbol{\ell}_t = T_t(\boldsymbol{\ell}_{t-1}, \boldsymbol{\epsilon}_t, \boldsymbol{x}; \boldsymbol{\phi}),\ \boldsymbol{\epsilon}_t \sim q_t(\boldsymbol{\epsilon}),\ t = 1, \ldots, M,\\
&\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\phi}) = f(\boldsymbol{\ell}_M, \boldsymbol{x}; \boldsymbol{\phi}),\quad \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{\phi}) = g(\boldsymbol{\ell}_M, \boldsymbol{x}; \boldsymbol{\phi}),\\
&q_{\boldsymbol{\phi}}(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{\phi}), \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{\phi})),
\end{aligned}$$

where $\boldsymbol{\ell}_0 = \emptyset$ and T_t, f, and g are all deterministic neural networks. Note given data $\boldsymbol{x}_i$, $\boldsymbol{\mu}(\boldsymbol{x}_i, \boldsymbol{\phi})$, $\boldsymbol{\Sigma}(\boldsymbol{x}_i, \boldsymbol{\phi})$ are now random variables rather than following vanilla VAE to assume deterministic values. This clearly moves the encoder variational distribution beyond a simple Gaussian form.

Semi-implicit variational autoencoder

Methods	$-\log p(\boldsymbol{x})$
Results below form Burda et al. (2015)	
VAE + IWAE	$= 86.76$
IWAE + IWAE	$= 84.78$
Results below form Salimans et al. (2015)	
DLGM + HVI (1 leapfrog step)	$= 88.08$
DLGM + HVI (4 leapfrog step)	$= 86.40$
DLGM + HVI (8 leapfrog steps)	$= 85.51$
Results below form Rezende & Mohamed (2015)	
DLGM+NICE (Dinh et al., 2014) (k = 80)	≤ 87.2
DLGM+NF (k = 40)	≤ 85.7
DLGM+NF (k = 80)	≤ 85.1
Results below form Gregor et al. (2015)	
DLGM	≈ 86.60
NADE	$= 88.33$
DBM 2hl	≈ 84.62
DBN 2hl	≈ 84.55
EoNADE-5 2hl (128 orderings)	$= 84.68$
DARN 1hl	≈ 84.13
Results below form Maaløe et al. (2016)	
Auxiliary VAE (L=1, IW=1)	≤ 84.59
Results below form Mescheder et al. (2017)	
VAE + IAF (Kingma et al., 2016)	$\approx 84.9 \pm 0.3$
Auxiliary VAE (Maaløe et al., 2016)	$\approx 83.8 \pm 0.3$
AVB + AC	$\approx 83.7 \pm 0.3$
SIVI (3 stochastic layers)	$= 84.07$
SIVI (3 stochastic layers)+ IW($\tilde{K} = 10$)	$= 83.25$

Summary for SIVI

- Uncertainty estimation is difficult but important in Variational Inference
- A key to achieve an accurate uncertainty estimation is to construct a flexible variational distribution that can capture the dependencies between latent variables
- Combining the advantages of having analytic point-wise evaluable density ratios and tractable computation via Monte Carlo estimation, semi-implicit variational inference (SIVI) can approach the accuracy of MCMC in quantifying posterior uncertainty, but often pays a lower computational cost and can generate independent posterior samples on the fly via the inferred stochastic variational inference network

Additional work

- Variational sampling
  - MCMC is able to accurately capture posterior uncertainty, and still hard to be replaced by variational inference in many challenging statistical inference problems
  - For a latent variable whose prior is not conjugate to the likelihood, the corresponding MCMC transition kernel may be difficult to design
  - Our idea is to use variational inference + deep neural network to learn MCMC transition kernels
- Augment-REINFORCE-merge gradient for discrete latent variable models
  - Backpropagate unbiased and low-variance gradient for discrete latent variables