
Semi-Implicit Variational Inference

Mingzhang Yin¹ Mingyuan Zhou²

Abstract

Semi-implicit variational inference (SIVI) is introduced to expand the commonly used analytic variational distribution family, by mixing the variational parameter with a flexible distribution. This mixing distribution can assume any density function, explicit or not, as long as independent random samples can be generated via reparameterization. Not only does SIVI expand the variational family to incorporate highly flexible variational distributions, including implicit ones that have no analytic density functions, but also sandwiches the evidence lower bound (ELBO) between a lower bound and an upper bound, and further derives an asymptotically exact surrogate ELBO that is amenable to optimization via stochastic gradient ascent. With a substantially expanded variational family and a novel optimization algorithm, SIVI is shown to closely match the accuracy of MCMC in inferring the posterior in a variety of Bayesian inference tasks.

1. Introduction

Variational inference (VI) is an optimization based method that is widely used for approximate Bayesian inference. It introduces variational distribution Q over the latent variables to approximate the posterior (Jordan et al., 1999), and its stochastic version is scalable to big data (Hoffman et al., 2013). VI updates the parameters of Q to move it closer to the posterior in each iteration, where the closeness is in general measured by the Kullback–Leibler (KL) divergence from the posterior to Q , minimizing which is shown to be the same as maximizing the evidence lower bound (ELBO) (Jordan et al., 1999). To make it simple to climb the ELBO to a local optimum, one often takes the

mean-field assumption that Q is factorized over the latent variables. The optimization problem is further simplified if each latent variable’s distribution is in the same exponential family as its prior, which allows exploiting conditional conjugacy to derive closed-form coordinate-ascent update equations (Bishop & Tipping, 2000; Blei et al., 2017).

Despite its popularity, VI has a well-known issue in underestimating the variance of the posterior, which is often attributed to the mismatch between the representation power of the variational family that Q is restricted to and the complexity of the posterior, and the use of KL divergence, which is asymmetric, to measure how different Q is from the posterior. This issue is often further amplified in mean-field VI (MFVI), due to the factorized assumption on Q that ignores the dependencies between different factorization components (Wainwright et al., 2008; Blei et al., 2017). While there exists a variety of methods that add some structure back to Q to partially restore the dependencies (Saul & Jordan, 1996; Jaakkola & Jordan, 1998; Hoffman & Blei, 2015; Giordano et al., 2015; Tran et al., 2015; 2016; Han et al., 2016; Ranganath et al., 2016; Maaløe et al., 2016; Gregor et al., 2015), it is still necessary for Q to have an analytic probability density function (PDF).

To further expand the variational family that Q belongs to, there has been significant recent interest in defining Q with an implicit model, which makes the PDF of Q become intractable (Huszár, 2017; Mohamed & Lakshminarayanan, 2016; Tran et al., 2017; Li & Turner, 2017; Mescheder et al., 2017; Shi et al., 2017). While using an implicit model could make Q more flexible, it makes it no longer possible to directly computing the log density ratio, as required for evaluating the ELBO. Thus, one often resorts to density ratio estimation, which, however, not only adds an additional level of complexity into each iteration of the optimization, but also is known to be a very difficult problem, especially in high-dimensional settings (Sugiyama et al., 2012).

To well characterize the posterior while maintaining simple optimization, we introduce semi-implicit VI (SIVI) that imposes a mixing distribution on the parameters of the original Q to expand the variational family with a semi-implicit hierarchical construction. The meaning of “semi-implicit” is twofold: 1) the original Q distribution is required to have an analytic PDF, but its mixing distribution is not subject to

¹Department of Statistics and Data Sciences, ²Department of IROM, McCombs School of Business, The University of Texas at Austin, Austin TX 78712, USA. Correspondence to: Mingzhang Yin <mzyin@utexas.edu>, Mingyuan Zhou <mingyuan.zhou@mcombs.utexas.edu>.

such a constraint; and 2) even if both the original Q and its mixing distribution have analytic PDFs, it is common that the marginal of the hierarchy is implicit, that is, having a non-analytic PDF. Our intuition behind SIVI is that even if this marginal is not tractable, its density can be evaluated with Monte Carlo estimation under the semi-implicit hierarchical construction, an expansion that helps model skewness, kurtosis, multimodality, and other characteristics that are exhibited by the posterior but failed to be captured by the original variational family. For MFVI, a benefit of this expansion is restoring the dependencies between its factorization components, as the resulted Q distribution becomes conditionally independent but marginally dependent.

SIVI makes three major contributions: 1) a reparameterizable implicit distribution can be used as a mixing distribution to effectively expand the richness of the variational family; 2) an analytic conditional Q distribution is used to sidestep the hard problem of density ratio estimation, and is not required to be reparameterizable in conditionally conjugate models; and 3) SIVI sandwiches the ELBO between a lower bound and an upper bound, and derives an asymptotically exact surrogate ELBO that is amenable to direct optimization via stochastic gradient ascent. With a flexible variational family and novel optimization, SIVI bridges the accuracy gap of posterior estimation between VI and Markov chain Monte Carlo (MCMC), which can accurately characterize the posterior using MCMC samples, as will be demonstrated in a variety of Bayesian inference tasks. Code is provided at <https://github.com/mingzhang-yin/SIVI>

2. Semi-Implicit Variational Inference

In VI, given observations \mathbf{x} , latent variables \mathbf{z} , model likelihood $p(\mathbf{x} | \mathbf{z})$, and prior $p(\mathbf{z})$, we approximate the posterior $p(\mathbf{z} | \mathbf{x})$ with variational distribution $q(\mathbf{z} | \boldsymbol{\psi})$ that is often required to be explicit. We optimize the variational parameter $\boldsymbol{\psi}$ to minimize $\text{KL}(q(\mathbf{z} | \boldsymbol{\psi}) || p(\mathbf{z} | \mathbf{x}))$, the KL divergence from $p(\mathbf{z} | \mathbf{x})$ to $q(\mathbf{z} | \boldsymbol{\psi})$. Since one may show $\log p(\mathbf{x}) = \text{ELBO} + \text{KL}(q(\mathbf{z} | \boldsymbol{\psi}) || p(\mathbf{z} | \mathbf{x}))$, where

$$\text{ELBO} = -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \boldsymbol{\psi})} [\log q(\mathbf{z} | \boldsymbol{\psi}) - \log p(\mathbf{x}, \mathbf{z})], \quad (1)$$

minimizing $\text{KL}(q(\mathbf{z} | \boldsymbol{\psi}) || p(\mathbf{z} | \mathbf{x}))$ is equivalent to maximizing the ELBO (Bishop & Tipping, 2000; Blei et al., 2017). Rather than treating $\boldsymbol{\psi}$ as the variational parameter to be inferred, SIVI regards $\boldsymbol{\psi} \sim q(\boldsymbol{\psi})$ as a random variable, as described below. Note that when $q(\boldsymbol{\psi})$ degenerates to a point mass density, SIVI reduces to vanilla VI.

2.1. Semi-Implicit Variational Family

Assuming $\boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi})$, where ϕ denotes the distribution parameter to be inferred, the semi-implicit variational distri-

bution for \mathbf{z} can be defined in a hierarchical manner as

$$\mathbf{z} \sim q(\mathbf{z} | \boldsymbol{\psi}), \quad \boldsymbol{\psi} \sim q_\phi(\boldsymbol{\psi}).$$

Marginalizing the intermediate variable $\boldsymbol{\psi}$ out, we can view \mathbf{z} as a random variable drawn from distribution family \mathcal{H} indexed by variational parameter ϕ , expressed as

$$\mathcal{H} = \{h_\phi(\mathbf{z}) : h_\phi(\mathbf{z}) = \int_{\boldsymbol{\psi}} q(\mathbf{z} | \boldsymbol{\psi}) q_\phi(\boldsymbol{\psi}) d\boldsymbol{\psi}\}.$$

Note $q(\mathbf{z} | \boldsymbol{\psi})$ is required to be explicit, but the mixing distribution $q_\phi(\boldsymbol{\psi})$ is allowed to be implicit. Moreover, unless $q_\phi(\boldsymbol{\psi})$ is conjugate to $q(\mathbf{z} | \boldsymbol{\psi})$, the marginal Q distribution $h_\phi(\mathbf{z}) \in \mathcal{H}$ is often implicit. These are the two reasons for referring to the proposed VI as semi-implicit VI (SIVI).

SIVI requires $q(\mathbf{z} | \boldsymbol{\psi})$ to be explicit, and also requires it to either be reparameterizable, which means $\mathbf{z} \sim q(\mathbf{z} | \boldsymbol{\psi})$ can be generated by transforming random noise $\boldsymbol{\varepsilon}$ via function $f(\boldsymbol{\varepsilon}, \boldsymbol{\psi})$, or allow the ELBO in (1) to be analytic. Whereas the mixing distribution $q(\boldsymbol{\psi})$ is required to be reparameterizable but not necessarily explicit. In particular, SIVI draws from $q(\boldsymbol{\psi})$ by transforming random noise $\boldsymbol{\varepsilon}$ via a deep neural network, which generally leads to an implicit distribution for $q(\boldsymbol{\psi})$ due to a non-invertible transform.

SIVI is related to the hierarchical variational model (Ranganath et al., 2016; Maaløe et al., 2016; Agakov & Barber, 2004) in using a hierarchical variational distribution, but, as discussed below, differs from it in allowing an implicit mixing distribution $q_\phi(\boldsymbol{\psi})$ and optimizing the variational parameter via an asymptotically exact surrogate ELBO. Note as long as $q_\phi(\boldsymbol{\psi})$ can degenerate to delta function $\delta_{\boldsymbol{\psi}_0}(\boldsymbol{\psi})$ for arbitrary $\boldsymbol{\psi}_0$, the semi-implicit variational family \mathcal{H} is a strict expansion of the original $\mathcal{Q} = \{q(\mathbf{z} | \boldsymbol{\psi}_0)\}$ family, that is, $\mathcal{Q} \subseteq \mathcal{H}$. For MFVI that assumes $q(\mathbf{z} | \boldsymbol{\psi}) = \prod_m q(z_m | \boldsymbol{\psi}_m)$, this expansion significantly helps restore the dependencies between z_m if $\boldsymbol{\psi}_m$ are not imposed to be independent between each other.

2.2. Implicit Mixing Distribution

While restricting $q(\mathbf{z} | \boldsymbol{\psi})$ to be explicit, SIVI introduces a mixing distribution $q_\phi(\boldsymbol{\psi})$ to enhance its representation power. In this paper, we construct $q_\phi(\boldsymbol{\psi})$ with an implicit distribution that generates its random samples via a stochastic procedure but may not allow a pointwise evaluable PDF. More specifically, an implicit distribution (Mohamed & Lakshminarayanan, 2016; Tran et al., 2017), consisting of a source of randomness $q(\boldsymbol{\varepsilon})$ for $\boldsymbol{\varepsilon} \in \mathbb{R}^g$ and a deterministic transform $T_\phi : \mathbb{R}^g \rightarrow \mathbb{R}^d$, can be constructed as $\boldsymbol{\psi} = T_\phi(\boldsymbol{\varepsilon})$, $\boldsymbol{\varepsilon} \sim q(\boldsymbol{\varepsilon})$, with PDF

$$q_\phi(\boldsymbol{\psi}) = \frac{\partial}{\partial \psi_1} \cdots \frac{\partial}{\partial \psi_d} \int_{T_\phi(\boldsymbol{\varepsilon}) \leq \boldsymbol{\psi}} q(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon}. \quad (2)$$

When T_ϕ is invertible and the integration is tractable, the PDF of $\boldsymbol{\psi}$ can be calculated with (2), but this is not the case

in general and hence $q_\phi(\psi)$ is often implicit. When $T_\phi(\cdot)$ is chosen as a deep neural network, thanks to its high modeling capacity, $q_\phi(\psi)$ can be highly flexible and the dependencies between the elements of ψ can be well captured.

Prevalently used in the study of thermodynamics, ecology, epidemiology, and differential equation systems, implicit distributions have only been recently introduced in VI to parameterize $q(z|\psi)$ (Li & Turner, 2017; Mescheder et al., 2017; Huszár, 2017; Tran et al., 2017). Using implicit distributions with intractable PDF increases flexibility but substantially complicates the optimization problem for VI, due to the need to estimate log density ratios involving intractable PDFs, which is particularly challenging in high dimensions (Sugiyama et al., 2012). By contrast, taking a semi-implicit construction, SIVI offers the best of both worlds: constructing a highly flexible variational distribution, without sacrificing the key benefit of VI in converting posterior inference into an optimization problem that is simple to solve. Below we develop a novel optimization algorithm that exploits SIVI’s semi-implicit construction.

3. Optimization for SIVI

To optimize the variational parameters of SIVI, below we first derive for the ELBO a lower bound, climbing which, however, could drive the mixing distribution $q_\phi(\psi)$ towards a point mass density. To prevent degeneracy, we add a non-negative regularization term, leading to a surrogate ELBO that is asymptotically exact, as can be further tightened by importance reweighting. To sandwich the ELBO, we also derive for the ELBO an upper bound, optimizing which, however, may lead to divergence. We further show that this upper bound can be corrected to a tighter upper bound that monotonically converges from above towards the ELBO.

3.1. Lower Bound of ELBO

Theorem 1 (Cover & Thomas (2012)). *The KL divergence from distribution $p(z)$ to distribution $q(z)$, expressed as $\text{KL}(q(z)||p(z))$, is convex in the pair $(q(z), p(z))$.*

Fixing the distribution $p(z)$ in Theorem 1, KL divergence can be viewed as a convex functional in $q(z)$. As in Appendix A, with Theorem 1 and Jensen’s inequality, we have

$$\text{KL}(\mathbb{E}_{\psi} q(z|\psi)||p(z)) \leq \mathbb{E}_{\psi} \text{KL}(q(z|\psi)||p(z)). \quad (3)$$

Thus, using $h_\phi(z) = \mathbb{E}_{\psi \sim q_\phi(\psi)} q(z|\psi)$ as the variational distribution, SIVI has a lower bound for its ELBO as

$$\begin{aligned} \underline{\mathcal{L}}(q(z|\psi), q_\phi(\psi)) &= \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{z \sim q(z|\psi)} \log \frac{p(\mathbf{x}, z)}{q(z|\psi)} \\ &= -\mathbb{E}_{\psi \sim q_\phi(\psi)} \text{KL}(q(z|\psi)||p(z|\mathbf{x})) + \log p(\mathbf{x}) \\ &\leq -\text{KL}(\mathbb{E}_{\psi \sim q_\phi(\psi)} q(z|\psi)||p(z|\mathbf{x})) + \log p(\mathbf{x}) \\ &= \mathcal{L} = \mathbb{E}_{z \sim h_\phi(z)} \log \frac{p(\mathbf{x}, z)}{h_\phi(z)}. \end{aligned} \quad (4)$$

The PDF of $h_\phi(z)$ is often intractable, especially if $q_\phi(\psi)$ is implicit, prohibiting a Monte Carlo estimation of the ELBO \mathcal{L} . By contrast, a Monte Carlo estimation of $\underline{\mathcal{L}}$ only requires $q(z|\psi)$ to have an analytic PDF and $q_\phi(\psi)$ to be convenient to sample from. It is this separation of evaluation and sampling that allows SIVI to combine an explicit $q(z|\psi)$ with an implicit $q_\phi(\psi)$ that is as powerful as needed, while maintaining tractable computation.

3.2. Degeneracy and Regularization

A direct optimization of the lower bound $\underline{\mathcal{L}}$ in (4), however, can suffer from degeneracy, as shown in the proposition below. All proofs are deferred to Appendix A.

Proposition 1 (Degeneracy). *Let us denote $\psi^* = \arg \max_{\psi} -\mathbb{E}_{z \sim q(z|\psi)} \log \frac{q(z|\psi)}{p(\mathbf{x}, z)}$, then*

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

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

Therefore, if optimizing the variational parameter by climbing $\underline{\mathcal{L}}(q(z|\psi), q_\phi(\psi))$, without stopping the optimization algorithm early, $q_\phi(\psi)$ could converge to a point mass density, making SIVI degenerate to vanilla VI.

To prevent degeneracy, we regularize $\underline{\mathcal{L}}$ by adding

$$B_K = \mathbb{E}_{\psi, \psi^{(1)}, \dots, \psi^{(K)} \sim q_\phi(\psi)} \text{KL}(q(z|\psi)||\tilde{h}_K(z)), \quad (5)$$

$$\text{where } \tilde{h}_K(z) = \frac{q(z|\psi) + \sum_{k=1}^K q(z|\psi^{(k)})}{K+1}.$$

Note that $B_K \geq 0$, with $B_K = 0$ if and only if $K = 0$ or $q_\phi(\psi)$ degenerates to a point mass density. Therefore, $\underline{\mathcal{L}}_0 = \underline{\mathcal{L}}$ and maximizing $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$ with $K \geq 1$ would encourage positive B_K and drive $q(\psi)$ away from degeneracy. Moreover, as $\lim_{K \rightarrow \infty} \tilde{h}_K(z) = \mathbb{E}_{\psi \sim q_\phi(\psi)} q(z|\psi) = h_\phi(z)$ by the strong law of large numbers and

$$\lim_{K \rightarrow \infty} B_K = \mathbb{E}_{\psi \sim q_\phi(\psi)} \text{KL}(q(z|\psi)||h_\phi(z)) \quad (6)$$

by interchanging two limiting operations, as discussed in detail in Appendix A, we have the following proposition.

Proposition 2. *Suppose \mathcal{L} and $\underline{\mathcal{L}}$ are defined as in (4) and B_K as in (5), 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 $\lim_{K \rightarrow \infty} \underline{\mathcal{L}}_K = \mathcal{L}$.*

3.3. Upper Bound of ELBO and Correction

Using the concavity of the logarithmic function, we have $\log h_\phi(z) \geq \mathbb{E}_{\psi \sim q_\phi(\psi)} \log q(z|\psi)$, and hence we can obtain an upper bound of SIVI’s ELBO as

$$\begin{aligned} \bar{\mathcal{L}}(q(z|\psi), q_\phi(\psi)) &= \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{z \sim h_\phi(z)} \log \frac{p(\mathbf{x}, z)}{q(z|\psi)} \\ &\geq \mathcal{L} = \mathbb{E}_{z \sim h_\phi(z)} \log \frac{p(\mathbf{x}, z)}{h_\phi(z)}. \end{aligned} \quad (7)$$

Comparing (4) and (7) shows that the lower bound $\underline{\mathcal{L}}$ and upper bound $\bar{\mathcal{L}}$ only differ from each other in whether the expectation involving \mathbf{z} is taken with respect to $q(\mathbf{z} | \psi)$ or $h_\phi(\mathbf{z})$. Moreover, one may show that $\bar{\mathcal{L}} - \underline{\mathcal{L}}$ is equal to

$$\mathbb{E}_{\psi \sim q(\psi)} [\text{KL}(q(\mathbf{z} | \psi) || h_\phi(\mathbf{z})) + \text{KL}(h_\phi(\mathbf{z}) || q(\mathbf{z} | \psi))].$$

Since $\bar{\mathcal{L}}$ may not be bounded above by the evidence $\log p(\mathbf{x})$ and $\bar{\mathcal{L}} - \underline{\mathcal{L}}$ is not bounded from above, there is no convergence guarantee if maximizing $\bar{\mathcal{L}}$. For this reason, we subtract $\bar{\mathcal{L}}$ by a correction term as

$$A_K = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\mathbf{z} \sim h_\phi(\mathbf{z})} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q_\phi(\psi)} \left[\log \left(\frac{1}{K} \sum_{k=1}^K q(\mathbf{z} | \psi^{(k)}) \right) - \log q(\mathbf{z} | \psi) \right]. \quad (8)$$

As $\mathbb{E}_{\psi^{(1)} \sim q_\phi(\psi)} \log q(\mathbf{z} | \psi^{(1)}) = \mathbb{E}_{\psi \sim q_\phi(\psi)} \log q(\mathbf{z} | \psi)$, we have $A_1 = 0$. The following proposition shows that the corrected upper bound $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$ monotonically converges from above towards the ELBO as $K \rightarrow \infty$.

Proposition 3. *Suppose $\underline{\mathcal{L}}$ and $\bar{\mathcal{L}}$ are defined as in (7) and A_K as in (8), then 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 $\lim_{K \rightarrow \infty} \bar{\mathcal{L}}_K = \underline{\mathcal{L}}$.*

The relationship between $\underline{\mathcal{L}}_K = \underline{\mathcal{L}} + B_K$ and $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$, two different asymptotically exact ELBOs, can be revealed by rewriting them as

$$\underline{\mathcal{L}}_K = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q_\phi(\psi)} \log \frac{p(\mathbf{x}, \mathbf{z})}{\frac{1}{K+1} [q(\mathbf{z} | \psi) + \sum_{k=1}^K q(\mathbf{z} | \psi^{(k)})]}, \quad (9)$$

$$\bar{\mathcal{L}}_K = \mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q_\phi(\psi)} \log \frac{p(\mathbf{x}, \mathbf{z})}{\frac{1}{K} \sum_{k=1}^K q(\mathbf{z} | \psi^{(k)})}. \quad (10)$$

Thus $\underline{\mathcal{L}}_K$ differs from $\bar{\mathcal{L}}_K$ in whether $q(\mathbf{z} | \psi)$ participates in computing the log density ratio, which is analytic thanks to the semi-implicit construction, inside the expectations. When K is small, using $\underline{\mathcal{L}}_K$ as the surrogate ELBO for optimization is expected to have better numerical stability than using $\bar{\mathcal{L}}_K$, as $\underline{\mathcal{L}}_0 = \underline{\mathcal{L}}$ relates to the ELBO as a lower bound while $\bar{\mathcal{L}}_1 = \bar{\mathcal{L}}$ does as an upper bound, but increasing K quickly diminishes the difference between $\underline{\mathcal{L}}_K$ and $\bar{\mathcal{L}}_K$, which are both asymptotically exact. It is also instructive to note that as $\mathbf{z} \sim q(\mathbf{z} | \psi)$ is used for Monte Carlo estimation, if assuming $\{q(\mathbf{z} | \psi), q(\mathbf{z} | \psi^{(1)}), \dots, q(\mathbf{z} | \psi^{(K)})\}$ has a dominant element, then it is most likely that $q(\mathbf{z} | \psi)$ dominates all $q(\mathbf{z} | \psi^{(k)})$. Therefore, maximizing $\underline{\mathcal{L}}_K$ in (9) would become almost the same as maximizing $\underline{\mathcal{L}}_0$, which would lead to degeneracy as in Proposition 1, which means $\psi = \psi^{(k)}$ and $q(\mathbf{z} | \psi) = q(\mathbf{z} | \psi^{(k)})$ for all k , contradicting the reasoning that $q(\mathbf{z} | \psi)$ dominates all $q(\mathbf{z} | \psi^{(k)})$.

Using the importance reweighting idea, Burda et al. (2015) provides a lower bound $L^{\tilde{K}} \geq \text{ELBO}$ that monotonically

converges from below to the evidence $\log p(\mathbf{x})$ as \tilde{K} increases. Using the same idea, we may also tighten the asymptotically exact surrogate ELBO in (9) using

$$\underline{\mathcal{L}}_K^{\tilde{K}} = \mathbb{E}_{(\mathbf{z}_1, \psi_1), \dots, (\mathbf{z}_{\tilde{K}}, \psi_{\tilde{K}}) \sim q(\mathbf{z} | \psi) q_\phi(\psi)} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q_\phi(\psi)} \log \frac{1}{\tilde{K}} \sum_{i=1}^{\tilde{K}} \frac{p(\mathbf{x}, \mathbf{z}_i)}{\frac{1}{K+1} [q(\mathbf{z}_i | \psi_i) + \sum_{k=1}^K q(\mathbf{z}_i | \psi^{(k)})]},$$

for which $\lim_{K \rightarrow \infty} \underline{\mathcal{L}}_K^{\tilde{K}} = L^{\tilde{K}} \geq \text{ELBO}$ and $\lim_{K, \tilde{K} \rightarrow \infty} \underline{\mathcal{L}}_K^{\tilde{K}} = \lim_{\tilde{K} \rightarrow \infty} L^{\tilde{K}} = \log p(\mathbf{x})$.

Using $\underline{\mathcal{L}}_{K_t}$ as the surrogate ELBO, where t indexes the number of iterations, $K_t \in \{0, 1, \dots\}$, and $K_{t+1} \geq K_t$, we describe the stochastic gradient ascent algorithm to optimize the variational parameter in Algorithm 1, in which we further introduce ξ as the variational parameter of the conditional distribution $q_\xi(\mathbf{z} | \psi)$ that is not mixed with another distribution. For Monte Carlo estimation in Algorithm 1, we use a single random sample for each $\psi^{(k)}$, J random samples for ψ , and a single sample of \mathbf{z} for each sample of ψ . One may further consult Rainforth et al. (2018) to help select K , J , and \tilde{K} for SIVI. We denote $\mathbf{z} = f(\varepsilon, \xi, \psi)$, $\varepsilon \sim p(\varepsilon)$ as the reparameterization for $\mathbf{z} \sim q_\xi(\mathbf{z} | \psi)$. As for ξ , if $\xi \neq \emptyset$, one may learn it as in Algorithm 1, set it empirically, or fix it at the value learned by another algorithm such as MFVI. In summary, SIVI constructs a flexible variational distribution by mixing a (potentially) implicit distribution with an explicit one, while maintaining tractable optimization via the use of an asymptotically exact surrogate ELBO.

3.4. Score Function Gradient in Conjugate Model

Let's consider the case that $q(\mathbf{z} | \psi)$ does not have a simple reparameterization but can benefit from conditional conjugacy. In particular, for a conditionally conjugate exponential family model, MFVI has an analytic ELBO, and its variational distribution can be directly used as the $q_\xi(\mathbf{z} | \psi)$ of SIVI. As in Appendix A, introducing a density ratio as

$$r_{\xi, \phi}(\mathbf{z}, \epsilon, \epsilon^{(1:K)}) = \frac{q_\xi(\mathbf{z} | T_\phi(\epsilon))}{\frac{q_\xi(\mathbf{z} | T_\phi(\epsilon)) + \sum_{k=1}^K q_\xi(\mathbf{z} | T_\phi(\epsilon^{(k)}))}{K+1}},$$

we approximate the gradient of $\underline{\mathcal{L}}_K$ with respect to ϕ as

$$\begin{aligned} \nabla_\phi \underline{\mathcal{L}}_K &\approx \frac{1}{J} \sum_{j=1}^J \left\{ -\nabla_\phi \mathbb{E}_{\mathbf{z} \sim q_\xi(\mathbf{z} | T_\phi(\epsilon_j))} [\log \frac{q_\xi(\mathbf{z} | T_\phi(\epsilon_j))}{p(\mathbf{x}, \mathbf{z})}] \right. \\ &+ \nabla_\phi \log r_{\xi, \phi}(\mathbf{z}_j, \epsilon_j, \epsilon^{(1:K)}) \\ &\left. + [\nabla_\phi \log q_\xi(\mathbf{z}_j | T_\phi(\epsilon_j))] \log r_{\xi, \phi}(\mathbf{z}_j, \epsilon_j, \epsilon^{(1:K)}) \right\}, \quad (11) \end{aligned}$$

where the first summation term is equivalent to the gradient of MFVI's ELBO, both the second and third terms correct the restrictions of $q_\xi(\mathbf{z} | T_\phi(\epsilon_j))$, and $\log r_{\xi, \phi}(\mathbf{z}, \epsilon, \epsilon^{(1:K)})$ 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; $\nabla_\xi \underline{\mathcal{L}}_K$ can be approximated in the same manner. For

non-conjugate models, we leave for future study the use of non-reparameterizable $q_{\xi}(z | \psi)$, for which one may apply customized variance reduction techniques (Paisley et al., 2012; Ranganath et al., 2014; Mnih & Gregor, 2014; Ruiz et al., 2016; Kucukelbir et al., 2017; Naesseth et al., 2017).

4. Related Work

There exists a wide variety of VI methods that improve on MFVI. Examples include adding dependencies between latent variables (Saul & Jordan, 1996; Hoffman & Blei, 2015), using a mixture of variational distributions (Bishop et al., 1998; Gershman et al., 2012; Salimans & Knowles, 2013; Guo et al., 2016; Miller et al., 2017), introducing a copula to capture the dependencies between univariate marginals (Tran et al., 2015; Han et al., 2016), handling non-conjugacy (Paisley et al., 2012; Titsias & Lázaro-Gredilla, 2014), and constructing a hierarchical variational distribution (Ranganath et al., 2016; Tran et al., 2017).

To increase the expressiveness of the PDF of a random variable, a simple but powerful idea is to transform it with complex deterministic and/or stochastic mappings. One successful application of this idea in VI is constructing the variational distribution with a normalizing flow, which transforms a simple random variable through a sequence of invertible differentiable functions with tractable Jacobians, to deterministically map a simple PDF to a complex one (Rezende & Mohamed, 2015; Kingma et al., 2016; Papamakarios et al., 2017). Normalizing flows help increase the flexibility of VI, but still require the mapping to be deterministic and invertible. Removing both restrictions, there have been several recent attempts to define highly flexible variational distributions with implicit models (Huszár, 2017; Mohamed & Lakshminarayanan, 2016; Tran et al., 2017; Li & Turner, 2017; Mescheder et al., 2017; Shi et al., 2017). A typical example is transforming random noise via a deep neural network, leading to a non-invertible highly nonlinear mapping and hence an implicit distribution.

While an implicit variational distribution can be made highly flexible, it becomes necessary in each iteration to address the problem of density ratio estimation, which is often transformed into a problem related to learning generative adversarial networks (Goodfellow et al., 2014). In particular, a binary classifier, whose class probability is used for density ratio estimation, is trained in each iteration to discriminate the samples generated by the model from those by the variational distribution (Mohamed & Lakshminarayanan, 2016; Uehara et al., 2016; Mescheder et al., 2017). Controlling the bias and variance in density ratio estimation, however, is in general a very difficult problem, especially in high-dimensional settings (Sugiyama et al., 2012).

SIVI is related to the hierarchical variational model (HVM)

(Ranganath et al., 2016; Maaløe et al., 2016) in having a hierarchical variational distribution, but there are two major distinctions: 1) the HVM restricts the mixing distribution in the hierarchy to have an explicit PDF, which can be constructed with a Markov chain (Salimans et al., 2015), a mixture model (Ranganath et al., 2016), or a normalizing flow (Ranganath et al., 2016; Louizos & Welling, 2017) but cannot come from an implicit model. By contrast, SIVI requires the conditional distribution $q(z | \psi)$ to have an explicit PDF, but does not impose such a constraint on the mixing distribution $q(\psi)$. In fact, any off-the-shelf reparameterizable implicit/explicit distribution can be used in SIVI, leading to considerably flexible $h_{\phi}(z) = \mathbb{E}_{\psi \sim q_{\phi}(\psi)} q(z | \psi)$. Moreover, SIVI does not require $q(z | \psi)$ to be reparameterizable for conditionally conjugate models. 2) the HVM optimizes on a lower bound of the ELBO, constructed by adding a recursively estimated variational distribution that approximates $q(\psi | z) = q(z | \psi)q(\psi)/h(z)$. By contrast, SIVI sandwiches the ELBO between two bounds, and directly optimizes on an asymptotically exact surrogate ELBO, which involves only simple-to-compute analytic density ratios.

5. Experiments

We implement SIVI in Tensorflow (Abadi et al., 2015) for a range of inference tasks. Note SIVI is a general purpose algorithm not relying on conjugacy, and has an inherent advantage over MCMC in being able to generate independent, and identically distributed (*iid*) posterior samples on the fly, this is, by feed-forward propagating *iid* random noises through the inferred semi-implicit hierarchy. The toy examples show SIVI captures skewness, kurtosis, and multimodality. A negative binomial model shows SIVI can accurately capture the dependencies between latent variables. A bivariate count distribution example shows for a conditionally conjugate model, SIVI can utilize a non-reparameterizable variational distribution, without being plagued by the high variance of score function gradient estimation. With Bayesian logistic regression, we demonstrate that SIVI can either work alone as a black-box inference procedure for correlated latent variables, or directly expand MFVI by adding a mixing distribution, leading to accurate uncertainty estimation on par with that of MCMC. Last but not least, moving beyond the canonical Gaussian based variational autoencoder (VAE), SIVI helps construct semi-implicit VAE (SIVAE) to improve unsupervised feature learning and amortized inference.

5.1. Expressiveness of SIVI

We first show the expressiveness of SIVI by approximating various target distributions. As listed in Table 1, the conditional layer of SIVI is chosen to be as simple as an isotropic Gaussian (or log-normal) distribution $\mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$. The implicit mixing layer is a multilayer perceptron (MLP), with

Table 1. Inference and target distributions

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

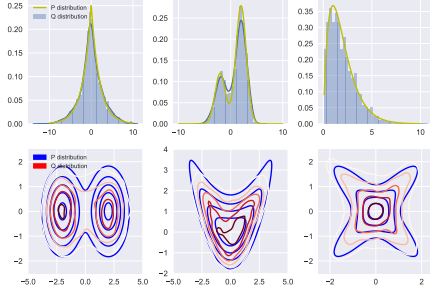


Figure 1. Approximating synthetic target distributions with SIVI

layer widths [30, 60, 30] and a ten dimensional isotropic Gaussian noise as its input. We fix $\sigma_0^2 = 0.1$ and optimize the implicit layer to minimize $\text{KL}(\mathbb{E}_{q_\psi(\psi)} q(z | \psi) || p(z))$. As shown in Figure 1, despite having a fixed purposely misspecified explicit layer, by training a flexible implicit layer, the random samples from which are illustrated in Figure 8 of Appendix D, SIVI infers a sophisticated marginal variational distribution that accurately captures the skewness, kurtosis, and/or multimodality exhibited by the target one.

5.2. Negative Binomial Model

We consider a negative binomial (NB) distribution with the gamma and beta priors ($a = b = \alpha = \beta = 0.01$) as

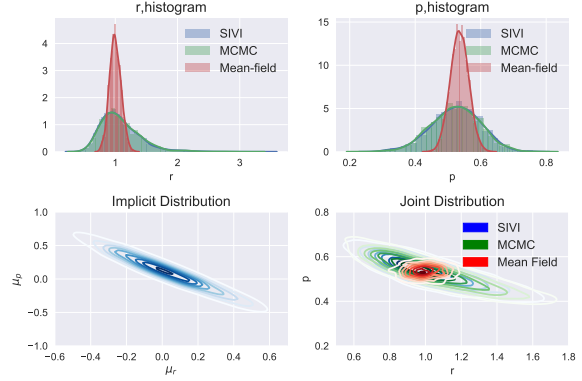
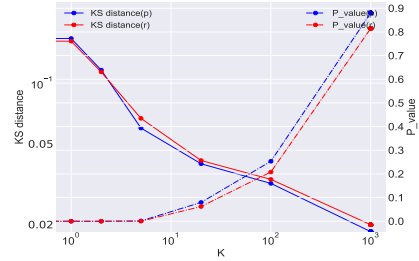
$$x_i \stackrel{iid}{\sim} \text{NB}(r, p), \quad r \sim \text{Gamma}(a, 1/b), \quad p \sim \text{Beta}(\alpha, \beta),$$

for which the posterior $p(r, p | \{x_i\}_{1,N})$ is not tractable. In comparison to Gibbs sampling, it is shown in Zhou et al. (2012) that MFVI, which uses $q(r, p) = \text{Gamma}(r; \tilde{a}, 1/\tilde{b})\text{Beta}(p; \tilde{\alpha}, \tilde{\beta})$ to approximate the posterior, notably underestimates the variance. This caveat of MFVI motivates a semi-implicit variational distribution as

$$q(r, p | \psi) = \text{Log-Normal}(r; \mu_r, \sigma_0^2) \text{Logit-Normal}(p; \mu_p, \sigma_0^2), \\ \psi = (\mu_r, \mu_p) \sim q(\psi), \sigma_0 = 0.1$$

where and an MLP based implicit $q(\psi)$, as in Section 5.1, is used by SIVI to capture the dependency between r and p .

We apply Gibbs sampling, MFVI, and SIVI to a real overdispersed count dataset of Bliss & Fisher (1953) that records the number of adult red mites on each of the 150 randomly selected apple leaves. With $K = 1000$, as shown in Figure 2, SIVI improves MFVI in closely matching the posterior inferred by Gibbs sampling. Moreover, the mixing distribution $q(\psi)$ helps well capture the negative correla-


 Figure 2. Top row: the marginal posteriors of r and p inferred by MFVI, SIVI, and MCMC. Bottom row: the inferred implicit mixing distribution $q(\psi)$ and joint posterior of r and p .

 Figure 3. 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. See Appendix D for the corresponding plots of marginal posteriors.

tion between r and p , as totally ignored by MFVI. The two-sample Kolmogorov-Smirnov (KS) distances, between 2000 posterior samples generated by SIVI and 2000 MCMC ones, are 0.0185 (p -value = 0.88) and 0.0200 (p -value = 0.81) for r and p , respectively. By contrast, for MFVI and MCMC, they are 0.2695 (p -value = 5.26×10^{-64}) and 0.2965 (p -value = 2.21×10^{-77}), which significantly reject the null hypothesis that the posterior inferred by MFVI matches that by MCMC. How the performance is related to K is shown in Figure 3 and Figures 9-10 of Appendix D, which suggests $K = 20$ achieves a nice compromise between complexity and accuracy, and as K increases, the posterior inferred by SIVI quickly approaches that inferred by MCMC.

5.3. Non-reparameterizable Variational Distribution

To show that SIVI can use a non-reparameterizable $q(z | \psi)$ in a conditionally conjugate model, as discussed in Section 3.4, we apply it to infer the two parameters of the Poisson-logarithmic bivariate count distribution as $p(n_i, l_i | r, p) = r^{l_i} p^{n_i} (1-p)^r / Z_i$, where $l_i \in \{0, \dots, n_i\}$ and Z_i are normalization constants not related to $r > 0$ and $p \in (0, 1)$ (Zhou & Carin, 2015; Zhou et al., 2016). With $r \sim \text{Gamma}(a, 1/b)$ and $p \sim \text{Beta}(\alpha, \beta)$, while the joint posterior $p(r, p | \{n_i, l_i\}_{1,N})$ is intractable, the conditional posteriors of r and p follow the gamma and beta distributions,

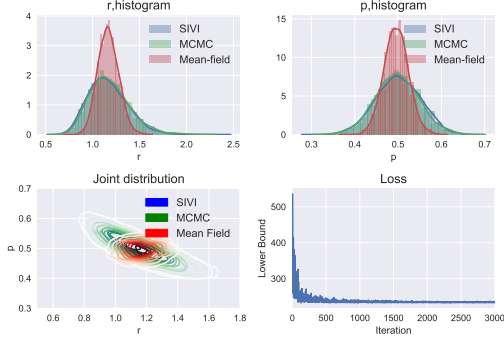


Figure 4. Top row: the marginal posteriors of r and p inferred by MFVI, SIVI, and MCMC. Bottom row: joint posteriors and the trace plots of $-\mathcal{L}_{K_t}$ (subject to the difference of a constant).

respectively. Although neither the gamma nor beta distribution is reparameterizable, SIVI multiplies them to construct a semi-implicit variational distribution as

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

where $\psi = (\psi_1, \psi_2, \psi_3, \psi_4) \sim q(\psi)$ is similarly constructed as in Section 5.1. We set $K = 200$ for SIVI.

As shown in Figure 4, despite the need of score function gradient estimation that is notorious for variance control, by utilizing conjugacy as in (11), SIVI well controls the variance of its gradient estimation, achieving accurate posterior estimation without requiring $q(z | \psi)$ to be reparameterizable. By contrast, MFVI, which uses only the first summation term in (11) for gradient estimation, ignores covariance structure and notably underestimates posterior uncertainty.

5.4. Bayesian Logistic Regression

We compare SIVI with MFVI, Stein variational gradient descent (SVGD) of Liu & Wang (2016), and MCMC on Bayesian logistic regression, expressed as

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

where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{iV})'$ are covariates, $y_i \in \{0, 1\}$ are binary response variables, and α is set as 0.01. With the Pólya-Gamma data augmentation of Polson et al. (2013), we collect posterior MCMC samples of β using Gibbs sampling. For MFVI, the variational distribution is chosen as a multivariate normal (MVN) $\mathcal{N}(\beta; \mu, \Sigma)$, with a diagonal or full covariance matrix. For SIVI, we treat Σ , diagonal or full, as a variational parameter, mix μ with an MLP based implicit distribution, and set $K = 500$. We consider three datasets: *waveform*, *spam*, and *nodal*. The details on datasets and inference are deferred to Appendix B. On *waveform*, the algorithm converges in about 500 iterations, which takes about 40 seconds on a 2.4 GHz CPU. Note the results of SIVI with $K = 100$ (or 50), which takes about 12 (or 8) seconds for 500 iterations, are almost identical to those

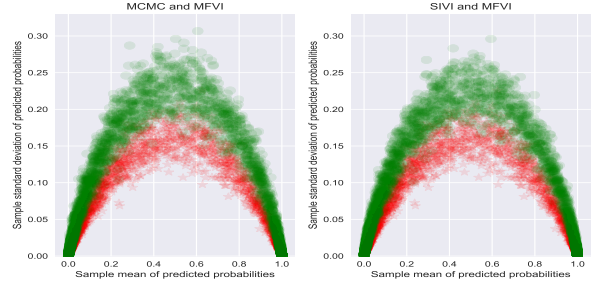


Figure 5. Comparison of MFVI (red) with a full covariance matrix, MCMC (green on left), and SIVI (green on right) with a full covariance matrix on quantifying predictive uncertainty for Bayesian logistic regression on *waveform*.

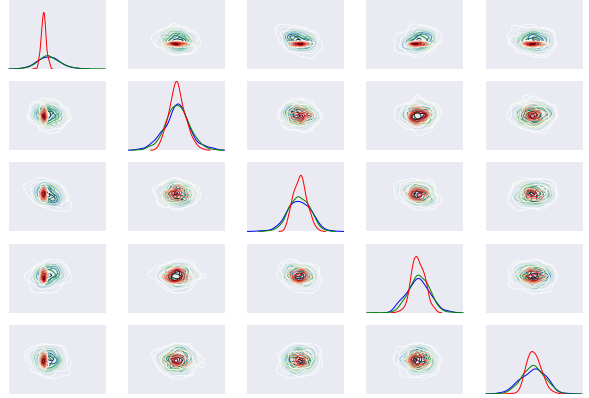


Figure 6. Marginal and pairwise joint posteriors for $(\beta_0, \dots, \beta_4)$ inferred by MFVI (red, full covariance matrix), MCMC (blue), and SIVI (green, full covariance matrix) on *waveform*.

shown in Figures 5-8 with $K = 500$. Given the posterior captured by the semi-implicit hierarchy, SIVI takes 0.92 seconds to generate 50,000 *iid* 22-dimensional β 's.

We collect β_j for $j = 1, \dots, 1000$ to represent the inferred posterior $p(\beta | \{\mathbf{x}_i, y_i\}_{1,N})$. For each test data \mathbf{x}_{N+i} , we calculate the predictive probabilities $1/(1 + e^{-\mathbf{x}_{N+i}^T \beta_j})$ for all j and compute its sample mean, and sample standard deviation that measures the uncertainty of the predictive distribution $p(y_{N+i} = 1 | \mathbf{x}_{N+i}, \{\mathbf{x}_i, y_i\}_{1,N})$. As shown in Figure 5, even with a full covariance matrix, the MVN variational distribution inferred by MFVI underestimates the uncertainty in out-of-sample prediction, let alone with a diagonal one, whereas SIVI, mixing the MVN with an MLP based implicit distribution, closely matches MCMC in uncertainty estimation. As shown in Figure 6, the underestimation of predictive uncertainty by MFVI can be attributed to variance underestimation for both univariate marginal and pairwise joint posteriors, which are, by contrast, well agreed on between SIVI and MCMC. Further examining the correlation coefficients of β , shown in Figure 7, all the univariate marginals, shown in Figure 11 of Appendix D, and additional results, show in Figures 12-17 of Appendix D, it is revealed that SIVI well characterizes the posterior distribution of β and is only slightly negatively affected if its

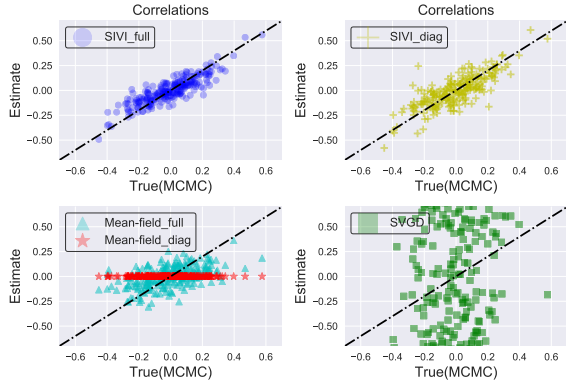


Figure 7. Comparing the correlation coefficients of β estimated from the posterior samples $\{\beta_i\}_{i=1:1000}$ of SIVI with that of MCMC on *waveform* for SIVI with a full/diagonal covariance matrix, MFVI with a full/diagonal covariance matrix, and SVGD.

explicit layer is restricted with a diagonal covariance matrix, whereas MFVI with a diagonal/full covariance matrix and SVGD all misrepresent uncertainty. Note we have also tried modifying the code of variational boosting (Guo et al., 2016; Miller et al., 2017) for Bayesian logistic regression, but failed to obtain satisfactory results. We attribute the success of SIVI to its ability in better capturing the dependencies between β_v and supporting a highly expressive non-Gaussian variational distribution by mixing an MVN with a flexible implicit distribution, whose parameters can be efficiently optimized via an asymptotically exact surrogate ELBO.

5.5. Semi-Implicit Variational Autoencoder

Variational Auto-encoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014), widely used for unsupervised feature learning and amortized inference, infers encoder parameter ϕ and decoder parameter θ to maximize the ELBO as

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

The encoder distribution $q_\phi(z | \mathbf{x})$ is required to be reparameterizable and analytically evaluable, which usually restricts it to a small exponential family. In particular, a canonical encoder is $q_\phi(z | \mathbf{x}) = \mathcal{N}(z | \mu(\mathbf{x}, \phi), \Sigma(\mathbf{x}, \phi))$, where the Gaussian parameters are deterministically transformed from the observations \mathbf{x} , via non-probabilistic deep neural networks parameterized by ϕ . Thus, given observation \mathbf{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 skewness, kurtosis, and multimodality.

To this end, rather than using a single-stochastic-layer encoder, we use SIVI that can add as many stochastic layers as needed, as long as the first stochastic layer $q_\phi(z | \mathbf{x})$ is reparameterizable and has an analytic PDF, and the layers added after are reparameterizable and simple to sample from. More specifically, we construct semi-implicit VAE (SIVAE)

by using a hierarchical encoder that injects random noise at M different stochastic layers as

$$\begin{aligned} \ell_t &= T_t(\ell_{t-1}, \epsilon_t, \mathbf{x}; \phi), \quad \epsilon_t \sim q_t(\epsilon), \quad t = 1, \dots, M, \\ \mu(\mathbf{x}, \phi) &= f(\ell_M, \mathbf{x}; \phi), \quad \Sigma(\mathbf{x}, \phi) = g(\ell_M, \mathbf{x}; \phi), \\ q_\phi(z | \mathbf{x}, \mu, \Sigma) &= \mathcal{N}(\mu(\mathbf{x}, \phi), \Sigma(\mathbf{x}, \phi)), \end{aligned} \quad (12)$$

where $\ell_0 = \emptyset$ and T_t , f , and g are all deterministic neural networks. Note given data \mathbf{x}_i , $\mu(\mathbf{x}_i, \phi)$, $\Sigma(\mathbf{x}_i, \phi)$ are now random variables rather than following vanilla VAE to assume deterministic values. This moves the encoder variational distribution beyond a simple Gaussian form.

To benchmark the performance of SIVAE, we consider the MNIST dataset that is stochastically binarized as in Salakhutdinov & Murray (2008). We use 55,000 for training and use the 10,000 observations in the testing set for performance evaluation. Similar to existing VAEs, we choose Bernoulli units, linked to a fully-connected neural network with two 500-unit hidden layers, as the decoder. Distinct from existing VAEs, whose encoders are often restricted to have a single stochastic layer, SIVAE allows SIVAE to use an MVN as its first stochastic layer, and draw the parameters of the MVN from $M = 3$ stochastic layers, whose structure is described in detail in Appendix C. As shown in Table 2 of Appendix C, SIVAE achieves a negative log evidence of 84.07, which is further reduced to 83.25 if choosing importance reweighing with $\tilde{K} = 10$. In comparison to other VAEs with a comparable single-stochastic-layer decoder, SIVAE achieves state-of-the-art performance by mixing an MVN with an implicit distribution defined as in (12) to construct a flexible encoder, whose marginal variational distribution is no longer restricted to the MVN distribution. We leave it for future study on further improving SIVAE by replacing the encoder MVN explicit layer with a normalizing flow, and adding convolution/autoregression to enrich the encoder’s implicit distribution and/or the decoder.

6. Conclusions

Combining the advantages of having analytic point-wise evaluable density ratios and tractable computation via Monte Carlo estimation, semi-implicit variational inference (SIVI) is proposed either as a black-box inference procedure, or to enrich mean-field variational inference with a flexible (implicit) mixing distribution. By designing a surrogate evidence lower bound that is asymptotically exact, SIVI establishes an optimization problem amenable to gradient ascent, without compromising the expressiveness of its semi-implicit variational distribution. Flexible but simple to optimize, SIVI approaches the accuracy of MCMC in quantifying posterior uncertainty in a wide variety of inference tasks, and is not constrained by conjugacy, often runs faster, and can generate *iid* posterior samples on the fly via the inferred stochastic variational inference network.

Acknowledgements

The authors thank Texas Advanced Computing Center (TACC) for computational support.

References

- Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dan, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, and Zheng, Xiaoqiang. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <http://tensorflow.org/>. Software available from tensorflow.org.
- Agakov, F. V. and Barber, D. An auxiliary variational method. In *International Conference on Neural Information Processing*, 2004.
- Bishop, Christopher M and Tipping, Michael E. Variational relevance vector machines. In *UAI*, pp. 46–53. Morgan Kaufmann Publishers Inc., 2000.
- Bishop, Christopher M, Lawrence, Neil D, Jaakkola, Tommi, and Jordan, Michael I. Approximating posterior distributions in belief networks using mixtures. In *NIPS*, pp. 416–422, 1998.
- Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- Bliss, Chester Ittner and Fisher, Ronald A. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200, 1953.
- Burda, Yuri, Grosse, Roger, and Salakhutdinov, Ruslan. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- Cover, Thomas M and Thomas, Joy A. *Elements of Information Theory*. John Wiley & Sons, 2012.
- Dinh, Laurent, Krueger, David, and Bengio, Yoshua. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- Gershman, Samuel J, Hoffman, Matthew D, and Blei, David M. Nonparametric variational inference. In *ICML*, pp. 235–242, 2012.
- Giordano, Ryan J, Broderick, Tamara, and Jordan, Michael I. Linear response methods for accurate covariance estimates from mean field variational bayes. In *NIPS*, pp. 1441–1449, 2015.
- Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.
- Gregor, Karol, Danihelka, Ivo, Graves, Alex, Rezende, Danilo, and Wierstra, Daan. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pp. 1462–1471, 2015.
- Guo, Fangjian, Wang, Xiangyu, Fan, Kai, Broderick, Tamara, and Dunson, David B. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.
- Habil, Eissa D. Double sequences and double series. *IUG Journal of Natural Studies*, 14(1), 2016.
- Han, Shaobo, Liao, Xuejun, Dunson, David, and Carin, Lawrence. Variational Gaussian copula inference. In *AISTATS*, pp. 829–838, 2016.
- Hoffman, Matthew and Blei, David. Stochastic structured variational inference. In *AISTATS*, pp. 361–369, 2015.
- Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- Huszár, Ferenc. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.
- Im, Daniel Jiwoong, Ahn, Sungjin, Memisevic, Roland, Bengio, Yoshua, et al. Denoising criterion for variational auto-encoding framework. In *AAAI*, pp. 2059–2065, 2017.
- Jaakkola, Tommi S and Jordan, Michael I. Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pp. 163–173. Springer, 1998.
- Jaakkola, Tommi S and Jordan, Michael I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.
- Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
- Kingma, Diederik P and Welling, Max. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Kingma, Diederik P, Salimans, Tim, Jozefowicz, Rafal, Chen, Xi, Sutskever, Ilya, and Welling, Max. Improved variational inference with inverse autoregressive flow. In *NIPS*, pp. 4743–4751, 2016.
- Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.
- Kurdila, Andrew J and Zabaranin, Michael. Convex functional analysis (systems and control: Foundations and applications). *Switzerland: Birkhäuser*, 2005.
- Li, Yingzhen and Turner, Richard E. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.
- Liu, Qiang and Wang, Dilin. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *NIPS*, pp. 2378–2386, 2016.
- Louizos, Christos and Welling, Max. Multiplicative normalizing flows for variational Bayesian neural networks. In *ICML*, pp. 2218–2227, 2017.

- Maaløe, Lars, Sønderby, Casper Kaae, Sønderby, Søren Kaae, and Winther, Ole. Auxiliary deep generative models. In *ICML*, pp. 1445–1453, 2016.
- Mescheder, Lars, Nowozin, Sebastian, and Geiger, Andreas. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.
- Miller, Andrew C., Foti, Nicholas J., and Adams, Ryan P. Variational boosting: Iteratively refining posterior approximations. In *ICML*, pp. 2420–2429, 2017.
- Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. In *ICML*, pp. 1791–1799, 2014.
- Mohamed, Shakir and Lakshminarayanan, Balaji. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.
- Naesseth, Christian, Ruiz, Francisco, Linderman, Scott, and Blei, David. Reparameterization gradients through acceptance-rejection sampling algorithms. In *AISTATS*, pp. 489–498, 2017.
- Paisley, John, Blei, David M, and Jordan, Michael I. Variational Bayesian inference with stochastic search. In *ICML*, pp. 1363–1370, 2012.
- Papamakarios, George, Murray, Iain, and Pavlakou, Theo. Masked autoregressive flow for density estimation. In *NIPS*, pp. 2335–2344, 2017.
- Polson, Nicholas G, Scott, James G, and Windle, Jesse. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504): 1339–1349, 2013.
- Rainforth, Tom, Kosiorek, Adam R, Le, Tuan Anh, Maddison, Chris J, Igl, Maximilian, Wood, Frank, and Teh, Yee Whye. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.
- Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *AISTATS*, pp. 814–822, 2014.
- Ranganath, Rajesh, Tran, Dustin, and Blei, David. Hierarchical variational models. In *ICML*, pp. 324–333, 2016.
- Rezende, Danilo and Mohamed, Shakir. Variational inference with normalizing flows. In *ICML*, pp. 1530–1538, 2015.
- Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pp. 1278–1286, 2014.
- Rudin, Walter. *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York, 1964.
- Ruiz, Francisco J. R., Titsias, Michalis K., and Blei, David M. The generalized reparameterization gradient. In *NIPS*, pp. 460–468, 2016.
- Salakhutdinov, Ruslan and Murray, Iain. On the quantitative analysis of deep belief networks. In *ICML*, pp. 872–879, 2008.
- Salimans, Tim and Knowles, David A. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.
- Salimans, Tim, Kingma, Diederik, and Welling, Max. Markov chain Monte Carlo and variational inference: Bridging the gap. In *ICML*, pp. 1218–1226, 2015.
- Saul, Lawrence K and Jordan, Michael I. Exploiting tractable substructures in intractable networks. In *NIPS*, pp. 486–492, 1996.
- Shi, Jiaxin, Sun, Shengyang, and Zhu, Jun. Implicit variational inference with kernel density ratio fitting. *arXiv preprint arXiv:1705.10119*, 2017.
- Sønderby, Casper Kaae, Raiko, Tapani, Maaløe, Lars, Sønderby, Søren Kaae, and Winther, Ole. Ladder variational autoencoders. In *NIPS*, pp. 3738–3746, 2016.
- Sugiyama, Masashi, Suzuki, Taiji, and Kanamori, Takafumi. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.
- Titsias, Michalis and Lázaro-Gredilla, Miguel. Doubly stochastic variational bayes for non-conjugate inference. In *ICML*, pp. 1971–1979, 2014.
- Tran, Dustin, Blei, David, and Airoldi, Edo M. Copula variational inference. In *NIPS*, pp. 3564–3572, 2015.
- Tran, Dustin, Ranganath, Rajesh, and Blei, David M. The variational Gaussian process. In *ICLR*, 2016.
- Tran, Dustin, Ranganath, Rajesh, and Blei, David. Hierarchical implicit models and likelihood-free variational inference. In *NIPS*, pp. 5529–5539, 2017.
- Uehara, Masatoshi, Sato, Issei, Suzuki, Masahiro, Nakayama, Kotaro, and Matsuo, Yutaka. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.
- Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- Zhou, Mingyuan. Softplus regressions and convex polytopes. *arXiv:1608.06383*, 2016.
- Zhou, Mingyuan and Carin, Lawrence. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015.
- Zhou, Mingyuan, Li, Lingbo, Dunson, David, and Carin, Lawrence. Lognormal and gamma mixed negative binomial regression. In *ICML*, pp. 859–866, 2012.
- Zhou, Mingyuan, Padilla, Oscar Hernan Madrid, and Scott, James G. Priors for random count matrices derived from a family of negative binomial processes. *J. Amer. Statist. Assoc.*, 111(515):1144–1156, 2016.

Semi-Implicit Variational Inference: Supplementary Material

Mingzhang Yin and Mingyuan Zhou

Algorithm 1 Semi-Implicit Variational Inference (SIVI)

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

output : Variational parameter ξ for the conditional distribution $q_\xi(\mathbf{z} | \psi)$, variational parameter ϕ for the mixing distribution $q_\phi(\psi)$

Initialize ξ and ϕ 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 $\psi^{(k)} = T_\phi(\epsilon^{(k)})$, $\epsilon^{(k)} \sim q(\epsilon)$ for $k = 1, \dots, K_t$; take sub-sample $\mathbf{x} = \{x_i\}_{i_1:i_M}$

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

Sample $\psi_j = T_\phi(\epsilon_j)$, $\epsilon_j \sim q(\epsilon)$

Sample $\mathbf{z}_j = f(\tilde{\epsilon}_j, \xi, \psi_j)$, $\tilde{\epsilon}_j \sim p(\epsilon)$

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

end

$t = t + 1$

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

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

end

A. Proofs

Proof of Inequility (3). To prove a functional form of Jensen's Inequality, let $h(\mathbf{z}) = \mathbb{E}_{\psi \sim q_\phi(\psi)} q(\mathbf{z} | \psi)$ and $\langle f, g \rangle_{L^2} = \int f(\mathbf{z})g(\mathbf{z})d\mathbf{z}$. From Theorem 1, we have convexity, and according to Theorem 6.2.1. of Kurdila & Zabaranin (2005), we have an equivalent first-order definition for convexity as

$$\text{KL}(q(\mathbf{z} | \psi) || p(\mathbf{z})) \geq \text{KL}(h(\mathbf{z}) || p) + \langle q(\mathbf{z} | \psi) - h(\mathbf{z}), \nabla_q \text{KL}(q || p)|_{h(\mathbf{z})} \rangle_{L^2}$$

Taking the expectation with respect to $\psi \sim q_\phi(\psi)$ on both sides, we have

$$\begin{aligned} & \mathbb{E}_{\psi \sim q_\phi(\psi)} \text{KL}(q(\mathbf{z} | \psi) || p(\mathbf{z})) \\ & \geq \text{KL}(h(\mathbf{z}) || p(\mathbf{z})) \\ & + \mathbb{E}_{\psi \sim q_\phi(\psi)} [\langle q(\mathbf{z} | \psi) - h(\mathbf{z}), \nabla_q \text{KL}(q || p)|_{h(\mathbf{z})} \rangle_{L^2}] \\ & = \text{KL}(h(\mathbf{z}) || p(\mathbf{z})) \\ & = \text{KL}(\mathbb{E}_{\psi \sim q_\phi(\psi)} q(\mathbf{z} | \psi) || p(\mathbf{z})). \end{aligned}$$

□

Proof of Proposition 1. We show that directly maximizing the lower bound $\underline{\mathcal{L}}$ of ELBO in (4) may drive $q(\psi)$ towards degeneracy. For VI that uses $q(\mathbf{z} | \psi)$ as its variational distribution, if supposing ψ^* is the optimum variational parameter, which means

$$\psi^* = \arg \max_{\psi} -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \log \frac{q(\mathbf{z} | \psi)}{p(\mathbf{x}, \mathbf{z})},$$

then we have

$$\begin{aligned} \underline{\mathcal{L}} &= -\mathbb{E}_{\psi \sim q_\phi(\psi)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \log \frac{q(\mathbf{z} | \psi)}{p(\mathbf{x}, \mathbf{z})} \\ &= \int q_\phi(\psi) [-\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \log \frac{q(\mathbf{z} | \psi)}{p(\mathbf{x}, \mathbf{z})}] d\psi \\ &\leq \int q_\phi(\psi) d\psi [-\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi^*)} \log \frac{q(\mathbf{z} | \psi^*)}{p(\mathbf{x}, \mathbf{z})}] \\ &= -\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi^*)} \log \frac{q(\mathbf{z} | \psi^*)}{p(\mathbf{x}, \mathbf{z})}. \end{aligned}$$

The equality in the above equation is reached if and only if $q(\psi) = \delta_{\psi^*}(\psi)$, which means the mixing distribution degenerates to a point mass density and hence SIVI degenerates to vanilla VI. □

Proof of Proposition 2. $B_0 = 0$ is trivial. Denote $\psi^{(0)} = \psi_v$. For iid samples $\psi^{(k)} \sim q_\phi(\psi)$, when $K \rightarrow \infty$, by the strong law of large numbers, $\tilde{h}_K(\mathbf{z}) = \frac{\sum_{k=0}^K q(\mathbf{z} | \psi^{(k)})}{K+1}$ converges almost surely to $\mathbb{E}_{q_\phi(\psi)} q(\mathbf{z} | \psi) = h_\phi(\mathbf{z})$. To prove (6), by the strong law of large numbers, we first rewrite it as the limit of a double sequence $S(K, J)$, where $K, J \in \{1, 2, \dots\}$, and check the condition for the interchange of iterated limits (Rudin, 1964; Habi1, 2016): i) The double limit exists; ii) Fixing one index of the double sequence, the one side limit exists for the other index .

$$\begin{aligned} & \lim_{K \rightarrow \infty} \mathbb{E}_{\psi^{(0)}, \psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \log \frac{\sum_{k=0}^K q(\mathbf{z} | \psi^{(k)})}{K+1} \\ &= \lim_{K \rightarrow \infty} \lim_{J \rightarrow \infty} \frac{1}{J} \sum_{j=1}^J \log \frac{1}{K+1} \sum_{k=0}^K q(\mathbf{z} | \psi_j^{(k)}) \\ &\triangleq \lim_{K \rightarrow \infty} \lim_{J \rightarrow \infty} S(K, J). \end{aligned}$$

Here $\psi_j^{(k)}$ are iid samples from $q(\psi)$. For i) we show double limit $\lim_{K, J \rightarrow \infty} S(K, J) = \log h(\mathbf{z})$. For $\forall \epsilon > 0$, $\exists N(\epsilon)$, when $K, J > N(\epsilon)$, $|\log \frac{1}{K+1} \sum_{k=0}^K q(\mathbf{z} | \psi_j^{(k)}) - \log h(\mathbf{z})| < \epsilon$ thanks to the law of large numbers, then

$$\begin{aligned} & \left| \sum_{j=1}^J \log \frac{1}{K+1} \sum_{k=0}^K q(\mathbf{z} | \psi_j^{(k)}) - J \log h(\mathbf{z}) \right| \\ & \leq \sum_{j=1}^J \left| \log \frac{1}{K+1} \sum_{k=0}^K q(\mathbf{z} | \psi_j^{(k)}) - \log h(\mathbf{z}) \right| \leq J\epsilon. \end{aligned}$$

Deviding both sides by J we get $|S(K, J) - \log h(\mathbf{z})| \leq \epsilon$ when $K, J > N(\epsilon)$. By definition, we have $\lim_{K, J \rightarrow \infty} S(K, J) = \log h(\mathbf{z})$.

ii) for each fixed $J \in \mathbb{N}$, $\lim_{K \rightarrow \infty} S(K, J) = \log h(\mathbf{z})$ exists; for each fixed $K \in \mathbb{N}$, $\lim_{J \rightarrow \infty} S(K, J) = \mathbb{E}_{\psi^{(0)}, \psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \log \frac{\sum_{k=0}^K q(\mathbf{z} | \psi^{(k)})}{K+1} \leq \log h(\mathbf{z})$ also exists. The limitation can then be interchanged and (6) is proved. Therefore, we have

$$\begin{aligned} \lim_{k \rightarrow \infty} \mathcal{L}_k &= \underline{\mathcal{L}} + \mathbb{E}_{\psi} \text{KL}(q(\mathbf{z} | \psi) || h_{\phi}(\mathbf{z})) \\ &= \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \left[\log \frac{q(\mathbf{z} | \psi)}{h_{\phi}(\mathbf{z})} - \log \frac{q(\mathbf{z} | \psi)}{p(x, \mathbf{z})} \right] \\ &= - \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \psi)} \log \frac{h_{\phi}(\mathbf{z})}{p(x, \mathbf{z})} \\ &= - \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \log \frac{h_{\phi}(\mathbf{z})}{p(x, \mathbf{z})} = \mathcal{L}. \end{aligned}$$

□

Proof of Proposition 3. Assume integer $K > M > 0$. Let \mathcal{I} be the set that consists of all the subsets of $\{1, \dots, K\}$ with cardinality M . Let I be a discrete uniform random variable and for element $\{i_1, \dots, i_M\} \in \mathcal{I}$, $P(I = \{i_1, \dots, i_M\}) = \frac{1}{\binom{K}{M}}$. We have $\mathbb{E}_I \frac{1}{M} \sum_{i \in I} q(\mathbf{z} | \psi^i) = \frac{1}{K} \sum_{i=1}^K q(\mathbf{z} | \psi^i)$. To show $\bar{\mathcal{L}}_K = \bar{\mathcal{L}} - A_K$ is monotonic decreasing, we only need to show A_K is monotonic increasing:

$$\begin{aligned} A_K &= \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \\ &\quad \log \frac{\frac{1}{K} \sum_{i=1}^K q(\mathbf{z} | \psi^{(i)})}{q(\mathbf{z} | \psi)} \\ &= \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \\ &\quad \log \mathbb{E}_I \left[\frac{\frac{1}{M} \sum_{i \in I} q(\mathbf{z} | \psi^{(i)})}{q(\mathbf{z} | \psi)} \right] \\ &\geq \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \\ &\quad \mathbb{E}_I \log \frac{\frac{1}{M} \sum_{i \in I} q(\mathbf{z} | \psi^{(i)})}{q(\mathbf{z} | \psi)} \\ &= \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(M)} \sim q(\psi)} \\ &\quad \log \frac{\frac{1}{M} \sum_{i=1}^M q(\mathbf{z} | \psi^{(i)})}{q(\mathbf{z} | \psi)} \\ &= A_M. \end{aligned}$$

We now show $\lim_{K \rightarrow \infty} \bar{\mathcal{L}}_K = \mathcal{L}$. Again, by the strong law of large numbers, $\frac{1}{K} \sum_{i=1}^K q(\mathbf{z} | \psi^{(i)})$ converges almost

surely to $\mathbb{E}_{\psi \sim q(\psi)} q(\mathbf{z} | \psi) = h_{\phi}(\mathbf{z})$ and hence

$$\begin{aligned} \lim_{K \rightarrow \infty} \bar{\mathcal{L}}_K &= \bar{\mathcal{L}} + \mathbb{E}_{\psi} \text{KL}(h_{\phi}(\mathbf{z}) || q(\mathbf{z} | \psi)) \\ &= - \mathbb{E}_{\mathbf{z} \sim h_{\phi}(\mathbf{z})} \mathbb{E}_{\psi \sim q(\psi)} \left[\log \frac{q(\mathbf{z} | \psi)}{p(x, \mathbf{z})} + \log \frac{h_{\phi}(\mathbf{z})}{q(\mathbf{z} | \psi)} \right] \\ &= \mathcal{L}. \end{aligned}$$

□

Proof of Equation (11). The gradient of B_K with respect to ϕ can be expressed as

$$\begin{aligned} \nabla_{\phi} B_K &= \nabla_{\phi} \mathbb{E}_{\psi \sim q(\psi)} \mathbb{E}_{\psi^{(1)}, \dots, \psi^{(K)} \sim q(\psi)} \left[\text{KL} \left(q(\mathbf{z} | \psi) \middle| \middle| \frac{q(\mathbf{z} | \psi) + \sum_{k=1}^K q(\mathbf{z} | \psi^{(k)})}{K+1} \right) \right] \\ &= \mathbb{E}_{\epsilon, \epsilon^{(1)}, \dots, \epsilon^{(K)} \sim p(\epsilon)} \nabla_{\phi} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | T_{\phi}(\epsilon))} \left[\log \frac{q(\mathbf{z} | T_{\phi}(\epsilon))}{\frac{q(\mathbf{z} | T_{\phi}(\epsilon)) + \sum_{k=1}^K q(\mathbf{z} | T_{\phi}(\epsilon^{(k)}))}{K+1}} \right] \\ &= \mathbb{E}_{\epsilon, \dots, \epsilon^{(K)}} \nabla_{\phi} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | T_{\phi}(\epsilon))} \log \left[r_{\phi}(\mathbf{z}, \epsilon, \epsilon^{(1:K)}) \right] \\ &= \mathbb{E}_{\epsilon, \dots, \epsilon^{(K)} \sim p(\epsilon)} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | T_{\phi}(\epsilon))} \left\{ q(\mathbf{z} | T_{\phi}(\epsilon)) \nabla_{\phi} \log \left[r_{\phi}(\mathbf{z}, \epsilon, \epsilon^{(1:K)}) \right] \right. \\ &\quad \left. + \left[\nabla_{\phi} \log q(\mathbf{z} | T_{\phi}(\epsilon)) \right] \log \left[r_{\phi}(\mathbf{z}, \epsilon, \epsilon^{(1:K)}) \right] \right\}. \end{aligned}$$

□

B. Bayesian Logistic Regression

We consider datasets *waveform* ($n = 5000$, $V = 21$, and $400/4600$ for training/testing), *spam* ($n = 3000$, $V = 2$, and $2000/1000$ for training/testing), and *nodal* ($n = 53$, $V = 5$, and $25/28$ for training/testing). The training-set-size to feature-dimension ratio n_{train}/V varies in these three datasets, and we expect the posterior uncertainty to be large if this ratio is small.

The contribution of observation i to the likelihood can be expressed as

$$P(y_i | \mathbf{x}_i, \beta) = \frac{e^{y \mathbf{x}'_i \beta}}{1 + e^{\mathbf{x}'_i \beta}} \propto e^{(y - \frac{1}{2}) \mathbf{x}'_i \beta} \mathbb{E}_{\omega_i} \left[e^{-\frac{\omega_i (\mathbf{x}'_i \beta)^2}{2}} \right],$$

where the expectation is taken respect to a Pólya-Gamma (PG) distribution (Polson et al., 2013) as $\omega_i \sim \text{PG}(1, 0)$, and hence we have an augmented likelihood as

$$P(y_i, \omega_i | \mathbf{x}_i, \beta) \propto e^{(y - \frac{1}{2}) \mathbf{x}'_i \beta - \frac{1}{2} \omega_i (\mathbf{x}'_i \beta)^2}.$$

B.1. Gibbs Sampling via Data Augmentation

Denoting $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, $\mathbf{y} = (y_1, \dots, y_N)'$, $\mathbf{A} = \text{diag}(\alpha_0, \dots, \alpha_V)'$, and $\mathbf{\Omega} = \text{diag}(\omega_1, \dots, \omega_N)$, we have

$$(\omega_i | -) \sim \text{PG}(1, \mathbf{x}'_i \boldsymbol{\beta}), \quad (\boldsymbol{\beta} | -) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = (\mathbf{A} + \mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}'(\mathbf{y} - 1/2)$. To sample from the Pólya-Gamma distribution, a random sample from which can be generated as a weighted sum of an infinite number of *iid* gamma random variables, we follow Zhou (2016) to truncate the infinite sum to the summation of M gamma random variables, where the parameters of the M th gamma random variable are adjusted to match the mean and variance of the finite sum with those of the infinite sum. We set $M = 5$ in this paper.

B.2. Mean-Field Variational Inference with Diagonal Covariance Matrix

We choose a fully factorized Q distribution as

$$Q = \left[\prod_i q(\omega_i) \right] \left[\prod_v q(\beta_v) \right].$$

To exploit conjugacy, defining

$$\begin{aligned} q(\omega_i) &= \text{PG}(1, \lambda_i), \\ q(\beta_v) &= \mathcal{N}(\mu_v, \sigma_v^2), \end{aligned}$$

we have closed-form coordinate ascent variational inference update equations as

$$\begin{aligned} \lambda_i &= \sqrt{\mathbb{E}[(\mathbf{x}'_i \boldsymbol{\beta})^2]} = \sqrt{\mathbf{x}'_i \mathbb{E}[\boldsymbol{\beta} \boldsymbol{\beta}'] \mathbf{x}_i}, \\ \sigma_v^2 &= \left(\mathbb{E}[\alpha_v] + \sum_i \mathbb{E}[\omega_i] x_{iv}^2 \right)^{-1} \\ \mu_v &= \sigma_v^2 \sum_i x_{iv} \left\{ y_i - 1/2 - \mathbb{E}[\omega_i] \sum_{\tilde{v} \neq v} x_{i\tilde{v}} \mathbb{E}[\beta_{\tilde{v}}] \right\}, \end{aligned}$$

where the expectations with respect to the q distributions can be expressed as $\mathbb{E}[\boldsymbol{\beta} \boldsymbol{\beta}'] = \boldsymbol{\mu} \boldsymbol{\mu}' + \text{diag}(\sigma_0^2, \dots, \sigma_V^2)$ and $\mathbb{E}[\omega_i] = \tanh(\lambda_i/2)/(2\lambda_i)$.

B.3. Mean-Field Variational Inference with Full Covariance Matrix

We choose a fully factorized Q distribution as

$$Q = \left[\prod_i q(\omega_i) \right] q(\boldsymbol{\beta}).$$

To exploit conjugacy, defining

$$\begin{aligned} q(\omega_i) &= \text{PG}(1, \lambda_i), \\ q(\boldsymbol{\beta}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{aligned}$$

we have closed-form coordinate ascent variational inference update equations as

$$\begin{aligned} \lambda_i &= \sqrt{\mathbb{E}[(\mathbf{x}'_i \boldsymbol{\beta})^2]} = \sqrt{\mathbf{x}'_i \mathbb{E}[\boldsymbol{\beta} \boldsymbol{\beta}'] \mathbf{x}_i}, \\ \boldsymbol{\Sigma} &= (\mathbb{E}[\mathbf{A}] + \mathbf{X}' \mathbb{E}[\boldsymbol{\Omega}] \mathbf{X})^{-1}, \quad \boldsymbol{\mu} = \boldsymbol{\Sigma} \mathbf{X}' (\mathbf{y} - 1/2), \end{aligned}$$

where the expectations with respect to the q distributions can be expressed as $\mathbb{E}[\boldsymbol{\beta} \boldsymbol{\beta}'] = \boldsymbol{\mu} \boldsymbol{\mu}' + \boldsymbol{\Sigma}$ and $\mathbb{E}[\omega_i] = \tanh(\lambda_i/2)/(2\lambda_i)$. Note the update equations shown above are identical to those shown in Jaakkola & Jordan (2000).

B.4. SIVI Configuration

For inputs in Algorithm 1, we choose a multi-layer perceptron with layer size [100, 200, 100] as T_ϕ for $\boldsymbol{\psi} = T_\phi(\boldsymbol{\epsilon})$, $\boldsymbol{\epsilon}$ as 50 dimensional isotropic Gaussian random variable and $K = 100$, $J = 50$. For the explicit layer, we choose an MVN as $q_\xi(\mathbf{z} | \boldsymbol{\psi}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\psi}, \boldsymbol{\xi})$. In this setting, $\boldsymbol{\psi}$ is the mean variable mixed with implicit distribution $q_\phi(\boldsymbol{\psi})$ while $\boldsymbol{\xi}$ is the covariance matrix which can be either diagonal or full. In the experiments, we update the neural network parameter ϕ by the Adam optimizer, with learning rate 0.01. We update $\boldsymbol{\xi}$ by gradient ascent, with step size $\eta_t = 0.001 * 0.9^{\text{iteration}/100}$. The implicit layer parameter ϕ and explicit layer parameter $\boldsymbol{\xi}$ are updated iteratively.

C. Experimental Settings and Results for SIVAE

We implement SIVI with $M = 3$ stochastic hidden layers, with the dimensions of hidden layers $[\ell_1, \ell_2, \ell_3]$ as [150, 150, 150] and with the dimensions of injected noises $[\epsilon_1, \epsilon_2, \epsilon_3]$ as [150, 100, 50]. Between two adjacent stochastic layers there is a fully connected deterministic layer with size 500 and *ReLU* activation function. We choose binary pepper and salt noise (Im et al., 2017) for $q_t(\boldsymbol{\epsilon})$. The model is trained for 2000 epochs with mini-batch size 200 and step-size $0.001 * 0.75^{\text{epoch}/100}$. K_t is gradually increased from 1 to 100 during the first 1500 epochs. The explicit and implicit layers are trained iteratively. Warm-up is used during the first 300 epochs as suggested by Sønderby et al. (2016) to gradually impose the prior regularization term $\text{KL}(q_\phi(\mathbf{z} | \mathbf{x}) || p(\mathbf{z}))$. The model is trained end-to-end using the Adam optimizer. After training process, as in Rezende et al. (2014) and Burda et al. (2015), we compute the marginal likelihood for test set by importance sampling with $S = 2000$:

$$\log p(\mathbf{x}) \approx \log \frac{1}{S} \sum_{s=1}^S \frac{p(\mathbf{x} | \mathbf{z}_s) p(\mathbf{z}_s)}{\hat{h}(\mathbf{z}_s | \mathbf{x})}, \quad \mathbf{z}_s \sim h(\mathbf{z}_s | \mathbf{x}),$$

where

$$\hat{h}(\mathbf{z}_s | \mathbf{x}) = \frac{1}{M} \sum_{k=1}^M q(\mathbf{z}_s | \boldsymbol{\psi}^{(k)}), \quad \boldsymbol{\psi}^{(k)} \stackrel{iid}{\sim} q_\phi(\boldsymbol{\psi} | \mathbf{x})$$

is used to estimate $h(\mathbf{z}_s | \mathbf{x})$; we set $M = 100$. The performance of SIVI and a comparison to reported results with other methods are provided in Table 2.

Table 2. Comparison of the negative log evidence between various algorithms.

Methods	$-\log p(\mathbf{x})$
<i>Results below form Burda et al. (2015)</i>	
VAE + IWAE	= 86.76
IWAE + IWAE	= 84.78
<i>Results below form Salimans et al. (2015)</i>	
DLGM + HVI (1 leapfrog step)	= 88.08
DLGM + HVI (4 leapfrog step)	= 86.40
DLGM + HVI (8 leapfrog steps)	= 85.51
<i>Results below form Rezende & Mohamed (2015)</i>	
DLGM+NICE (Dinh et al., 2014) (k = 80)	≤ 87.2
DLGM+NF (k = 40)	≤ 85.7
DLGM+NF (k = 80)	≤ 85.1
<i>Results below form Gregor et al. (2015)</i>	
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
<i>Results below form Maaløe et al. (2016)</i>	
Auxiliary VAE (L=1, IW=1)	≤ 84.59
<i>Results below form Mescheder et al. (2017)</i>	
VAE + IAF (Kingma et al., 2016)	≈ 84.9 ± 0.3
Auxiliary VAE (Maaløe et al., 2016)	≈ 83.8 ± 0.3
AVB + AC	≈ 83.7 ± 0.3
SIVI (3 stochastic layers)	= 84.07
SIVI (3 stochastic layers)+ IW($\tilde{K} = 10$)	= 83.25

D. Additional Figures

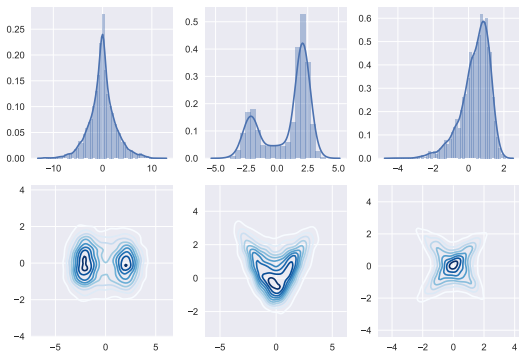


Figure 8. Visualization of the MLP based implicit distributions $\psi \sim q(\psi)$, which are mixed with isotropic Gaussian (or Log-Normal) distributions to approximate the target distributions shown in Figure 1.

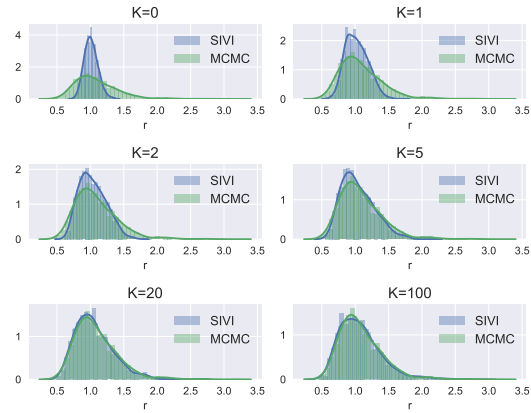


Figure 9. The marginal posterior distribution of the negative binomial dispersion parameter r inferred by SIVI becomes more accurate as K increases

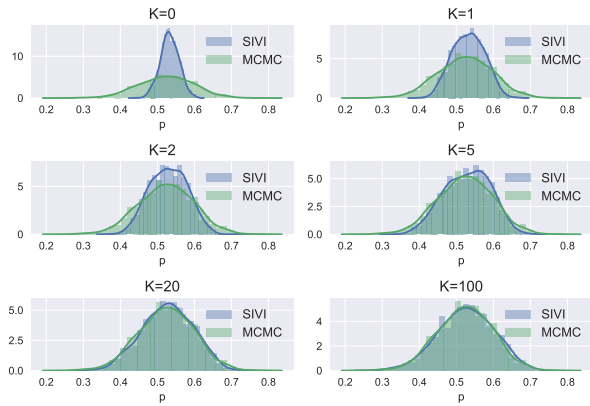


Figure 10. The marginal posterior distribution of the negative binomial probability parameter p inferred by SIVI becomes more accurate as K increases.

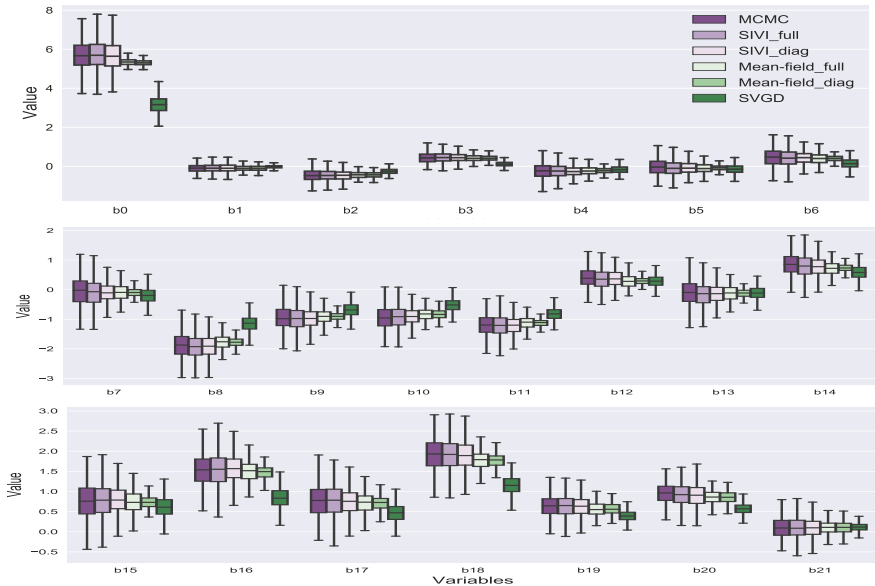


Figure 11. Comparison of all marginal posteriors of β_v inferred by various methods for Bayesian logistic regression on *waveform*.

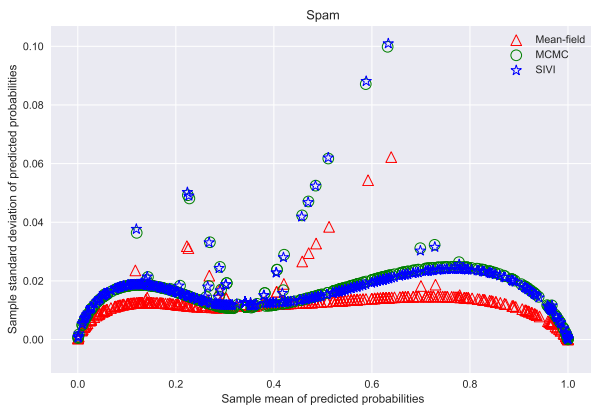


Figure 12. Sample means and standard deviations of predictive probabilities for dataset *spam*.

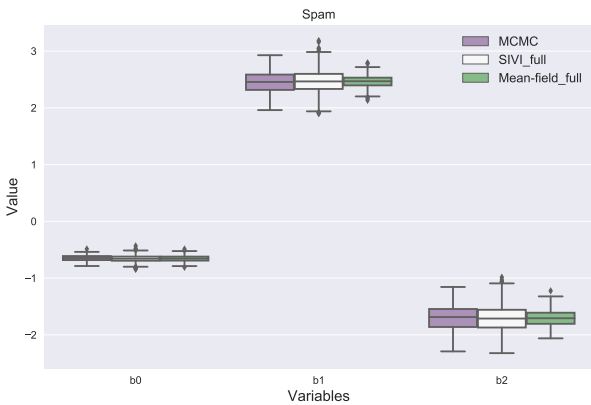


Figure 13. Boxplot of marginal posteriors inferred by MCMC, SIVI, and MFVI for dataset *spam*.

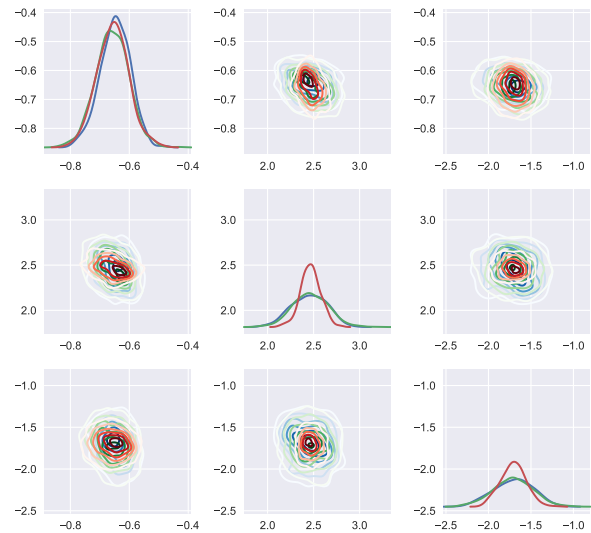


Figure 14. Univariate marginal and pairwise joint posteriors for dataset *spam*. Blue, green, and red are for MCMC, SIVI with a full covariance matrix, and MFVI with a full covariance matrix, respectively.

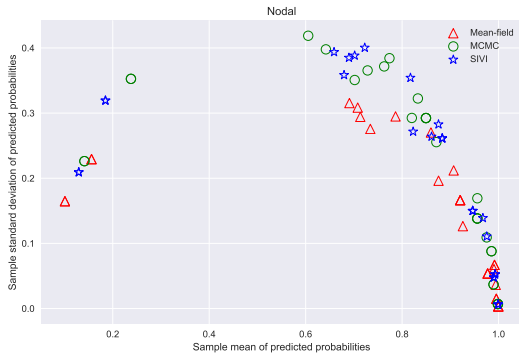


Figure 15. Sample means and standard deviations of predictive probabilities for dataset *nodal*.

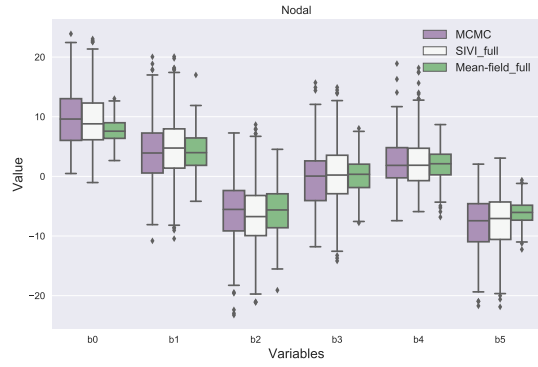


Figure 16. Boxplot of marginal posteriors inferred by MCMC, SIVI, and MFVI for dataset *nodal*.

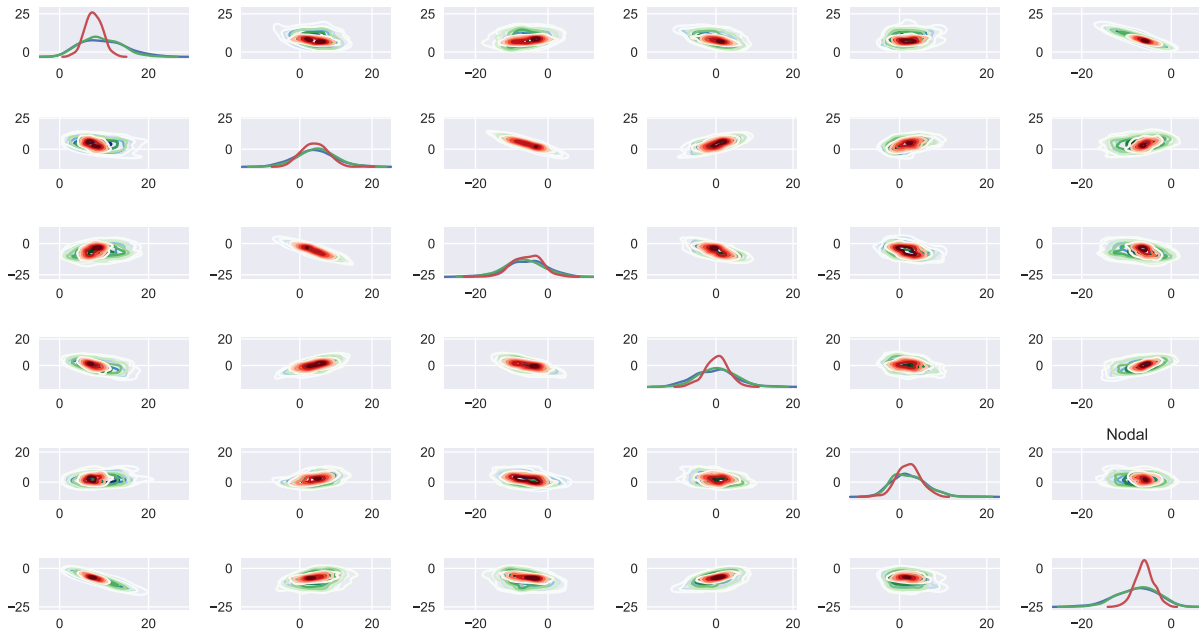


Figure 17. Univariate marginal and pairwise joint posteriors for dataset *nodal*. Blue, green, and red are for MCMC, SIVI with a full covariance matrix, and MFVI with a full covariance matrix, respectively.