Variational Inference based on Robust Divergences

Futoshi Futami[1,2] **Issei Sato**[1,2] **Masashi Sugiyama**[2,1]

[1]The University of Tokyo, [2]RIKEN

Abstract

Robustness to outliers is a central issue in real-world machine learning applications. While replacing a model to a heavy-tailed one (e.g., from Gaussian to Student-t) is a standard approach for robustification, it can only be applied to simple models. In this paper, based on Zellner's optimization and variational formulation of Bayesian inference, we propose an outlier-robust pseudo-Bayesian variational method by replacing the Kullback-Leibler divergence used for data fitting to a robust divergence such as the β- and γ-divergences. An advantage of our approach is that superior but complex models such as deep networks can also be handled. We theoretically prove that, for deep networks with ReLU activation functions, the *influence function* in our proposed method is bounded, while it is unbounded in the ordinary variational inference. This implies that our proposed method is robust to both of input and output outliers, while the ordinary variational method is not. We experimentally demonstrate that our robust variational method outperforms ordinary variational inference in regression and classification with deep networks.

1 Introduction

Robustness is a fundamental topic in machine learning and statistics. Although specific definitions of robustness may be problem-dependent, a commonly shared notion is "*an insensitivity to small deviations from the assumptions*", according to the seminal book by Huber and Ronchetti [2011]. Robustness to outliers is getting more important these days since recent advances in sensor technology give a vast amount of data with spiky noise and crowd-annotated data is full of human errors (Raykar et al. [2010], Zhang et al. [2016], Liu et al. [2012], Bonald and Combes [2017]).

A standard approach to robust machine learning is a *model-based* method, which uses a heavier-tailed distribution such as the Student-t distribution instead of the Gaussian distribution as a likelihood function (Murphy [2012]). However, as pointed out in Wang et al. [2017], the model-based method is applicable only to a simple modeling setup.

To handle more complex models, we employ the optimization and variational formulation of Bayesian inference by Zellner [1988]. In this formulation, the posterior model is optimized to fit data under the Kullback-Leibler (KL) divergence, while it is regularized to be close to the prior. In this paper, we propose replacing the KL divergence for data fitting to a robust divergence, such as the β-divergence (Basu et al. [1998]) and the γ-divergence (Fujisawa and Eguchi [2008]).

Another robust Bayesian inference method proposed by Ghosh and Basu [2016] follows a similar line to our method, which adopts the β-divergence for pseudo-Bayesian inference. They rigorously analyzed the statistical efficiency and robustness of the method, and numerically illustrated its behavior for the Gaussian distribution.

Our work can be regarded as an extension of their work to variational inference so that more complex models such as deep networks can be handled. For deep networks with ReLU activation functions, we prove that the *influence function* (IF) (Huber and Ronchetti [2011]) of our proposed inference method is bounded, while it is unbounded in the ordinary variational inference. This implies that our proposed method is robust to both input and output outliers, while the ordinary variational method is not.

In Wang et al. [2017], another robust Bayesian inference method based on a *weighted likelihood* was proposed, where weights are drawn from their prior distribution. They also conducted IF analysis and showed that IF is bounded *asymptotically.* On the other hand,

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

our method is guaranteed to have a bounded IF for finite samples. In addition, by using IF, we numerically show that influence to the predictive distribution by outliers is also bounded in our proposed method.

Finally, we experimentally demonstrate that our robust variational method outperforms ordinary variational inference in regression and classification with neural networks.

2 Preliminaries

In this section, we review preliminary materials on statistical inference.

2.1 Maximum Likelihood Estimation and Its Robust Variants

Let us consider the problem of estimating an unknown probability distribution[1] $p^*(x)$ from its independent samples $x_{1:N} = \{x_i\}_{i=1}^N$. To this end, we consider a parametric model $p(x;\theta)$ with parameter θ, and minimize the generalization error measured by the KL divergence D_{KL} from $p^*(x)$ to $p(x;\theta)$:

$$D_{\mathrm{KL}}\left(p^*(x)\|p(x;\theta)\right) = \int p^*(x)\log\left(\frac{p^*(x)}{p(x;\theta)}\right)dx. \quad (1)$$

However, since $p^*(x)$ is unknown in practice, it is replaced with

$$D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x;\theta)\right) = \text{Const.} - \frac{1}{N}\sum_{i=1}^{N}\ln p(x_i;\theta), \quad (2)$$

where $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N \delta(x, x_i)$, is the empirical distribution and δ is the Dirac delta function. Minimizing this empirical Kullback-Leibler divergence is equivalent to *maximum likelihood estimation.* Equating the partial derivative of Eq.(2) with respect to θ to zero, we obtain the following estimating equation:

$$0 = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_i;\theta). \quad (3)$$

Maximum likelihood estimation is sensitive to outliers because it treats all data points equally. To circumvent this problem, outlier-robust divergence estimation has been developed in statistics.

The *density power divergence*, which is also known as the β-divergence, is a vital example (Basu et al. [1998]). The β-divergence from functions g to f is defined as

$$\begin{aligned} D_\beta\left(g\|f\right) &= \frac{1}{\beta}\int g(x)^{1+\beta}dx \\ &-\frac{\beta+1}{\beta}\int g(x)f(x)^\beta dx + \int f(x)^{1+\beta}dx. \end{aligned} \quad (4)$$

The γ-divergence (Fujisawa and Eguchi [2008]) is another family of robust divergences:

$$\begin{aligned} D_\gamma\left(g\|f\right) &= \frac{1}{\gamma(1+\gamma)}\ln\int g(x)^{1+\gamma}dx \\ &-\frac{1}{\gamma}\ln\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\ln\int f(x)^{1+\gamma}dx. \end{aligned} \quad (5)$$

In the limit of $\beta \to 0$ and $\gamma \to 0$, both the β- and γ-divergences converge to the KL divergence:

$$\lim_{\beta\to 0} D_\beta\left(g\|f\right) = \lim_{\gamma\to 0} D_\gamma\left(g\|f\right) = D_{\mathrm{KL}}(g\|f). \quad (6)$$

Similarly to maximum likelihood estimation, minimizing the β-divergence (or the γ-divergence) from empirical distribution $\hat{p}(x)$ to $p(x;\theta)$ gives an empirical estimator:

$$\arg\min_\theta D_\beta\left(\hat{p}(x)\|p(x;\theta)\right). \quad (7)$$

This yields the following estimating equation:

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N}p(x_i;\theta)^\beta\frac{\partial}{\partial\theta}\ln p(x_i;\theta) \\ &- \mathbb{E}_{p(x;\theta)}\left[p(x;\theta)^\beta\frac{\partial}{\partial\theta}\ln p(x_i;\theta)\right], \end{aligned} \quad (8)$$

where the second term assures the unbiasedness of the estimator. The first term in Eq.(8) is the likelihood weighted according to the power of the probability density for each data point. Since the probability densities of outliers are usually much smaller than those of inliers, those weights effectively suppress the likelihood of outliers.

When $\beta = 0$, all weights become one and thus Eq.(8) is reduced to Eq.(3). Therefore, adjusting β corresponds to controlling the trade-off between robustness and efficiency. See Appendices A and B for more details.

Eqs.(3) and (8) are called an M-estimator, and Eq.(8) is also called a Z-estimator (Huber and Ronchetti [2011], Basu et al. [1998], Van der Vaart [1998]). In various machine learning applications, those methods showed superior performance (Narayan et al. [2015], Samek et al. [2013], Cichocki et al. [2011]).

2.2 Bayesian Inference and Variational Methods

In Bayesian inference, parameter θ is regarded as a random variable, having prior distribution $p(\theta)$. With

[1] Although we focus on estimating density $p^*(x)$, almost the same discussion is possible for estimating conditional density $p^*(y|x)$, as explained in Section 3.

Bayes' theorem, the Bayesian posterior distribution $p(\theta|x_{1:N})$ can be obtained as

$$p(\theta|x_{1:N}) = \frac{p(x_{1:N}|\theta)p(\theta)}{p(x_{1:N})}. \tag{9}$$

Zellner [1988] showed that $p(\theta|x_{1:N})$ can also be obtained by solving the following optimization problem: [2]

$$\mathop{\arg\min}_{q(\theta)\in\mathcal{P}} L(q(\theta)), \tag{10}$$

where $\mathcal{P}$ is the set of all probability distributions, $-L(q(\theta))$ is the *evidence lower-bound* (ELBO),

$$\begin{aligned} L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) \\ - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta, \end{aligned}$$

and $d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)$ denotes the *cross-entropy*:

$$d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) = -\frac{1}{N}\sum_{i=1}^{N} \ln p(x_i|\theta). \tag{11}$$

Note that Bayes posterior (9) can be expressed as

$$p(\theta|x_{1:N}) = \frac{e^{-N d_{\mathrm{KL}}(\hat{p}(x)\|p(x|\theta))}p(\theta)}{p(x_{1:N})}. \tag{12}$$

In practice, the optimization problem of Eq.(10) is often intractable analytically, and thus we need to use some approximation method. A popular approach is to restrict the domain of the optimization problem to a set of analytically tractable probability distributions $\mathcal{Q}$. Let us denote such a tractable distribution as $q(\theta;\lambda) \in \mathcal{Q}$, where λ is a parameter. Then the optimization problem is expressed as

$$\mathop{\arg\min}_{q(\theta;\lambda)\in\mathcal{Q}} L(q(\theta;\lambda)). \tag{13}$$

This optimization problem is called *variational inference*(VI).

3 Robust Variational Inference based on Robust Divergences

In this section, we propose a robust variational inference method based on robust divergences.

As detailed in Appendix C, Eq.(10) can be equivalently expressed as

$$\mathop{\arg\min}_{q(\theta)\in\mathcal{P}} \mathbb{E}_{q(\theta)}[D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)] + \frac{1}{N} D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right). \tag{14}$$

The first term can be regarded as the expected likelihood (see Eq.(2)), while the second term "regularizes" $q(\theta)$ to be close to prior $p(\theta)$.

To enhance robustness to data outliers, let us replace the KL divergence in the expected likelihood term with the β-divergence:

$$\mathop{\arg\min}_{q(\theta)\in\mathcal{P}} \mathbb{E}_{q(\theta)}[D_{\beta}\left(\hat{p}(x)\|p(x|\theta)\right)] + \frac{1}{N} D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right). \tag{15}$$

Note that Eq.(15) can be equivalently expressed as

$$\mathop{\arg\min}_{q(\theta)\in\mathcal{P}} L_{\beta}(q(\theta)), \tag{16}$$

where $-L_{\beta}(q(\theta))$ is the β-ELBO defined as

$$\begin{aligned} L_{\beta}(q(\theta) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) \\ - \int q(\theta)\left(-N d_{\beta}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta, \end{aligned} \tag{17}$$

and $d_{\beta}(\hat{p}(x)\|p(x|\theta))$ denotes the β-cross-entropy:

$$\begin{aligned} d_{\beta}(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N} p(x_i|\theta)^{\beta} \\ + \int p(x|\theta)^{1+\beta} dx. \end{aligned}$$

For its solution, we have the following theorem (its proof is available in Appendix D):

Theorem 1 *The solution of Eq.(15) is given by*

$$q(\theta) = \frac{e^{-N d_{\beta}(\hat{p}(x)\|p(x|\theta))}p(\theta)}{\int e^{-N d_{\beta}(\hat{p}(x)\|p(x|\theta))}p(\theta)d\theta}. \tag{18}$$

Interestingly, the above expression of $q(\theta)$ is the same as the *pseudo posterior* proposed in Ghosh and Basu [2016]. Although the pseudo posterior is not equivalent to the *posterior distribution* derived by Bayes' theorem, the spirit of updating prior information by observed data is inherited (Ghosh and Basu [2016]). For this reason, we refer to Eq.(18) simply as a *posterior* in this paper. We discuss how prior information is updated in pseudo-Bayes-posteriors in Appendix E.

The optimization problem (15) is generally intractable. Following the same line as the discussion in Section 2.2, let us restrict the set of all probability distributions to a set of analytically tractable parametric distributions, $q(\theta;\lambda) \in \mathcal{Q}$. Then the optimization problem yields

$$\mathop{\arg\min}_{q(\theta;\lambda)\in\mathcal{Q}} L_{\beta}(q(\theta;\lambda)).$$

We call this method *β-variational inference* (β-VI).

[2] Zellner's formulation of Bayesian inference was also used for extending variational inference to constrained methods (Zhu et al. [2014], Koyejo and Ghosh [2013]).

Table 1: Cross-entropies for robust variational inference.

	Unsupervised	Supervised
β	$-\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N}p(x_i\|\theta)^{\beta}+\int p(x\|\theta)^{1+\beta}dx$	$-\frac{\beta+1}{\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}p(y_i\|x_i,\theta)^{\beta}\right\}+\left\{\frac{1}{N}\sum_{i=1}^{N}\int p(y\|x_i,\theta)^{1+\beta}dy\right\}$
γ	$-\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^{N}\frac{p(x_i\|\theta)^{\gamma}}{\{\int p(x\|\theta)^{1+\gamma}dx\}^{\frac{\gamma}{1+\gamma}}}$	-$\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^{N}\frac{p(y_i\|x_i,\theta)^{\gamma}}{\{\int p(y\|x_i,\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$

We optimize objective function L_β by black-box variational inference method and re-parameterization trick (Ranganath et al. [2014]). In our implementation, we estimate the gradient of the objective function (17) by Monte Carlo sampling.

So far, we focused on the unsupervised learning case and the β-divergence. Actually, we can easily generalize the above discussion to the supervised learning case and also to the γ-divergence, by simply replacing the cross-entropy with a corresponding one shown in Table 1. We denote the objective function for the γ-divergence as L_γ in the same way as Eq.(17). Note that, there are several choices for the γ-cross-entropy, as detailed in Appendix H. Explicit expression of L, L_β, and L_γ are summarized in Appendix F.

4 Influence Function Analysis

In this section, we analyze the robustness of our proposed method based on the *influence function* (IF) (Huber and Ronchetti [2011]). IFs have been used in robust statistics to study how much contamination affects estimated statistics.

4.1 Influence Function

First, we review the notion of IFs. Let G be an empirical cumulative distribution of $\{x_i\}_{i=1}^n$:

$$G\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}\Delta_{x_i}\left(x\right), \tag{19}$$

where Δ_x stands for the point-mass 1 at x. Let $G_{\varepsilon,z}$ be a contaminated version of G at z:

$$G_{\varepsilon,z}(x)=(1-\varepsilon)G(x)+\varepsilon\Delta_z(x), \tag{20}$$

where ε is a contamination proportion. For a statistic T and empirical distribution G, IF at point z is defined as follows (Huber and Ronchetti [2011]):

$$\begin{aligned}\mathrm{IF}\left(z,T,G\right)&=\frac{\partial}{\partial\varepsilon}T\left(G_{\varepsilon,z}(x)\right)\Big|_{\varepsilon=0}\\&=\lim_{\varepsilon\to0}\frac{T\left(G_{\varepsilon,z}(x)\right)-T\left(G(x)\right)}{\varepsilon}.\end{aligned} \tag{21}$$

Intuitively, IF is a relative bias of a statistic caused by contamination at z.

4.2 Derivation of Influence Functions

Now we analyze how posterior distributions derived by VI are affected by contamination. In ordinary VI, we derive a posterior by minimizing Eq.(13). Let us consider an approximate posterior $q(\theta;m)$ parametrized by m. Then the objective function given by Eq.(13) can be regarded as a function of m whose first-order optimality condition yields

$$0=\frac{\partial}{\partial m}L\Big|_{m=m^*}. \tag{22}$$

For notational simplicity, we denote $q(\theta;m^*)$ by $q^*(\theta)$.

Referring to Eq.(21), T corresponds to m^*, and G is approximated empirically by the training dataset in VI. Then substituting Eq.(20) into Eq.(13) and using Eq.(21) and Eq.(22) yield the following theorem (its proof is available in Appendix F):

Theorem 2 *When data contamination is given by Eq.(20), IF of ordinary VI is given by*

$$\left(\frac{\partial^2L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta))+Nl(z)\right], \tag{23}$$

IF of β-VI is given by

$$\left(\frac{\partial^2L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta))+Nl_\beta(z)\right], \tag{24}$$

and IF of γ-VI is given by

$$\left(\frac{\partial^2L_\gamma}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta))+Nl_\gamma(z)\right], \tag{25}$$

where $l(z)$, $l_\beta(z)$, and $l_\gamma(z)$ are defined in Table 2.

Using these expressions, we analyze how estimated variational parameters can be perturbed by outliers. In practice, it is important to calculate $\sup_z|\mathrm{IF}(z,\theta,G)|$, because if it diverges, the model can be sensitive to small contamination of data.

4.3 Influence Function Analysis for Specific Models

In our analysis, we consider two types of outliers—outliers related to input x and outliers related to output y. For true data generating distributions $p^*(x)$ and

Table 2: Influence functions for robust variational inference.

	Unsupervised	Supervised z=(x',y')
$l(z)$	$\ln p\,(z\vert\theta)$	$\ln p\,(y'\vert x',\theta)$
$l_\beta(z)$	$\frac{\beta+1}{\beta}p(z\vert\theta)^\beta - \int p(x\vert\theta)^{1+\beta}dx$	$\frac{\beta+1}{\beta}p(y'\vert x',\theta)^\beta - \int p(y\vert x',\theta)^{1+\beta}dy$
$l_\gamma(z)$	$\frac{\gamma+1}{\gamma}\frac{p(z\vert\theta)^\gamma}{\{\int p(x\vert\theta)^{1+\gamma}dx\}^{\frac{\gamma}{1+\gamma}}}$	$\frac{\gamma+1}{\gamma}\frac{p(y'\vert x',\theta)^\gamma}{\{\int p(y\vert x',\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$

Table 3: Behavior of $\sup_z |\mathrm{IF}(z,W,G)|$ in neural networks, "Regression" and "Classification" indicate the cases of ordinary VI, while "β- and γ-Regression" and "β- and γ-Classification" mean that we used β-VI or γ-VI. "Activation function" means the type of activation functions used. "Linear" means that there is no nonlinear transformation, inputs are just multiplied W and added b. $(x_\mathrm{o}:U, y_\mathrm{o}:U)$ means that IF is unbounded while $(x_\mathrm{o}:B, y_\mathrm{o}:U)$ means that IF is bounded for input related outliers, but unbounded for output related outliers.

Activation function	Regression	β- and γ-Regression	Classification	β- and γ-Classification
Linear	$(x_\mathrm{o}:U, y_\mathrm{o}:U)$	$(x_\mathrm{o}:B, y_\mathrm{o}:B)$	$(x_\mathrm{o}:U)$	$(x_\mathrm{o}:B)$
ReLU	$(x_\mathrm{o}:U, y_\mathrm{o}:U)$	$(x_\mathrm{o}:B, y_\mathrm{o}:B)$	$(x_\mathrm{o}:U)$	$(x_\mathrm{o}:B)$
tanh	$(x_\mathrm{o}:B, y_\mathrm{o}:U)$	$(x_\mathrm{o}:B, y_\mathrm{o}:B)$	$(x_\mathrm{o}:B)$	$(x_\mathrm{o}:B)$

$p^*(y|x)$, input-related outlier x_o does not obey $p^*(x)$ and output-related outlier y_o does not obey $p^*(y|x)$. Below we investigate whether IFs are bounded even when $x_\mathrm{o}\to\infty$ or $y_\mathrm{o}\to\infty$.

Although general IF analysis has been extensively carried out in statistics (Huber and Ronchetti [2011]), few works exist focusing on specific models that we often use in recent machine learning applications. Based on this, we consider neural network models for regression and classification (logistic regression). In neural networks, there are parameters $\theta=\{W,b\}$ where outputs of hidden units are calculated by multiplying W to input and then adding b. Our analysis shows that $\sup_z|\mathrm{IF}(z,b,G)|$ is always bounded (see Appendix I for details), and our exemplary analysis results for $\sup_z|\mathrm{IF}(z,W,G)|$ are summarized in Table 3.

From Table 3, we can confirm that ordinary VI is always non-robust to output-related outliers. As for input-related outliers, ordinary VI is robust for the "tanh"-activation function, but not for the ReLU and linear activation functions. On the other hand, IFs of our proposed method are bounded for all three activation functions including ReLU. We have further conducted IF analysis for the Student-t likelihood, which is summarized in Appendix I.

Actually, in Bayesian inference, what we really want to know in the end is the *predictive distribution* at test point x_test:

$$p(x_\mathrm{test}|x_{1:N}) = \int p(\theta|x_{1:N})p(x_\mathrm{test}|\theta)d\theta \\ \approx \int q^*(\theta)p(x_\mathrm{test}|\theta)d\theta.$$

Therefore, it is important to investigate how the predictive distribution is affected by outliers. If the training dataset is contaminated at a rate of ϵ at point z, we can analyze the effect of such data contamination on the predictive distribution by using IFs of the posterior distribution:

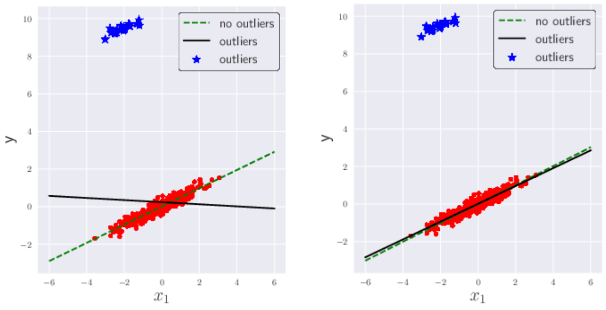
$$\frac{\partial}{\partial\epsilon}\mathbb{E}_{q^*(\theta)}\left[p(x_\mathrm{test}|\theta)\right] = \frac{\partial\mathbb{E}_{q^*(\theta)}\left[p(x_\mathrm{test}|\theta)\right]}{\partial m}\frac{\partial m^*\left(G_{\varepsilon,z}(x)\right)}{\partial\varepsilon}, \tag{26}$$

where $\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial\varepsilon}$ can be analyzed with the IFs derived above. Since analytical discussion on this expression is difficult, we numerically examined its behavior in Section 5.2.

The above expression looks similar to the ones derived in Giordano et al. [2015] and Koh and Liang [2017]. However, discussion in Giordano et al. [2015] focused on prior perturbation and expression in Koh and Liang [2017] is applicable only to maximum likelihood estimation. To our knowledge, ours is the first work to derive IFs of variational inference for data contamination.

5 Experiments

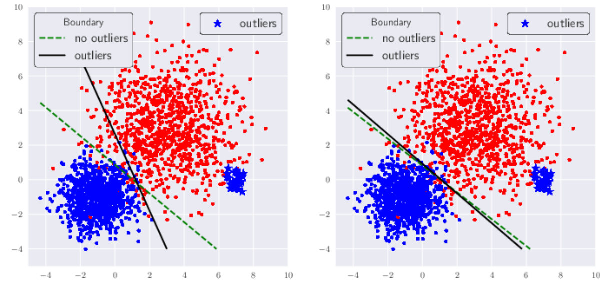
In this section, we report the experimental results of our proposed method on toy and benchmark datasets. In all the experiments, we used mean-field black-box VI combined with the Adam (Kingma and Ba [2014]) optimizer and assumed that the prior and approximated posterior are both Gaussian. Detailed experimental setups can be found in Appendix L.



(a) Ordinary VI with outliers

(b) Proposed VI (β = 0.1) with outliers

Figure 1: Linear regression. Predictive distributions are derived by variational inference (VI).



(a) Ordinary VI with outliers

(b) Proposed VI (β=0.4) with outliers

Figure 2: Boundaries of logistic regression using ordinary VI and the proposed method

5.1 Toy Data Experiment

We performed a toy dataset experiment for both regression and classification tasks to analyze the performance of the proposed method. We used a two-dimensional toy data and observed how the performance and the predictive distribution are affected by outliers when using ordinary VI and our method. The linear regression and logistic regression models are used. The detailed experimental setup is given in Appendix L.1.

For regression, the toy data and predictive distribution are shown in Fig. 1, where the horizontal axis indicates the first input feature x_1 and the vertical axis indicates the output y. As outliers, we considered input related outliers, which are caused by measurement error. The result of ordinary VI is heavily affected by outliers when there exist outliers, while the result of the proposed method is less affected by outliers.

For classification, we considered the situation where some of the labels are wrongly specified, as shown in Fig. 2. We also illustrated obtained decision boundaries in Fig. 2(a), which shows that the ordinary VI based method is heavily affected by outliers and Fig. 2(b) shows that our method with $\beta = 0.4$ is less affected by outliers.

5.2 Influence to Predictive Distribution

Based on Eq.(26), we numerically studied the influence of outliers on the predictive distribution. In this study, we used a two-hidden-layer neural network with 20 units in each hidden layer for regression and for classification with logistic loss.

Regression

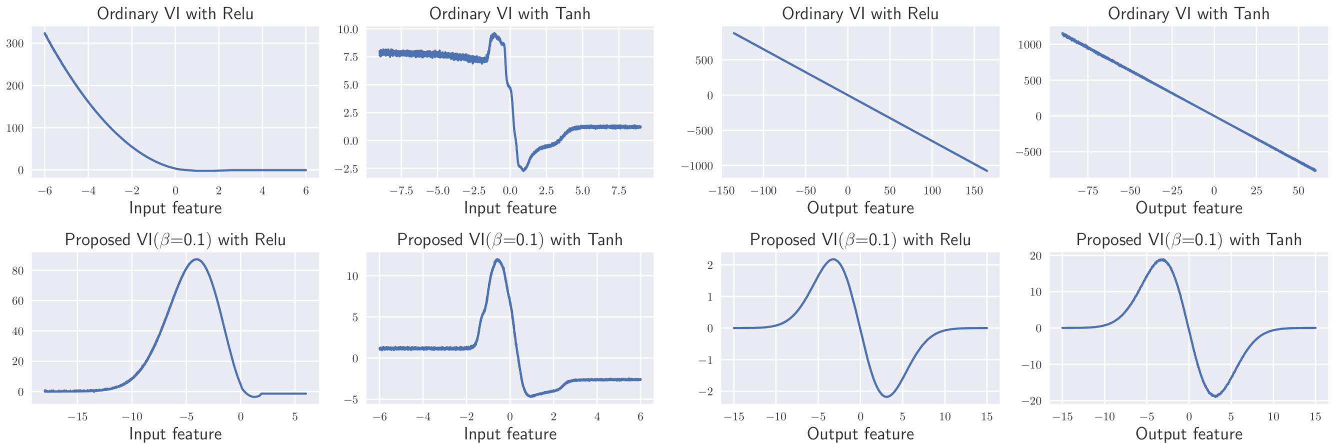
We used the powerplant dataset in UCI (Lichman [2013]) which has four features for each input. Since it is difficult to visualize the behavior of the influence of predictive distributions, instead, we plot how the log-likelihood of a test point is influenced by an outlier. We compared the influence of ordinary VI based method and proposed method (β=0.1). To calculate Eq.(26), we have to specify an outlier and a test data point. As an input related outlier, we randomly chose a single data point from the training data and moved the first feature of the chosen data from $-\infty$ to $+\infty$. Similarly, as an output related outlier, we moved randomly chosen output y from $-\infty$ to $+\infty$. As the test data point, we randomly chose a single data point from the test data. For the detailed experimental setting, see Appendix L.2.

The results are shown in Fig. 3, where the horizontal axis indicates the value of the perturbed feature, and the vertical axis indicates the value of $\frac{\partial}{\partial \epsilon}\mathbb{E}_{q^*(\theta)}\left[\ln p(x_{\text{test}}|\theta)\right]$.

The results in Fig. 3 show that the model using the ReLU activation inferred by ordinary VI can be affected infinitely by input related outliers, while the influence is bounded in our method. As for output related outliers, models inferred by ordinary VI are infinitely influenced, while influence in our method is bounded. From those results, we can see that our method is robust for both input and output related outliers in the sense that test point prediction is not perturbed infinitely by contaminating a single training point.

A notable difference from the IF analysis in Section. 4.3 is that for the perturbation by input related outliers for the tanh activation function, the value of of $\frac{\partial}{\partial \epsilon}\mathbb{E}_{q^*(\theta)}\left[\ln p(x_{\text{test}}|\theta)\right]$, does not converge to zero even for the proposed method in the limit that the absolute value of the input related outlier goes to ∞.

This might be due to the fact that in the limit, the input to the next layer goes to ±1 when the tanh activation function is used. For the next layer, an input



(a) Influence by input related outlier

(b) Influence by output related outlier

Figure 3: Influence on the test log-likelihood for neural net regression. The horizontal axis indicates the value of the perturbed feature, while the vertical axis indicates the value of $\frac{\partial}{\partial \epsilon}\mathbb{E}_{q^*(\theta)}\left[\ln p(x_{\text{test}}|\theta)\right]$.

which has value ±1 might not be so strange compared to regular data, and thus it is not regarded as an outlier. Therefore, during the optimization process, the likelihood of input related outliers is not downweighted so much in the robust divergence and the influence of outliers remains non-zero. If we use the ReLU activation function, in the limit, the input to the next layer becomes much larger than the regular data, and thus it is regarded as an outlier.

Classification

We used the eeg dataset in UCI which has 14 features as input. In the same way as the regression experiment, as an input related outlier, we randomly chose a single data point from the training data and moved the third feature of the chosen data from $-\infty$ to $+\infty$. The result of how the test log-likelihood is influenced is given in Fig. 4. For ordinary VI, using the ReLU activation function causes unbounded influence, while our method keeps the influence bounded. We can also confirm that the influence in our method converges to smaller value than that in ordinary VI in the limit even in the case of tanh.

As an output related outlier, we investigated the influence of label misspecification. We flipped one of the labels in the training data and observed how the test log-likelihood changes. By assuming $\epsilon = \frac{1}{N}$, where N is the number of training data, we calculated $\frac{1}{N}\frac{1}{N}\sum_i \frac{1}{N_{\text{test}}}\sum_j \frac{\partial}{\partial \epsilon_i}\mathbb{E}_{q^*(\theta)}\left[\ln p(y^j_{\text{test}}|x^j_{\text{test}},\theta)\right]$, which represents the averaged amount of change in the test log-likelihood, and the term inside the sum over j means the change in the log-likelihood for the jth test data caused by flipping the label of the ith training data. Without IF, this is difficult to calculate because we have to retrain a neural network with flipped data and this is extremely demanding .

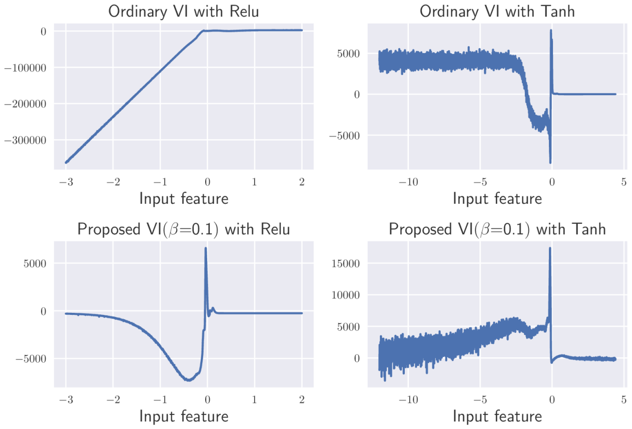


Figure 4: Influence on the test log-likelihood by input related outlier for neural net classification with logistic loss.

Table 4: Average change in the test log-likelihood

	Ordinary VI	Proposed VI ($\beta = 0.1$)
ReLU	-1.65e-3	-3.29e-5
tanh	-2.3e-3	-3.49e-4

Table 4 shows that the change in the test log-likelihood in our method is smaller than that in ordinary VI. This implies that our method is robust against label misspecification.

From these case studies, we confirmed that our method is robust for both input and output related outliers in both regression and classification settings in the sense that the prediction is less influenced by outliers.

Table 5: Test regression accuracy in RMSE

Dataset	Outliers	KL(G)	KL(St)	WL	Rényi	BB-α	β	γ
concrete	0%	8.87(2.57)	**7.34(0.41)**	7.89(0.77)	7.62(0.44)	**7.34(0.31)**	7.58(0.38)	**7.34(0.76)**
N=1030	10%	15.7(2.50)	8.94(2.65)	12.3(2.41)	14.2(1.74)	11.4(2.69)	**8.11(0.89)**	8.26(0.98)
D=8	20%	16.8(0.70)	11.1(3.78)	14.3(2.91)	15.6(1.90)	11.9(2.64)	**8.15(0.99)**	9.25(1.27)
powerplant	0%	4.41(0.13)	4.43(0.15)	4.46(0.17)	4.48(0.15)	4.38(0.83)	**4.37(0.15)**	4.45(0.17)
N=9568	10%	6.44(1.88)	4.54(0.14)	5.12(0.41)	5.49(0.45)	5.91(1.63)	**4.39(0.14)**	4.47(0.16)
D=4	20%	9.97(4.7)	4.56(1.45)	6.44(0.52)	6.87(1.09)	5.52(1.31)	**4.41(0.15)**	4.53(1.46)
protein	0%	5.61(0.38)	**4.79(0.05)**	5.50(0.62)	5.62(0.25)	4.89(0.05)	4.86(0.05)	**4.79(0.04)**
N=45730	10%	6.13(0.02)	4.92(0.05)	6.13(0.03)	6.11(0.03)	6.13(0.03)	4.91(0.04)	**4.90(0.06)**
D=9	20%	6.14(0.03)	4.98(0.07)	6.14(0.03)	6.12(0.03)	6.10(0.28)	4.96(0.05)	**4.95(0.06)**

Table 6: Test classification accuracy

Dataset	Outliers	KL	KL(ϵ)	WL	Rényi	BB-α	β	γ
spam	0%	90.9(5.8)	91.2(4.4)	89.2(5.7)	90.0(0.7)	92.9(1.5)	**93.3(1.3)**	92.2(0.8)
N=4601	10%	76.5(37.6)	90.0(5.1)	89.1(5.7)	**92.6(1.4)**	91.6(1.4)	92.4(1.2)	92.1(1.1)
D=57	20%	60.6(48.3)	89.8(5.5)	88.3(5.3)	91.6(1.6)	91.6(1.6)	**92.2(1.3)**	91.6(1.4)
eeg	0%	72.8(2.9)	77.7(3.2)	**81.3(2.4)**	68.4(7.9)	77.5(3.3)	75.9(5.5)	80.2(3.4)
N=14890	10%	56.0(2.6)	62.7(0.09)	56.0(2.4)	57.5(9.6)	67.9(8.2)	60.8(8.1)	**72.5(2.6)**
D=14	20%	56.0(2.7)	60.0(7.1)	56.0(2.4)	57.7(2.4)	67.4(8.8)	56.0(2.4)	**72.2(6.4)**
covertype	0%	65.2(8.8)	73.1(6.2)	**73.4(6.3)**	72.0(6.6)	73.2(4.8)	70.5(5.9)	**73.4(6.1)**
N=581012	10%	60.2(16.9)	**74.4(6.2)**	73.7(5.5)	65.4(8.5)	70.6(5.9)	65.7(9.0)	72.4(7.7)
D=54	20%	56.4(18.7)	71.4(10.4)	71.2(7.2)	67.6(9.7)	67.1(8.1)	66.2(9.6)	**72.3(5.9)**

5.3 How to Determine β and γ

Finally we show that by choosing parameters β and γ by cross validation, our method can achieve even better performance compared to ordinary VI and other existing robust methods on several benchmark datasets in UCI. The detailed experimental setup is described in Appendix L.3.

Regression

We used a neural net which has two hidden layers each with 20 units and the ReLU activation function. As outliers, we added both input and output related outliers. The experimental results are summarized in Table 5. In Table 5, "Outliers" means the percentage of outliers in the training dataset we contamined artificially. KL(G) means ordinary VI with the Gaussian likelihood, KL(St) is ordinary VI with the Student-t likelihood, WL means the method proposed in Wang et al. [2017], Rényi is the Rényi divergence minimization method proposed in Li and Turner [2016] and BB-α is the black-box α divergence minimization method proposed in Hernandez-Lobato et al. [2016] and Li and Gal [2017].

Our method compares favorably with ordinary VI and existing robust methods for all the datasets.

Classification

We used a neural net which has two hidden layers each with 20 units except for the covertype dataset. For the covertype dataset, we used a neural net which has one hidden layer with 50 units. We used the ReLU activation function for all the networks. As outliers, we considered both input and output related outliers. The experimental results are in Table 6. In Table 6, KL(ϵ) means that we used the robust loss function which is $p(y=1|g(x,\theta)) = \epsilon+(1-2\epsilon)\sigma(g(x,\theta))$, where σ is the sigmoid function, $g(x,\theta)$ is the input to the final layer and ϵ is the hyperparameter.

Our method performs equally to or better than ordinary VI and other existing methods for all the datasets.

6 Conclusions

In this work, we proposed outlier-robust variational inference based on robust divergences. We can make our estimation robust against outliers without changing models. We also theoretically compared our proposed method and the ordinary variational inference by using the influence function. By using the influence function, we can evaluate how much outliers affect our predictions. The analysis showed that the influence of outliers is bounded in our model, but is unbounded by ordinary variational inference in many cases. Further, experiments demonstrated that our method is robust for both input and output related outliers in both regression and classification settings. In addition, our method outperforms ordinary VI on benchmark datasets.

In our future work, we would like to extend the method to be applicable to more complex models, such as time series or structured data. Another way is to combine other approximation methods such as MCMC or expectation propagation.

Acknowledgements

This work was supported by KAKENHI 17H00757 and JST CREST JPMJCR1403.

References

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Thomas Bonald and Richard Combes. A minimax optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 4355–4363, 2017.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9): 2053–2081, 2008.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68 (2):413–437, Apr 2016.

Ryan Giordano, Tamara Broderick, and Michael Jordan. Robust inference with variational bayes. *arXiv preprint arXiv:1512.02578*, 2015.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, USA, 20–22 Jun 2016.

P.J. Huber and E.M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2011.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, Sydney, Australia, 06–11 Aug 2017.

Oluwasanmi Koyejo and Joydeep Ghosh. *Constrained Bayesian inference for low rank multitask learning*, pages 341–350. 2013.

Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2052–206, Sydney, Australia, 06–11 Aug 2017.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 692–700. Curran Associates, Inc., 2012.

Kevin P Murphy. Machine learning: A probabilistic perspective. 2012.

Karthik Narayan, Ali Punjani, and Pieter Abbeel. Alpha-beta divergences discover micro and macro structures in data. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning*, pages 796–804, 2015.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

Wojciech Samek, Duncan Blythe, Klaus-Robert Müller, and Motoaki Kawanabe. Robust spatial filtering with beta divergence. In *Advances in Neural Information Processing Systems*, pages 1007–1015, 2013.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with Bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3646–3655, Sydney, Australia, 06–11 Aug 2017.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4): 278–280, 1988.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(1):3537–3580, 2016.

Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research*, 15:1799–1847, 2014.