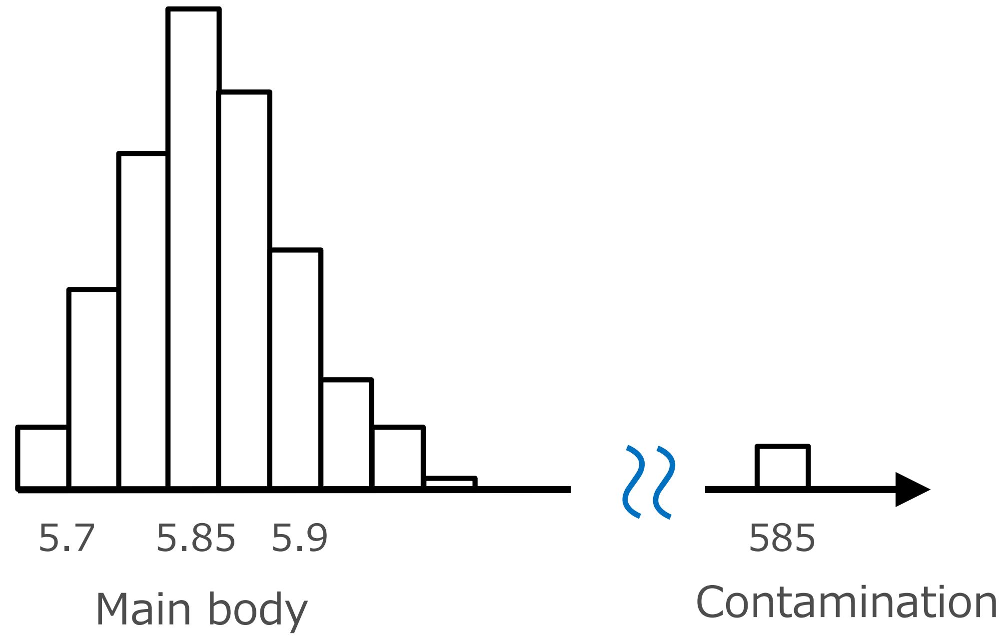
Variational Inference based on Robust Divergences

Futoshi Futami (1, 2) Issei Sato (1, 2) Masashi Sugiyama (2, 1)
(1) The University of Tokyo (2) RIKEN

Advances in Approximate Bayesian Inference
December 8, 2017

What is the outlier-robust inference?



Samples are generated from some unknown distribution.

$$\{x_i\}_{i=1}^{N} \sim p^*(x) \qquad p^*(x) = (1-\varepsilon)p_0^*(x) + \varepsilon\delta(x)$$

Main body Contamination

We aim at placing an estimated probability distribution close to the main body of the unknown distribution.

Maximum likelihood estimation

- Estimate $p^*(x)$ by using $p(x;\theta)$.
- Generalization error is measured by KL divergence:

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

Empirical approximation

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x, x_i),$$

$$\arg\min_{\theta} D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x;\theta)\right) \Rightarrow 0 = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_i;\theta).$$

Maximum likelihood estimation is sensitive to outliers because it treats all data points equally.

Robust divergences

β divergence (density power divergence)

$$D_\beta\left(g \| f\right) = \frac{1}{\beta} \int g(x)^{1+\beta} dx + \frac{\beta+1}{\beta} \int g(x) f(x)^{\beta} dx + \int f(x)^{1+\beta} dx$$

(Basu et al. [1998])

γ divergence

$$D_\gamma\left(g \| f\right) = \frac{1}{\gamma(1+\gamma)} \ln \int g(x)^{1+\gamma} dx - \frac{1}{\gamma} \ln \int g(x) f(x)^{\gamma} dx + \frac{1}{1+\gamma} \ln \int f(x)^{1+\gamma} dx$$

(Fujisawa and Eguchi. [2008])

Robust divergence minimization

Minimizing empirical β or γ divergence instead of KL

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

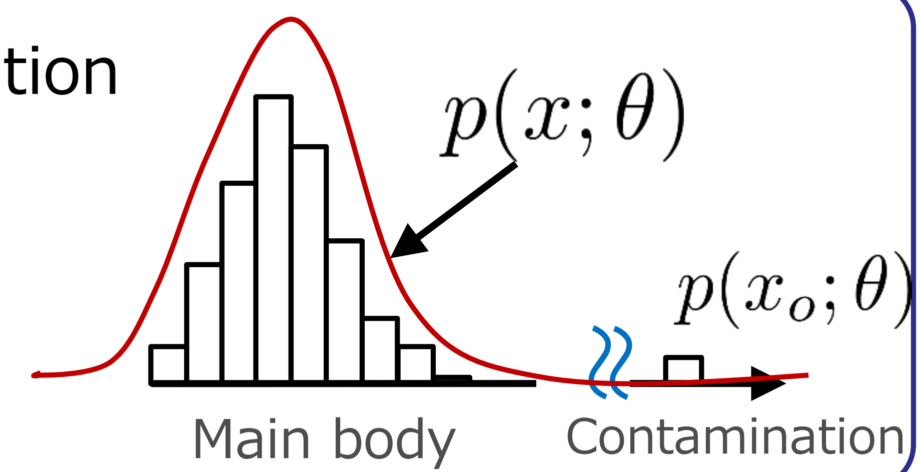
$$\Rightarrow \quad 0 = \frac{1}{N}\sum_{i=1}^{N} \underline{p(x_i;\theta)^{\beta}} \frac{\partial}{\partial\theta} \ln p(x_i;\theta) - \mathbb{E}_{p(x;\theta)}\left[p(x;\theta)^{\beta}\frac{\partial}{\partial\theta}\ln p(x_i;\theta)\right]$$

Density power weights

The likelihood weighted according to the power of the probability for each data point.

We want to model the distribution of the main body of data.

Outliers x_o have small $p(x_o;\theta)$.



Bayesian inference (reformulation)

θ : random variable

$p(\theta)$:prior distribution

Bayes' theorem

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

Reformulation of Bayesian inference

Zellner[1988], Zhu et al. [2014]

$$\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta))$$

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

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

Solution

$$q(\theta) = \frac{e^{-N d_{\mathrm{KL}}(\hat{p}(x)\|p(x|\theta))} p(\theta)}{\int e^{-N d_{\mathrm{KL}}(\hat{p}(x)\|p(x|\theta))} p(\theta) d\theta}$$

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

Variational inference

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

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

This is often intractable analytically, we need some approximation method.

Restrict the domain of the optimization problem to analytically tractable distributions

$$q(\theta; m) \in \mathcal{Q}$$

$$\mathop{\arg\min}_{q(\theta;m)\in\mathcal{Q}} L(q(\theta; m))$$

- This method is called variational inference.
- $-L(q(\theta;\lambda))$ is called the evidence lower-bound (ELBO).

Summary up to now

Maximum Likelihood estimation

$$\arg\min_{\theta} D_{\mathrm{KL}}\left(\hat{p}(x) \| p(x;\theta)\right)$$

Robust estimation

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

Bayesian Inference

$$\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

Variational inference

$$\arg\min_{q(\theta;m)\in\mathcal{Q}} L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

?

How to incorporate robust divergence property ?

Interpretation

Bayesian Inference

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

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

Maximum Likelihood estimation

$$\underset{\theta}{\arg\min}\, D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x;\theta)\right)$$

seems like...

$\underset{q(\theta)\in\mathcal{P}}{\arg\min}$ Expected likelihood + Regularization: close to prior

Robust inference

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

$$\Downarrow$$

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

$$\Downarrow$$

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, L_\beta(q(\theta)),$$

$$L_\beta(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right).$$

β Cross entropy: $d_\beta(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N}p(x_i|\theta)^\beta + \int p(x|\theta)^{1+\beta}dx.$

Variational inference based on robust divergence

$$\arg\min_{q(\theta)\in\mathcal{P}} L_\beta(q(\theta)),$$

$$L_\beta(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta.$$

β Cross entropy: $d_\beta(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N} p(x_i|\theta)^\beta + \int p(x|\theta)^{1+\beta} dx.$

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

Kinds of "pseudo posterior"

Conjugate relation is broken in this formulation.
Analytical solution is intractable.

Variational inference based on robust divergence

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

$$L_\beta(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta.$$

β Cross entropy: $d_\beta(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N} p(x_i|\theta)^\beta + \int p(x|\theta)^{1+\beta} dx.$

Let us use variational inference by restricting the domain of the optimization.

$$q(\theta; m) \in \mathcal{Q}$$

Robust variational inference

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

Proposing method

Maximum Likelihood estimation

$$\arg\min_{\theta} D_{\mathrm{KL}}\left(\hat{p}(x) \| p(x;\theta)\right)$$

θ : random variable

Bayesian Inference

$$\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

Variational inference

$$\arg\min_{q(\theta;\lambda)\in\mathcal{Q}} L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

Robust estimation

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

θ : random variable

Robust Inference

$$\arg\min_{q(\theta)\in\mathcal{P}} L_{\beta}(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\beta}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

Variational inference

$$\arg\min_{q(\theta;\lambda)\in\mathcal{Q}} L_{\beta}(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\beta}\left(\hat{p}(x)\|p(x|\theta)\right)\right) d\theta$$

Analysis based on influence function (IF)

(Huber and Ronchetti [2011])

- We can analyze the robustness through IFs.
- IFs represent relative bias of a estimated static caused by outliers.

Empirical distribution : $G(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x, x_i)$

Contaminated version of G at z : $G_{\varepsilon,z}(x) = (1-\varepsilon)G(x) + \varepsilon\delta(x,z)$

ε :contamination proportion

For a static T and empirical distribution G , IF at point z is defined as:

$$\mathrm{IF}(z, T, G) = \frac{\partial}{\partial \varepsilon} T(G_{\varepsilon,z}(x))\Big|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{T(G_{\varepsilon,z}(x)) - T(G(x))}{\varepsilon}.$$

How to use IFs ?

1 . Investigate whether $\underline{\sup_z |\mathrm{IF}\,(z, m, G)\,| < \infty}$ or not.

If it diverges, the model can be sensitive to small contamination of data.

2. How much is the predictive distribution affected by outliers ?

What we want to know is predictive distribution.

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

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

IF of variational inference

m^* : satisfies first order condition $0 = \left.\frac{\partial}{\partial m} L \right|_{m=m^*}$

$q^*(\theta) := q(\theta; m^*)$

T corresponds to the variational parameter m

IF of variational inference

- For usual variational inference,

$$\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial \varepsilon} = \left(\frac{\partial^2 L}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q^*(\theta)} \left[D_{\mathrm{KL}}(q^*(\theta) \| p(\theta)) + N \ln p\left(z|\theta\right)\right]$$

- For β-variational inference,

$$\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial \varepsilon} = \left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q^*(\theta)} \left[D_{\mathrm{KL}}(q^*(\theta) \| p(\theta)) + N \frac{\beta+1}{\beta} p(z|\theta)^\beta - \int p(x|\theta)^{1+\beta} dx\right]$$

IF of some specific models

- Let us investigate whether $\sup_z |\mathrm{IF}(z, m, G)| < \infty$.
- Consider regression and logistic regression for Bayesian neural networks.

Input related outlier : $x_o \not\sim p^*(x)$

Output related outlier : $y_o \not\sim p^*(y|x)$

Behavior of $\sup_z |\mathrm{IF}(z, W, G)|$

Activation function	Regression	β- and γ-Regression	Classification	β- and γ-Classification
Linear	$(x_o : U, y_o : U)$	$(x_o : B, y_o : B)$	$(x_o : U)$	$(x_o : B)$
ReLU	$(x_o : U, y_o : U)$	$(x_o : B, y_o : B)$	$(x_o : U)$	$(x_o : B)$
tanh	$(x_o : B, y_o : U)$	$(x_o : B, y_o : B)$	$(x_o : B)$	$(x_o : B)$

$(x_o : U, y_o : U)$: IF is unbounded.

$(x_o : B, y_o : U)$: IF is bounded for input related outliers, but unbounded for output related outliers.

IF of our proposed method is always bounded.

Experiments on the UCI dataset

- Neural net which has two hidden layers each with 20 units and the ReLU activation function.
- We used the re-parameterization trick with 10 MC samples.
- We determine β or γ by cross-validation. (from 0.1 to 0.9 for the experiment. We found that range from 0.1 to 0.5 is enough.)
- We added outliers to training data with proportion increased from 0% to 20%.

Dataset	Outliers	KL(G)	KL(St)	WL	Réyni	BB-α	β	γ
concrete	0%	7.46(0.34)	7.36(0.4)	8.04(1.01)	7.16(0.39)	7.18(0.30)	7.27(0.28)	**5.53(0.48)**
N=1030	10%	8.58(0.46)	7.63(0.52)	10.37(1.16)	8.04(0.43)	7.37(0.38)	7.58(0.25)	**6.20(0.74)**
D=8	20%	9.40(1.01)	8.37(0.70)	11.46(0.93)	8.63(0.52)	7.81(0.51)	8.50(0.87)	**6.85(1.15)**
powerplant	0%	4.49(0.15)	4.46(0.16)	4.46(0.18)	4.49(0.14)	4.41(0.13)	4.36(0.11)	**4.28(0.14)**
N=9568	10%	4.71(0.17)	4.59(0.15)	4.81(0.23)	4.66(0.19)	4.56(0.17)	4.41(0.16)	**4.33(0.15)**
D=4	20%	5.12(0.26)	4.65(0.10)	5.04(0.25)	4.82(0.23)	4.70(0.13)	4.52(0.15)	**4.38(0.15)**
protein	0%	5.88(0.50)	**4.78(0.07)**	5.77(0.56)	4.82(0.04)	4.81(0.04)	4.87(0.05)	**4.78(0.05)**
N=45730	10%	6.14(0.03)	**4.84(0.06)**	6.14(0.028)	4.88(0.04)	4.86(0.04)	4.96(0.06)	4.86(0.07)
D=9	20%	6.14(0.03)	4.90(0.08)	6.14(0.031)	4.90(0.05)	**4.86(0.05)**	4.97(0.06)	**4.86(0.07)**

Conclusions

1. We proposed an outlier-robust pseudo-Bayesian variational method by replacing the KL divergence used for data fitting to a robust divergence.

2. We analyzed our proposed method by using influence functions analytically and numerically.

3. We confirmed usefulness of our proposed method on the UCI datasets by using Bayesian neural nets.