Robust Bayesian inference via γ-divergence

Tomoyuki Nakagawa[1] Shintaro Hashimoto[2]

[1] Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Noda, Chiba, 278–8510, Japan

[2] Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Hiroshima, 739–8526, Japan

(Last update June 12, 2018)

Abstract

This paper presents the robust Bayesian inference based on the γ-divergence which is the same divergence as "type 0 divergence" in Jones et al. (2001) on the basis of Windham (1995). It is known that the minimum γ-divergence estimator works well to estimate the probability density for heavily contaminated data, and to estimate the variance parameters. In this paper, we propose a robust posterior distribution against outliers based on the γ-divergence and show the asymptotic properties of the proposed estimator. We also discuss some robustness properties of the proposed estimator and illustrate its performances in some simulation studies.

Keywords: Bayes estimator; density power divergence; γ-divergence; quasi-posterior; robust estimation

Mathematics Subject Classification: Primary 62F15; Secondary 62F35

1 Introduction

It is well-known that the maximum likelihood estimator (MLE) has a large bias in the presence of outliers. The robustness against outliers has been investigated in many aspects such as influence function and breakdown point (for the details, see Hampel et al. (1986) and Huber (1981)). As another approach, the robust estimators based on the divergences have been developed to reduce the effect of outliers. A pioneering work of the robust estimation based on the divergence was given by Basu et al. (1998). They proposed the minimum density power divergence estimator and showed its asymptotic properties and robustness. On the other hand, Jones et al. (2001) proposed another class of estimators in a similar spirit where the identity function was replaced by the logarithmic function (see also Windham (1995)). They referred to the new class of divergences as the class of "type 0" divergences as opposed to the density power divergence being the class of "type 1" divergences, and also compared the properties of the corresponding estimators with those of the minimum density power divergence estimators. Fujisawa and Eguchi (2008) dealt with the same divergence as "type 0" divergence (they call it γ-divergence) and showed that the corresponding estimator has a small latent bias for heavily contaminated data under some conditions.

In the Bayesian context, it is also well-known that the ordinary Bayes estimator under the quadratic loss function (i.e., the posterior mean) is not robust against outliers. The theory of the Bayesian robustness against outliers has been developed in terms of the heaviness of the tails of distributions (see e.g., Dawid (1973), Andrade and O'Hagan (2006), Desgagné (2015)). However, this approach may lead to a loss of precision when the contamination is not present. Recently, Ghosh and Basu (2016) proposed the robust Bayes estimation based on the density power divergence. They discussed in details the mean parameter estimation when the variance parameter was known. They also showed the asymptotic property of the Bayes estimators, and characterized the robustness in terms of the influence function. However, it is known that the estimators based on the density power divergence does not work well the estimation for the variance parameter, and are unstable when the ratio of contamination is not small. These facts were discussed by Fujisawa and Eguchi (2008) in a frequentist viewpoint. The robust estimations based on the γ-divergence have been developed in various models (see e.g., Hirose et al. (2017) and Kawashima and Fujisawa (2017)).

In this paper, we propose a robust posterior distribution based on the γ-divergence, and derive a different property on the estimation of the variance parameter and conclude that the robust Bayesian estimation via the γ-divergence is superior to that via the density power divergence in the sense of the estimation of the variance parameter. Furthermore, in simulation studies, we show that the proposed method is also robust under heavy contamination. This paper is organized as follows: In Section 2, we propose a new robust posterior distribution via the γ-divergence which is called the "γ-posterior" in this paper. The γ-posterior is derived by replacing the likelihood function with the γ-divergence in a similar way to the Ghosh and Basu (2016). In Section 3, it is shown that the γ-posterior and its posterior mean have the asymptotic normalities under some regularity conditions. In Section 4, the influence function for the posterior mean based on the γ-posterior is obtained, and we compare our influence function with that of Ghosh and Basu (2016) in the normal model. The robustness of the prior perturbation in the term of the local sensitivity measure which is proposed by Gustafon and Wasserman (1996) is also discussed. In Section 5, we show that the posterior means based on the γ-posterior numerically outperform those of based on the density power and ordinary posteriors in terms of empirical biases of estimators. Further, by making a comparison of the posterior distributions with or without outliers based on the posterior samples generated by Markov chain Monte Carlo (MCMC) algorithm, our proposal posterior is better than the competitors in terms of the rejection of outliers.

2 Construction of robust posterior distribution

First, we give the fundamental setting for the parameter estimation and introduce the γ-divergence and γ-cross entropy. Then we propose a robust posterior distribution based on the γ-divergence.

Let $X_1, \dots, X_n$ be independent and identically distributed (i.i.d.) random variables according to the probability density function $g(x)$. Let $f(x)$ be underlying probability density function and $\delta(x)$ be the contamination probability density function. Suppose that $g(x)$ is the contaminated probability density function given by $g(x) = (1 - \varepsilon)f(x) + \varepsilon\delta(x)$, where ε is the ratio of contamination. We often use a point mass at x as $\delta(x)$. In this paper, we consider only the case which the contamination density $\delta(x)$ mostly lies on the tail of the underlying density $f(x)$. In other words, for an outlier x_o, it holds that $f(x_o) \approx 0$. It is known that some divergences might have some problems of instability for sampled values very close to zero (for example, the case of exponential distribution with mean 1 in the example of Jones et al. (2001)), and corresponding minimum divergence estimators are statistically useless. However, we suppose that we do not consider such cases in this paper. We now consider the parametric model $f_\theta(x) = f(x;\theta)$ $(\theta \in \Theta)$ as a candidate model, where $\Theta \subset \mathbb{R}^p$ is a parameter space of $\theta = (\theta_1, \dots, \theta_p)^\top$. We assume that the target density is included in the candidate model $\{f_\theta |\ \theta \in \Theta\}$, that is, $f(x)$ is expressed by $f(x) = f_{\theta_0}(x)$ using a parameter $\theta_0 \in \Theta$. Hereafter, we will often omit arguments of functions for simplicity. Jones et al. (2001) proposed the divergence between the probability densities f_θ and g as follows:

$$\begin{aligned} D_\gamma(g, f_\theta) &= \frac{1}{\gamma(\gamma+1)} \log \int g^{1+\gamma} dx - \frac{1}{\gamma} \log \int g f_\theta^\gamma dx + \frac{1}{\gamma+1} \log \int f_\theta^{1+\gamma} dx \\ &= -d_\gamma(g, g) + d_\gamma(g, f_\theta) \quad \text{(say)}, \end{aligned}$$

where $\gamma > 0$ is a tuning parameter on robustness and $d_\gamma(g, f_\theta)$ is called the γ-cross entropy. This divergence is also called "γ-divergence" in Fujisawa and Eguchi (2008). In order to derive the minimum γ-divergence estimator for θ, we may consider the minimization problem

$$\min_{\theta\in\Theta} d_\gamma(g, f_\theta) = \min_{\theta\in\Theta} \left\{ -\frac{1}{\gamma} \log \int g f_\theta^\gamma dx + \frac{1}{\gamma+1} \log \int f_\theta^{1+\gamma} dx \right\}$$

with respect to θ. Though the true density $g(x)$ is unknown, we note that the γ-cross entropy is empirically estimable by $d_\gamma(\bar{g}, f_\theta)$, where $\bar{g}$ is the empirical probability density of $\boldsymbol{X}_n = (X_1, \dots, X_n)$. Then the robust estimator of θ is given by $\arg\min_{\theta\in\Theta} d_\gamma(\bar{g}, f_\theta)$ (see Jones et al. (2001) and Fujisawa and Eguchi (2008)).

Now, we consider the following monotone transformation of the γ-cross entropy

$$\tilde{d}_\gamma(g, f_\theta) = -\frac{1}{\gamma} \left\{ \exp\left(-\gamma d_\gamma(g, f_\theta)\right) - 1 \right\} = -\frac{\frac{1}{\gamma} \int g f_\theta^\gamma dx}{\{\int f_\theta^{1+\gamma} dx\}^{\gamma/(1+\gamma)}} + \frac{1}{\gamma}. \tag{1}$$

Remark 2.1. We note that the second term of the right-hand side of the (1), i.e., "$+1/\gamma$" is necessary to prove (iii) in Proposition 2.1. However, this term is canceled out in the denominator and numerator when we calculate the posterior distribution.

This transformation is essential in this paper (for the details, we will discuss later). Then we give some properties of the transformed γ-cross entropy $\tilde{d}_\gamma(g, f_\theta)$.

Proposition 2.1. *Let g and f be the probability density functions and let κ_1, κ_2 and κ be the positive constants. Then $\tilde{d}_\gamma(g, f)$ have the following properties*

(i) $\tilde{d}_\gamma(\kappa_1 g, \kappa_2 f) = \kappa_1 \tilde{d}_\gamma(g, f) + (1 - \kappa_1)/\gamma$.

(ii) $\tilde{d}_\gamma(g, f) = \tilde{d}_\gamma(g, g)$ *holds if and only if* $f = \kappa g$. *In particular,* $g = f$ *if* g *and* f *are the density functions.*

(iii) $\lim_{\gamma\to 0} \tilde{d}_\gamma(g, f) = -\int g \log f dx = d_{KL}(g, f)$, *where* $d_{KL}(g, f)$ *is the Kullback-Leibler cross entropy between* g *and* f.

Proof. The proof of (i) and (ii) are omitted because they are proved straightforward from the definition of the γ-cross entropy. We only prove (iii). The proof of (iii) be proved by using the Taylor expansion $f^\gamma = 1 + \gamma \log f + O(\gamma^2)$ $(\gamma \to 0)$. Then

$$\frac{1}{\gamma}\int g f^\gamma dx - 1 = \frac{1}{\gamma}\int g(f^\gamma - 1)dx = \int g \log f dx + O(\gamma) \quad (\gamma \to 0),$$
$$\left(\int f^{1+\gamma} dx\right)^{\gamma/(1+\gamma)} = 1 + O(\gamma^2) \quad (\gamma \to 0).$$

Therefore, we have

$$\begin{aligned}\tilde{d}_\gamma(g,f) &= -\frac{1}{\gamma}\left[\frac{\int g f^\gamma dx}{\{\int f^{1+\gamma}dx\}^{\gamma/(1+\gamma)}} - 1\right] \\ &= -\frac{1}{\gamma}\left[\frac{\int g f^\gamma dx - 1}{\{\int f^{1+\gamma}dx\}^{\gamma/(1+\gamma)}} + \frac{1}{\{\int f^{1+\gamma}dx\}^{\gamma/(1+\gamma)}} - 1\right] \\ &\to -\int g \log f dx \quad (\gamma \to 0)\end{aligned}$$

This completes the proof. □

Replacing the true density g with the empirical density $\bar{g}$ of $\boldsymbol{X}_n$, we have

$$\begin{aligned}-n\tilde{d}_\gamma(\bar{g}, f_\theta) &= \sum_{i=1}^n \frac{\frac{1}{\gamma} f_\theta(X_i)^\gamma}{\{\int f_\theta(x)^{1+\gamma}dx\}^{\gamma/(1+\gamma)}} - \frac{n}{\gamma} \\ &= \sum_{i=1}^n q_\theta^{(\gamma)}(X_i) - \frac{n}{\gamma} = Q_n^{(\gamma)}(\theta) \quad \text{(say)},\end{aligned} \tag{2}$$

where

$$q_\theta^{(\gamma)}(x) = q^{(\gamma)}(\theta; x) = \frac{1}{\gamma} f_\theta(x)^\gamma \left\{\int f_\theta(x)^{1+\gamma}dx\right\}^{-\gamma/(1+\gamma)}.$$

We refer to $Q_n^{(\gamma)}(\theta)$ as the γ-*likelihood* which is a kind of quasi-likelihoods (or weighted likelihoods). From (iii) in Proposition 2.1, we have

$$\lim_{\gamma \to 0} Q_n^{(\gamma)}(\theta) = \sum_{i=1}^n \log f_\theta(X_i).$$

Hence, the γ-likelihood is the generalization of the log-likelihood. Here, we show a simple example of the parameter estimation based on the γ-likelihood.

Example 2.1 (Normal distribution)**.** Let $X_1, \dots, X_n$ be i.i.d. random variables from the normal distribution with mean μ and variance σ^2, that is, $\theta = (\mu, \sigma^2)^\top$. In this case, we have $f_\theta(x_i)^\gamma = (1/\sqrt{2\pi\sigma^2})^\gamma \exp\{-\gamma(x_i - \mu)^2/2\sigma^2\}$ and $\{\int f_\theta^{1+\gamma}dx\}^{-\gamma/(1+\gamma)} = \{(2\pi\sigma^2)^{\gamma/2}(1+$

$\gamma)^{1/2}\}^{\gamma/(1+\gamma)}$. Then the γ-likelihood $Q_n^{(\gamma)}$ is given by

$$Q_n^{(\gamma)}(\theta) = \frac{1}{\gamma}\left\{(2\pi\sigma^2)^{\gamma/2}(1+\gamma)^{1/2}\right\}^{\gamma/(1+\gamma)} \sum_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\gamma} \exp\left(-\frac{\gamma(x_i-\mu)^2}{2\sigma^2}\right) - \frac{n}{\gamma}$$

for $\gamma > 0$. To obtain the minimum γ-divergence estimator, we may solve the maximization problem $\max_{\theta\in\Theta} Q_n^{(\gamma)}(\theta)$ with respect to θ. Fujisawa and Eguchi (2008) provided an iterative algorithm to solve this problem. As we can see, the γ-likelihood $Q_n^{(\gamma)}(\theta)$ is slightly different from the log-likelihood function given by $\ell_n(\theta) = -(n\log(2\pi\sigma^2)/2) - \sum_{i=1}^n (x_i-\mu)^2/2\sigma^2$.

In the Bayesian context, the parameter estimation is based on the posterior distribution of parameter θ given by the data $\boldsymbol{X}_n$. The ordinary posterior density is given by

$$\pi(\theta|\boldsymbol{X}_n) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)d\theta} \propto \exp\left\{-nd_{KL}(\bar{g}, f_\theta)\right\}\pi(\theta), \tag{3}$$

where $L_n(\theta) = \prod_{i=1}^n f(x_i;\theta)$ is the likelihood function and $\pi(\theta)$ is the prior density of θ. Then we propose the γ-posterior which is the posterior density based on the γ-likelihood.

Definition 2.1 (γ-posterior)**.** *Let $Q_n^{(\gamma)}(\theta)$ be the γ-likelihood given by* (2)*. We define the γ-posterior by*

$$\begin{aligned}\pi^{(\gamma)}(\theta|\boldsymbol{X}_n) &= \frac{\exp\{-n\tilde{d}_\gamma(\bar{g}, f_\theta)\}\pi(\theta)}{\int \exp\{-n\tilde{d}_\gamma(\bar{g}, f_\theta)\}\pi(\theta)d\theta} \\ &= \frac{\exp\{Q_n^{(\gamma)}(\theta)\}\pi(\theta)}{\int \exp\{Q_n^{(\gamma)}(\theta)\}\pi(\theta)d\theta} = \frac{\prod_{i=1}^n \exp(q_\theta^{(\gamma)}(X_i))\pi(\theta)}{\int \prod_{i=1}^n \exp(q_\theta^{(\gamma)}(X_i))\pi(\theta)d\theta},\end{aligned} \tag{4}$$

where $\pi(\theta)$ is the prior density of θ and $\gamma > 0$ is a tuning parameter on robustness.

The γ-posterior is a kind of quasi-posterior distributions. The applications of quasi-posterior distributions have been developed in recent years (see also Hooker and Vidyashankar (2014) and Ghosh and Basu (2016)). We note that $\pi^{(\gamma)}(\theta|\boldsymbol{X}_n)$ is close to the ordinary posterior density (3) as $\gamma \to 0$. Since the monotone transformed γ-cross entropy $\tilde{d}_\gamma(\bar{g}, f_\theta)$ is additive in $\bar{g}$ for i.i.d. random variables, we can update the posterior density for new data $\boldsymbol{X}_{new} = (X_{n+1}, \ldots, X_m)$. In fact, we have

$$\begin{aligned}\pi^{(\gamma)}(\theta|\boldsymbol{X}_{all}) &\propto \left\{\prod_{i=n+1}^{m} \exp(q_\theta^{(\gamma)}(X_i))\right\}\pi^{(\gamma)}(\theta|\boldsymbol{X}_{old}) \\ &\propto \left\{\prod_{i=n+1}^{m} \exp(q_\theta^{(\gamma)}(X_i))\right\}\left\{\prod_{i=1}^{n} \exp(q_\theta^{(\gamma)}(X_i))\right\}\pi(\theta) \\ &= \left\{\prod_{i=1}^{m} \exp(q_\theta^{(\gamma)}(X_i))\right\}\pi(\theta),\end{aligned}$$

where $\boldsymbol{X}_{old} = (X_1, \ldots, X_n)$ and $\boldsymbol{X}_{all} = (\boldsymbol{X}_{old}, \boldsymbol{X}_{new})$.

Remark 2.2. If we do not use the monotone transformed γ-cross entropy, then we have

$$\pi^{(\gamma)}(\theta|\boldsymbol{X}_{all}) \neq \left\{\prod_{i=1}^{m} \exp(q_\theta^{(\gamma)}(X_i))\right\}\pi(\theta).$$

In other words, we can not carry out the usual Bayesian update.

If we assume the uniform prior distribution for $\pi(\theta)$, the maximum a posteriori (MAP) estimator $\hat{\theta}_{MAP}$ is given by

$$\hat{\theta}_{MAP}^{(\gamma)} = \arg\max_{\theta\in\Theta}\left[\left\{\prod_{i=1}^{n} \exp(q_\theta^{(\gamma)}(X_i))\right\}\pi(\theta)\right].$$

Since the posterior density under the uniform prior is proportional to the likelihood function, the MAP estimator $\hat{\theta}_{MAP}^{(\gamma)}$ is the same as the minimum type 0 or γ-divergence estimator given by Jones et al. (2001) and Fujisawa and Eguchi (2008).

Let $L(\theta, d)$ be the loss function for the decision $d \in \mathcal{D}$, where $\mathcal{D}$ is the decision space. Then the Bayes estimator based on the γ-posterior under the loss function $L(\theta, d)$ is defined by

$$\hat{\theta}_n^{(\gamma)L} = \arg\min_{\delta} \int L(\theta, \delta)\pi^{(\gamma)}(\theta|\boldsymbol{X}_n)d\theta.$$

In particular, if we consider the quadratic loss function $L(\theta, d) = \|\theta - d\|^2$ with Euclidean norm $\|\cdot\|$, then the Bayes estimator corresponds to the posterior mean based on γ-posterior given by

$$\hat{\theta}_n^{(\gamma)} = E^{\pi^{(\gamma)}}(\theta|\boldsymbol{X}_n) = \int \theta\pi^{(\gamma)}(\theta|\boldsymbol{X}_n)d\theta.$$

Note that when $\gamma \to 0$ the Bayes estimator based on γ-posterior corresponds to the usual Bayes estimator under the quadratic loss function.

3 Asymptotic properties of estimators

In this section, we show some asymptotic properties of the estimation based on the γ-posterior. We define θ_g by $\theta_g = \arg\min_{\theta\in\Theta} d_\gamma(g, f_\theta)$. We assume the following regularity conditions on the density $f_\theta(x) = f(x;\theta)$ $(\theta \in \Theta \subset \mathbb{R}^p)$.

(A1) The support of the density does not depend on unknown parameter θ and f_θ is thrice differentiable with respect to θ in neighborhood U of θ_g.

(A2) Interchange of the order of integration with respect to x and differentiation as θ_g is justified. The expectations $E_g[\partial_i q^{(\gamma)}(\theta_g; X_1)]$ and $E_g[\partial_i\partial_j q^{(\gamma)}(\theta_g; X_1)]$ are all finite and there exists $M_{ijk}(x)$ such that

$$\sup_{\theta\in U}\left|\partial_i\partial_j\partial_k q^{(\gamma)}(\theta; X_1)\right| \leq M_{ijk}(x) \text{ and } E_g\left[M_{ijk}(X_1)\right] < \infty$$

for all $i, j, k = 1, \ldots, p$, where $\partial_i = \partial/\partial\theta_i$ and $\partial = \partial/\partial\theta$.

(A3) For any $\delta > 0$, with probability one

$$\sup_{\|\theta-\theta_g\|>\delta} n^{-1}\left\{Q_n^{(\gamma)}(\theta) - Q_n^{(\gamma)}(\theta_g)\right\} < -\varepsilon$$

for some $\varepsilon > 0$ and for all sufficiently large n.

Moreover, the maximum γ-likelihood estimator $\tilde{\theta}_n^{(\gamma)}$ and the posterior mean under the γ-posterior $\hat{\theta}_n^{(\gamma)}$ are defined by

$$\tilde{\theta}_n^{(\gamma)} = \arg\max_{\theta\in\Theta} Q_n^{(\gamma)}(\theta), \quad \hat{\theta}_n^{(\gamma)} = \int_\Theta \theta\pi^{(\gamma)}(\theta|\boldsymbol{X}_n)d\theta,$$

respectively. The matrices $I^{(\gamma)}(\theta)$ and $J^{(\gamma)}(\theta)$ are defined by

$$I^{(\gamma)}(\theta) = E_{\theta_g}\left[\partial q^{(\gamma)}(\theta; X_1)\partial^\top q^{(\gamma)}(\theta; X_1)\right], \quad J^{(\gamma)}(\theta) = -E_{\theta_g}\left[\partial\partial^\top q^{(\gamma)}(\theta; X_1)\right],$$

respectively. We assume that both $I^{(\gamma)}(\theta)$ and $J^{(\gamma)}(\theta)$ are positive definite matrices. Then we have the following theorem concerning the asymptotic normality of the γ-posterior.

Theorem 3.1. *Under the conditions* (A1)-(A3)*, we assume that $\tilde{\theta}_n^{(\gamma)}$ is a consistent solution of the γ-likelihood equation, that is $\partial Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)}) = 0$ and $\tilde{\theta}_n^{(\gamma)} \xrightarrow{p} \theta_g$ as $n \to \infty$. Then for any prior density $\pi(\theta)$ which is continuous and positive at θ_g, it holds*

$$\int \left|\pi^{*(\gamma)}(t|\boldsymbol{X}_n) - (2\pi)^{-p/2}\left|J^{(\gamma)}(\theta_g)\right|^{1/2}\exp\left(-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right)\right| dt \xrightarrow{p} 0 \tag{5}$$

as $n \to \infty$, where $\pi^{*(\gamma)}(t|\boldsymbol{X}_n)$ is the γ-posterior density of $t = \sqrt{n}(\theta - \tilde{\theta}_n^{(\gamma)})$ given $\boldsymbol{X}_n$.

The proof of this theorem is similar to that of Theorem 4.2 in Ghosh et al. (2006).

Proof. Putting $t = \sqrt{n}(\theta - \tilde{\theta}_n^{(\gamma)})$ in (4), we have

$$\begin{aligned}\pi^{*(\gamma)}(t|\boldsymbol{X}_n) &= \frac{\exp\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\}\pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)}{\int \exp\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\}\pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)dt} \\ &= C_n^{-1}\pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\exp\left\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right\},\end{aligned}$$

where

$$C_n = \int \pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\exp\left\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right\} dt.$$

We put

$$\begin{aligned}g_n(t) =& \pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\exp\left\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right\} \\ &- \pi(\theta_0)\exp\left\{-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right\}.\end{aligned}$$

Then, it suffices to show

$$\int |g_n(t)|dt \xrightarrow{p} 0 \tag{6}$$

as $n \to \infty$. If (6) holds, $C_n \xrightarrow{p} \pi(\theta_g)(2\pi)^{p/2}|J^{(\gamma)}(\theta_g)|^{-1/2}$ and therefore, the integral in (5), which is dominated by

$$\begin{aligned} &C_n^{-1}\int |g_n(t)|dt \\ &+\int \left| C_n^{-1}\pi(\theta_g)\exp\left\{-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right\} - (2\pi)^{-p/2}|J^{(\gamma)}(\theta_g)|^{1/2}\exp\left\{-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right\}\right| dt \end{aligned}$$

also convergence in probability to zero. In order to show (6), we consider the two regions $\mathcal{R}_1 = \{t| \ \|t\| > \delta_0\sqrt{n}\}$ and $\mathcal{R}_2 = \{t| \ \|t\| \leq \delta_0\sqrt{n}\}$, for a small positive constant δ_0. We will separately show $\int_{\mathcal{R}_i} |g(t)|dt \xrightarrow{p} 0$ for $i = 1, 2$. For $i = 1$, we note that it holds

$$\begin{aligned} \int_{\mathcal{R}_1} |g_n(t)|dt \leq &\int_{\mathcal{R}_1} \pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\exp\left\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right\} dt \\ &+ \int_{\mathcal{R}_1} \pi(\theta_0)\exp\left\{-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right\} dt. \end{aligned}$$

It is easy to see that the second integral goes to zero by the usual tail estimates for a normal distribution. For the first integral, from (A3), we note that it holds

$$n^{-1}\left\{Q_n^{(\gamma)}(\theta) - Q_n^{(\gamma)}(\theta_g)\right\} < -\varepsilon$$

for all $t \in \mathcal{R}_1$ and sufficiently large n. Therefore, the first integral is expressed by

$$\begin{aligned} &\int_{\mathcal{R}_1} \pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)\exp\left\{Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right\} dt \\ &\leq e^{-n\varepsilon}\int_{\mathcal{R}_1} \pi(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t)dt. \end{aligned}$$

Since the prior density π is integrable, we have

$$\int_{\mathcal{R}_1} |g_n(t)|dt \xrightarrow{p} 0$$

as $n \to \infty$. Next, we consider the case of $i = 2$. By the Taylor expansion of $Q_n^{(\gamma)}(\theta)$ at $\theta = \tilde{\theta}_n^{(\gamma)}$, we have

$$Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)} + n^{-1/2}t) - Q_n^{(\gamma)}(\tilde{\theta}_n^{(\gamma)}) = -\frac{1}{2}t^\top \left[\hat{J}^{(\gamma)}(\tilde{\theta}_n^{(\gamma)})\right] t + R_n(t),$$

where $\hat{J}^{(\gamma)}(\theta) = -n^{-1}\sum_{i=1}^n \partial\partial^\top q_\theta(X_i)$ and the remainder term $R_n(t)$ is

$$R_n(t) = \frac{1}{6n\sqrt{n}}\sum_{i,j,k} \partial_i\partial_j\partial_k Q_n^{(\gamma)}(\theta'_n)t_it_jt_k,$$

where $\theta'_n = \tilde{\theta}^{\gamma}_n + hn^{-1/2}t$ for some h such that $0 < h < 1$. By assumption (A2), it holds $R_n(t) \xrightarrow{p} 0$ and $\hat{J}^{(\gamma)}(\tilde{\theta}^{(\gamma)}_n) \xrightarrow{p} J^{(\gamma)}(\theta_g)$ as $n \to \infty$. Therefore, we have $g_n(t) \xrightarrow{p} 0$ as $n \to \infty$. For suitably chosen δ_0 and any $t \in \mathcal{R}_2$ we have

$$|R_n(t)| < \frac{1}{4} t^\top \hat{J}^{(\gamma)}(\tilde{\theta}^{(\gamma)}_n)t$$

for sufficiently large n so that it holds

$$\begin{aligned}\exp\left\{Q^{(\gamma)}_n(\tilde{\theta}^{(\gamma)}_n + n^{-1/2}t) - Q^{(\gamma)}_n(\tilde{\theta}^{(\gamma)}_n)\right\} &< \exp\left\{-\frac{1}{4}t^\top \hat{J}^{(\gamma)}(\tilde{\theta}^{(\gamma)}_n)t\right\} \\ &< \exp\left\{-\frac{1}{8}t^\top J^{(\gamma)}(\theta_g)t\right\}.\end{aligned}$$

Then, for suitably chosen small $\delta_0 > 0$, $|g_n(t)|$ is dominated by an integrable function on the region $\mathcal{R}_2$. Thus, we have

$$\int_{\mathcal{R}_2} |g_n(t)| dt \xrightarrow{p} 0$$

as $n \to \infty$. This completes the proof. □

Remark 3.1. By the definition of C_n and $C_n \xrightarrow{p} \pi(\theta_g)(2\pi)^{p/2}|J^{(\gamma)}(\theta_g)|^{-1/2}$ as $n \to \infty$, the log-marginal likelihood is expressed by

$$\begin{aligned}\log \int \prod_{i=1}^n \exp(q^{(\gamma)}(X_i))\pi(\theta)d\theta =& Q^{(\gamma)}_n(\tilde{\theta}^{(\gamma)}_n) - \frac{p}{2}\log n + \frac{p}{2}\log 2\pi - \frac{1}{2}\log|J^{(\gamma)}(\theta_g)| \\ &+ \log \pi(\theta_g) + o_p(1) \end{aligned} \tag{7}$$

as $n \to \infty$. By ignoring the terms which stay bounded as the $n \to \infty$ in (7), the Bayesian information criterion based on the γ-likelihood (BIC$_\gamma$) can be defined as

$$\mathrm{BIC}_\gamma = Q^{(\gamma)}_n(\tilde{\theta}^{(\gamma)}_n) - \frac{p}{2}\log n,$$

where p is a dimension of θ. We note that the original Bayesian information criterion (BIC) which is based on the log-likelihood is proposed by Schwarz (1978). However, we do not discuss the details in this paper.

Theorem 3.2. *In addition to assumptions of Theorem 3.1, assume that the prior density $\pi(\theta)$ has a finite expectation. Then it holds $\sqrt{n}(\hat{\theta}^{(\gamma)}_n - \tilde{\theta}^{(\gamma)}_n) \xrightarrow{p} 0$ as $n \to \infty$.*

Proof. Proceeding as in the proof of Theorem 3.1 and using the assumption of finite expectation for π, (5) can be strengthened to

$$\int \|t\| \left|\pi^{*(\gamma)}(t|\boldsymbol{X}_n) - (2\pi)^{-p/2}\left|J^{(\gamma)}(\theta_g)\right|^{1/2}\exp\left(-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right)\right| dt \xrightarrow{p} 0 \tag{8}$$

as $n \to \infty$. From this, we have

$$\int t\pi^{*(\gamma)}(t|\boldsymbol{X}_n)dt \xrightarrow{p} \int t(2\pi)^{-p/2}\left|J^{(\gamma)}(\theta_g)\right|^{1/2}\exp\left(-\frac{1}{2}t^\top J^{(\gamma)}(\theta_g)t\right)dt = 0.$$

Therefore, $\sqrt{n}(\hat{\theta}_n^{(\gamma)} - \tilde{\theta}_n^{(\gamma)}) = \int t\pi^{*(\gamma)}(t|\boldsymbol{X}_n)dt \xrightarrow{p} 0$ as $n \to \infty$. This completes the proof. □

From Theorem 3.2, the posterior mean based on the γ-posterior $\hat{\theta}_n^{(\gamma)}$ and the maximum γ-likelihood estimator $\tilde{\theta}_n^{(\gamma)}$ are asymptotically equivalent up to the first order.

Corollary 3.1. *Suppose the conditions of Theorem 3.2 hold. If it holds* $\sqrt{n}(\tilde{\theta}_n^{(\gamma)} - \theta_g) \xrightarrow{d} N_p(0, V^{(\gamma)}(\theta_g))$ *as* $n \to \infty$ *for some positive definite* $V^{(\gamma)}(\theta_g) = J^{(\gamma)}(\theta_g)^{-1}I^{(\gamma)}(\theta_g)J^{(\gamma)}(\theta_g)^{-1}$, *then we have* $\sqrt{n}(\hat{\theta}_n^{(\gamma)} - \theta_g) \xrightarrow{d} N_p(0, V^{(\gamma)}(\theta_g))$ *as* $n \to \infty$.

Proof. From Theorem 3.2, the proof is straightforward to apply the Slutzky's theorem. □

We note that the asymptotic normality of the minimum γ-divergence estimator $\tilde{\theta}_n^{(\gamma)}$ has also proved by Jones et al. (2001) and Fujisawa and Eguchi (2008).

4 Bayesian robustness of estimators

4.1 Influence function

In a similar way to Ghosh and Basu (2016), we consider the robustness of the posterior mean based on the γ-posterior in terms of the influence function. Let $X_1, \ldots, X_n$ be generated from the true distribution G with the density g, and we consider the parametric family $\{F_\theta : \theta \in \Theta\}$ with the density f_θ. Let $\pi(\theta)$ be the prior density for θ. The γ-posterior density as a function of G and θ is defined by

$$\pi^{(\gamma)}(\theta; G) = \frac{\exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)}{\int \exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)d\theta},$$

where

$$Q^{(\gamma)}(\theta; G, F_\theta) = \frac{1}{\gamma}\left[\int f_\theta(x)^\gamma dG(x)\right]\left(\int f_\theta(x)^{1+\gamma}dx\right)^{-\gamma/(1+\gamma)}.$$

For a fixed sample size n, the γ-Bayes functional under the general loss function $L(\cdot,\cdot)$ is defined by

$$T_n^{(\gamma)L}(G) = \arg\min_t \frac{\int L(\theta, t)\exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)d\theta}{\int \exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)d\theta}.$$

Under the quadratic loss function, the γ-Bayes functional $T_n^{(\gamma)L}(G)$ is the γ-posterior mean functional

$$T_n^{(\gamma)e}(G) = \frac{\int \theta\exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)d\theta}{\int \exp(nQ^{(\gamma)}(\theta; G, F_\theta))\pi(\theta)d\theta}.$$

Hereafter, we consider the case of the quadratic loss function. We now consider the contaminated model $F_\varepsilon = (1-\varepsilon)G + \varepsilon\Delta_y$, where ε is the contamination ratio and Δ_y is the contaminating distribution degenerated at y. Then the influence function of the γ-posterior mean functional for a fixed n at G is defined by

$$IF_n(y, T_n^{(\gamma)e}, G) = \frac{\partial}{\partial\varepsilon} T_n^{(\gamma)e}(F_\varepsilon)\bigg|_{\varepsilon=0} = n\mathrm{Cov}_{\pi^{(\gamma)}(\theta;G)}(\theta, k_\gamma(\theta; y, g)), \tag{9}$$

where $\mathrm{Cov}_{\pi^{(\gamma)}(\theta;G)}$ is the covariance under the γ-posterior $\pi^{(\gamma)}(\theta; G)$ and

$$k_\gamma(\theta; y, g) = \frac{\partial}{\partial\varepsilon} Q^{(\gamma)}(\theta; F_\varepsilon, F_\theta) = \frac{1}{\gamma}\left[f_\theta(y)^\gamma - \int f_\theta(x)^\gamma g(x)dx\right]\left(\int f_\theta(x)^{1+\gamma}dx\right)^{-\gamma/(1+\gamma)} \tag{10}$$

for $\gamma > 0$. For $\gamma = 0$, we have $k_0(\theta; y, g) = \log f_\theta(y) - \int g(x)\log f_\theta(x)dx$ which is the influence function of the usual posterior mean for a fixed n. Ghosh and Basu (2016) proposed the posterior distribution based on the density power divergence (we call it the "density power posterior" in this paper). From the result of Ghosh and Basu (2016), the influence function of the posterior mean under the density power posterior is given by

$$IF_n(y, T_n^{(\alpha)e}, G) = \frac{\partial}{\partial\varepsilon} T_n^{(\alpha)e}(F_\varepsilon)\bigg|_{\varepsilon=0} = n\mathrm{Cov}_{\pi^{(\alpha)}(\theta;G)}(\theta, k_\alpha(\theta; y, g)), \tag{11}$$

where $\mathrm{Cov}_{\pi^{(\alpha)}(\theta;G)}$ is the covariance under the the density power posterior $\pi^{(\alpha)}(\theta; G)$ and

$$k_\alpha(\theta; y, g) = \frac{\partial}{\partial\varepsilon} Q^{(\alpha)}(\theta; F_\varepsilon, F_\theta) = \frac{1}{\alpha}\left[f_\theta(y)^\alpha - \int f_\theta(x)^\alpha g(x)dx\right] \tag{12}$$

for $\alpha > 0$, where

$$Q^{(\alpha)}(\theta; G, F_\theta) = \frac{1}{\alpha}\left[\int f_\theta(x)^\alpha dG(x) - \frac{1}{1+\alpha}\int f_\theta(x)^{1+\alpha}dx\right].$$

From (10) and (12), we can find that the former is the form of the ratio, while the later is the subtraction. We show the influence curves for different values of γ and α. Consider two cases.

(a) The posterior mean for the mean parameter μ in $N(\mu, 1)$ under the prior $\pi(\mu) \propto 1$ (Figure 1).

(b) The posterior mean for the variance parameter σ^2 in $N(0, \sigma^2)$ under the prior $\pi(\sigma^2) \propto \sigma^{-2}$ (Figure 2).

We assume that the true densities are the standard normal distribution $N(0, 1)$ in both cases. In this setting, we can obtain $k_\gamma(\theta; y, g)$ and $k_\alpha(\theta; y, g)$ as closed forms. We note that

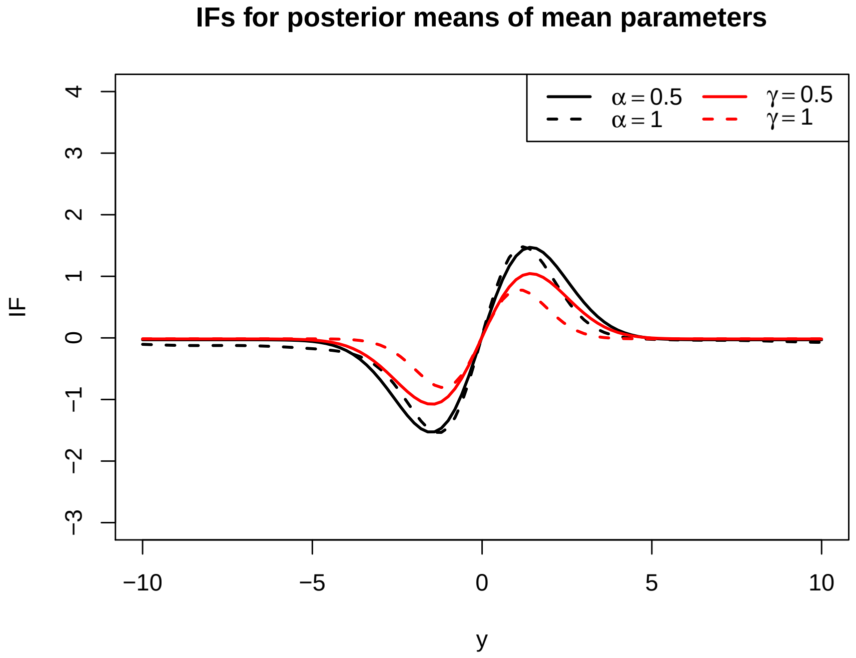

Figure 1: Influence functions for the posterior means of the mean parameter. The black curves are the influence functions based on the density power posterior, and the red curves are the influence functions based on the γ-posterior.

$k_\alpha(\mu;\theta,g)$ has already derived by pp. 428 in Ghosh and Basu (2016). Then we have

$$\begin{aligned}
k_\gamma(\mu;y,g) &= \frac{1}{\gamma(\sqrt{2\pi})^\gamma}\left(\left(\frac{1}{\sqrt{2\pi}}\right)^\gamma\frac{1}{\sqrt{1+\gamma}}\right)^{-\gamma/(1+\gamma)}\left\{e^{-\gamma(y-\mu)^2/2}-\frac{1}{\sqrt{1+\gamma}}e^{-\gamma\mu^2/(2(1+\gamma))}\right\},\\
k_\gamma(\sigma^2;y,g) &= \frac{1}{\gamma(\sqrt{2\pi})^\gamma}\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^\gamma\frac{1}{\sqrt{1+\gamma}}\right)^{-\gamma/(1+\gamma)}\left\{e^{-\gamma y^2/(2\sigma^2)}-\sqrt{\frac{\sigma^2}{\gamma+\sigma^2}}\right\},\\
k_\alpha(\mu;y,g) &= \frac{1}{\alpha(\sqrt{2\pi})^\alpha}\left\{e^{-\alpha(y-\mu)^2/2}-\frac{1}{\sqrt{1+\alpha}}e^{-\alpha\mu^2/(2(1+\alpha))}\right\},\\
k_\alpha(\sigma^2;y,g) &= \frac{1}{\alpha(\sqrt{2\pi\sigma^2})^\alpha}\left\{e^{-\alpha y^2/(2\sigma^2)}-\sqrt{\frac{\sigma^2}{\alpha+\sigma^2}}\right\}.
\end{aligned}$$

However, it is not easy to calculate the covariances in (9) and (11) as closed forms. We give the influence curves by using MCMC samples with 10,000 iterations from $\pi^{(\gamma)}(\theta;G)$ and $\pi^{(\alpha)}(\theta;G)$.

From Figure 1, we can find that the influence functions for the posterior means of the normal mean parameter are bounded for both cases. From Figure 2, it is also shown that the influence functions for the posterior means of the normal variance parameter are bounded for both cases. In Figure 2, we note that the influence functions of the posterior means for the variance parameter based on the γ-divergence seem to have the redescending properties as $y\to\infty$.

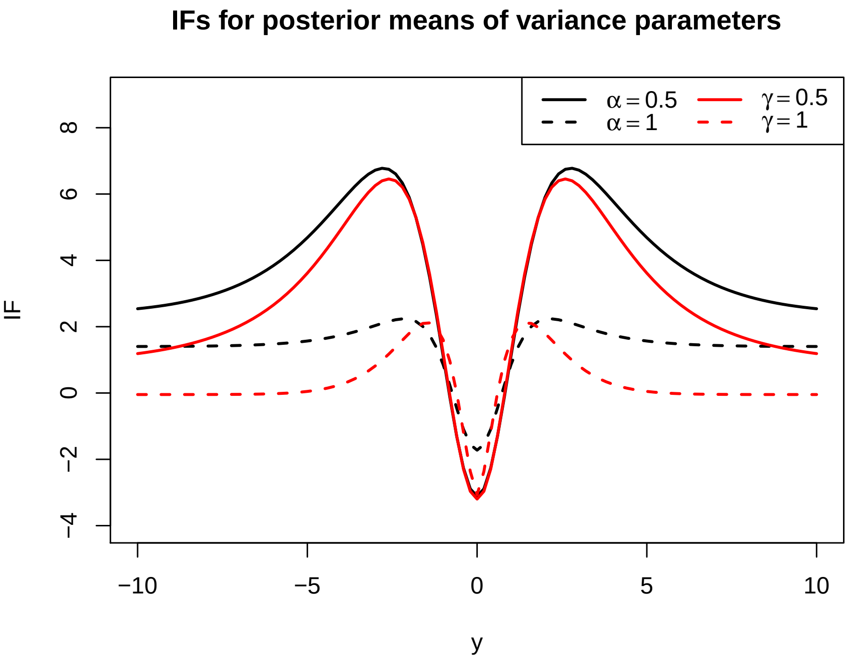

Figure 2: Influence functions for the posterior means of the variance parameter. The black curves are the influence functions based on the density power posterior, and the red curves are the influence functions based on the γ-posterior.

4.2 Prior robustness in the view of local sensitivity

In the previous section, we considered the robustness for outliers. However, the robustness of the selection of the prior distribution is also an important problem in Bayesian inference. There are several criteria to measure the robustness of the prior distribution, for example, minimax method, local sensitivity, global sensitivity and so on (see Ghosh et al. (2006)). In this section, we consider the local measure of sensitivities with small perturbations on the prior distribution.

First, we consider the set $\mathcal{P}$ of all probability densities over the parameter space Θ and a distance $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ to quantify the changes between original and contaminated densities. Let ν_ε be a data generated density which includes the true prior π and contaminated prior ν. For example, we often use the following two type perturbations

$$\nu_\varepsilon = (1-\varepsilon)\pi + \varepsilon\nu \quad \text{(linear perturbation)},$$
$$\nu_\varepsilon = c(\varepsilon)\pi^{1-\varepsilon}\nu^\varepsilon \quad \text{(geometric perturbation)},$$

where $0 < \varepsilon < 1$. We note that ε is the prior perturbation ratio, not the contamination ratio of outlier. Gustafon and Wasserman (1996) defined the local sensitivity of $\mathcal{P}$ with the contaminated prior ν as

$$s(\pi, \nu; \boldsymbol{X}_n) = \lim_{\varepsilon \downarrow 0} \frac{d(\pi(\theta|\boldsymbol{X}_n), \nu_\varepsilon(\theta|\boldsymbol{X}_n))}{d(\pi(\theta), \nu_\varepsilon(\theta))},$$

where $\pi(\theta|\boldsymbol{X}_n)$ and $\nu_\varepsilon(\theta|\boldsymbol{X}_n)$ are posterior densities under priors π and ν_ε, respectively. As distance function d, we often use the total variation or ϕ-divergence (Dey and Birmiwal (1994)). We can extend this measure to the γ-posterior density straightforward. The local sensitivity of $\mathcal{P}$ for the γ-posterior with the contaminated prior ν is defined by

$$s_\gamma(\pi, \nu; \boldsymbol{X}_n) = \lim_{\varepsilon \downarrow 0} \frac{d(\pi^{(\gamma)}(\theta|\boldsymbol{X}_n), \nu_\varepsilon^{(\gamma)}(\theta|\boldsymbol{X}_n))}{d(\pi(\theta), \nu_\varepsilon(\theta))},$$

where $\pi^{(\gamma)}(\theta|\boldsymbol{X}_n)$ and $\nu_\varepsilon^{(\gamma)}(\theta|\boldsymbol{X}_n)$ are the γ-posterior densities with $\gamma > 0$ under priors π and ν_ε, respectively. Using this measure, we may be able to construct the posterior distribution which is robust against both outliers and the selection of priors.

5 Simulation studies

In this section, We now show the performance of the γ-posterior and its posterior mean by comparing with other types of posterior distributions. We suppose that the parametric model is the normal distribution with mean μ and variance σ^2, and we put $\theta = (\mu, \sigma^2)^\top$. Let $\theta_0 = (0, 1)^\top$ be the true value of θ. We assume that the contamination density is the normal distribution with mean 6 and variance 1 and we set the ratio of contamination is 0.00, 0.05 or 0.20.

While Ghosh and Basu (2016) assume that the variance parameter σ^2 is known, we assume that σ^2 is unknown and consider the joint estimation problem for $\theta = (\mu, \sigma^2)^\top$ under the uniform and non-informative priors. In order to compare the empirical biases of posterior means, we use the posterior densities based on the three types of likelihood functions, that is, the ordinary log-likelihood function which is based on the Kullback Leibler cross entropy, the density power likelihood and γ-likelihood functions.

- ordinary log-likelihood:

$$\ell_n(\theta) = -\frac{n}{2}(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x-\mu)^2}{2\sigma^2}.$$

- density power likelihood (Basu et al. (1998)):

$$Q_n^{(\alpha)}(\theta) = \frac{1}{\alpha}\sum_{i=1}^{n} r_\theta^{(\alpha)}(x_i) - n(2\pi\sigma^2)^{(-\alpha/2)}(1+\alpha)^{-3/2}$$

  for $\alpha > 0$, where

$$r_\theta^{(\alpha)}(x_i) = \frac{1}{(2\pi\sigma^2)^{\alpha/2}} \exp\left(-\frac{\alpha(x_i-\mu)^2}{2\sigma^2}\right).$$

- γ-likelihood (Jones et al. (2001) and Fujisawa and Eguchi (2008)):

$$Q_n^{(\gamma)}(\theta) = \frac{1}{\gamma}\{(2\pi\sigma^2)^{-\gamma/2}(1+\gamma)^{-1/2}\}^{-\gamma/(1+\gamma)} \sum_{i=1}^{n} r_\theta^{(\gamma)}(x_i) - \frac{n}{\gamma}$$

for $\gamma > 0$, where

$$r_\theta^{(\gamma)}(x_i) = \frac{1}{(2\pi\sigma^2)^{\gamma/2}} \exp\left(-\frac{\gamma(x_i-\mu)^2}{2\sigma^2} \right).$$

Since exact calculations of the posterior means are not easy, we use the importance sampling Monte Carlo algorithm using a proposal distribution $N(0,1)$ for μ and χ_5^2 for σ^2 (for the details of importance sampling, see Robert and Casella (2004)). We carry out the importance sampling with 10,000 steps and we compute the empirical bias of the posterior means $(\hat{\mu}, \hat{\sigma}^2)$ for (μ, σ^2) by 10,000 iterations. The simulation results are shown in Tables 1 to 4.

Tables 1 and 2 are the results of the joint estimation for μ and σ^2 under the uniform prior $\pi(\mu,\sigma) \propto 1$. We compute the empirical biases of the posterior means based on the ordinary posterior, density power posterior and γ-posterior. Similarly, Tables 3 and 4 are the results of the joint estimation for μ and σ^2 under the non-informative prior $\pi(\mu,\sigma) \propto 1/\sigma^2$.

From Tables 1 and 3, the empirical biases of the posterior means for the mean parameter μ based on the γ-posterior are similar behaviors to the case of the density power posterior. On the other hand, it is seen that the empirical biases of the posterior means for the variance parameter σ^2 based on the γ-posterior is much smaller than these of based on the ordinary and density power posteriors in Tables 2 and 4.

Table 1: The empirical biases of the posterior means for the mean parameter under the uniform prior

		ordinary	density power posterior				γ-posterior			
		α, γ	α				γ			
ε	n	0.00	0.30	0.50	0.70	1.00	0.30	0.50	0.70	1.00
0.00	20	0.001	0.001	0.004	0.005	0.005	0.001	0.002	0.003	0.004
0.00	50	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001
0.00	100	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.002
0.05	20	0.302	0.094	0.076	0.059	0.028	0.083	0.059	0.053	0.046
0.05	50	0.300	0.026	0.015	0.026	0.039	0.022	0.008	0.006	0.007
0.05	100	0.302	0.013	0.005	0.004	0.013	0.012	0.004	0.002	0.002
0.20	20	1.191	0.724	0.490	0.328	0.166	0.730	0.507	0.390	0.292
0.20	50	1.194	0.678	0.413	0.342	0.242	0.643	0.280	0.164	0.131
0.20	100	1.202	0.614	0.209	0.154	0.218	0.574	0.091	0.021	0.011

Table 2: The empirical biases of the posterior means for the variance parameter under the uniform prior

		ordinary	density power posterior				γ-posterior			
		α, γ	α				γ			
ε	n	0.00	0.30	0.50	0.70	1.00	0.30	0.50	0.70	1.00
0.00	20	0.266	1.249	4.444	9.155	12.755	0.955	2.187	4.175	7.280
0.00	50	0.090	0.259	0.515	1.447	7.275	0.230	0.361	0.547	1.020
0.00	100	0.041	0.111	0.192	0.336	1.292	0.101	0.150	0.208	0.319
0.05	20	2.441	2.766	6.143	10.225	13.075	2.126	3.367	5.338	8.173
0.05	50	1.943	0.523	0.821	2.366	8.700	0.425	0.464	0.668	1.275
0.05	100	1.833	0.233	0.288	0.489	2.125	0.189	0.189	0.236	0.354
0.20	20	7.504	9.700	11.373	12.993	13.934	8.793	8.933	9.702	11.014
0.20	50	6.333	6.017	5.944	8.379	12.484	5.358	3.285	2.844	3.682
0.20	100	6.054	4.631	2.423	3.003	7.895	4.180	0.957	0.496	0.596

Table 3: The empirical biases of the posterior means for the mean parameter under the non-informative prior

		ordinary	density power posterior				γ-posterior			
		α, γ	α				γ			
ε	n	0.00	0.30	0.50	0.70	1.00	0.30	0.50	0.70	1.00
0.00	20	-0.004	-0.004	-0.004	-0.005	-0.008	-0.004	-0.004	-0.004	-0.004
0.00	50	-0.001	-0.001	-0.001	-0.002	-0.002	-0.001	-0.001	-0.002	-0.002
0.00	100	0.000	0.000	0.001	0.001	0.001	0.000	0.001	0.001	0.001
0.05	20	0.293	0.046	0.029	0.028	0.017	0.039	0.017	0.012	0.009
0.05	50	0.298	0.018	0.007	0.007	0.015	0.016	0.004	0.002	0.001
0.05	100	0.299	0.011	0.003	0.002	0.004	0.010	0.003	0.001	0.001
0.20	20	1.189	0.614	0.350	0.238	0.139	0.599	0.297	0.180	0.119
0.20	50	1.197	0.578	0.241	0.170	0.153	0.542	0.140	0.051	0.029
0.20	100	1.205	0.549	0.121	0.064	0.090	0.507	0.048	0.010	0.005

Table 4: The empirical biases of the posterior means for the variance parameter under the non-informative prior

		ordinary	density power posterior				γ-posterior			
		α, γ	α				γ			
ε	n	0.00	0.30	0.50	0.70	1.00	0.30	0.50	0.70	1.00
0.00	20	0.116	0.410	1.051	2.500	4.971	0.334	0.556	0.871	1.434
0.00	50	0.045	0.123	0.218	0.440	1.811	0.107	0.150	0.195	0.276
0.00	100	0.018	0.053	0.089	0.149	0.389	0.046	0.063	0.080	0.105
0.05	20	2.031	1.002	1.716	3.265	5.520	0.792	0.883	1.200	1.759
0.05	50	1.811	0.277	0.341	0.677	2.478	0.219	0.188	0.225	0.312
0.05	100	1.757	0.148	0.159	0.246	0.631	0.113	0.084	0.095	0.121
0.20	20	6.535	5.775	5.246	5.948	6.936	5.221	3.543	3.027	3.010
0.20	50	6.041	4.392	2.756	3.219	5.554	3.940	1.363	0.782	0.761
0.20	100	5.910	3.846	1.294	1.220	2.799	3.452	0.440	0.180	0.177

Finally, we compare the γ-posterior for the mean and variance parameters given data with/without outliers with those of the other types of posteriors. The setting of statistical model is the same as the above. We set $n = 200$, $\alpha, \gamma = 0.5$ and $\varepsilon = 0.20$. We show the histograms of the posterior samples given by a random walk Metropolis algorithm which is a kind of MCMC methods with 10,000 iterations. From Figure 3, the ordinary posterior distribution with outliers are so far from the case of without outliers. On the other hand, in Figure 4, it seems that the density power posterior slightly improves the ordinary posterior. However, the effect of the outliers in the case of estimation for σ^2 is not completely removed. Hence, we can find that the ordinary posterior is not robust against outliers and the density power posterior is partially robust against outliers. On the other hand, it seems that the γ-posterior almost removes the effect of the outliers for both μ and σ^2.

6 Concluding remarks

In this paper, a robust posterior distribution based on the γ-divergence was proposed. Some asymptotic properties for estimator were also shown. Further, we showed that the influence functions for the posterior means based on the γ-posterior are bounded for some tuning parameters in the cases of the estimation for both mean and variance parameters in the

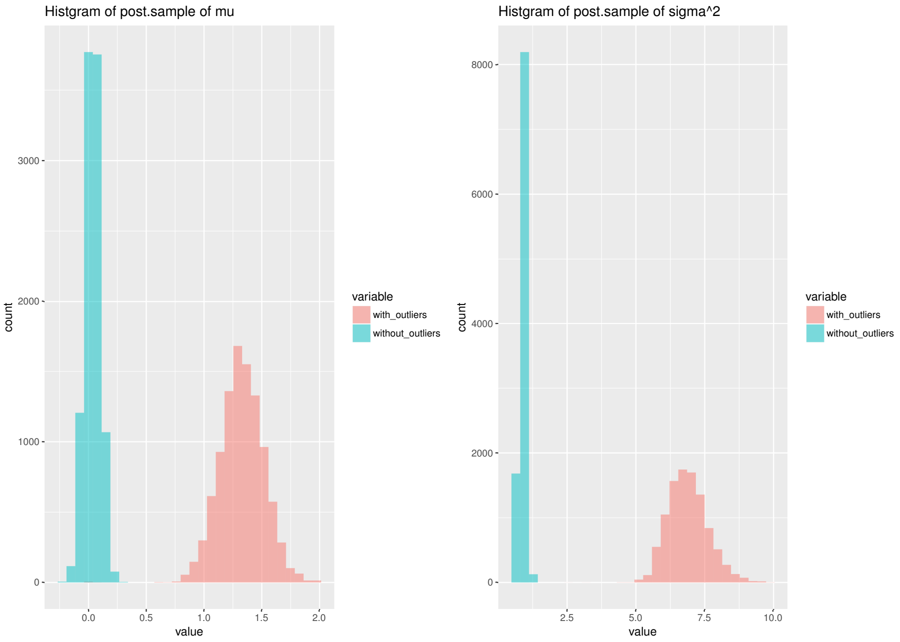

Figure 3: Histograms of the posterior samples based on the ordinary posterior for μ (left) and σ^2 (right) given by data with/without outliers.

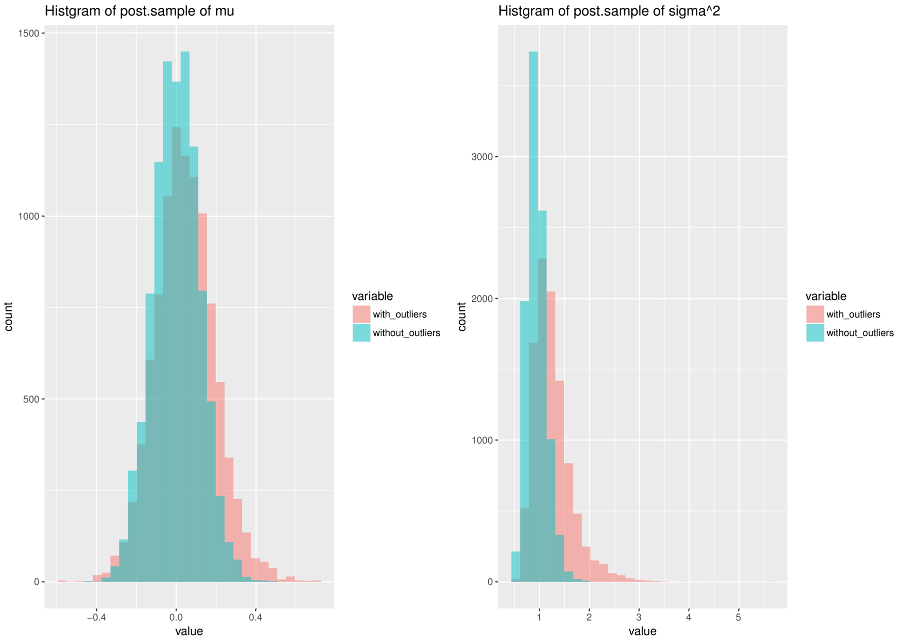

Figure 4: Histograms of the posterior samples based on the density power posterior for μ (left) and σ^2 (right) given by data with/without outliers.

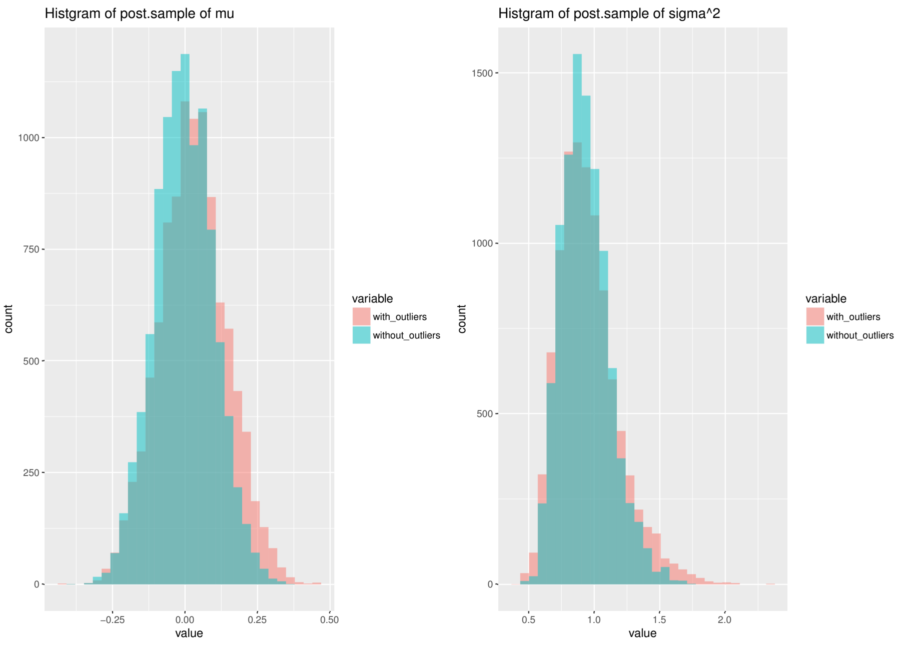

Figure 5: Histograms of the posterior samples based on the γ-posterior for μ (left) and σ^2 (right) given by data with/without outliers.

normal model. In simulation studies, we showed that the empirical biases of the posterior means based on the γ-posterior is smaller than other competitors for the joint estimation of mean and variance parameters in the normal model.

As future works, it is necessary to consider the selection of the tuning parameter γ to balance the efficiency and robustness of the estimator. Also, this paper may be extended to the Bayesian linear regression and generalized linear regression models.

References

[1] Andrade, J. A. A. and O'Hagan, A. (2006). Bayesian robustness modelling using regularly varying distributions. *Bayesian Anal.*, **1**: 169–188.

[2] Basu, A., Harris, I., Hjort, N. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**: 549–559.

[3] Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, **60**: 664–667.

[4] Desgagné, A. (2015). Robustness to outliers in location-scale parameter model using log-regularly varying distributions. *Ann. Statist.*, **43**: 1568–1595.

[5] Dey, D.K. and Birmiwal, L. (1994). Robust Bayesian analysis using entropy and divergence measures. *Statist. Prob. Lett.*, **20**: 287–294.

[6] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.*, **99** (9): 2053-2081.

[7] Ghosh, A. and Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Ann. Inst. Statist. Math.*, **68**: 413–437.

[8] Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis.* Springer, New York.

[9] Gustafon, P. and Wasserman, L. (1996). Local sensitivity diagnostics for Bayesian inference. *Ann. Statist.*, **23**: 2153–2167.

[10] Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. M. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* Wiley, New York.

[11] Hirose, K., Fujisawa, H. and Sese, J. (2017). Robust sparse Gaussian graphical modeling. *J. Multivar. Anal.*, **161**: 172-190.

[12] Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, **23**(3): 556–584.

[13] Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

[14] Jones, M. C., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88**: 865–873.

[15] Kawashima, T. and Fujisawa, H. (2017). Robust and sparse regression via gamma-divergence. *Entropy*, **19**: 608.

[16] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**(2): 461–464.

[17] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 3rd ed. Springer.

[18] Windham, M. P. (1995). Robustifying model fitting. *J. R. Statisti. Soc.* B **57**: 599–609.