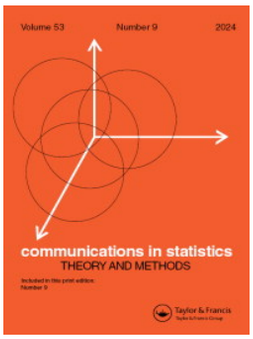
# On robustness of model selection criteria based on divergence measures: Generalizations of BHHJ divergence-based method and comparison

Kurata, Sumino
Graduate School of Information Science and Technology, The University of Tokyo



出版情報：Communications in Statistics - Theory and Methods, pp.1-18, 2022-12-15. Taylor & Francis
バージョン：




Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lsta20

# On robustness of model selection criteria based on divergence measures: Generalizations of BHHJ divergence-based method and comparison

Sumito Kurata

**To cite this article:** Sumito Kurata (15 Dec 2022): On robustness of model selection criteria based on divergence measures: Generalizations of BHHJ divergence-based method and comparison, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2022.2155788

**To link to this article:** https://doi.org/10.1080/03610926.2022.2155788

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2022.2155788



# On robustness of model selection criteria based on divergence measures: Generalizations of BHHJ divergence-based method and comparison

Sumito Kurata

Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

**ABSTRACT**

In model selection problems, robustness is one important feature for selecting an adequate model from the candidates. We focus on statistical divergence-based selection criteria and investigate their robustness. We mainly consider BHHJ divergence and related classes of divergence measures. BHHJ divergence is a representative robust divergence measure that has been utilized in, for example, parametric estimation, hypothesis testing, and model selection. We measure the robustness against outliers of a selection criterion by approximating the difference of values of the criterion between the population with outliers and the non-contaminated one. We derive and compare the conditions to guarantee robustness for model selection criteria based on BHHJ and related divergence measures. From the results, we find that conditions for robust selection differ depending on the divergence families, and that some expanded classes of divergence measures require stricter conditions for robust model selection. Moreover, we prove that robustness in estimation does not always guarantee robustness in model selection. Through numerical experiments, we confirm the advantages and disadvantages of each divergence family, asymptotic behavior, and the validity for employing criteria on the basis of robust divergence. Especially, we reveal the superiority of BHHJ divergence in robust model selection for extensive cases.

**ARTICLE HISTORY**
Received 5 January 2022
Accepted 30 November 2022

**KEYWORDS**
Model selection; BHHJ divergence; JHHB divergence; C divergence; robustness

## 1. Introduction

Divergence measures, indices of farness (dissimilarity) between two probability distributions, have been applied to such tasks as parametric estimation, hypothesis testing, and model selection. In model selection, two representative methods for evaluating statistical models are Akaike's information criterion (AIC, (Akaike 1973; Akaike 1974)) and Bayesian information criterion (BIC, (Schwarz 1978)). AIC is an approximation of expected log-likelihood (e.g., (Akaike 1973; Akaike 1978)), and we can regard AIC as

**CONTACT** Sumito Kurata kurata@imi.kyushu-u.ac.jp Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Tokyo 113-8656, Japan.
Current affiliation of Sumito Kurata: Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/03610926.2022.2155788.

being built upon the Kullback-Leibler divergence (KL divergence, (Kullback and Leibler 1951)) between the probability distribution of the true data-generating structure and that of a model. The derivation process of BIC was not the same as that of AIC: BIC selects a model that maximizes the Bayesian posterior probability from among candidates, and BIC was derived on the basis of the logarithm of marginal likelihood. A notable advantage of BIC over AIC is the property called consistency of model selection, that the probability of selecting the 'adequate' model (the model coinciding with the true data-generating structure) tends to one as the sample size goes to infinity. AIC and BIC each have an additive form of $(-2)$ times the maximum log-likelihood and a penalty term, and we use the maximum likelihood estimator (MLE), which minimizes the KL divergence between the empirical distribution based on observations and the distribution of a model, as the estimator of unknown parameters in a statistical model.

To date, many families of divergence measures that generalize KL divergence have been introduced, such as power divergence (Cressie and Read 1984), $f$-divergence (Csiszár 1967; Morimoto 1963; Pardo 2005), and alpha divergence (Amari 2009; Amari and Nagaoka 2000). In this paper, we mainly focus on the BHHJ divergence (or density power divergence) family proposed by (Basu et al. 1998), a generalization of KL divergence specialized on robustness in parametric estimation. BHHJ divergence achieves robust estimation by down-weighting outliers (distinctive, unusual, or mistaken data that are distant from other data) in the observation set. Each divergence in the BHHJ family is characterized by a tuning parameter that controls the tradeoff between efficiency and robustness in estimation. Some studies have applied this divergence family to hypothesis testing (e.g., (Basu et al. 2013; Toma 2010)) and model selection (e.g., (Kurata and Hamada 2019; Mattheou, Lee, and Karagrigoriou 2009)), and confirmed that BHHJ divergence-based methods also performed robustly in testing and model selection problems. (Kurata and Hamada 2018) derived an AIC-type family of model selection criteria from the BHHJ divergence family, and showed that their proposed method selects the adequate model even if some outliers are mixed in the observation set. Moreover, (Kurata and Hamada 2020) proposed RCC, a BIC-type family of criteria based on the BHHJ divergence family that have both consistency and robustness of model selection.

In the present paper, we consider the question of whether criteria based on divergence measures other than KL and BHHJ divergence have robustness. There are some wide families of divergence measures that include not only KL divergence but also BHHJ divergence. (Jones et al. 2001) proposed the JHHB divergence family, which generalizes BHHJ divergence by an additional tuning parameter. The JHHB family includes the gamma divergence family, that has been used in the field of robust parametric estimation (Fujisawa and Eguchi 2008). Moreover, (Vonta, Mattheou, and Karagrigoriou 2012) and (Maji et al. 2019) independently expanded the BHHJ divergence family via a class of convex functions. This wide family, referred to as the C divergence family, also includes the $f$-divergence family.

We investigate the robustness of model selection problems of criteria derived from such generalized divergence measures. Since outliers are distant from other observations, they often have a bad influence on values of estimates and model selection criteria. Thus, we need to evaluate the perturbation of the estimator and criterion. In parametric estimation, influence functions and gross error sensitivity provide measures of damage caused by an outlier to an estimator (e.g., (Hampel et al. 1986; Huber 1983)). Here we consider the

sensitivity of a criterion to outliers by exploring the difference between populations with and without outliers. We generally assume that observations are drawn from a (true) population distribution, and we can interpret that outliers are drawn from a probability distribution differing from the true distribution. Therefore, we investigate the robustness of a criterion by evaluating the difference of its values between contaminated and non-contaminated data-generating distributions. If the difference of values of the criterion is not bounded for a perturbation of the data-generating distribution, we can regard the corresponding divergence measure as not having robustness of model selection. We introduce the sufficient conditions to have robustness for each divergence family.

This paper is organized as follows. We review divergence families related to BHHJ divergence in Section 2, and introduce a general form of the model selection criteria considered in this study, including AIC, BIC, and RCC, in Section 3. We then investigate the robustness of the criterion based on each divergence family, from the viewpoint of when the data-generating distribution is contaminated, in Section 4. We verify the performance of criteria by numerical experiments in Section 5. We conclude this paper in Section 6. Proofs of theorems and additional results of numerical experiments are presented in the supplemental material.

## 2. Divergence measures and model selection criteria

In this section, we first introduce some families of divergence measures. We mainly focus on BHHJ divergence, which has robustness against outliers in parametric estimation. (Kurata and Hamada 2018) and (Kurata and Hamada 2019) showed that an AIC-type criterion based on BHHJ divergence performs robustly in model selection problems, and (Kurata and Hamada 2020) proposed a BIC-type criterion that has consistency of model selection as well as robustness. We also introduce wide families that include BHHJ divergence. Let $G$ be a probability distribution, and $F^{\boldsymbol{\theta}}$ be a parametric model with respect to $G$, with a $p$-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We assume that $G$ and $F^{\boldsymbol{\theta}}$ have probability (density) functions $g(\cdot)$ and $f(\cdot \mid \boldsymbol{\theta})$, respectively.

### 2.1. BHHJ divergence family and KL divergence

As a measure of 'farness' (not mathematical distance) between two probability distributions $G$ and $F^{\boldsymbol{\theta}}$, (Basu et al. 1998) introduced the BHHJ divergence family,

$$D_{\alpha}^{\mathrm{BHHJ}}(G\,;\,F^{\boldsymbol{\theta}}) = \int \left\{ f(y\,|\,\boldsymbol{\theta})^{\alpha+1} - \frac{\alpha+1}{\alpha}\, f(y\,|\,\boldsymbol{\theta})^{\alpha}\, g(y) + \frac{1}{\alpha}\, g(y)^{\alpha+1} \right\} dy, \tag{1}$$

where $\alpha$ is a positive tuning parameter that controls the tradeoff between efficiency and robustness. As $\alpha \to 0$, Equation (1) tends to the following form:

$$D^{\mathrm{KL}}(G\,;\,F^{\boldsymbol{\theta}}) = D_{0}^{\mathrm{BHHJ}}(G\,;\,F^{\boldsymbol{\theta}}) = \int g(y)\ \log \frac{g(y)}{f(y\,|\,\boldsymbol{\theta})}\, dy,$$

corresponding to KL divergence (Kullback and Leibler 1951), which has efficiency in parametric estimation. On the other hand, when $\alpha$ is large, this divergence has robustness in estimation.

### 2.2. JHHB divergence family

(Jones et al. 2001) generalized the BHHJ divergence family via an additional tuning parameter $\varphi \geq 0$ as follows:

$$D^{\mathrm{JHHB}}_{\alpha,\varphi}(G\,;\,F^{\boldsymbol{\theta}}) = \frac{1}{\varphi}\left\{\int f(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy\right\}^{\varphi} - \frac{\alpha+1}{\varphi\ \alpha}\left\{\int f(y\,|\,\boldsymbol{\theta})^{\alpha}\ g(y)\,dy\right\}^{\varphi} + \frac{1}{\varphi\ \alpha}\left\{\int g(y)^{\alpha+1}\,dy\right\}^{\varphi}. \tag{2}$$

We refer to this as the JHHB divergence family. Obviously, this family coincides with the BHHJ divergence family when $\varphi = 1$. Additionally, Equation (2) has the limit

$$D^{\mathrm{JHHB}}_{\alpha,0}(G\,;\,F^{\boldsymbol{\theta}}) = \log\left\{\int f(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy\right\} - \frac{\alpha+1}{\alpha}\ \log\left\{\int f(y\,|\,\boldsymbol{\theta})^{\alpha}\ g(y)\,dy\right\} + \frac{1}{\alpha}\ \log\left\{\int g(y)^{\alpha+1}\,dy\right\}, \tag{3}$$

as $\varphi$ goes to 0. This divergence is also known as logarithmic density power divergence or gamma divergence (e.g., (Fujisawa and Eguchi 2008; Jones et al. 2001)). The JHHB family has robustness in parametric estimation for any $\alpha > 0$ and $\varphi \geq 0$, but it only has exact unbiasedness when $\varphi$ is equal to 1 or 0 (Jones et al. 2001). Additionally, the parametric estimation based on JHHB divergence is the M-estimation only for $\varphi = 0$ or 1.

### 2.3. C divergence family

(Vonta, Mattheou, and Karagrigoriou 2012) and (Maji et al. 2019) independently derived a wide class of divergence measures including BHHJ divergence. Let $N$ be an element of $\mathcal{N}$, the set of strictly convex functions on $[-1, +\infty)$ that are three times continuous differentiable and satisfy $N(0) = 0$, $N'(0) = 0$, and $N''(0) > 0$. For $\alpha > 0$ and the above function $N \in \mathcal{N}$, the C divergence family is defined as follows:

$$D^{\mathrm{C}}_{\alpha,N}(G\,;\,F^{\boldsymbol{\theta}}) = \int N\left(\frac{g(y)}{f(y\,|\,\boldsymbol{\theta})} - 1\right) f(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy\,. \tag{4}$$

When we use

$$N(z) = 1 - \frac{(\alpha+1)\ (z+1)}{\alpha} + \frac{(z+1)^{\alpha+1}}{\alpha}$$

as the function $N$, Equation (4) coincides with the BHHJ divergence family. This family also contains generalized power divergence (Maji et al. 2019; Vonta, Mattheou, and Karagrigoriou 2012) and S divergence (Ghosh et al. 2017), which have been utilized in robust parametric estimation. In addition, C divergence becomes $f$-divergence (e.g., (Csiszár 1967; Morimoto 1963; Pardo 2005)), which includes power divergence (Cressie and Read 1984) and alpha divergence (Amari 2009; Amari and Nagaoka 2000), as $\alpha \to 0$. Note that, although both the JHHB and C divergence families contain BHHJ divergence, these two families have no inclusion relation.

Since parametric estimation based on C divergence is generally not M-estimation and each element in the C divergence family is generally not a decomposable pseudo-distance (Broniatowski, Toma, and Vajda 2012; Broniatowski and Vajda 2012), we cannot estimate parameters by replacing the true distributions $G$ with either empirical distributions or relative frequencies if we consider continuous distributions. In such cases, we need to estimate $G$ via histogram density estimation, kernel density estimation, or the like (Maji et al. 2019). We hereafter denote the estimated distribution of $G$ by $\hat{G}$, regardless of the method (e.g., empirical distribution or kernel density). We assume that $\hat{G}$ uniformly converges to the true cumulative distribution function.

## 3. Model selection criteria based on divergence measures

Hereafter, we consider the case where not all of the observations have the same distribution. Let $\boldsymbol{G} = (G_1, ..., G_K)$ be the set consisting of the 'true' distributions. We assume that $n_k$ observations follow $G_k$ independently for each $k = 1, ..., K$, and let $\hat{\boldsymbol{G}} = (\hat{G}_1, ..., \hat{G}_K)$ be the estimated distributions with respect to $\boldsymbol{G}$ (such as empirical distributions) based on $n_1 + \cdots + n_K = n$ observations. When $K = 1$, this situation is the independent and identically distributed (i.i.d.) setting, and when $n_k = 1$ for all $k$ (i.e., $K = n$), this is the non-homogeneous setting, which includes the generalized linear model (see, e.g., (Ghosh and Basu 2016; Kurata and Hamada 2018)). In this paper, we consider the following measure of overall farness between $\boldsymbol{G}$ and the model distributions $\boldsymbol{F}^{\boldsymbol{\theta}} = (F_1^{\boldsymbol{\theta}}, ..., F_K^{\boldsymbol{\theta}})$:

$$\bar{D}(\boldsymbol{G}\,;\,\boldsymbol{F}^{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \frac{n_k}{n}\; D(G_k\,;\,F_k^{\boldsymbol{\theta}})\,, \tag{5}$$

where $D$ is a divergence measure such as KL and BHHJ divergence introduced in Section 2. When estimating unknown parameter $\boldsymbol{\theta}$, we obtain the divergence-based estimates by minimizing $\bar{D}(\hat{\boldsymbol{G}}\,;\,\boldsymbol{F}^{\boldsymbol{\theta}})$. We denote the minimum divergence estimator by $\hat{\boldsymbol{\theta}}_D$, and define the 'best fitting parameter' as the value of $\boldsymbol{\theta}$ that minimizes Equation (5).

In this paper, we discuss the following form of model selection criteria based on divergence measures:

$$2\; n\; \bar{H}(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_D)\;\; +\;\; B\,, \tag{6}$$

where the main term,

$$\bar{H}(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_D) = \sum_{k=1}^{K} \frac{n_k}{n}\; H_k(\hat{G}_k\,;\,F_k^{\hat{\boldsymbol{\theta}}_D})\,,$$

is composed of $H_k$ $(k = 1, ..., K)$ obtained by subtracting parts independent of the model from $D(\hat{G}_k\,;\,F_k^{\hat{\boldsymbol{\theta}}_D})$ (see also Section 2 and Equation (5)), and $B$ is a bias term. For example, $H_k$ $(k = 1, ..., K)$ for BHHJ divergence (including KL divergence) are as follows:

$$H_k(G_k\,;\,F_k^{\boldsymbol{\theta}}) = \begin{cases} -\int g_k(y)\ \ \log f_k(y\,|\,\boldsymbol{\theta})\,dy & (\alpha = 0)\,, \\ \int f_k(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \dfrac{\alpha+1}{\alpha}\displaystyle\int f_k(y\,|\,\boldsymbol{\theta})^{\alpha}\ g_k(y)\,dy & (\alpha > 0)\,. \end{cases}$$

The form of criteria given in Equation (6) includes most of the information criteria such as AIC, BIC, GIC (Konishi and Kitagawa 1996), GBIC (Konishi, Ando, and Imoto 2004), criteria for the regularization method (e.g., (Ninomiya and Kawano 2016; Umezu et al. 2019)), and divergence-based criteria (e.g., (Kurata and Hamada 2020; Mattheou, Lee, and Karagrigoriou 2009; Toma 2014)).

Hereinafter, we assume the following conditions:

**(A1)** The support of the model $\mathcal{Y} = \{y \mid f_k(y\,|\,\boldsymbol{\theta}) > 0\}$ does not depend on either $k = 1, \ldots, K$ or $\boldsymbol{\theta} \in \Theta$, and the probability (density) functions $\{g_k\}$ also have the same support $\mathcal{Y}$.

**(A2)** $H_k(\hat{G}_k\,;\,F_k^{\boldsymbol{\theta}})$ is six times continuously differentiable with respect to $\boldsymbol{\theta} \in \Theta$ for any $k$.

**(A3)** Positive constants $c_1$, $c_2$, and $c_3$ exist such that, if $\boldsymbol{\theta} \in \mathcal{B}_{c_1}(\hat{\boldsymbol{\theta}}_D)$,

$$\left|\frac{\partial^q H_k(\hat{G}_k\,;\,F_k^{\boldsymbol{\theta}})}{\partial\theta_{j_1}\cdots\partial\theta_{j_q}}\right| \le c_2 \ \text{ a.s.}$$

holds for any $k$ and $j_1, \ldots, j_q \in \{1, \ldots, p\}$ with $0 \le q \le 6$, and

$$\boldsymbol{\theta} \in \Theta \setminus \mathcal{B}_{c_1}(\hat{\boldsymbol{\theta}}_D) \ \Rightarrow\ \bar{H}(\boldsymbol{y}\,;\,\boldsymbol{\theta}) - \bar{H}(\boldsymbol{y}\,;\,\hat{\boldsymbol{\theta}}_D) \ge c_3 \ \text{ a.s.}\,,$$

for any $\boldsymbol{y} \in (\mathcal{Y})^n$, where $\mathcal{B}_{c_1}(\hat{\boldsymbol{\theta}}_D)$ is the open ball of radius $c_1$ centered at the minimum divergence estimator $\hat{\boldsymbol{\theta}}_D$.

**(A4)** There exists an open set $\Theta_\star \subset \Theta$ containing the best fitting parameter such that, for any $k$, the probability (density) function $f_k(y\,|\,\boldsymbol{\theta})$ is bounded with respect to $y \in \mathcal{Y}$ and $\boldsymbol{\theta} \in \Theta_\star$ (i.e., $\sup_{y\in\mathcal{Y}} \sup_{\boldsymbol{\theta}\in\Theta_\star} |f_k(y\,|\,\boldsymbol{\theta})| \le \exists\ c_4 < +\infty$), and independent of $n$.

**(A5)** The Hessian matrix of $\bar{H}$, $\frac{\partial^2 \bar{H}(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\ \partial\boldsymbol{\theta}^\top}$, is nonsingular for any $\boldsymbol{y} \in (\mathcal{Y})^n$ and $\boldsymbol{\theta} \in \Theta_\star$.

**(A6)** For each $k = 1, \ldots, K$, there exists a constant value $q_k \in (0, 1)$ such that $n_k/n \to q_k$ as $n \to +\infty$.

(A1)–(A6) are conditions to establish the asymptotic properties of the minimum divergence estimators and to derive the model selection criteria described in the latter half of this section. In the conventional investigations regarding statistical divergence and model selection, similar or related conditions have been discussed (e.g., (Basu et al. 1998; Ghosh and Basu 2013; Kass, Tierney, and Kadane 1990; Kurata and Hamada 2020)). Conditions on robust selection will be discussed in the next section.

Now, we derive BIC-type criteria based on a wide class of divergence measures. We consider a divergence measure having the form given in Equation (5), and define a quasi-marginal distribution based on a divergence measure $D$, $m_D$, as follows:

$$m_D(\boldsymbol{Y}) = \int \exp\left\{-n\ \bar{H}(\boldsymbol{Y}\,;\,\boldsymbol{\theta})\right\}\ \pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}\,.$$

Let $\pi(\boldsymbol{\theta})$ be a proper, independent of the sample size, and four times continuously differentiable prior probability (density) function. In the same way as (Kurata and Hamada 2020), we can prove

$$-2\ \log m_D(\boldsymbol{Y}) = 2\ n\ \bar{H}(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_D) + p\ \log n + O_P(1)\,,$$

by using the Laplace approximation (e.g., (Kass, Tierney, and Kadane 1990)). Thus, we can obtain a divergence-based Bayesian criterion (DBBC) by ignoring small-order terms with respect to $n$:

$$\mathrm{DBBC}_D = 2\ n\ \bar{H}(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_D) + p\ \log n\,. \tag{7}$$

Clearly, this has the form of Equation (6) with $B = p\log n$. (Kurata and Hamada 2020) showed that, when we use BHHJ divergence with $\alpha > 0$ as $D$, the corresponding DBBC-type criterion (named RCC) has consistency and robustness of model selection simultaneously. In analogy with RCC, we can prove that DBBC-type criteria for other divergence measures have consistency of model selection under the conditions (A1)–(A6).

## 4. Robustness of model selection

An influence function has often been used to assess the damage caused by a perturbation in data-generating distribution (e.g., (Hampel et al. 1986; Huber 1983)). When using an influence function, we regard an estimator as a functional of the probability distribution that the observations follow. We can find that a model selection criterion is also a functional of the data-generating distribution since it depends on observations and the estimator. When the data-generating distribution is $\boldsymbol{G} = (G_1, ..., G_K)$, we denote the functional form of the best fitting parameter and estimator by $\boldsymbol{T}(\boldsymbol{G})$ and $\boldsymbol{T}(\hat{\boldsymbol{G}})$, respectively. Additionally, we define the functional form of a criterion having the form in Equation (6) as follows:

$$\bar{\mathcal{H}}(\boldsymbol{G}) = \sum_{k=1}^{K} \frac{n_k}{n}\, H_k\left(G_k\,;\,F_k^{\boldsymbol{T}(\boldsymbol{G})}\right).$$

We assume that, for each $k = 1, ..., K\ (1 \le K \le n)$, the data-generating distribution (hereafter, denoted by $\Omega_k$) is a mixture of the true population distribution $G_k$ and $\Upsilon_k$, a distribution differing from $G_k$:

$$\Omega_k = \Omega_k(G_k,\ \Upsilon_k,\ r) = (1-r)\ G_k + r\ \Upsilon_k\,, \tag{8}$$

where $r > 0$ is the contamination rate. Let $\boldsymbol{\Omega} = (\Omega_1, ..., \Omega_K)$ be the data-generating distributions, and let $\upsilon_k$ be the probability (density) function corresponding to distribution $\Upsilon_k$ $(k = 1, ..., K)$. To measure the sensitivity of a model selection criterion, we now evaluate the difference between contaminated distribution $\boldsymbol{\Omega}$ and non-contaminated distribution $\boldsymbol{G}$:

$$\mathcal{I}(\boldsymbol{G},\ \boldsymbol{\Upsilon},\ r) = \bar{\mathcal{H}}(\boldsymbol{\Omega}) - \bar{\mathcal{H}}(\boldsymbol{G})\,. \tag{9}$$

If a criterion fluctuates intensely due to the influence of contamination, the absolute value of $\mathcal{I}$ will be unboundedly large. By contrast, if a criterion is hardly affected, $\mathcal{I}$ will be bounded for the outlier-generating distribution $\boldsymbol{\Upsilon}$. Hence, we can utilize Equation (9) as an indicator of the sensitivity of the criterion. We consider evaluating $\mathcal{I}$ to assess robustness of a criterion.

### 4.1. Regularity conditions for robustness of model selection

In this subsection, we list conditions for proving robustness of model selection criteria. Note that the sufficient conditions are different according to the underlying divergence family: each divergence-based criterion requires some of the following conditions.

**(B1)** For some $\alpha > 0$, $\int f_k(y\,|\,\boldsymbol{\theta})^{\alpha}\ \upsilon_k(y)\,dy$ is finite for any $k$ and any $\boldsymbol{\theta} \in \Theta_\star$.
**(B2)** For some $\alpha > 0$, $\int \left\| \frac{\partial \log f_k(y\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| f_k(y\,|\,\boldsymbol{\theta})^{\alpha}\ \upsilon_k(y)\,dy$ is finite for any $k$ and any $\boldsymbol{\theta} \in \Theta_\star$.
**(B3)** For some $\alpha > 1$, the following terms are finite for any $k$ and any $\boldsymbol{\theta} \in \Theta_\star$ :

$$\int f_k(y\,|\,\boldsymbol{\theta})^{\alpha-1}\ \upsilon_k(y)^2\,dy\,, \qquad \int f_k(y\,|\,\boldsymbol{\theta})^{\alpha-1}\ g_k(y)\ \upsilon_k(y)\,dy\,,$$
$$\int \left\| \frac{\partial \log f_k(y\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| f_k(y\,|\,\boldsymbol{\theta})^{\alpha-1}\ g_k(y)\ \upsilon_k(y)\,dy\,.$$

### 4.2. Measure of influence from contamination of data-generating distribution

Since the mixture distributions $\Omega_1, \ldots, \Omega_K$ depend on contamination rate $r$, we represent $\bar{\mathcal{H}}(\boldsymbol{\Omega})$ as follows:

$$\bar{\mathcal{H}}(\boldsymbol{\Omega}) = \sum_{k=1}^{K} \frac{n_k}{n}\, H_k\left(\Omega_k\,;\, F_k^{T(\boldsymbol{\Omega})}\right) =: \sum_{k=1}^{K} \frac{n_k}{n}\ \mathcal{M}_k(r)\,.$$

Using the first-order Taylor expansion of $\bar{\mathcal{H}}(\boldsymbol{\Omega})$ around $r = 0$, we obtain

$$\mathcal{I}(\boldsymbol{G},\ \boldsymbol{\Upsilon},\ r) = r \sum_{k=1}^{K} \frac{n_k}{n}\ \mathcal{M}_k'(0) + o(r)\,.$$

Now, we evaluate $\mathcal{M}_k'(0)$ for each of the divergence measures introduced in Section 2, to examine robustness of model selection criteria from the viewpoint of whether the first-order approximation term of $\mathcal{I}$ is finite in the presence of contamination of the data-generating distribution.

**Theorem 1.** *Assume conditions (A1)–(A6), (B1), and boundedness of the influence function for the corresponding divergence. Then, criteria based on BHHJ divergence with $\alpha > 0$, JHHB divergence with $\alpha > 0$ and $\varphi \geq 0$ (including gamma divergence), and C divergence with $\alpha > 0$ satisfy that $\left|\mathcal{M}_k'(0)\right|$ is finite for arbitrary $\boldsymbol{\Upsilon}$.*

Furthermore, we examine the second-order approximation of $\mathcal{I}$ in Equation (9) to verify the robustness of a model selection criterion in more detail. Using the second-order Taylor expansion of $\bar{\mathcal{H}}$ around $r = 0$, we obtain

$$\mathcal{I}(\boldsymbol{G},\ \boldsymbol{\Upsilon},\ r) = r \sum_{k=1}^{K} \frac{n_k}{n}\ \mathcal{M}_k'(0)\ +\ \frac{r^2}{2} \sum_{k=1}^{K} \frac{n_k}{n}\ \mathcal{M}_k''(0)\ +\ o(r^2)\,.$$

The following theorem gives the conditions for the boundedness of the second-order approximation term of $\mathcal{I}$ for each divergence family.

**Theorem 2.** *Assume conditions (A1)–(A6), (B1), (B2), and boundedness of the influence function for the corresponding divergence. Then, criteria based on BHHJ divergence with $\alpha > 0$ and JHHB divergence with $\alpha > 0$ and $\varphi \geq 0$ (including gamma divergence) satisfy that $\left|\mathcal{M}_k''(0)\right|$ is finite for arbitrary $\boldsymbol{\Upsilon}$. This holds also for C divergence if we further assume (B3).*

The proofs of Theorems 1 and 2 are found in the supplemental material. To guarantee the boundedness of the second-order approximation for all elements of the C divergence family, we need to assume $\alpha > 1$. Note that, without assuming $\alpha > 1$ and the boundedness of $\upsilon_k^2$ for all $k$, we can not generally prove the boundedness of the first term in (B3), even if $f_k = g_k$ almost surely. However, in general, making $\alpha$ in the C divergence family larger increases the asymptotic variance of the estimator, and thus a too-large $\alpha$ is not desirable from the viewpoint of efficiency of estimation (e.g., (Basu et al. 1998; Maji et al. 2019)). BHHJ divergence is an exception, in that it requires only $\alpha > 0$ to guarantee the boundedness of the second-order term, as well as the first-order term. Therefore, Theorem 2 shows the superiority of BHHJ divergence in the C divergence family. Additionally, for the JHHB family including BHHJ and gamma divergence, we need not assume $\alpha > 1$ and specify the form of probability function of the outlier-generating distribution ($\upsilon_k$) for robust model selection.

### 4.3. On JHHB divergence family: Case of non-homogeneous setting

In the previous subsection, we assumed $K$ groups of distributions where the $k$-th group has $n_k$ i.i.d. observations. Under condition (A6), even if some observations in a group are outliers, we can conduct robust estimation and model selection based on robust divergence families, using other (non-outlying) observations drawn from the true distribution $G_k$. However, we often face situations that observations are independent but not identically distributed (non-homogeneous setting, see also, e.g., (Ghosh and Basu 2013)). In the non-homogeneous setting, each group has only one observation (i.e., $K = n$ and $n_k = 1$), and (A6) no longer holds since $n_k/n \to 0$ as $n \to +\infty$. Thus, if the observation is an outlier (generated from $\Upsilon_k$), this observation can have a bad influence on model selection.

We here focus on the JHHB divergence family (including BHHJ and gamma divergence). Using our definitions (Equations (2) and (3)), we replace the true unknown distribution $G_k$ by the empirical distribution based on only one data point $Y_k$ to estimate parameters and to derive model selection criteria as in (Ghosh and Basu 2013) and (Kurata and Hamada 2018). To evaluate robustness of model selection criteria, we consider the following term:

$$\bar{H}(\boldsymbol{Y}\,;\,\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n} H_k(Y_k\,;\,\boldsymbol{\theta})$$

$$= \begin{cases} \frac{1}{n}\ \sum_k \left[\frac{1}{\varphi}\left\{\int f_k(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy\right\}^{\varphi} - \frac{\alpha+1}{\varphi\ \alpha}\ f_k(Y_k\,|\,\boldsymbol{\theta})^{\varphi\alpha}\right] & (\varphi > 0)\,, \\ \frac{1}{n}\ \sum_k \left[\log\left\{\int f_k(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy\right\} - (\alpha+1)\ \log f_k(Y_k\,|\,\boldsymbol{\theta})\right] & (\varphi = 0)\,. \end{cases}$$

Since observation $Y_k$ may be an outlier, we now consider the expectation of the above formula based on the mixture distribution, $\boldsymbol{\Omega} = (\Omega_1, ..., \Omega_K)$, defined in Equation (8):

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{E}_{Y\sim\Omega_k}[H_k(Y\,;\,\boldsymbol{T}(\boldsymbol{\Omega}))]\,.$$

Using this, we can evaluate $\mathcal{I}$ (Equation (9)) in the non-homogeneous setting via Taylor expansion, as in the previous subsection.

**Theorem 3.** *Assume conditions (A1)–(A5), (B1), (B2), and boundedness of the influence function for the corresponding divergence. Then, criteria based on BHHJ divergence with $\alpha > 0$ and JHHB divergence with $\alpha > 0$ and $\varphi \geq 0$ satisfy that both the first- and second-order approximations of $\mathcal{I}$ are finite, but criteria based on gamma divergence do not.*

We describe the proof in the supplemental material. It is remarkable that criteria based on gamma divergence have robustness when condition (A6) holds, but lose their robustness in the non-homogeneous settings. By contrast, criteria based on BHHJ divergence maintain robustness of model selection whether (A6) holds or not.

## 5. Numerical results

In this section, we verify the performance of model selection criteria based on BHHJ and its related divergence families through some numerical experiments. As $N \in \mathcal{N}$ of the C divergence family excluding BHHJ divergence, we use the following function:

$$N_\lambda^{\mathrm{GP}}(z) = \frac{(z+1)^{\lambda+1}}{\lambda\ (\lambda+1)} - \frac{z+1}{\lambda\ (\lambda+1)} - \frac{z}{\lambda+1}\,, \quad \lambda \neq -1,\ 0. \tag{10}$$

This function $N_\lambda^{\mathrm{GP}}$ is used for generalized power (GP) divergence (e.g., (Maji et al. 2019; Vonta, Mattheou, and Karagrigoriou 2012)). When $\alpha \to 0$, GP divergence coincides with power divergence as proposed by (Cressie and Read 1984). We here consider the DBBC-type criteria defined in Equation (7).

### 5.1. Selection of normal models using real dataset

In the first experiment, we consider a selection problem of continuous distributions using a real dataset consisting of the parallax of the sun (Short 1763). I randomly split the dataset with sample size $n = 158$ into $K = 2$ groups ($n_1 = n_2 = 79$) and assumed that the observations independently followed normal distributions: $Y_1^1, ..., Y_{n_1}^1 \overset{i.i.d.}{\sim} G_1 = N(\mu_1,\ s_1)$ and

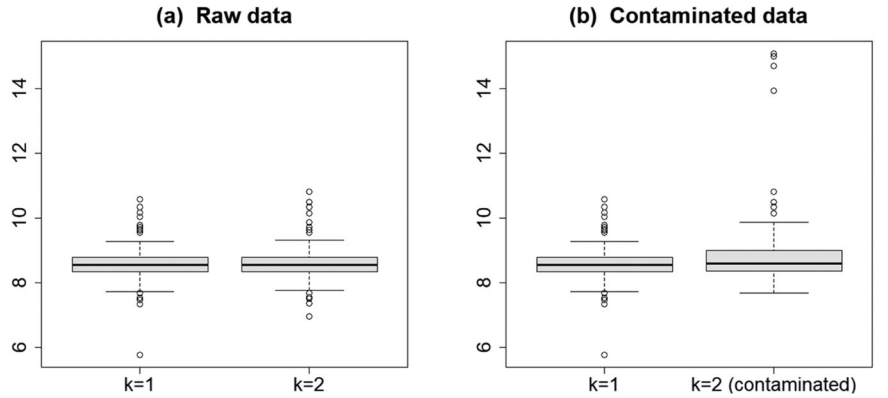



**Figure 1.** (a) Boxplots of the two groups of raw data of the parallax of the sun (in seconds of a degree). (b) Boxplots of contaminated data. The $k = 1$ group (left) is the same as that in (a), while the $k = 2$ group (right) is contaminated.

$Y_1^2, \ldots, Y_{n_2}^2 \overset{i.i.d.}{\sim} G_2 = N(\mu_2,\ s_2)$. Figure 1a shows the boxplots of the two groups. For this dataset, we compare the following four models:

$$\begin{cases} \text{Model I} & : \quad \mu_1 = \mu_2,\ s_1 = s_2, \\ \text{Model II} & : \quad \mu_1 = \mu_2,\ s_1 \neq s_2, \\ \text{Model III} & : \quad \mu_1 \neq \mu_2,\ s_1 = s_2, \\ \text{Model IV} & : \quad \mu_1 \neq \mu_2,\ s_1 \neq s_2. \end{cases}$$

I employed DBBC-type criteria based on KL, BHHJ, JHHB (including gamma), and GP divergence. As the tuning parameter $\alpha$, I used 0.50 and 1.25 in this experiment, to verify the difference between $\alpha < 1$ and $\alpha > 1$. I also used other tuning parameters as follows: $\varphi = 0.50$ for JHHB divergence (Equation (2)), and $\lambda = 1.25$ for GP divergence (Equation (10)). As we need to estimate the probability density functions of the $G$'s for GP divergence, I utilized the Gaussian kernel density (e.g., (Silverman 2018)). As a consequence, all of DBBC-type criteria selected Model I. This result agreed with the boxplots in Figure 1a.

Next, I replaced 5% (4 of 79) in the second population with twice their original value, and conducted the selection problem for the synthetic contaminated dataset. Figure 1b shows the boxplots of the contaminated observations. As a result, BIC selected Model IV, and criteria based on GP divergence with $\alpha = 0.50$ selected Model III. By contrast, the criterion based on GP divergence with $\alpha = 1.25$ selected Model I. Criteria based on BHHJ, JHHB, and gamma divergence with both $\alpha = 0.50$ and 1.25 selected Model I. These results supported Theorem 2: criteria based on the C divergence family (excepting BHHJ divergence) require $\alpha > 1$ to guarantee the boundedness of the second-order approximation of $\mathcal{I}$, the difference of the value of a criterion between the contaminated and non-contaminated data-generating distributions. On the other hand, those of the JHHB divergence family (including BHHJ and gamma divergence) require $\alpha > 0$. In this experiment, I also verified for another values of the tuning parameter of BHHJ and gamma divergence: $\alpha = 0.001$, 0.01, 0.10, 0.20, ..., 1.40, and 1.50, consequently, BHHJ

divergence-based criteria with $\alpha \geq 0.01$ and gamma divergence-based ones with $\alpha \geq 0.10$ selected Model I in both of the experiment.

Note that, the optimal value or range of the tuning parameter has previously been investigated. For example, (Mantalos, Mattheou, and Karagrigoriou 2010) showed that $\alpha = 0.25$ is relatively suitable as the tuning parameter of their proposed model selection criterion based on BHHJ divergence, in auto-regressive models. (Kurata and Hamada 2018) and (Kurata and Hamada 2020) pointed out that AIC- and BIC-type criteria based on BHHJ divergence with some large $\alpha$ (for example, $\alpha \geq 0.50$) perform well in heavily contaminated cases. (Riani et al. 2020) proposed an estimation procedure via the approach of S-estimation. (Basak, Basu, and Jones 2021) developed an iterated algorithm to determine one optimal value of $\alpha$ for parametric estimation from the observations, by expanding the methods in (Warwick and Jones 2005) and (Ghosh and Basu 2015). It has been pointed out that parametric estimation and model selection with large $\alpha$ can become unstable when the sample size is not sufficiently large (e.g., (Basu et al. 1998; Ghosh and Basu 2013; Kurata and Hamada 2019)). Actually, asymptotic variance of an estimator increases with increasing $\alpha$ (e.g., (Jones et al. 2001; Ghosh and Basu 2016)). For the above reasons, several studies have assumed $\alpha \leq 1$ in order to conduct stable estimation (e.g., (Basak, Basu, and Jones 2021; Fujisawa and Eguchi 2006; Maji et al. 2019)). Thus, from the viewpoint of efficiency, it is preferable to employ $\alpha < 1$ as the tuning parameter of C divergence; however, use of $\alpha > 1$ is required to guarantee robustness in model selection problems (Theorem 2).

### 5.2. Selection of discrete distributions

Next, we consider the problem of selecting an adequate probability distribution from three discrete distributions, the geometric, Poisson, and negative binomial distributions, which have the following probability functions:

$$\begin{aligned} f^{\mathrm{Ge}}(y) &= (1-\rho)^{y}\ \rho\,, \\ f^{\mathrm{Po}}(y) &= \frac{\mu^{y}}{\Gamma(y+1)}\ \mathrm{e}^{-\mu}\,, \\ f^{\mathrm{Nb}}(y) &= \frac{\Gamma(y+\kappa)}{\Gamma(y+1)\ \Gamma(\kappa)}\ \left(\frac{\mu}{\mu+\kappa}\right)^{y}\ \left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa}, \end{aligned} \tag{11}$$

where $\rho \in (0,\ 1)$, $\mu > 0$, and $\kappa > 0$ are the parameters in the respective probability distributions, and $\Gamma$ is the gamma function. Hereafter, we use the notation $Y \sim \mathrm{Ge}(\rho)$, $Y \sim \mathrm{Po}(\mu)$, and $Y \sim \mathrm{Nb}(\mu, \kappa)$ to indicate that a random variable $Y$ has $f^{\mathrm{Ge}}$, $f^{\mathrm{Po}}$, and $f^{\mathrm{Nb}}$, respectively, in Equation (11) as its probability function. It can be confirmed immediately that the negative binomial distribution includes the other two distributions: when $Y \sim \mathrm{Nb}(\mu, \kappa)$, $Y$ has a geometric distribution $\mathrm{Ge}(1/(\mu+1))$ if $\kappa = 1$, and $\mathrm{Nb}(\mu, \kappa)$ converges to a Poisson distribution $\mathrm{Po}(\mu)$ if $\kappa \to +\infty$. In this experiment, I assumed that $n = 10000$ samples were Poisson distributed with parameter $\mu = 50$: $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathrm{Po}(50)$. This is an i.i.d. setting ($K = 1$ in Equation (5)). To generate outliers, I chose some of the observations randomly and doubled their values. As the

**Table 1.** Selection rates (%) of the DBBC-type criteria based on KL, BHHJ, JHHB, gamma, and GP divergence for different contamination rates in the selection problem of discrete distributions. 'Ge', 'Po', and 'Nb' indicate the geometric, Poisson, and negative binomial distribution, respectively.

| | 0 % | | | 1 % | | | 10 % | | |
|---|---|---|---|---|---|---|---|---|---|
| Contamination rate | Ge | Po | Nb | Ge | Po | Nb | Ge | Po | Nb |
| KL | 0 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 100 |
| BHHJ ($\alpha = 0.50$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 88 | 12 |
| BHHJ ($\alpha = 1.25$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 100 | 0 |
| JHHB ($\alpha = 0.50$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 82 | 18 |
| JHHB ($\alpha = 1.25$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 97 | 3 |
| Gamma ($\alpha = 0.50$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 93 | 7 |
| Gamma ($\alpha = 1.25$) | 0 | 100 | 0 | 0 | 99 | 1 | 0 | 99 | 1 |
| GP ($\alpha = 0.50$) | 0 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 100 |
| GP ($\alpha = 1.25$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 100 | 0 |

tuning parameters of each divergence measure, I used the same values as in the previous experiment (Subsection 5.1).

Table 1 shows the selection rates of 1000 trials of the DBBC-type criteria derived from each of the divergence measures. From Table 1, we confirm that criteria based on KL divergence selected the adequate model (Po) in non-contaminated cases, but they did not cope well with outliers. By contrast, criteria based on BHHJ, JHHB, and gamma divergence performed well in all three settings regardless of whether $\alpha < 1$ or $\alpha > 1$ was used. Additionally, the criterion based on the C divergence family (GP) with $\alpha < 1$ was negatively influenced by outliers, but that with $\alpha > 1$ maintained accuracy even in contaminated cases. As with the previous experiment, these results agreed with Theorems 1 and 2.

### 5.3. Case of non-homogeneous setting: Negative binomial regression model

To verify the performance in cases of the non-homogeneous setting (see also Subsection 4.3), we finish by showing the result of an experiment using the negative binomial regression model, which is a generalized linear model and is suitable for cases where the response variable is a non-negative integer (e.g., (Allison and Waterman 2002; Kurata, Kuroda, and Komaki 2022; Lawless 1987)). Let $y_k$ and $\boldsymbol{x}_k = (1,\ x_{k,1}, \ldots, x_{k,p})$ be the response variable and the (standardized) explanatory variable vector related to the $k$-th individual, respectively ($k = 1, \ldots, K$). I assumed some sample sizes $n$, and $Y_k \overset{indep.}{\sim} \mathrm{Nb}(\mu_k,\ 1)$ with

$$\mu_k = \mu_k(\boldsymbol{x}_k\,;\,\boldsymbol{\beta}) = \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_k\right).$$

The expectation, $\mu_k$, depends on the explanatory variables and regression coefficient vector $\boldsymbol{\beta} = (\beta_0,\ \beta_1, \ldots, \beta_p)$. Since the value of $\mu_k$ differs by $k$, this is the non-homogeneous setting, i.e., $n_k = 1$ for all $k$. In this experiment, I supposed four candidate explanatory variables generated by random numbers, and further assumed that two of the four explanatory variables were unnecessary. I set the true coefficient of two of the four explanatory variables equal to zero: $\beta_0^{\mathrm{true}} = 1.50$, $\beta_1^{\mathrm{true}} = 0$, $\beta_2^{\mathrm{true}} = 0$, $\beta_3^{\mathrm{true}} = 0.50$, and $\beta_4^{\mathrm{true}} = -0.50$. In this subsection, we refer to models with some of the necessary

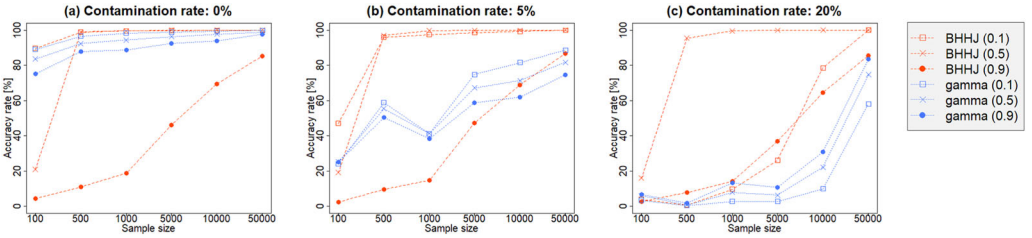



**Figure 2.** Accuracy rates of DBBC-type criteria (Equation (7)) based on BHHJ and gamma divergence (with $\alpha = 0.1, 0.5, 0.9$) in the selection problem of negative binomial regression models for different sample sizes and contamination rates.

**Table 2.** Selection rates (%) of the DBBC-type criteria based on KL, BHHJ, and gamma divergence for different contamination rates in the selection problem of negative binomial regression models with sample size $n = 500$. 'Under', 'Adequate', and 'Over' indicate the selection rates of under-specified models, the adequate model, and over-specified models, respectively.

| | 0 % | | | 5 % | | | 20 % | | |
|---|---|---|---|---|---|---|---|---|---|
| Contamination rate | Under | Adequate | Over | Under | Adequate | Over | Under | Adequate | Over |
| KL | 0 | 98 | 2 | 25 | 60 | 16 | 100 | 0 | 0 |
| BHHJ ($\alpha = 0.01$) | 0 | 98 | 2 | 22 | 63 | 15 | 100 | 0 | 0 |
| BHHJ ($\alpha = 0.10$) | 0 | 99 | 1 | 0 | 96 | 4 | 99 | 1 | 0 |
| BHHJ ($\alpha = 0.30$) | 0 | 100 | 0 | 0 | 100 | 0 | 0 | 100 | 0 |
| BHHJ ($\alpha = 0.50$) | 0 | 99 | 0 | 2 | 97 | 1 | 4 | 95 | 1 |
| BHHJ ($\alpha = 0.70$) | 41 | 55 | 5 | 45 | 51 | 4 | 57 | 40 | 3 |
| BHHJ ($\alpha = 0.90$) | 84 | 11 | 5 | 85 | 10 | 5 | 90 | 8 | 2 |
| Gamma ($\alpha = 0.01$) | 0 | 98 | 2 | 24 | 60 | 16 | 100 | 0 | 0 |
| Gamma ($\alpha = 0.10$) | 0 | 96 | 4 | 23 | 59 | 19 | 100 | 0 | 0 |
| Gamma ($\alpha = 0.30$) | 0 | 95 | 5 | 18 | 58 | 24 | 99 | 1 | 0 |
| Gamma ($\alpha = 0.50$) | 0 | 92 | 8 | 16 | 55 | 29 | 99 | 1 | 0 |
| Gamma ($\alpha = 0.70$) | 0 | 90 | 10 | 14 | 53 | 33 | 98 | 1 | 0 |
| Gamma ($\alpha = 0.90$) | 0 | 88 | 12 | 11 | 50 | 38 | 98 | 2 | 0 |

explanatory variables missing as under-specified models, and models employing unnecessary variables as over-specified models.

Figure 2 shows the accuracy rates of 1000 trials of the DBBC-type criteria based on BHHJ and gamma divergence, for some sample sizes and contamination rates. Table 2 also shows the detailed selection rates for $n = 500$. Additionally, in the supplemental material, detailed results of model selection for various sample sizes are shown. We can see that most criteria adequately selected the adequate model in the non-contaminated cases when the sample size is large, due to the consistency of DBBC-type criteria, however, criteria based on KL and gamma divergence failed to select the correct model in the contaminated cases. We can also confirm that BHHJ divergence-based criteria with too small $\alpha$ (e.g., 0.01 and 0.10) were negatively influenced by outliers. Moreover, when the sample size is not sufficiently large, criteria with large $\alpha$ (e.g., 0.70 and 0.90) did not perform well because parametric estimation with large $\alpha$ generally results in lack of efficiency, as mentioned in Section 4. Criteria based on BHHJ divergence for middle $\alpha$ recorded high accuracy in cases from non-contaminated to heavily contaminated. In the case of $n = 500$, we can confirm that BHHJ divergence with $\alpha = 0.30$ and 0.50 performed stably for every contamination rate (Table 2).

## 6. Conclusion

In this paper, we discussed robustness of model selection criteria having form of Equation (6) based on divergence measures as evaluated by the approximate difference between contaminated and non-contaminated data-generating distributions. We investigated sufficient conditions to establish robustness of model selection for several divergence families, and thereby showed that the conditions are different depending on the family. Although criteria derived from non-robust divergence measures (such as KL and $f$-divergence) perform well in non-contaminated settings, such methods tend to exhibit decreased accuracy if observations are contaminated by outliers. From the viewpoint of the stability of the value of a criterion, we proved that the BHHJ divergence family and most of the related divergence families have robustness against contamination of the data-generating distribution under appropriate conditions. The C divergence family, one generalization of BHHJ divergence and $f$-divergence, inherits the robustness of BHHJ divergence in parametric estimation and hypothesis testing; nevertheless, to guarantee robustness of the C divergence-based methods in model selection, we must assume stricter conditions in addition to those for BHHJ divergence (Theorems 1, 2, and experiments in Subsections 5.1, 5.2). The conditions for the C divergence family require $\alpha > 1$, but this restriction may induce unstable results since parametric estimation via large $\alpha$ is generally inefficient. Additionally, when using C divergence in continuous populations, we must replace the true (unknown) distribution with some kind of estimated distribution such as kernel density, whereas BHHJ divergence-based criteria do not require such replacement. When employing the JHHB divergence family, another generalized family of BHHJ divergence including the gamma divergence family, we can prove robustness under fewer conditions than those for the C divergence family; however, members of the JHHB family excepting BHHJ and gamma divergence are disadvantageous in that they do not have exact unbiasedness in parametric estimation. Moreover, we proved that criteria based on gamma divergence are sensitive to outliers in the non-homogeneous setting (Theorem 3 and Subsection 5.3). Through the discussion and experiments in this paper, we identified that (i) generalized divergence families do not necessarily inherit the robust features and conditions for achieving robustness in model selection problems; (ii) not all divergence measures having estimation robustness have selection robustness; and (iii) criteria based on BHHJ divergence can perform robustly under relatively weak restrictions in a broad range of settings.

## Acknowledgements

I deeply thank the editors and reviewers for their valuable comments and suggestions. This work was supported by JSPS KAKENHI Grant Number JP20K19753. R, a software environment for statistical computing and graphics (R Core Team 2020), was used for the data analysis.

## Declaration of interest statement

The author has no conflicts of interest to declare.

## ORCID

Sumito Kurata http://orcid.org/0000-0002-3001-6594

## References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium*, ed. B. N. Petrox and F. Caski, 267–81. Budapest. Akadémiai Kiadó.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6):716–23. doi:10.1109/TAC.1974.1100705.

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* 30 (1):9–14. doi:10.1007/BF02480194.

Allison, P. D., and R. P. Waterman. 2002. Fixed-effects negative binomial regression models. *Sociological Methodology* 32 (1):247–65. doi:10.1111/1467-9531.00117.

Amari, S. 2009. α-Divergence is unique, belonging to both *f*-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory* 55 (11):4925–31. doi:10.1109/TIT.2009.2030485.

Amari, S., and H. Nagaoka. 2000. *Methods of information geometry*. New York: AMS and Oxford University Press.

Basak, S., A. Basu, and M. C. Jones. 2021. On the 'optimal' density power divergence tuning parameter. *Journal of Applied Statistics* 48 (3):536–56. doi:10.1080/02664763.2020.1736524.

Basu, A., A. Mandal, N. Martin, and L. Pardo. 2013. Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics* 65 (2):319–48. doi:10.1007/s10463-012-0372-y.

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones. 1998. Robust and efficient estimation by minimizing a density power divergence. *Biometrika* 85 (3):549–59. doi:10.1093/biomet/85.3.549.

Broniatowski, M., A. Toma, and I. Vajda. 2012. Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference* 142 (9):2574–85. doi:10.1016/j.jspi.2012.03.019.

Broniatowski, M., and I. Vajda. 2012. Several applications of divergence criteria in continuous families. *Kybernetika* 48 (4):600–36.

Cressie, N., and T. R. C. Read. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 46 (3):440–64. doi:10.1111/j.2517-6161.1984.tb01318.x.

Csiszár, I. 1967. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica* 2:229–318.

Fujisawa, H., and S. Eguchi. 2006. Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference* 136 (11):3989–4011. doi:10.1016/j.jspi.2005.03.008.

Fujisawa, H., and S. Eguchi. 2008. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* 99 (9):2053–81. doi:10.1016/j.jmva.2008.02.004.

Ghosh, A., and A. Basu. 2013. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics* 7 (1):2420–56. doi:10.1214/13-EJS847.

Ghosh, A., and A. Basu. 2015. Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: The density power divergence approach. *Journal of Applied Statistics* 42 (9):2056–72. doi:10.1080/02664763.2015.1016901.

Ghosh, A., and A. Basu. 2016. Robust estimation in generalized linear models: The density power divergence approach. *Test* 25 (2):269–90. doi:10.1007/s11749-015-0445-3.

Ghosh, A., I. R. Harris, A. Maji, A. Basu, and L. Pardo. 2017. A generalized divergence for statistical inference. *Bernoulli* 23 (4A):2746–83. doi:10.3150/16-BEJ826.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust statistics: The approach based on influence functions*. New York: Wiley.

Huber, P. J. 1983. Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association* 78 (381):66–72. doi:10.1080/01621459.1983.10477928.
Jones, M. C., N. L. Hjort, I. R. Harris, and A. Basu. 2001. A comparison of related density-based minimum divergence estimators. *Biometrika* 88 (3):865–73. doi:10.1093/biomet/88.3.865.
Kass, R. E., L. Tierney, and J. B. Kadane. 1990. The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics* 7:473–88.
Konishi, S., and G. Kitagawa. 1996. Generalised information criteria in model selection. *Biometrika* 83 (4):875–90. doi:10.1093/biomet/83.4.875.
Konishi, S., T. Ando, and S. Imoto. 2004. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* 91 (1):27–43. doi:10.1093/biomet/91.1.27.
Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22 (1):79–86. doi:10.1214/aoms/1177729694.
Kurata, S., and E. Hamada. 2018. A robust generalization and asymptotic properties of the model selection criterion family. *Communications in Statistics - Theory and Methods* 47 (3):532–47. doi:10.1080/03610926.2017.1307405.
Kurata, S., and E. Hamada. 2019. A discrete probabilistic model for analyzing pairwise comparison matrices. *Communications in Statistics - Theory and Methods* 48 (15):3801–15. doi:10.1080/03610926.2018.1481975.
Kurata, S., and E. Hamada. 2020. On the consistency and the robustness in model selection criteria. *Communications in Statistics - Theory and Methods* 49 (21):5175–95. doi:10.1080/03610926.2019.1615093.
Kurata, S., R. Kuroda, and F. Komaki. 2022. Statistical modeling for temporal dominance of sensations data incorporating individual characteristics of panelists: An application to data of milk chocolate. *Journal of Food Science and Technology* 59 (6):2420–8. doi:10.1007/s13197-021-05260-9.
Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* 15 (3):209–25. doi:10.2307/3314912.
Maji, A., A. Ghosh, A. Basu, and L. Pardo. 2019. Robust statistical inference based on the C-divergence family. *Annals of the Institute of Statistical Mathematics* 71 (5):1289–322. doi:10.1007/s10463-018-0678-5.
Mantalos, P., K. Mattheou, and A. Karagrigoriou. 2010. An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics - Simulation and Computation* 39 (5):865–79. doi:10.1080/03610911003650391.
Mattheou, K., S. Lee, and A. Karagrigoriou. 2009. A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference* 139 (2):228–35. doi:10.1016/j.jspi.2008.04.022.
Morimoto, T. 1963. Markov processes and the H-theorem. *Journal of the Physical Society of Japan* 18 (3):328–31. doi:10.1143/JPSJ.18.328.
Ninomiya, Y., and S. Kawano. 2016. AIC for the Lasso in generalized linear models. *Electronic Journal of Statistics* 10 (2):2537–60. doi:10.1214/16-EJS1179.
Pardo, L. 2005. *Statistical inference based on divergence measures*. Boca Raton, FL: Chapman Hall/CRC Press.
R Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria.
Riani, M., A. C. Atkinson, A. Corbellini, and D. Perrotta. 2020. Robust regression with density power divergence: Theory, comparisons, and data analysis. *Entropy* 22 (4):399. doi:10.3390/e22040399.
Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461–4. doi:10.1214/aos/1176344136.
Short, J. 1763. Second paper concerning the parallax of the Sun determined from the observations of the late transit of Venus, in which this subject is treated of more at length, and the quantity of the parallax more fully ascertained. *Philosophical Transactions of the Royal Society of London* 53:300–45.
Silverman, B. W. 2018. *Density estimation for statistics and data analysis*. New York: Routledge.
Toma, A. 2010. Robust tests based on density power divergence estimators and saddlepoint approximations. *Mathematical Reports* 12:383–92.

Toma, A. 2014. Model selection criteria using divergences. *Entropy* 16 (5):2686–98. doi:10.3390/e16052686.

Umezu, Y., Y. Shimizu, H. Masuda, and Y. Ninomiya. 2019. AIC for the non-concave penalized likelihood method. *Annals of the Institute of Statistical Mathematics* 71 (2):247–74. doi:10.1007/s10463-018-0649-x.

Vonta, F., K. Mattheou, and A. Karagrigoriou. 2012. On properties of the $(\Phi, a)$-power divergence family with applications in goodness of fit tests. *Methodology and Computing in Applied Probability* 14 (2):335–56.

Warwick, J., and M. C. Jones. 2005. Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation* 75 (7):581–8. doi:10.1080/00949650412331299120.