A robust generalization and asymptotic properties of the model selection criterion family

Kurata, Sumito
Graduate School of Engineering Science, Osaka University

Hamada, Etsuo
Graduate School of Engineering Science, Osaka University

出版情報：Communications in Statistics - Theory and Methods. 47 (3), pp.532-547, 2017-09-11. Taylor and Francis
バージョン：
権利関係：

Journal: Communications in Statistics - Theory and Methods

A ROBUST GENERALIZATION AND ASYMPTOTIC PROPERTIES OF THE MODEL SELECTION CRITERION FAMILY

Sumito Kurata and Etsuo Hamada
Graduate School of Engineering Science
Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, JAPAN.
kurata@sigmath.es.osaka-u.ac.jp

Key Words: Model selection; BHHJ divergence; Nonhomogeneous data; Robustness; Polynomial regression.

Abstract

When selecting a model, robustness is a desirable property. However, most model selection criteria that are based on the Kullback-Leibler divergence tend to have reduced performance when the data are contaminated by outliers. In this paper, we derive and investigate a family of criteria that generalize the Akaike information criterion (AIC). When applied to a polynomial regression model, in the noncontaminated case, the performance of this family of criteria is asymptotically equal to that of the AIC. Moreover, the proposed criteria tend to maintain sufficient levels of performance even in the presence of outliers.

1 Introduction

To evaluate a statistical model of a phenomenon, we often use a measure of the statistical divergence to measure the "farness" (not the mathematical distance) between the true distribution and that of the parametric model. This is the basis of the various model selection criteria that are based on the divergence.

A representative criterion is the Akaike information criterion (AIC), which was proposed by [Akaike (1974)]. The AIC has been further developed in many studies; for example, [Sugiura (1978)] and [Hurvich and Tsai (1989)] corrected the bias of the AIC so that it could be used with small samples, [Takeuchi (1976)] suggested an alternative of the AIC with weakened restrictions (the TIC), and [Konishi and Kitagawa (1996)] generalized the TIC so that it could be used for any estimation method (the GIC). The AIC and many related criteria originate from the divergence that was established by [Kullback and Leibler (1951)], which is known as the Kullback-Leibler (KL) divergence. Also, The Bayesian information criterion (BIC), which was proposed by [Schwarz (1978)], has the same main term as does the AIC. [Nishii (1984)] examined the asymptotic consistency of the criteria based on the KL divergence.

Many divergence measures have been proposed (for example, see [Read and Cressie (1988)], [Pardo (2005)]). One such example is the BHHJ divergence, which was defined by [Basu, *et al.* (1998)] (for more information, see [Basu, *et al.* (2011)]; it is also known as the density power divergence or the beta divergence). This family of measures is characterized by a nonnegative parameter α and converges to the KL divergence as α goes to zero. An advantage of the BHHJ divergence family is that the parameter estimation is robust when $\alpha > 0$; this is due to the weighting assigned to outliers. Under appropriate assumptions, in terms of the influence function, the gross error sensitivity, and the breakdown point, the methods based on the BHHJ divergence achieve better estimations than that of the maximum likelihood estimation (see [Basu, *et al.* (1998)]).

In general, when using a divergence measure, it is assumed that the observed data are independent and identically distributed (i.i.d.). However, in actual applications, the data often follow different distributions. For instance, in a polynomial regression model, the various response variables may have different distributions because they have separate explanatory variables, even if all of the error terms have a common distribution. [Ghosh and Basu (2013)] adapted the BHHJ divergence to independent but not identically distributed data.

We propose a family of model selection criteria based on the non-identically distributed

version of the BHHJ divergence, and we will analyze its properties in detail. This family of criteria can be regarded as a generalization of the AIC, but a different one than the GIC family. As with the original divergence measures, these criteria depend on a tuning parameter α, and the selection result converges to that of the AIC as α goes to zero. This family of criteria tend to maintain selection accuracy even when there are outliers in the observed data.

Our paper is arranged as follows. In Section 2, we define a family of criteria based on the BHHJ divergence (the BHHJ-C), and its asymptotic selection probability is shown in Section 3. An important advantage of the BHHJ-C is that it produces a robust estimate; therefore, in Section 4, we evaluate the robustness of the model selection when the distribution is contaminated by an outlier. In Section 5, we present numerical simulations to compare the model selection of the BHHJ-C to that of conventional methods. Our conclusions are presented in Section 6. Some derivations and proofs are presented in the Appendix.

2 A family of BHHJ divergence-based criteria

Let G be a probability distribution, and let $F_{\boldsymbol{\theta}}$ be a statistical model with respect to G. The BHHJ divergence family is a measure of the farness between G and $F_{\boldsymbol{\theta}}$ and is defined by

$$d_\alpha(G\,;\,F_{\boldsymbol{\theta}}) \;=\; \int f(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \frac{\alpha+1}{\alpha}\int f(y\,|\,\boldsymbol{\theta})^{\alpha}\,dG(y) + \frac{1}{\alpha}\int g(y)^{\alpha+1}\,dy\,, \tag{1}$$

where g and f are the probability (density) functions of G and $F_{\boldsymbol{\theta}}$. The family has a tuning parameter α (≥ 0). In particular, (1) converges to $d_0(G\,;\,F_{\boldsymbol{\theta}}) = \int \log \frac{g(y)}{f(y\,|\,\boldsymbol{\theta})}\,dG(y)$ as α goes to zero, which is the KL divergence. Although it is usually assumed that the data are i.i.d., for example, a polynomial regression model breaks the assumption of homogeneity. Each explanatory variable $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ $(i = 1, \ldots, n)$ has a different distribution, even if the error terms $\epsilon_1\,,\,\ldots\,,\,\epsilon_n$ are i.i.d. with $N(0, s)$.

Let Y_i be independently distributed as G_i $(i = 1, \ldots, n)$, let $F_{i,\boldsymbol{\theta}}$ be a parametric model with respect to G_i, and let $\boldsymbol{\theta}(\in \Theta)$ be a common unknown parameter. [Ghosh and Basu (2013)] adapted the BHHJ divergence for several independent but non-identical cases. They pro-

posed a measure of the overall farness between $\boldsymbol{G} = (G_1\,,\,\ldots\,,\,G_n)^T$ and $\boldsymbol{F}_{\boldsymbol{\theta}} = (F_{1,\boldsymbol{\theta}}\,,\,\ldots\,,\,F_{n,\boldsymbol{\theta}})^T$ as

$$d_\alpha(\boldsymbol{G}\,;\,\boldsymbol{F}_{\boldsymbol{\theta}}) \;=\; \frac{1}{n}\sum_{i=1}^{n} d_0(G_i\,;\,F_{i,\boldsymbol{\theta}}). \tag{2}$$

With respect to (2), we define $\boldsymbol{\theta}_\alpha = \arg\min_{\boldsymbol{\theta}} d_\alpha(\boldsymbol{G}\,;\,\boldsymbol{F}_{\boldsymbol{\theta}})$ as the *best fitting parameter*. Suppose that the parameter space Θ of the candidate models is a subset of $\mathbf{R}^p$, and there is an open subset $\Theta_O \subset \Theta$ that contains the best fitting parameter for an arbitrary $\alpha \geq 0$.

The minimum BHHJ divergence estimator (*BHHJ-MDIVE*) $\hat{\boldsymbol{\theta}}_\alpha$ is defined as the argument of the minimum of $H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, where we have the following definition:

$$H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}) \;=\; \frac{1}{n}\,\sum_{i=1}^{n}\left\{\int f_i(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \frac{\alpha+1}{\alpha}\,f_i(Y_i\,|\,\boldsymbol{\theta})^\alpha\right\}. \tag{3}$$

This is interpreted as the farness between the observed data $\boldsymbol{Y} = (Y_1\,,\,\ldots\,,\,Y_n)^T$ and the model $\boldsymbol{F}_{\boldsymbol{\theta}}$. Additionally, since the log-likelihood can be regarded as a part of $-d_0(\boldsymbol{G}\,;\,\boldsymbol{F}_{\boldsymbol{\theta}})$, and since $-H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ is a section of $d_\alpha(\boldsymbol{G}\,;\,\boldsymbol{F}_{\boldsymbol{\theta}})$ that converges to $d_0(\boldsymbol{G}\,;\,\boldsymbol{F}_{\boldsymbol{\theta}})$ as α goes to zero, we can consider that $-H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ is a quasi-likelihood. The expectation of (3) with respect to $\boldsymbol{G}$ is

$$H^\star_\alpha(\boldsymbol{\theta}) \;=\; \frac{1}{n}\,\sum_{i=1}^{n}\left\{\int f_i(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \frac{\alpha+1}{\alpha}\int f_i(y\,|\,\boldsymbol{\theta})^\alpha\,dG_i(y)\right\}, \tag{4}$$

where this is the part of the BHHJ divergence (2) that is related to the model $\boldsymbol{F}_{\boldsymbol{\theta}}$.

An important advantage of the BHHJ-MDIVE is that it is robust. It is estimated by

$$\sum_{i=1}^{n}\left\{\int \boldsymbol{u}_i(y\,;\,\boldsymbol{\theta})\,f_i(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \boldsymbol{u}_i(Y_i\,;\,\boldsymbol{\theta})\,f_i(Y_i\,|\,\boldsymbol{\theta})^\alpha\right\} \;=\; \mathbf{0}\,, \tag{5}$$

where $\boldsymbol{u}_i(y\,;\,\boldsymbol{\theta}) \;=\; \frac{\partial \log f_i(y\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is a score function. Equation (5) is a generalization of the MLE. The second term in (5) depends on the observed data and is the mean of the score functions weighted by the probability (density) functions. Compared to the MLE, the BHHJ-MDIVE has better gross error sensitivity and a better breakdown point ([Basu, *et al.* (1998)], [Ghosh and Basu (2013)]).

By using Theorem 3.1 of [Ghosh and Basu (2013)], the BHHJ-MDIVE $\hat{\boldsymbol{\theta}}_\alpha$ is a consistent estimator for the best fitting parameter, and $\sqrt{n}\,\hat{\boldsymbol{\theta}}_\alpha$ is asymptotically distributed with the

normal distribution and the asymptotic variance-covariance matrix $\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha)\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}$.

$$\begin{aligned}\boldsymbol{J}_\alpha(\boldsymbol{\theta}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{J}_\alpha^{(i)}(\boldsymbol{\theta}) = \tfrac{1}{n}\sum \mathbf{E}_{G_i}\left[\tfrac{\partial^2 d_\alpha(\hat{G}_i\,;F_{i,\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right] = \tfrac{\partial^2 H_\alpha^\star(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\,,\\ \boldsymbol{K}_\alpha(\boldsymbol{\theta}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{K}_\alpha^{(i)}(\boldsymbol{\theta}) = \tfrac{1}{n}\sum \mathbf{E}_{G_i}\left[\tfrac{\partial d_\alpha(\hat{G}_i\,;F_{i,\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\tfrac{\partial d_\alpha(\hat{G}_i\,;F_{i,\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}^T}\right]\,.\end{aligned} \tag{6}$$

Note that in (6), $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ and $\boldsymbol{K}_\alpha(\boldsymbol{\theta})$ have the same limiting one, which is the Fisher information matrix, as α goes to zero. The asymptotic variance increases monotonically with α; hence, a smaller α implies better efficiency, although a larger α implies greater robustness.

We make the following regularity conditions $(i = 1, \ldots, n)$:

(C0) The full model contains the true distribution $\boldsymbol{G}$.

(C1) The support $\mathcal{Y}_S = \{y \mid f_i(y\,|\,\boldsymbol{\theta}) > 0\}$ does not depend on either i or $\boldsymbol{\theta}$, and the true probability (density) functions $\{g_i\}$ also have the support $\mathcal{Y}_S$.

(C2) The probability (density) functions $\{f_i(y\,|\,\boldsymbol{\theta})\}$ are in the C^3 class for almost all $y \in \mathcal{Y}_S$ and for all $\boldsymbol{\theta} \in \Theta_O$.

(C3) For arbitrary i, the integrals $\int f_i(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy$ and $\int f_i(y\,|\,\boldsymbol{\theta})^\alpha\,dG_i(y)$ are three times differentiable with respect to $\boldsymbol{\theta} \in \Theta_O$, and the differential and the integral are exchangeable.

(C4) For $\boldsymbol{\theta} \in \Theta_O$, the matrices $\boldsymbol{J}_\alpha^{(1)}(\boldsymbol{\theta})\,,\ \ldots\ ,\ \boldsymbol{J}_\alpha^{(n)}(\boldsymbol{\theta})$ are positive definite. The minimum eigenvalue υ_n^0 of $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ satisfies $\inf_n \upsilon_n^0 > 0$.

(C5) For arbitrary i, $y \in \mathcal{Y}_S$, and $\boldsymbol{\theta} \in \Theta_O$, there exists $Q_{jkl}^{(i)}$ $(j, k, l = 1\,,\ \ldots\ ,\ p)$ such that

$$\left|\frac{\partial^3 V_\alpha^{(i)}(y\,;\boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_l}\right| \le Q_{jkl}^{(i)}(y)\,,\quad \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{G_i}\left[Q_{jkl}^{(i)}(Y)\right] = O(1)\,,$$

where $V_\alpha^{(i)}(y\,;\boldsymbol{\theta}) = \int f_i(y\,|\,\boldsymbol{\theta})^{\alpha+1}\,dy - \frac{\alpha+1}{\alpha}\,f_i(y\,|\,\boldsymbol{\theta})^\alpha$.

(C6) For arbitrary i and $\boldsymbol{\theta} \in \Theta_O$, $\boldsymbol{J}_\alpha^{(i)}(\boldsymbol{\theta}) = O(1)$, $\boldsymbol{K}_\alpha^{(i)}(\boldsymbol{\theta}) = O(1)$, and

$$\int \{\alpha\,u_j^{(i)}(y\,;\boldsymbol{\theta})\,u_k^{(i)}(y\,;\boldsymbol{\theta}) - \Xi_{jk}^{(i)}(y\,;\boldsymbol{\theta})\}^2 f_i(y\,|\,\boldsymbol{\theta})^{2\alpha}\,dG_i(y) = O(1)\,,$$

where $u_j^{(i)}$ is the j-th component of the score function $\boldsymbol{u}_i$, and $\Xi_{jk}^{(i)}$ is the $(j,\,k)$-th element of the observed information matrix $\boldsymbol{\Xi}_i$.

(C7) For arbitrary $\epsilon > 0$ and $\boldsymbol{\theta} \in \Theta_O$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_{G_i} \left[\left\| \boldsymbol{\psi}_\alpha^{(i)}(Y\,;\,\boldsymbol{\theta}) \right\|^2 1\left(\left\| \boldsymbol{\psi}_\alpha^{(i)}(Y\,;\,\boldsymbol{\theta}) \right\| > \epsilon\sqrt{n} \right) \right] \;=\; 0\,,$$

where $\boldsymbol{\psi}_\alpha^{(i)}(y\,;\,\boldsymbol{\theta}) = \boldsymbol{K}_\alpha(\boldsymbol{\theta})^{-\frac{1}{2}} \frac{\partial V_\alpha^{(i)}(y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

(C8) The matrix $\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\boldsymbol{K}_\alpha(\boldsymbol{\theta})$ is continuous for arbitrary $\boldsymbol{\theta} \in \Theta_O$.

(C9) For arbitrary $\boldsymbol{\theta} \in \Theta_O$,

$$\begin{aligned}
&\textstyle\sum_{i=1}^{n} \mathbf{P}\left\{ \left\| \frac{\partial V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| > n \right\} \;=\; o(1)\,,\\
&\textstyle\frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_{G_i} \left[\left\| \frac{\partial V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|^2 1\left(\left\| \frac{\partial V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| > n \right) \right] \;=\; o(1)\,,\\
&\textstyle\frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E}_{G_i} \left[\left\| \frac{\partial V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|^2 1\left(\left\| \frac{\partial V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \le n \right) \right] \;=\; o(1)\,,
\end{aligned}$$

and

$$\begin{aligned}
&\textstyle\sum_{i=1}^{n} \mathbf{P}\left\{ \left\| \frac{\partial^2 V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \right\| > n \right\} \;=\; o(1)\,,\\
&\textstyle\frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_{G_i} \left[\left\| \frac{\partial^2 V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \right\|^2 1\left(\left\| \frac{\partial^2 V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \right\| > n \right) \right] \;=\; o(1)\,,\\
&\textstyle\frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E}_{G_i} \left[\left\| \frac{\partial^2 V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \right\|^2 1\left(\left\| \frac{\partial^2 V_\alpha^{(i)}(Y\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T} \right\| \le n \right) \right] \;=\; o(1)\,.
\end{aligned}$$

Condition (C0) is a fundamental principle of many model selection criteria, such as the AIC. Conditions (C1) to (C7) are necessary to determine the consistency and the asymptotic normality of the estimator; these are almost the same as those listed in [Ghosh and Basu (2013)]. Conditions (C8) and (C9) are used to derive and investigate the model selection criterion based on the BHHJ divergence when the data are non-identically distributed. Note that (C6), (C7) and (C9) are fulfilled immediately if the probability (density) functions of the model are independent with the sample size n (for example, the polynomial regression model).

Under these conditions, we can derive a family of model selection criteria that are based on the BHHJ divergence by approximating $H_\alpha^\star(\hat{\boldsymbol{\theta}}_\alpha)$, in a way similar to that used to derive the AIC. Note that, because the data are inhomogeneous, additional conditions are required, compared to the i.i.d. case. An overview of the derivation is shown in the Appendix.

Definition 1 (BHHJ-C). *For $\alpha > 0$, we define the BHHJ-C as*

$$\text{BHHJ-C}_\alpha \;=\; H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha) + \frac{1}{n}\,B_\alpha(\hat{\boldsymbol{\theta}}_\alpha)\,, \tag{7}$$

where $B_\alpha(\hat{\boldsymbol{\theta}}_\alpha) = \sum_{j=1}^{m} \lambda_j(\hat{\boldsymbol{\theta}}_\alpha)$, m is the rank of $\boldsymbol{K}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)\boldsymbol{J}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)^{-1}\boldsymbol{K}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)$, and $\lambda_1(\hat{\boldsymbol{\theta}}_\alpha)\,,\,\ldots\,,\,\lambda_m(\hat{\boldsymbol{\theta}}_\alpha)$ are the nonzero eigenvalues of $\boldsymbol{J}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)^{-1}\boldsymbol{K}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)$.

We note that, in the i.i.d. case (i.e., $G_1 = \cdots = G_n$), (7) is a network information criterion [Murata, *et al.* (1994)] that uses the BHHJ divergence as the discrepancy risk, and it corresponds to the divergence information criterion [Mattheou, *et al.* (2009)] if we use the variance of an appropriate normal distribution for the bias term $B_\alpha(\boldsymbol{\theta})$. We note that the bias term $B_\alpha(\hat{\boldsymbol{\theta}}_\alpha) = \sum_j \lambda_j(\hat{\boldsymbol{\theta}}_\alpha) = \mathrm{tr}\{\boldsymbol{J}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)^{-1}\boldsymbol{K}_\alpha(\hat{\boldsymbol{\theta}}_\alpha)\}$ converges to p (the dimension of the parameter $\boldsymbol{\theta}$) as α goes to zero, because the two matrices $\boldsymbol{J}_\alpha$ and $\boldsymbol{K}_\alpha$ have the same limiting one. On the other hand, since the main term $H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha)$ is the part of the BHHJ divergence that depends on the model, and the log-likelihood function is also a part of the KL divergence, the $\lim_{\alpha\to 0}\text{BHHJ-C}_\alpha$ is not always equal to the AIC. Nevertheless, the parameter selection of the AIC and that of the BHHJ-C converge as α goes to zero. Incidentally, the bias of the AIC does not depend on i since $\boldsymbol{J}_0 = \boldsymbol{K}_0$, so any inhomogeneity is not relevant to the AIC.

3 Asymptotic properties of the BHHJ-C

For the BHHJ-C, we consider the selection probability of each of two parametric models. The larger model has p unknown parameters, and the smaller one is the same as the larger model but is restricted by r equality constraints. We write these constraints as $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}_r$, and let $\boldsymbol{M}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{h}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}$, where $\boldsymbol{M}(\boldsymbol{\theta})$ is a $p \times r$ matrix of rank r. Note that, by Condition (C0), the larger model includes the true distribution.

Let $\boldsymbol{\theta}_\alpha = \arg\min_{\boldsymbol{\theta}} H^\star_\alpha(\boldsymbol{\theta})$ be the best fitting parameter, and let $\hat{\boldsymbol{\theta}}_\alpha = \arg\min_{\boldsymbol{\theta}} H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ be the BHHJ-MDIVE for the larger model; further, let $\boldsymbol{\theta}^{\mathrm{c}}_\alpha = \arg\min_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{0}} H^\star_\alpha(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}^{\mathrm{c}}_\alpha = \arg\min_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{0}} H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta})$ be the respective quantities for the smaller model. Because of the constraints, it is obvious that $H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha) \le H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}^{\mathrm{c}}_\alpha)$.

Under this setting, there are two cases that we can consider. The first situation is that the smaller model maintains "adequacy" (i.e., the ability to include the true distribution) in spite of the constraints, that is, $\boldsymbol{\theta}_\alpha^{\mathrm{c}} = \boldsymbol{\theta}_\alpha$. In this case, the larger model should not be chosen, since it will overfit the data. In the second situation, the smaller model is no longer adequate because of the constraints, and thus we should choose the larger model.

It has been shown that the likelihood ratio follows the chi-square distribution with r degrees of freedom (for example, see [Inagaki (2003)]). In this section, we investigate the asymptotic selection probability of the BHHJ-C by generalizing this proposition.

3.1 Selection probability for BHHJ-C

Now, we will consider the case in which the smaller model is adequate. The BHHJ-Cs for the larger and smaller models are respectively

$$\text{BHHJ-C}_\alpha = H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha) + \tfrac{1}{n}B_\alpha(\hat{\boldsymbol{\theta}}_\alpha),\ \text{BHHJ-C}_\alpha^{\mathrm{c}} = H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) + \tfrac{1}{n}B_\alpha^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})\,.$$

Note that the bias terms are different, since they depend on the model; also, note that $H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha)$, $B_\alpha(\hat{\boldsymbol{\theta}}_\alpha)$, $H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})$, and $B_\alpha^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})$ are $O_P(1)$ with respect to the sample size n.

We define the *overfitting probability* as $\mathbf{P}\left\{\text{BHHJ-C}_\alpha^{\mathrm{c}} - \text{BHHJ-C}_\alpha\ >\ 0\right\}$, that is, the probability of selecting the larger model.

Theorem 1. *As n goes to infinity, the asymptotic distribution of*

$\mathbf{P}\left\{2\,n\,(\text{BHHJ-C}_\alpha^{\mathrm{c}} - \text{BHHJ-C}_\alpha)\ >\ 0\right\}$ *is equivalent to*

$$\sum_{j=1}^{r} \rho_j(\boldsymbol{\theta}_\alpha)\, Z_j^2 - 2\,\left\{B_\alpha(\boldsymbol{\theta}_\alpha) - B_\alpha^{\mathrm{c}}(\boldsymbol{\theta}_\alpha)\right\}\,, \tag{8}$$

where $Z_1\,,\,\ldots\,,\,Z_r$ are independently distributed as $N(0,1)$, $\rho_1(\boldsymbol{\theta}_\alpha)\,,\,\ldots\,,\,\rho_r(\boldsymbol{\theta}_\alpha)$ are nonzero eigenvalues of $\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha)$, and

$$\boldsymbol{S}_\alpha(\boldsymbol{\theta})\ =\ \boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\boldsymbol{M}(\boldsymbol{\theta})\,\left\{\boldsymbol{M}(\boldsymbol{\theta})^T\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\boldsymbol{M}(\boldsymbol{\theta})\right\}^{-1}\boldsymbol{M}(\boldsymbol{\theta})^T\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\,.$$

Proof. We apply the method of Lagrange multipliers:

$$\text{maximize}\ :\ -H_\alpha^{\mathrm{c}}(\boldsymbol{Y}\,;\,\boldsymbol{\theta})\ =\ -H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}) - \boldsymbol{h}(\boldsymbol{\theta})^T\boldsymbol{\kappa},$$

where $\boldsymbol{\kappa} \in \mathbf{R}^r$. The BHHJ-MDIVE of the smaller model fulfills

$$-\frac{\partial H_\alpha^{\mathrm{c}}(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})}{\partial \boldsymbol{\theta}} \;=\; -\frac{\partial H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})}{\partial \boldsymbol{\theta}} - \boldsymbol{M}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})\,\hat{\boldsymbol{\kappa}}_\alpha^{\mathrm{c}} \;=\; \mathbf{0}_p\,, \quad -\frac{\partial H_\alpha^{\mathrm{c}}(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})}{\partial \boldsymbol{\kappa}} \;=\; -\boldsymbol{h}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) \;=\; \mathbf{0}_r\,,$$

where $\hat{\boldsymbol{\kappa}}_\alpha^{\mathrm{c}}$ is the BHHJ-MDIVE of $\boldsymbol{\kappa}$. Remark that $\hat{\boldsymbol{\kappa}}_\alpha^{\mathrm{c}} \xrightarrow{\mathrm{P}} \boldsymbol{\kappa}_\alpha = \mathbf{0}_r$, because $\frac{\partial H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})}{\partial \boldsymbol{\theta}} - \frac{\partial H_\alpha^{\star}(\boldsymbol{\theta}_\alpha)}{\partial \boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \mathbf{0}_p$ (by Condition (C9) and the weak law of large numbers [Chung (2001)]), and $\boldsymbol{M}(\cdot)$ is not the zero matrix. Thus,

$$\begin{aligned}
\mathbf{0}_{p+r} &= \sqrt{n}\begin{pmatrix} -\frac{\partial H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})}{\partial \boldsymbol{\theta}} - \boldsymbol{M}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}})\,\hat{\boldsymbol{\kappa}}_\alpha^{\mathrm{c}} \\ -\boldsymbol{h}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) \end{pmatrix} \\
&= \sqrt{n}\begin{pmatrix} -\frac{\partial H_\alpha(\boldsymbol{Y}\,;\boldsymbol{\theta}_\alpha)}{\partial \boldsymbol{\theta}} \\ \mathbf{0}_r \end{pmatrix} + \sqrt{n}\begin{pmatrix} -\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha) & -\boldsymbol{M}(\boldsymbol{\theta}_\alpha) \\ -\boldsymbol{M}(\boldsymbol{\theta}_\alpha)^T & \mathcal{O} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}} - \boldsymbol{\theta}_\alpha \\ \hat{\boldsymbol{\kappa}}_\alpha - \boldsymbol{\kappa}_\alpha \end{pmatrix} + o_P(1)\,,
\end{aligned}$$

and we obtain that:

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}} - \boldsymbol{\theta}_\alpha \\ \hat{\boldsymbol{\kappa}}_\alpha - \boldsymbol{\kappa}_\alpha \end{pmatrix} = \begin{pmatrix} \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha) & \boldsymbol{M}(\boldsymbol{\theta}_\alpha) \\ \boldsymbol{M}(\boldsymbol{\theta}_\alpha)^T & \mathcal{O} \end{pmatrix}^{-1}\begin{pmatrix} -\sqrt{n}\,\frac{\partial H_\alpha(\boldsymbol{Y}\,;\boldsymbol{\theta}_\alpha)}{\partial \boldsymbol{\theta}} \\ \mathcal{O} \end{pmatrix} + o_P(1)\,.$$

Therefore,

$$\sqrt{n}\,\left(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}} - \boldsymbol{\theta}_\alpha\right) \;=\; -\sqrt{n}\,\boldsymbol{L}_\alpha(\boldsymbol{\theta}_\alpha)\,\frac{\partial H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}_\alpha)}{\partial \boldsymbol{\theta}} + o_P(1)\,, \tag{9}$$

where $\boldsymbol{L}_\alpha(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1} - \boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\boldsymbol{M}(\boldsymbol{\theta})\,\left\{\boldsymbol{M}(\boldsymbol{\theta})^T\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\boldsymbol{M}(\boldsymbol{\theta})\right\}^{-1}\boldsymbol{M}(\boldsymbol{\theta})^T\boldsymbol{J}_\alpha(\boldsymbol{\theta})^{-1}\,.$$

To evaluate the BHHJ-MDIVE for the larger model, we use the Taylor expansion, as follows:

$$\sqrt{n}\,\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right) \;=\; -\sqrt{n}\,\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\,\frac{\partial H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}_\alpha)}{\partial \boldsymbol{\theta}} + o_P(1). \tag{10}$$

Using (9), (10), and the weak law of large numbers, we obtain

$$\begin{aligned}
2\,n\,\left\{H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha) - H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}_\alpha)\right\} &= \boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha)^T\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\,\boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha) + o_P(1)\,, \\
2\,n\,\left\{H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) - H_\alpha(\boldsymbol{Y}\,;\,\boldsymbol{\theta}_\alpha)\right\} &= \boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha)^T\boldsymbol{L}_\alpha(\boldsymbol{\theta}_\alpha)\,\boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha) + o_P(1),
\end{aligned}$$

with $\boldsymbol{z}_\alpha(\boldsymbol{\theta}) = \sqrt{n}\frac{\partial H_\alpha(\boldsymbol{Y}\,;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. From these equations, we obtain

$$2\,n\,\left\{H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) - H_\alpha(\boldsymbol{Y}\,;\,\hat{\boldsymbol{\theta}}_\alpha)\right\} \;=\; \boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha)^T\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)\,\boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha) + o_P(1)\,. \tag{11}$$

Note that $\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)$ on the right-hand side of (11) is a nonnegative definite symmetric matrix, and $\boldsymbol{z}_\alpha(\boldsymbol{\theta}_\alpha)$ is asymptotically distributed with $N_p(\boldsymbol{0}, \boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha))$. Thus, the asymptotic distribution of the quadratic form on the right-hand side of (11) is the same as the distribution of $\sum_{j=1}^{r} \rho_j(\boldsymbol{\theta}_\alpha) Z_j^2$; see [Dik and Gunst (1985)]. From Condition (C8), we also obtain $B_\alpha(\hat{\boldsymbol{\theta}}_\alpha) \xrightarrow{\mathrm{P}} B_\alpha(\boldsymbol{\theta}_\alpha)$ and $B_\alpha^{\mathrm{c}}(\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) \xrightarrow{\mathrm{P}} B_\alpha^{\mathrm{c}}(\boldsymbol{\theta}_\alpha)$ because of the continuity of the bias term. Therefore, by using Slutsky's theorem, we obtain the required result. □

Corollary 1 (The underfitting probability of the BHHJ-C)**.** *If the smaller model cannot express the true distribution, the difference* $H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha^{\mathrm{c}}) - H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha)$ *is positive and* $O_P(1)$ *almost surely. Therefore, the probability of underfitting* $\mathbf{P}\left\{\text{BHHJ-C}_\alpha^{\mathrm{c}} - \text{BHHJ-C}_\alpha < 0\right\}$ *tends to zero as n goes to infinity; that is, if there is at least one adequate model, in the limit as n goes to infinity, the BHHJ-C does not choose an underfitting model.*

3.2 An example of the asymptotic selection probability of the BHHJ-C

If the nonzero eigenvalues of $\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha)$ are all the same, the first term in (8) ($\sum_{j=1}^{r} \rho_j(\boldsymbol{\theta}_\alpha) Z_j^2$) can be represented by the chi-square distribution with r degrees of freedom. In particular, in the polynomial regression model and under an appropriate setting, the asymptotic overfitting probability is constant, regardless of the value of α. We prove this proposition below.

Corollary 2 (The polynomial regression model)**.** *Let* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *be fixed explanatory variables, and let the response variables* Y_i $(i = 1, \ldots, n)$ *be normally distributed. We let the larger model be* $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \epsilon_i$*, where the* ϵ_i*'s are independently distributed as* $N(0, s)$*. Additionally, some* r $(1 \le r \le p)$ *constraints are given as* $\beta_{p+1-r} = \cdots = \beta_p = 0$*. Then, it holds that*

$$\rho_1(\boldsymbol{\theta}_\alpha) \;=\; \cdots \;=\; \rho_r(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)\,\tfrac{\zeta_{2\alpha}}{\zeta_\alpha}\,,$$

where

$$\begin{aligned}
\zeta_\alpha \;&=\; (2\,\pi)^{-\frac{\alpha}{2}}\,(\alpha+1)^{-\frac{3}{2}}\,s_\alpha^{-\frac{\alpha}{2}-1}\,,\\
\tau_\alpha \;&=\; \tfrac{1}{4}\,(2\,\pi)^{-\frac{\alpha}{2}}\,(\alpha+1)^{-\frac{5}{2}}\,(\alpha^2+2)\,s_\alpha^{-\frac{\alpha}{2}-2}\,,
\end{aligned}$$

in which s_α is the best fitting value of the scale parameter s (the variance of the error terms). In this case, the overfitting probability is equivalent to $\mathbf{P}\{\chi^2_r > 2r\}$ for any $\alpha \geq 0$.

The proof of Corollary 2 is shown in the Appendix. Note that the asymptotic overfitting probability is not always independent of α. For instance, in the previous example, if we add the $(r+1)$-th constraint to the variance parameter as $s = s^*$ with some positive value s^*, the nonzero eigenvalues are not uniform. The asymptotic behavior of BHHJ-C is generally related to α.

4 Sensitivity of the divergence-based model selection criteria

The estimation based on the BHHJ divergence is robust when $\alpha > 0$. When assessing robustness, an influence function is often used to assess the response to a perturbation in the population distribution. The estimator can be regarded as a functional of the probability distribution of the observed data. Since the model selection criterion depends on data and estimators, it is also a functional of the distribution function. In this section, we consider the sensitivity of the model selection criterion.

An *outlier* is a data point that does not fall within the expected range of the true population distribution. We will denote as δ_{z_i} the distribution that takes the value z_i with probability one. Let $\Omega^\nu_{z_i}$ be the mixture distribution $(1-\nu)G_i + \nu\delta_{z_i}$ $(i = 1, \ldots, n)$ for some small $\nu > 0$, and let $\boldsymbol{\Omega}^\nu_{\boldsymbol{z}} = (\Omega^\nu_{z_1}, \ldots, \Omega^\nu_{z_n})^T$. The influence function (IF) of $\boldsymbol{T}(\cdot)$, the functional form of the estimator, is defined by $\boldsymbol{T}^{(1)}_\alpha(\boldsymbol{z}; \boldsymbol{T}_\alpha(\boldsymbol{G})) = \lim_{\nu\to 0} \frac{\boldsymbol{T}_\alpha(\boldsymbol{\Omega}^\nu_{\boldsymbol{z}}) - \boldsymbol{T}_\alpha(\boldsymbol{G})}{\nu}$, where $\boldsymbol{z} = (z_1, \ldots, z_n)^T$. For the IF, the gross error sensitivity (GES) is given as $\sup_{\boldsymbol{z}} \boldsymbol{T}^{(1)}_\alpha(\boldsymbol{z}; \boldsymbol{T}_\alpha(\boldsymbol{G}))$ ([Huber (1983)], [Hampel, *et al.* (1986)]).

Let $\boldsymbol{T}_\alpha(\hat{\boldsymbol{G}})$ be a functional form of the BHHJ-MDIVE family (which includes the MLE) $\hat{\boldsymbol{\theta}}_\alpha$ for $\alpha \geq 0$. The best fitting parameter $\boldsymbol{\theta}_\alpha$ can be expressed as $\boldsymbol{T}_\alpha(\boldsymbol{G})$. Since the bias term in BHHJ-C, $\frac{1}{n}B_\alpha$, is a function of the parameter and it is continuous, due to (C8), the behavior of the bias is determined by $\boldsymbol{T}_\alpha$. The main term H_α depends on $\boldsymbol{T}_\alpha$ and the data $\boldsymbol{Y}$,

so the influence of an outlier is not always finite, even if the estimator is robust. Therefore, we will primarily consider the main term in the BHHJ-C. We will use the functional form of the main term:

$$\mathcal{H}_0(\boldsymbol{G}) \,=\, -\frac{1}{n}\sum_{i=1}^{n}\int \log f_i(y\,|\,\boldsymbol{T}_0(\boldsymbol{G}))\,dG_i(y)\,,$$

and for $\alpha > 0$,

$$\mathcal{H}_\alpha(\boldsymbol{G}) \,=\, \frac{1}{n}\sum_{i=1}^{n}\left\{\int f_i(y\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha+1}\,dy - \frac{\alpha+1}{\alpha}\int f_i(y\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha}\,dG_i(y)\right\}$$

for the distribution $\boldsymbol{G} = (G_1\,,\,\ldots\,,\,G_n)^T$.

The idea of the IF is diverted to many studies (e.g. for test statistics in [Ghosh, *et al.* (2015)]). Now, we attempt to measure the sensitivity of a model selection criterion by applying this idea.

Definition 2. *We define*

$$\mathcal{H}_\alpha^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G})) \,=\, \lim_{\nu\to 0}\frac{\mathcal{H}_\alpha(\boldsymbol{\Omega}_{\boldsymbol{z}}^{\nu}) - \mathcal{H}_\alpha(\boldsymbol{G})}{\nu} \,=\, \left[\frac{\partial\mathcal{H}_\alpha(\boldsymbol{\Omega}_{\boldsymbol{z}}^{\nu})}{\partial\nu}\right]_{\nu=0} \tag{12}$$

to measure the sensitivity of the BHHJ-C.

Equation (12) can be regarded as the change in the BHHJ-C due to outliers in the data. When estimating parameters, it is undesirable that the existence of an outlier should have a large effect on the value of the estimates. Similarly, the sensitivity of a model selection criterion (12) should be finite with respect to outliers in the data.

When $\alpha = 0$ (i.e., the criterion is based on the KL divergence), the sensitivity is

$$\begin{aligned}\mathcal{H}_0^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_0(\boldsymbol{G})) \;&=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\int \boldsymbol{u}_i(y\,;\,\boldsymbol{T}_0(\boldsymbol{G}))^T\,dG_i(y)\,\boldsymbol{T}_0^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_0(\boldsymbol{G}))\right.\\ &\qquad\left.+\int \log f_i(y\,|\,\boldsymbol{T}_0(\boldsymbol{G}))\,dG_i(y) - \log f_i(z_i\,|\,\boldsymbol{T}_0(\boldsymbol{G}))\right\}.\end{aligned} \tag{13}$$

In contrast, when $\alpha > 0$, we obtain

$$\begin{aligned}\mathcal{H}_\alpha^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G})) \;&=\; \frac{\alpha+1}{n}\sum_{i=1}^{n}\left[\int f_i(y\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha+1}\boldsymbol{u}_i(y\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^T\,dy\,\boldsymbol{T}_\alpha^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G}))\right.\\ &\qquad-\int f_i(y\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha}\boldsymbol{u}_i(y\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^T\,dG_i(y)\,\boldsymbol{T}_\alpha^{(1)}(\boldsymbol{z}\,;\,\boldsymbol{T}_\alpha(\boldsymbol{G}))\\ &\qquad\left.+\frac{1}{\alpha}\left\{\int f_i(y\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha}\,dG_i(y) - f_i(z_i\,|\,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha}\right\}\right].\end{aligned} \tag{14}$$

Whether the suprema of (13) and (14) are finite is determined by several factors, such as the probability density functions g_i and f_i, the adequacy of the model, and the GES of the estimator.

Now, we will consider the polynomial regression model. The portion of $\boldsymbol{T}_\alpha^{(1)}$ (the IF of the BHHJ-MDIVE) that is related to the contaminated data $\boldsymbol{z}$ is $\left(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 \exp\{-\alpha(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2/(2s)\}$, so the GES is finite if and only if $\alpha > 0$. When $\alpha = 0$, since an extreme outlier has a probability density that is exceedingly close to 0, $-\log f_i(z_i \,|\, \boldsymbol{T}_0(\boldsymbol{G}))$ can become infinite, even if the model is adequate (i.e., $\boldsymbol{G} = \boldsymbol{F}_{\boldsymbol{T}_\alpha(\boldsymbol{G})}$). Whereas, the IF is finite, and the value $f_i(z_i \,|\, \boldsymbol{T}_\alpha(\boldsymbol{G}))^\alpha$ has an upper bound for any $\alpha > 0$. Therefore, (13) is infinite, and (14) is finite even if $\boldsymbol{z}$ is an extreme outlier. This leads to the following theorem.

Theorem 2. *In the polynomial regression model, the sensitivity (12) is infinite if $\alpha = 0$, and it is finite if $\alpha > 0$.*

5 Simulation results

In this section, we present some numerical simulations with a polynomial regression model in order to examine the results described in Sections 3 and 4. We also investigate the accuracy of the model selection criteria in the presence of outliers.

Our simulations were designed as follows. The set of the models of the response variables $\boldsymbol{Y} = (Y_1\,,\,\ldots\,,\,Y_n)^T$ were specified as

$$\begin{aligned} &Y_i \,=\, \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i\,, \quad \boldsymbol{x}_i \,=\, (1\,,\,x_i\,,\,\ldots\,,\,x_i^p)^T\,, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0,\,s)\,, \\ &\boldsymbol{\beta} = (\beta_0\,,\,\beta_1\,,\,\ldots\,,\,\beta_p)^T\,, \quad s \in (0,\,+\infty)\,, \quad i = 1, \ldots, n\,, \end{aligned}$$

for the nonrandom explanatory variables $\{\boldsymbol{x}_i\}$. In this simulation, we assigned the values $x_1\,,\,\ldots\,,\,x_n$ such that they equally divided the interval $[-2,\,2]$.

We give the "true model" that generates the observed data as follows:

$$Y_i \,=\, \eta_i(\boldsymbol{x}_i) + \epsilon_i \,=\, 0.5\,x_i - 1.5\,x_i^2 + 0.5\,x_i^4 + \epsilon_i\,,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independently distributed as $N(0, 1)$. We suppose the candidates $p = 0, 1, \ldots, 5$ to observe the tendency of the selection (underfitting, true, or overfitting model) of the criteria. Since this is included in the candidate models, the criteria will preferentially select the fourth-order model ($p = 4$).

We used the following values for the tuning parameter: $\alpha = 0.1, 0.25, 0.5, 0.75, 1, 1.25,$ and 1.5.

5.1 Selection probability of the BHHJ-C

For the polynomial regression models, the BHHJ-C selection probabilities are asymptotically uniform and independent of $\alpha \geq 0$, by Corollary 2; here, we confirm that. The sample size n was 5000. Table 1 shows the relative frequencies of selecting the true (4th-order) model, an underfitting (3rd-order) model, and an overfitting (5th-, 6th-, and 14th-order) model. The results for the BHHJ-C are close to the theoretical percentages given in Corollaries 1 and 2.

Table 1: The selection frequency of the BHHJ-C

	4	3	**4**	5	**4**	6	**4**	14
(theoretical)	100	0	84.3	15.7	86.5	13.5	97.1	2.9
AIC	100	0	84.5	15.5	86.0	14.0	97.0	3.0
BHHJ-$C_{0.10}$	100	0	84.4	15.6	86.3	13.7	97.1	2.9
BHHJ-$C_{0.25}$	100	0	84.6	15.4	86.3	13.7	96.9	3.1
BHHJ-$C_{0.50}$	100	0	84.2	15.8	86.4	13.6	96.9	3.1
BHHJ-$C_{0.75}$	100	0	84.3	15.7	86.4	13.6	96.9	3.1
BHHJ-$C_{1.00}$	100	0	84.1	15.9	86.2	13.8	97.1	2.9
BHHJ-$C_{1.25}$	100	0	84.1	15.9	86.3	13.7	96.9	3.1
BHHJ-$C_{1.50}$	100	0	84.1	15.9	86.0	14.0	97.1	2.9

5.2 Robustness of the model selection criteria

We used the same candidate models as in the simulations presented in the previous subsection. Here, we assume that some outliers are contained in the observed data. Let $\epsilon_1, \ldots, \epsilon_n$

be independently distributed error terms:

$$\epsilon_i \sim \begin{cases} N(0,\,1), & (\text{w.p. } 1-\nu), \\ U\left(\min_i\{\eta_i(\boldsymbol{x}_i)\}-10\,,\ \max_i\{\eta_i(\boldsymbol{x}_i)\}+10\right), & (\text{w.p. } \nu), \end{cases}$$

where "w.p." means with probability. We examined the selection probability of the BHHJ-C for $n = 100$ samples. For comparison, we also use the criteria that have the log-likelihood function (a part of KL divergence) as the main term: the BIC, the GIC, and the AIC. The GIC's estimator is equivalent to the BHHJ-MDIVE, so it can be regarded as another generalization of the AIC using the BHHJ-MDIVE.

We show the results of $10,000$ simulations for each of four patterns with the following contamination rates: $\nu = 0\%, 5\%, 10\%$, and 20% in Tables 2, 3, 4, and 5, respectively.

Table 2: contamination rate: 0%

	0–3	**4**	5
BIC	0.1	97.0	2.9
TIC	0.0	94.1	5.9
$\mathrm{GIC}_{0.10}$	0.0	93.9	6.1
$\mathrm{GIC}_{0.25}$	0.0	93.8	6.2
$\mathrm{GIC}_{0.50}$	0.0	94.0	6.0
$\mathrm{GIC}_{0.75}$	0.2	93.6	6.1
$\mathrm{GIC}_{1.00}$	0.8	92.8	6.5
$\mathrm{GIC}_{1.25}$	1.6	91.0	7.3
$\mathrm{GIC}_{1.50}$	2.9	89.3	7.7
AIC	0.0	85.2	14.8
BHHJ-$\mathrm{C}_{0.10}$	0.0	85.4	14.6
BHHJ-$\mathrm{C}_{0.25}$	0.0	85.5	14.5
BHHJ-$\mathrm{C}_{0.50}$	0.0	85.8	14.2
BHHJ-$\mathrm{C}_{0.75}$	0.0	86.7	13.3
BHHJ-$\mathrm{C}_{1.00}$	0.1	88.3	11.6
BHHJ-$\mathrm{C}_{1.25}$	0.2	89.9	9.8
BHHJ-$\mathrm{C}_{1.50}$	0.8	91.5	7.6

Table 3: contamination rate: 5%

	0–3	**4**	5
BIC	49.6	49.3	1.2
TIC	67.9	31.0	1.1
$\mathrm{GIC}_{0.10}$	85.0	13.8	1.3
$\mathrm{GIC}_{0.25}$	88.2	10.8	1.1
$\mathrm{GIC}_{0.50}$	88.2	10.8	1.1
$\mathrm{GIC}_{0.75}$	87.8	11.1	1.1
$\mathrm{GIC}_{1.00}$	87.1	11.7	1.3
$\mathrm{GIC}_{1.25}$	87.3	11.4	1.4
$\mathrm{GIC}_{1.50}$	87.5	11.1	1.4
AIC	11.5	75.3	13.3
BHHJ-$\mathrm{C}_{0.10}$	0.1	86.9	13.0
BHHJ-$\mathrm{C}_{0.25}$	0.0	85.0	15.0
BHHJ-$\mathrm{C}_{0.50}$	0.0	85.4	14.5
BHHJ-$\mathrm{C}_{0.75}$	0.0	86.6	13.3
BHHJ-$\mathrm{C}_{1.00}$	0.3	88.1	11.5
BHHJ-$\mathrm{C}_{1.25}$	0.7	89.7	9.5
BHHJ-$\mathrm{C}_{1.50}$	1.8	90.8	7.5

Table 4: contamination rate: 10%

	0–3	**4**	5
BIC	80.4	19.1	0.6
TIC	88.1	11.6	0.3
$GIC_{0.10}$	96.3	3.2	0.4
$GIC_{0.25}$	98.8	1.0	0.1
$GIC_{0.50}$	99.0	0.9	0.1
$GIC_{0.75}$	98.7	1.1	0.2
$GIC_{1.00}$	98.4	1.4	0.2
$GIC_{1.25}$	98.3	1.5	0.2
$GIC_{1.50}$	98.2	1.6	0.3
AIC	33.2	55.9	10.9
BHHJ-$C_{0.10}$	5.2	85.1	9.7
BHHJ-$C_{0.25}$	0.1	86.6	13.3
BHHJ-$C_{0.50}$	0.0	86.4	13.5
BHHJ-$C_{0.75}$	0.2	87.2	12.6
BHHJ-$C_{1.00}$	0.5	88.6	10.9
BHHJ-$C_{1.25}$	1.1	89.8	9.0
BHHJ-$C_{1.50}$	2.3	90.8	6.8

Table 5: contamination rate: 20%

	0–3	**4**	5
BIC	97.3	2.7	0.1
TIC	95.1	4.6	0.2
$GIC_{0.10}$	97.0	2.6	0.3
$GIC_{0.25}$	99.6	0.1	0.3
$GIC_{0.50}$	99.9	0.1	0.1
$GIC_{0.75}$	99.9	0.1	0.1
$GIC_{1.00}$	99.9	0.1	0.0
$GIC_{1.25}$	99.8	0.2	0.1
$GIC_{1.50}$	99.9	0.1	0.1
AIC	67.5	26.9	5.7
BHHJ-$C_{0.10}$	46.5	47.7	5.8
BHHJ-$C_{0.25}$	7.0	84.0	9.0
BHHJ-$C_{0.50}$	2.0	88.0	10.1
BHHJ-$C_{0.75}$	1.9	88.8	9.4
BHHJ-$C_{1.00}$	2.7	89.6	7.9
BHHJ-$C_{1.25}$	3.9	89.9	6.2
BHHJ-$C_{1.50}$	6.2	89.0	4.7

In the noncontaminated case (Table 2), the BIC had the best performance, and the other criteria also achieved high accuracies. However, when the data were contaminated with outliers, the accuracies of the BIC and GIC family (including the TIC) were reduced by more than half, even when the contamination rate was only 5% (Table 3). When the contamination rate was 20% (Table 5), the BIC and GIC family had performance rates of nearly zero; that is, the corresponding criteria are unable to select the true model when the data contain outliers. In contrast, however, the BHHJ-C for $\alpha \geq 0.25$ maintained a high rate of performance. When α was very small or zero, the performance deteriorated, because the BHHJ-C became similar to the AIC.

The suitable choice of the α is a difficult issue. In many previous studies, $\alpha \leq 1$ is assumed (for example, [Basu, *et al.* (1998)] and [Ghosh and Basu (2015)]). However, the

BHHJ-C_α with $\alpha > 1$ achieved the best accuracy in the contaminated cases. Taking the instability of the estimation of the large α into account, we consider that [1, 1.5] is also a preferable range of α from the viewpoint of the robust model selection in the polynomial regression. Although the "optimal α" depends on the setting and purpose, this simulation result suggests that we should consider that it can be more than 1.

Since the outliers were not normally distributed, the contaminated cases did not satisfy Condition (C0), but in spite of this, the BHHJ-C with $\alpha \geq 0.25$ achieved almost the same accuracy rate as in the noncontaminated case. One reason for this is that the supremum of $\mathcal{H}_\alpha^{(1)}$ is finite, as shown in Theorem 2. The boxplots of the values obtained by the AIC and BHHJ-$C_{1.50}$ for the 4th-order models with contamination rates of 0% and 5% (corresponding to Tables 2 and 3, respectively) are shown in Figures 1 and 2, respectively.

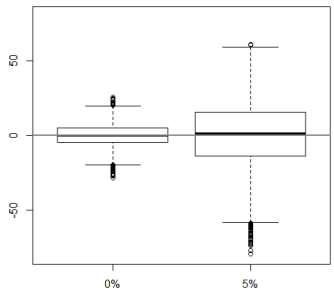

Figure 1: AIC

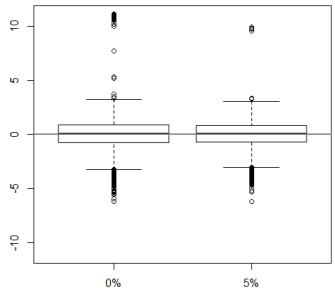

Figure 2: BHHJ-C ($\alpha = 1.5$)

In the two figures, the data are centered. As it can be seen in Figure 1, the values of the AIC are destabilized by the presence of outliers. The standard deviation of the AIC in the noncontaminated case is 7.33, and in the 5%-contaminated case, it increases to 21.79. However, with the BHHJ-$C_{1.50}$ the standard deviations of the two cases are very similar: 1.41 and 1.21. Therefore, we have shown that the distribution of BHHJ-$C_{1.50}$ is nearly independent of the presence of outliers. This shows that the BHHJ-C reduces the effect of outliers without reducing the accuracy, and this is a distinct advantage of using the BHHJ-C.

6 Conclusion

In this paper, we derived a family of the robust model selection criteria, the BHHJ-C. When used with the polynomial regression model, although when α was large, the estimator based on the BHHJ divergence achieved robustness at the cost of efficiency, as $\alpha \to 0$, the asymptotic selection probability of the BHHJ-C was equivalent to that of the AIC. Moreover, for large α, the BHHJ-C tends to maintain its level of performance by controlling the weight assigned to the observed data points, and this reduces the effect of outliers.

The results presented in Section 5 show that in the presence of outliers, the model selection criteria based on the log-likelihood, i.e. KL divergence (the AIC, the GIC family, and the BIC) had poorer performance, but the BHHJ-C with large α was still able to select the correct model. This shows the important advantage of our proposed criterion.

A Derivation of the BHHJ-C (7)

We rewrite (4) as $H^{\star}_{\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} U^{(i)}_{\alpha}(\boldsymbol{\theta})$. The Taylor expansion of $U^{(i)}_{\alpha}(\hat{\boldsymbol{\theta}}_{\alpha})$ around the best fitting parameter $\boldsymbol{\theta}_{\alpha}$ is

$$\begin{aligned} U^{(i)}_{\alpha}(\hat{\boldsymbol{\theta}}_{\alpha}) &= U^{(i)}_{\alpha}(\hat{\boldsymbol{\theta}}_{\alpha}) + \frac{\partial U^{(i)}_{\alpha}(\boldsymbol{\theta}_{\alpha})}{\partial \boldsymbol{\theta}^{T}} \left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right) \\ &\quad + \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right)^{T} \frac{\partial^{2} U^{(i)}_{\alpha}(\boldsymbol{\theta}_{\alpha})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{T}} \left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right) + o_{P}\left(\frac{1}{n}\right), \end{aligned} \tag{15}$$

for arbitrary $i = 1, \ldots, n$. Since the best fitting parameter minimizes $H_{\alpha}(\boldsymbol{Y}\,;\boldsymbol{\theta})$ and $\frac{\partial^{2} H^{\star}_{\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{T}} = \boldsymbol{J}_{\alpha}(\boldsymbol{\theta})$, the expectation is as follows:

$$\mathbf{E}\left[H^{\star}_{\alpha}(\hat{\boldsymbol{\theta}}_{\alpha})\right] = H^{\star}_{\alpha}(\boldsymbol{\theta}_{\alpha}) + \frac{1}{2}\,\mathbf{E}\left[\left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right)^{T} \boldsymbol{J}_{\alpha}(\boldsymbol{\theta}_{\alpha}) \left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right)\right] + o\left(\frac{1}{n}\right), \tag{16}$$

this is obtained by summing (15) over i. In a similar way, we have

$$\begin{aligned} & H^{\star}_{\alpha}(\boldsymbol{\theta}_{\alpha}) \\ &= \mathbf{E}\left[H_{\alpha}(\boldsymbol{Y}\,;\boldsymbol{\theta}_{\alpha})\right] \\ &= \mathbf{E}\left[H_{\alpha}(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_{\alpha})\right] + \frac{1}{2}\,\mathbf{E}\left[\left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right)^{T} \frac{\partial^{2} H_{\alpha}(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_{\alpha})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{T}} \left(\hat{\boldsymbol{\theta}}_{\alpha} - \boldsymbol{\theta}_{\alpha}\right)\right] + o\left(\frac{1}{n}\right). \end{aligned}$$

The weak law of large numbers when the random variables are not identically distributed (see [Chung (2001)]) and Condition (C9) imply $\frac{\partial^2 H_\alpha(\boldsymbol{Y}\,;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T} - \boldsymbol{J}_\alpha(\boldsymbol{\theta}) \xrightarrow{\mathrm{P}} \boldsymbol{O}$ for arbitrary $\boldsymbol{\theta} \in \Theta_O$ as $n \to +\infty$, so we have

$$H^\star_\alpha(\boldsymbol{\theta}_\alpha) \,=\, \mathbf{E}\left[H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha)\right] + \frac{1}{2}\,\mathbf{E}\left[\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)^T \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)\,\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)\right] + o\left(\frac{1}{n}\right)\,. \tag{17}$$

By (16) and (17), we have

$$\mathbf{E}\left[H^\star_\alpha(\hat{\boldsymbol{\theta}}_\alpha)\right] \,=\, \mathbf{E}\left[H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha)\right] + \mathbf{E}\left[\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)^T \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)\,\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)\right] + o\left(\frac{1}{n}\right)\,.$$

Therefore, we obtain a valid approximation of $H^\star_\alpha(\hat{\boldsymbol{\theta}}_\alpha)$ as follows:

$$\mathbf{E}\left[H_\alpha(\boldsymbol{Y}\,;\hat{\boldsymbol{\theta}}_\alpha)\right] + \mathbf{E}\left[\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)^T \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)\,\left(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha\right)\right]\,. \tag{18}$$

Using the asymptotic normality, the second term of (18) is asymptotically equivalent to $\frac{1}{n}\sum_{j=1}\lambda_j(\boldsymbol{\theta}_\alpha)$. Thus we can approximate it by $\frac{1}{n}B_\alpha(\hat{\boldsymbol{\theta}}_\alpha)$, as described in Definition 1.

B Proof of Corollary 2

Under the statement in Corollary 2, the constraint vector $\boldsymbol{h}(\boldsymbol{\theta})$ and the matrix $\boldsymbol{M}(\boldsymbol{\theta})$ can be written as

$$\boldsymbol{h}(\boldsymbol{\theta}) \,=\, \begin{pmatrix} \beta_{p+1-r} \\ \vdots \\ \beta_p \end{pmatrix}\,, \quad \boldsymbol{M}(\boldsymbol{\theta}) \,=\, \frac{\partial \boldsymbol{h}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \,=\, \begin{pmatrix} \boldsymbol{O}_{(p+1-r)\times r} \\ \boldsymbol{I}_r \\ \mathbf{0}_r^T \end{pmatrix}\,,$$

where $\boldsymbol{\theta} = (\beta_0\,,\,\beta_1\,,\,\ldots\,,\,\beta_p\,,\,s)^T$ (note that $\boldsymbol{\theta}$ is $p+2$ dimensional).

Here, we partition the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1\,,\,\ldots\,,\,\boldsymbol{x}_n)^T \in \mathbf{R}^{n\times(p+1)}$ as

$$\boldsymbol{X} \,=\, (\boldsymbol{X}_\bullet\,\boldsymbol{X}_\circ)\,, \quad \boldsymbol{X}_\bullet \in \mathbf{R}^{n\times(p+1-r)}\,, \quad \boldsymbol{X}_\circ \in \mathbf{R}^{n\times r}\,,$$

where $\boldsymbol{X}_\bullet$ is the part that is the same in the two models, and $\boldsymbol{X}_\circ$ is the part that is unique

to the larger model. The matrices $\boldsymbol{J}_\alpha$ and $\boldsymbol{K}_\alpha$ defined by (6) can be calculated as follows:

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)\begin{pmatrix} \zeta_\alpha \frac{\boldsymbol{X}_\bullet^T \boldsymbol{X}_\bullet}{n} & \zeta_\alpha \frac{\boldsymbol{X}_\bullet^T \boldsymbol{X}_\circ}{n} & \mathcal{O} \\ \zeta_\alpha \frac{\boldsymbol{X}_\circ^T \boldsymbol{X}_\bullet}{n} & \zeta_\alpha \frac{\boldsymbol{X}_\circ^T \boldsymbol{X}_\circ}{n} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} & \tau_\alpha \end{pmatrix},$$

$$\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)^2\begin{pmatrix} \zeta_{2\alpha} \frac{\boldsymbol{X}_\bullet^T \boldsymbol{X}_\bullet}{n} & \zeta_{2\alpha} \frac{\boldsymbol{X}_\bullet^T \boldsymbol{X}_\circ}{n} & \mathcal{O} \\ \zeta_{2\alpha} \frac{\boldsymbol{X}_\circ^T \boldsymbol{X}_\bullet}{n} & \zeta_{2\alpha} \frac{\boldsymbol{X}_\circ^T \boldsymbol{X}_\circ}{n} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} & \tau_{2\alpha} - \frac{\alpha^2}{4}\zeta_\alpha^2 \end{pmatrix}.$$

The inverse of $\boldsymbol{J}_\alpha$ is

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1} \;=\; (\alpha+1)^{-1}\begin{pmatrix} \frac{n}{\zeta_\alpha}\boldsymbol{j}^{11} & \frac{n}{\zeta_\alpha}\boldsymbol{j}^{12} & \mathcal{O} \\ \frac{n}{\zeta_\alpha}\boldsymbol{j}^{21} & \frac{n}{\zeta_\alpha}\boldsymbol{j}^{22} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} & \frac{1}{\tau_\alpha} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{j}^{11} &= (\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1} - (\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T\boldsymbol{X}_\circ\boldsymbol{\Upsilon}^{-1}\boldsymbol{X}_\circ\boldsymbol{X}_\bullet(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}, \\ \boldsymbol{j}^{12} &= -(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T\boldsymbol{X}_\circ\,\boldsymbol{\Upsilon}^{-1}, \quad \boldsymbol{j}^{21} = \left(\boldsymbol{j}^{12}\right)^T, \quad \boldsymbol{j}^{22} = \boldsymbol{\Upsilon}^{-1}, \\ \boldsymbol{\Upsilon} &= \boldsymbol{X}_\circ^T\boldsymbol{X}_\circ - \boldsymbol{X}_\circ^T\boldsymbol{X}_\bullet(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T\boldsymbol{X}_\circ = \boldsymbol{X}_\circ^T(\boldsymbol{I}_n - \boldsymbol{\mathcal{P}}_\bullet)\,\boldsymbol{X}_\circ. \end{aligned}$$

Note that $\boldsymbol{\mathcal{P}}_\bullet$ is the proposition matrix $\boldsymbol{X}_\bullet(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T$.

We have the following relationships:

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)^{-1}\frac{n}{\zeta_\alpha}\begin{pmatrix} -(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T\boldsymbol{X}_\circ\,\boldsymbol{\Upsilon}^{-1} \\ \boldsymbol{\Upsilon}^{-1} \\ \mathcal{O} \end{pmatrix},$$

$$\left\{\boldsymbol{M}(\boldsymbol{\theta}_\alpha)^T\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_\alpha)\right\}^{-1} \;=\; (\alpha+1)\frac{\zeta_\alpha}{n}\boldsymbol{\Upsilon},$$

$$\boldsymbol{M}(\boldsymbol{\theta}_\alpha)^T\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)\frac{\zeta_{2\alpha}}{\zeta_\alpha}(\mathcal{O}\;\boldsymbol{I}_r\;\mathcal{O}),$$

and from them, we obtain

$$\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)\,\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha) \;=\; (\alpha+1)\frac{\zeta_{2\alpha}}{\zeta_\alpha}\begin{pmatrix} \mathcal{O} & -(\boldsymbol{X}_\bullet^T\boldsymbol{X}_\bullet)^{-1}\boldsymbol{X}_\bullet^T\boldsymbol{X}_\circ & \mathcal{O} \\ \mathcal{O} & \boldsymbol{I}_r & \mathcal{O} \\ \mathcal{O} & \mathcal{O} & \mathcal{O} \end{pmatrix}.$$

It is obvious that the nonzero eigenvalues of $\boldsymbol{S}_\alpha(\boldsymbol{\theta}_\alpha)\boldsymbol{K}_\alpha(\boldsymbol{\theta}_\alpha)$ are $(\alpha+1)\frac{\zeta_{2\alpha}}{\zeta_\alpha}$, and the number of eigenvalues is r.

The convergence values of the bias terms B_α and B_α^{c} are

$$(\alpha+1)\left\{\frac{\zeta_{2\alpha}}{\zeta_\alpha}(p+1)+\frac{\tau_{2\alpha}}{\tau_\alpha}-\frac{\alpha^2}{4}\frac{\zeta_\alpha^2}{\tau_\alpha}\right\},\ (\alpha+1)\left\{\frac{\zeta_{2\alpha}}{\zeta_\alpha}(p+1-r)+\frac{\tau_{2\alpha}}{\tau_\alpha}-\frac{\alpha^2}{4}\frac{\zeta_\alpha^2}{\tau_\alpha}\right\},$$

respectively, and their difference is $(\alpha+1)\frac{\zeta_{2\alpha}}{\zeta_\alpha}r$. Therefore, the asymptotic probability of overfitting is

$$\mathbf{P}\left\{(\alpha+1)\,\frac{\zeta_{2\alpha}}{\zeta_\alpha}\,\chi_r^2-2\,(\alpha+1)\,\frac{\zeta_{2\alpha}}{\zeta_\alpha}\,r\,>\,0\right\}\,=\,\mathbf{P}\left\{\chi_r^2\,>\,2\,r\right\}\,.$$

We have obtained the required result.

Acknowledgement

The authors would like to express their gratitude to the reviewer and the editor in chief for their valuable comments, which have considerably improved the earlier version of the article. This work was partly supported by JSPS KAKENHI Grant Number 16J04579.

References

[Akaike (1974)] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19** (6), 716–723.

[Basu, *et al.* (1998)] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85** (3), 549–559.

[Basu, *et al.* (2011)] Basu, A., Shioya, H., Park, C. (2011). *Statistical inference: the minimum distance approach*. CRC Press.

[Chung (2001)] Chung, K. L. (2001). *A course in probability theory*. Academic Press.

[Dik and Gunst (1985)] Dik, J. J. and de Gunst, M. C. M. (1985). The distribution of general quadratic forms in normal variables. *Statistica Neerlandica*, **39** (1), 14–26.

[Ghosh and Basu (2013)] Ghosh, A. and Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics*, **7**, 2420–2456.

[Ghosh and Basu (2015)] Ghosh, A., Basu, A. (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**, 2056–2072.

[Ghosh, *et al.* (2015)] Ghosh, A., Basu, A., Pardo, L. (2015). On the robustness of a divergence based test of simple statistical hypotheses, *Journal of Statistical Planning and Inference*, **161**, 91–108.

[Hampel, *et al.* (1986)] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. Wiley.

[Huber (1983)] Huber, P. J. (1983). Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, **78** (381), 66–72.

[Hurvich and Tsai (1989)] Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76** (2), 297–307.

[Inagaki (2003)] Inagaki, N. (2003). *Statistical mathematics (revised edition)*. Shokabo (in Japanese).

[Konishi and Kitagawa (1996)] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83** (4), 875–890.

[Kullback and Leibler (1951)] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22** (1), 79–86.

[Mattheou, *et al.* (2009)] Mattheou, K., Lee, S., Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, **139** (2), 228–235.

[Murata, *et al.* (1994)] Murata, N., Yoshizawa, S., Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, **5** (6), 865–872.

[Nishii (1984)] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12** (2), 758–765.

[Pardo (2005)] Pardo, L. (2005). *Statistical inference based on divergence measures*, CRC Press.

[Read and Cressie (1988)] Read, T. R. C. and Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer.

[Schwarz (1978)] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** (2), 461–464.

[Sugiura (1978)] Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*, **7** (1), 13–26.

[Takeuchi (1976)] Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models. *Mathematical Science*, **153**, 12–18 (in Japanese).