Journal of Computational and Applied Mathematics 440 (2024) 115630

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Model Selection for independent not identically distributed observations based on Rényi's pseudodistances

Angel Felipe, Maria Jaenada, Pedro Miranda *, Leandro Pardo

Department of Statistics and O.R., Complutense University of Madrid, Spain

ARTICLE INFO

Article history:
Received 19 April 2023
Received in revised form 20 September 2023

Keywords:
Rényi's pseudodistance
Robustness
Restricted model
Multiple linear regression model

ABSTRACT

Model selection criteria are rules used to select the best statistical model among a set of candidate models, striking a trade-off between goodness of fit and model complexity. Most popular model selection criteria measures the goodness of fit trough the model log-likelihood function, yielding to non-robust criteria. This paper presents a new family of robust model selection criteria for independent but not identically distributed observations (i.n.i.d.o.) based on the Rényi's pseudodistance (RP). The RP-based model selection criterion is indexed with a tuning parameter α controlling the trade-off between efficiency and robustness. Some theoretical results about the RP criterion are derived and the theory is applied to the multiple linear regression model, obtaining explicit expressions of the model selection criterion. Moreover, restricted models are considered and explicit expressions under the multiple linear regression model with nested models are accordingly derived. Finally, a simulation study empirically illustrates the robustness advantage of the method.

© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

Consider a set of real-life observations coming from an unknown distribution to be statistically modeled. Different candidate models may be assumed to fit the data and so a natural question arises as to how to choose the model that best fits the data. If the assumed model is too simple, with few number of parameters, it may not capture some important patterns and relationships in the data. In contrast, if the assumed model is too complex with large number of parameters, the estimated model parameters may over-fit the observed data (including possible sample noise), then resulting in a poor performance when the model is applied to new data. A model selection criterion is a rule used to select a statistical model among a set of candidates based on the observed data. It defines an objective criterion function quantifying the compromise between goodness of fit and model complexity, typically measured through an expected dissimilarity or divergence. Then, the dissimilarity measure needs to be minimized to select the model with the best trade-off. In other words, model selection criteria rely on a measure of fairness between a candidate model and the true model (i.e., the probability distribution generating the data).

The Akaike information criterion (AIC) is one of the most widely known and used in statistical practice model selection criterion. It was developed by Akaike [1,2] as the first model selection criterion in the statistical literature. The AIC estimates the expected Kullback–Leibler divergence [3] between the true model underlying the data and a fitted candidate model, and selects the model with minimum AIC. Of course, the true model underlying the data is generally unknown and so an empirical estimate obtained from the observed data is used.

* Corresponding author.
E-mail address: pmiranda@ucm.es (P. Miranda).

https://doi.org/10.1016/j.cam.2023.115630
0377-0427/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Following similar ideas than the AIC, several other model selection criteria have been proposed in the literature. For example, Schwarz in [4] developed the "Bayesian information criterion" (BIC), which imposes a stronger penalty for model complexity than AIC. Also derived from AIC, Hurvich and Tsai [5–7] studied the bias problem of the AIC and corrected it with a new criterion called "Corrected Akaike information criterion" (AIC_C). This criterion tries to cope with the fact that the AIC is only asymptotically unbiased and hence, the bias may be important when the sample size is not large enough and the number of parameters is large. Indeed, under small samples sizes the AIC tends to overfitting the observed data. Konishi and Kitagawa [8] extended the framework in which AIC has been developed to a general framework, including other estimation methods than maximum likelihood to fit the assumed candidate model. The resulting model selection criterion was called the "generalized information criterion" (GIC). The penalty term of GIC reduces to that of "Takeuchi information criterion" (TIC) developed by Takeuchi in [9] when the fitting method is maximum likelihood. Finally, Bozdogon [10] proposed another variant of AIC, called CAIC, that corrected its lack of consistency. Interesting surveys about model selection criteria can be found in [11,12].

Most of the previous procedures measure the fairness in terms of the Kullback–Leibler divergence. However, some other divergence measures have been explored, extending the methods with better robustness properties. For example, [13] considered the density power divergence (DPD) [14] to define a robust model selection criterion. Similarly, Toma et al. [15] introduced another robust criterion for model selection based on the Rényi pseudodistance (RP) [16]. Related to the problem of selecting the best model in regression, i.e. addressing the problem of selecting the most appropriate variables, we can consider the recent proposals based on divergence measures appearing in [17,18]. See also [19].

All the previous criteria assume that the observations are independent and identically distributed. A new problem appears if the observations are independent but not identically distributed (i.n.i.d.o.). In this context, Kurata and Hamada [20] considered a criterion based on DPD, extending the theory of [13]. The main purpose of this paper is to introduce a new robust model selection criterion in the context of i.n.i.d.o. based on RP, thus extending the methods of [15].

The rest of the paper goes as follows. In Section 2 we introduce RP for i.n.i.d.o. and we present some theoretical results necessary for next sections. The criterion based on RP is considered in Section 3 and an application to multiple linear regression model (MLRM) is presented. Section 4 studies the restricted case, where some additional conditions on the parameter space are imposed. The corresponding explicit expressions for the MLRM comparing a model with many parameters to other with a reduced number of parameters are derived. In Section 5 we deal with the Influence Function of this criterion. In Section 6 a simulation study illustrates the robustness of the proposed criterion and compare it with other model selection criteria. Section 7 deals with a real data example. Some final conclusions are presented in Section 8. The proofs of the main results in the paper appear in an Appendix.

2. Rényi's pseudodistance for independent but not identically distributed observations

Let $Y_1, \ldots, Y_n$ be i.n.i.d.o. observations, where each Y_i has true probability distribution function $G_i, i = 1, \ldots, n$, and probability density function $g_i, i = 1, \ldots, n$, respectively. For inferential purposes, it is assumed that the true density function g_i could belong to a parametric family of densities, $f_i(y, \boldsymbol{\theta}), i = 1, \ldots, n$, with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ a common model parameter for all the density functions. In the following, we shall denote by $F_i(y, \boldsymbol{\theta})$ the distribution function associated to the density function $f_i(y, \boldsymbol{\theta}), i = 1, \ldots, n$.

The value of $\boldsymbol{\theta}$ that best fits the original distributions $g_1, \ldots, g_n$, would naturally minimize some kind of distance between the true and assumed densities, $(g_1(y), \ldots, g_n(y))$ and $(f_1(y, \boldsymbol{\theta}), \ldots, f_n(y, \boldsymbol{\theta}))$. Here, we will use the family of RP divergence measures defined in [16] as measure of closeness between both sets of densities.

Definition 1. Consider $f(\cdot, \boldsymbol{\theta}), g(\cdot)$ two probability density functions. The **Rényi's pseudodistance** (RP) between f and g of tuning parameter $\alpha > 0$ is defined by

$$R_\alpha\left(f(\cdot, \boldsymbol{\theta}), g(\cdot)\right) = \frac{1}{\alpha+1} \log\left(\int f(y, \boldsymbol{\theta})^{\alpha+1} dy\right) - \frac{1}{\alpha} \log\left(\int f(y, \boldsymbol{\theta})^{\alpha} g(y) dy\right) + \frac{1}{\alpha\,(\alpha+1)} \log\left(\int g(y)^{\alpha+1} dy\right). \tag{1}$$

The RP divergence defined in Eq. (1) is always positive and it only reaches the zero when both densities coincide. Then, the best model parameter value approximating the underlying distribution would naturally minimize Eq. (1) in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Indeed, if the true distribution g belongs to the assumed parametric model with true parameter $\boldsymbol{\theta}_0$, the global minimizer of the RP is necessarily $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

At $\alpha = 0$, the corresponding **Rényi's pseudodistance** between f and g can be defined by taking continuous limits as follows

$$\begin{aligned} R_0\left(f(\cdot, \boldsymbol{\theta}), g(\cdot)\right) &= \lim_{\alpha \downarrow 0} R_\alpha\left(f(y, \boldsymbol{\theta}), g(y)\right) = \int g(y) \log \frac{g(y)}{f(y, \boldsymbol{\theta})} dy \\ &= \int g(y) \log g(y) dy - \int g(y) log f(y, \boldsymbol{\theta}) dy. \end{aligned} \tag{2}$$

Hence, $R_0\,(f(\cdot,\boldsymbol{\theta}), g(\cdot))$ coincides with the Kullback–Leibler divergence measure between g and f. For more results about Kullback–Leibler divergence measure, see [21]. The RP have been applied in many different statistical models with very promising results in terms of robustness with a small loss of efficiency. For example, [22] considered the RP divergence under the name of γ-cross entropy. Additionally, Toma and Leoni-Aubin [23] defined new robust and efficient measures based on RP. In [24], Wald-type tests based on RP were developed in the context of MLRM, and were extended later in [25] for the generalized multiple regression model. In [26], Wald-type tests based on RP for two dependent normal populations were developed. Moreover, in [25] a robust approach for comparing two dependent normal populations via a Wald-type test based on RP was carried out. In [27] the restricted MRPE was considered and their asymptotic properties studied; moreover, an application to Rao-type tests based on the restricted RP was there developed.

Note that the last term in Eq. (1) does not depend on $\boldsymbol{\theta}$. Hence, the minimizer of the RP measure can be obtained, for $\alpha > 0$, by minimizing the surrogate function

$$\frac{1}{\alpha+1}\log\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right) - \frac{1}{\alpha}\log\left(\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy\right). \tag{3}$$

The above expression can be rewritten using logarithm properties as

$$-\frac{1}{\alpha}\log\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}},$$

and thus minimizing $R_\alpha(f(\cdot,\boldsymbol{\theta}), g(\cdot))$ in $\boldsymbol{\theta}$, for $\alpha > 0$, is equivalent to minimize

$$V_\alpha^*(\boldsymbol{\theta}) = -\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}}. \tag{4}$$

Similarly, for $\alpha = 0$, we have that the first term in Eq. (2) does not depend on $\boldsymbol{\theta}$ and hence, minimizing $R_0\,(f(\cdot,\boldsymbol{\theta}), g(\cdot))$ is equivalent to minimizing

$$V_0^*(\boldsymbol{\theta}) = -\int g(y)\log f(y,\boldsymbol{\theta})dy. \tag{5}$$

However, now Expression (4) does not tend to Expression (5) when $\alpha \to 0$. In order to recover such convergence, and then extend the classical results based on Kullback–Leibler divergence, we slightly modify Expression (4) as

$$V_\alpha(\boldsymbol{\theta}) = -\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\alpha\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}, \tag{6}$$

where the value of $\boldsymbol{\theta}$ minimizing (4) is the same as for minimizing (6). Next lemma proves the required convergence of the objective functions.

Lemma 2. *For any two density function $f(\cdot,\boldsymbol{\theta})$ and $g(\cdot)$, the following convergence holds*

$$\lim_{\alpha\to 0} V_\alpha(\boldsymbol{\theta}) = V_0(\boldsymbol{\theta}).$$

Proof. First, note that

$$\lim_{\alpha\to 0}\left(-\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\alpha\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}\right) \tag{7}$$

leads to an indeterminate $(0/0)$. Let us denote

$$z(\alpha) = \left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}.$$

Taking derivatives on its logarithm

$$\log z(\alpha) = \frac{\alpha}{\alpha+1}\log\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right),$$

we obtain, after some algebra, that $\frac{\partial \log z(\alpha)}{\partial\alpha} = \frac{1}{z(\alpha)}\frac{\partial z(\alpha)}{\partial\alpha}$. On the other hand, the derivative of the function $\log z(\alpha)$ is given by

$$\frac{\partial \log z(\alpha)}{\partial\alpha} = \frac{1}{(\alpha+1)^2}\log\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right) + \frac{\alpha}{\alpha+1}\frac{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}\log f(y,\boldsymbol{\theta})dy\right)}{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)},$$

and solving the above equation we have that

$$\frac{\partial z(\alpha)}{\partial \alpha} = \left[\frac{1}{(\alpha+1)^2} \log\left(\int f(y,\boldsymbol{\theta})^{\alpha+1} dy \right) + \frac{\alpha}{\alpha+1} \frac{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1} \log f(y,\boldsymbol{\theta}) dy \right)}{\left(\int f(y,\boldsymbol{\theta})^{\alpha+1} dy \right)} \right] \times \left(\int f(y,\boldsymbol{\theta})^{\alpha+1} dy \right)^{\frac{\alpha}{\alpha+1}}.$$

Hence, applying L'Hôpital rule in (7), we obtain that

$$\lim_{\alpha\to 0} -\frac{\int f(y,\boldsymbol{\theta})^{\alpha} g(y) dy}{\alpha \left(\int f(y,\boldsymbol{\theta})^{\alpha+1} dy \right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha} = \lim_{\alpha\to 0} \frac{-\int f(y,\boldsymbol{\theta})^{\alpha} g(y) \log f(y,\boldsymbol{\theta}) dy + \frac{\partial z(\alpha)}{\partial \alpha}}{z - \alpha \frac{\partial z(\alpha)}{\partial \alpha}}.$$

Finally,

- $\lim_{\alpha\to 0} \int f(y,\boldsymbol{\theta})^{\alpha} g(y) \log f(y,\boldsymbol{\theta}) dy = \int g(y) \log f(y,\boldsymbol{\theta}) dy$.
- $\lim_{\alpha\to 0} \frac{\partial z(\alpha)}{\partial \alpha} = \frac{1}{1} \log 1 + \frac{0}{1} \frac{\int f(y,\boldsymbol{\theta}) \log f(y,\boldsymbol{\theta}) dy}{1} = 0$.
- $\lim_{\alpha\to 0} z = 1^0 = 1$.

Hence, the result holds. ■

Now, let us denote $V_{i,\alpha}(\boldsymbol{\theta})$ the corresponding objective functions for each pair of distributions $(f_i(y,\boldsymbol{\theta}), g_i(y))$, $i = 1,\dots,n$, as given in (6). As all densities $f_i(y,\boldsymbol{\theta})$ share a common parameter, the model parameter that best approximates the different underlying densities should minimize the weighted objective function, giving equal weights to all functions $V_{i,\alpha}(\boldsymbol{\theta})$. Hence, we consider

$$H_\alpha(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} V_{i,\alpha}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[-\frac{\int f_i(y,\boldsymbol{\theta})^{\alpha} g_i(y) dy}{\alpha \left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1} dy \right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha} \right]. \tag{8}$$

Definition 3. Consider $(g_1(y),\dots,g_n(y))$ and $(f_1(y,\boldsymbol{\theta}),\dots,f_n(y,\boldsymbol{\theta}))$, n pairs of true and assumed densities for i.n.i.d.o. random variables Y_i, $i = 1,\dots,n$. For any $\alpha \geq 0$, the value $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$ satisfying

$$\boldsymbol{\theta}_{\boldsymbol{g},\alpha} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \left[-\frac{\int f_i(y,\boldsymbol{\theta})^{\alpha} g_i(y) dy}{\alpha \left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1} dy \right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha} \right] = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} V_{i,\alpha}(\boldsymbol{\theta}).$$

is called the **best-fitting parameter according to RP**.

In the following we shall assume that there exists an open subset $\boldsymbol{\Theta_0} \subset \boldsymbol{\Theta}$ that contains the best-fitting parameter $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$.

For any fixed $i = 1,\dots,n$, the true distribution g_i of the random variable Y_i is usually unknown in practice and thus $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$ must be empirically estimated. As we only have one observation of each variable Y_i, the best way to estimate g_i based on the observation y_i is assuming that the distribution is degenerate in y_i. We will denote this degenerate distribution by $\widehat{g}_i$. Therefore, the empirical estimate of the RP divergence with $\alpha > 0$, given in Eq. (1) is

$$R_\alpha\left(f_i(Y_i,\boldsymbol{\theta}), \widehat{g}_i\right) = \frac{1}{\alpha+1} \log\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1} dy \right) - \frac{1}{\alpha} \log f_i(Y_i,\boldsymbol{\theta})^{\alpha} + k, \tag{9}$$

and similarly the empirical estimate of the RP for $\alpha = 0$, stated in (2), yields to

$$R_0\left(f_i(Y_i,\boldsymbol{\theta}), \widehat{g}_i\right) = -\log f_i(Y_i,\boldsymbol{\theta}) + k, \tag{10}$$

where k in (9) and (10) denotes a constant that does not depend on $\boldsymbol{\theta}$. As discussed earlier, the best estimator of the model parameter $\boldsymbol{\theta}$, based on the RP divergence should minimize its empirical estimate. But again, minimizing the estimated RP, $R_\alpha\left(f_i(Y_i,\boldsymbol{\theta}), \widehat{g}_i\right)$, for $\alpha > 0$, is equivalent to minimizing

$$\widehat{V}_{i,\alpha}\left(Y_i,\boldsymbol{\theta}\right) = -\frac{f_i(Y_i,\boldsymbol{\theta})^{\alpha}}{\alpha \left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1} dy \right)^{\frac{\alpha}{\alpha+1}}} + \frac{1}{\alpha}. \tag{11}$$

and for $\alpha = 0$, we can proceed the same way and conclude that minimizing $R_0\left(f_i(Y_i,\boldsymbol{\theta}), \widehat{g}_i\right)$ in $\boldsymbol{\theta}$, is equivalent to minimizing

$$\widehat{V}_{0,\alpha}\left(Y_i,\boldsymbol{\theta}\right) = -\log f(Y_i,\boldsymbol{\theta}). \tag{12}$$

Now, all the available information about the true value of the parameter comes from the set observed data, and so to obtain the best estimation fitting jointly all the observations we should consider the weighted objective function given

for $\alpha > 0$ as

$$\begin{aligned} H_{n,\alpha}(\boldsymbol{\theta}) =& \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{f_i(Y_i,\boldsymbol{\theta})^{\alpha}}{\alpha L_{\alpha}^{i}(\boldsymbol{\theta})}+\frac{1}{\alpha}\right] \\ =& \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta}). \end{aligned} \tag{13}$$

with

$$L_{\alpha}^{i}(\boldsymbol{\theta}) = \left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}},$$

and correspondingly,

$$H_{n,0}(\boldsymbol{\theta}) = \lim_{\alpha\to 0} H_{n,\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{i,0}(Y_i,\boldsymbol{\theta}). \tag{14}$$

Remark at this point that the expected values of the estimates are indeed the theoretical objective functions

$$V_{i,\alpha}(\boldsymbol{\theta}) = E_{Y_i}\left[\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})\right], \quad H_{\alpha}(\boldsymbol{\theta}) = E_{Y_1,\ldots,Y_n}\left[H_{n,\alpha}(\boldsymbol{\theta})\right].$$

Definition 4. Given $Y_1,\ldots,Y_n$ be i.n.i.d.o. and $\alpha > 0$, the **minimum RP estimator (MRPE)**, $\widehat{\boldsymbol{\theta}}_{\alpha}$, is given by

$$\widehat{\boldsymbol{\theta}}_{\alpha} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} H_{n,\alpha}(\boldsymbol{\theta}), \tag{15}$$

with $H_{n,\alpha}(\boldsymbol{\theta})$ defined in (13) for $\alpha > 0$ and in (14) for $\alpha = 0$.

Note that at $\alpha = 0$, we recover the maximum likelihood estimator (MLE) of the model and so the MRPE family includes the classical estimator as a particular case.

As the MRPE, $\widehat{\boldsymbol{\theta}}_{\alpha}$, is a minimum of a differentiable function, it must annul the first derivatives of the function $H_{n,\alpha}(\boldsymbol{\theta})$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\theta})}{\partial\theta_j} = 0, \quad j = 1,\ldots,p.$$

That is, the estimation equations of the MRPE are

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha L_{\alpha}^{i}(\boldsymbol{\theta})^{2}}\left(\alpha f_i(Y_i,\boldsymbol{\theta})^{\alpha}u_j(Y_i,\boldsymbol{\theta})L_{\alpha}^{i}(\boldsymbol{\theta}) - \frac{\partial L_{\alpha}^{i}(\boldsymbol{\theta})}{\partial\theta_j}f_i(Y_i,\boldsymbol{\theta})^{\alpha}\right) = 0, \quad j = 1,\ldots,p,$$

with

$$u_j(y,\boldsymbol{\theta}) = \frac{\partial\log(f_i(y,\boldsymbol{\theta}))}{\partial\theta_j},$$

and

$$\begin{aligned} \frac{\partial L_{\alpha}^{i}(\boldsymbol{\theta})}{\partial\theta_j} =& \frac{\alpha}{\alpha+1}\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}-1}(\alpha+1)\int f_i(y,\boldsymbol{\theta})^{\alpha+1}u_j(y,\boldsymbol{\theta})dy \\ =& \alpha\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}-1}\int f_i(y,\boldsymbol{\theta})^{\alpha+1}u_j(y,\boldsymbol{\theta})dy, \; i = 1,\ldots,n. \end{aligned}$$

It is interesting to observe that if $Y_1,\ldots,Y_n$ are independent and identically distributed (i.i.d.) random variables, the MRPE $\widehat{\boldsymbol{\theta}}_{\alpha}$ coincides with the estimator $\widehat{\boldsymbol{\theta}}_{\alpha}^{*}$ proposed in [28].

We next study the asymptotic distribution of the MRPE, $\widehat{\boldsymbol{\theta}}_{\alpha}$. For notation simplicity, let us define the matrices $\boldsymbol{\Psi}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ as follows:

$$\boldsymbol{\Psi}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right), \tag{16}$$

with

$$\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) = \left(E_{Y_i}\left[\frac{\partial^2\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\right)_{j,k=1,\ldots,p}, i = 1,\ldots,n,$$

and

$$\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)=\frac{1}{n}\sum_{i=1}^{n}Var_{Y_i}\left[\left(\frac{\partial\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\theta})}{\partial\theta_j}\right)_{j=1,\ldots,p}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}},\ i=1,\ldots,n. \tag{17}$$

Additionally, let $\lambda_1,\ldots,\lambda_n$ be the eigenvalues of $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$. From now on, we will assume that $\inf_n\lambda_n>0$, so that $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ can be inverted.

Consider the conditions **C1–C7** that are given in Appendix. Now, the following result, whose proof can be seen in [26], holds.

Theorem 5. *Suppose the previous regularity conditions* **C1–C7** *hold. Then,*

$$\sqrt{n}\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-\frac{1}{2}}\ \boldsymbol{\Psi}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\left(\widehat{\boldsymbol{\theta}}_{\alpha}-\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\underset{n\to\infty}{\overset{L}{\to}}N(\mathbf{0}_p,\boldsymbol{I}_p), \tag{18}$$

being $\boldsymbol{I}_p$ the p-dimensional identity matrix.

Remark 6. In [26], a similar study is done for the estimation problem. The difference between this paper and [26] is that in order to extend ML as established in Lemma 2, we have changed the signs of some expressions and the term $\frac{1}{\alpha}$ is added. More concretely, we have considered

$$V_{\alpha}(\boldsymbol{\theta})=-\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\alpha\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}}+\frac{1}{\alpha},$$

while the corresponding expression in [26] is

$$V_{\alpha}(\boldsymbol{\theta})=\frac{\int f(y,\boldsymbol{\theta})^{\alpha}g(y)dy}{\alpha\left(\int f(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}}}.$$

Hence, the expressions for $H_{\alpha}(\boldsymbol{\theta})$ are opposite in this paper and in [26] up to an additive term. Note however that this term does not depend on $\boldsymbol{\theta}$, so that the estimation equations are the same. Moreover, the expressions for $\boldsymbol{J}_{\alpha}^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and $\boldsymbol{\Omega}_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ are the same, so that the convergence of Theorem 5 applies in our case, too.

2.1. Example: The MPRE under the MLRM

Consider $(Y_1,\ldots,Y_n)$ a set of random variables, related to the explanatory variables $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ through the MLRM,

$$Y_i=\boldsymbol{X}_i^T\boldsymbol{\beta}+\varepsilon_i,\quad i=1,\ldots,n, \tag{19}$$

where the errors $\varepsilon_i's$ are i.i.d. normal random variables with mean zero and variance σ^2, $\boldsymbol{X}_i^T=(X_{i1},\ldots,X_{ip})$ is the vector of independent variables corresponding to the ith condition and $\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_p\right)^T$ is the vector of regression coefficients to be estimated. We will consider that, for each i, $\boldsymbol{X}_i$ is fixed, yielding to i.n.i.d.o. $Y_i's$, with $Y_i\sim\mathcal{N}(\boldsymbol{X}_i^T\boldsymbol{\beta},\sigma^2)$.

We next derive the explicit expression of the MRPE for the parameters $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma)$. With the previous notation, the assumed density functions are $f_i(y,\boldsymbol{\beta},\sigma)\equiv\mathcal{N}(\boldsymbol{X}_i^T\boldsymbol{\beta},\sigma^2)$ and then, using Eq. (6), we have that for $\alpha>0$,

$$\begin{aligned}\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\beta},\sigma)&=-\frac{\frac{1}{(2\pi)^{\alpha/2}\sigma^{\alpha}}\exp\left(\frac{-\alpha(Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)}{\alpha\left((2\pi)^{\alpha/2}\sigma^{\alpha}\sqrt{1+\alpha}\right)^{-\frac{\alpha}{\alpha+1}}}+\frac{1}{\alpha}\\&=-\frac{1}{\alpha}\left(\frac{1+\alpha}{2\pi}\right)^{\frac{\alpha}{2(\alpha+1)}}\sigma^{-\frac{\alpha}{\alpha+1}}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)+\frac{1}{\alpha}.\end{aligned} \tag{20}$$

and thus, the MRPE for $\alpha>0$ is obtained minimizing the averaged objective function

$$\begin{aligned}H_{n,\alpha}(\boldsymbol{\beta},\sigma)&=\frac{1}{n}\sum_{i=1}^{n}\widehat{V}_{i,\alpha}(Y_i;\boldsymbol{\beta},\sigma)\\&=-\frac{1}{\alpha}\left(\frac{1+\alpha}{2\pi}\right)^{\frac{\alpha}{2(\alpha+1)}}\frac{1}{n}\sum_{i=1}^{n}\sigma^{-\frac{\alpha}{\alpha+1}}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)+\frac{1}{\alpha}.\end{aligned}$$

Ignoring all constant terms, we have that the MRPE for the MLRM is given, for $\alpha>0$, as

$$\left(\widehat{\boldsymbol{\beta}}_{\alpha},\widehat{\sigma}_{\alpha}\right)=\arg\min_{\boldsymbol{\beta},\sigma}\sum_{i=1}^{n}-\sigma^{-\frac{\alpha}{\alpha+1}}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right).$$

Moreover, taking derivatives with respect to $\boldsymbol{\beta}$ and σ, the estimation equations of $\widehat{\boldsymbol{\beta}}_\alpha$ and $\widehat{\sigma}_\alpha$ are

$$\begin{aligned}&\sum_{i=1}^{n}\exp\left(-\tfrac{\alpha}{2}\left(\tfrac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left(\tfrac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{X}_i=\mathbf{0}_p\\&\sum_{i=1}^{n}\exp\left(-\tfrac{\alpha}{2}\left(\tfrac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left\{\left(\tfrac{Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2-\tfrac{1}{1+\alpha}\right\}=0\end{aligned},\tag{21}$$

which is exactly the same system as the one obtained in [26]. For $\alpha = 0$, if we denote $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T_{n\times p}$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, we get the MLE of $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\sigma}_0$, i.e.

$$\widehat{\boldsymbol{\beta}}_0 = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \text{ and } \widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_0\right)^2.$$

Finally, from the results in [26], it can be seen that matrices $\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\beta},\sigma)$ and $\boldsymbol{\Omega}_{n,\alpha}(\boldsymbol{\beta},\sigma)$ are given by

$$\begin{aligned}\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\beta},\sigma) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{J}^{(i)}\left(\boldsymbol{\beta},\sigma^2\right)\\&= k\sigma^{-\frac{3\alpha+2}{\alpha+1}}(\alpha+1)^{-\frac{3}{2}}\begin{bmatrix}\frac{1}{n}\mathbb{X}^T\mathbb{X} & 0\\ 0 & \frac{2}{\alpha+1}\end{bmatrix}\\&= K_1(\alpha+1)^{-\frac{3}{2}}\begin{bmatrix}\frac{1}{n}\mathbb{X}^T\mathbb{X} & 0\\ 0 & \frac{2}{\alpha+1}\end{bmatrix},\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Omega}_{n,\alpha}(\boldsymbol{\beta},\sigma) &= \frac{1}{n}\sum_{i=1}^{n}Var_{Y_i}\left[\left(\frac{\partial V_{i,\alpha}(Y_i;\boldsymbol{\beta},\sigma^2)}{\partial\theta_j}\right)_{j=1,\ldots,k}\right]\\&= K_1^2\sigma^2\frac{1}{(2\alpha+1)^{3/2}}\begin{bmatrix}\frac{1}{n}\mathbb{X}^T\mathbb{X} & \mathbf{0}\\ \mathbf{0} & \frac{(3\alpha^2+4\alpha+2)}{(\alpha+1)^2(2\alpha+1)}\end{bmatrix}.\end{aligned}$$

with

$$k = \frac{1}{\alpha}\left(\frac{1+\alpha}{2\pi}\right)^{\frac{\alpha}{2(\alpha+1)}}, \quad K_1 = k\alpha\sigma^{-\frac{3\alpha+2}{\alpha+1}}.\tag{22}$$

These are the same matrices appearing in [26] up to a multiplicative constant. For the previous matrices and taking $\alpha = 0$, we get the Fisher information matrix for $(\boldsymbol{\beta},\sigma)$ in both cases, i.e.

$$\boldsymbol{\Psi}_{n,0}(\boldsymbol{\beta},\sigma) = \begin{bmatrix}\frac{1}{\sigma^2}\frac{1}{n}\mathbb{X}^T\mathbb{X} & 0\\ 0 & \frac{2}{\sigma^2}\end{bmatrix},$$

and

$$\boldsymbol{\Omega}_{n,0}(\boldsymbol{\beta},\sigma) = \begin{bmatrix}\frac{1}{\sigma^2}\frac{1}{n}\mathbb{X}^T\mathbb{X} & \mathbf{0}\\ \mathbf{0} & \frac{2}{\sigma^2}\end{bmatrix}.$$

3. Model selection criterion based on RP

In this section we present the model selection criterion based on RP. Let us consider a collection of l candidate models

$$\left\{\boldsymbol{M}^{(s)} = \left(M_1^{(s)}, \ldots, M_n^{(s)}\right)\right\}_{s\in\{1,\ldots,l\}}\tag{23}$$

such that each $\boldsymbol{M}^{(s)}$ is characterized by the parametric density functions

$$\boldsymbol{f}(\cdot,\boldsymbol{\theta}_s) = (f_1(\cdot,\boldsymbol{\theta}_s), \ldots, f_n(\cdot,\boldsymbol{\theta}_s)), \ \boldsymbol{\theta}_s \in \boldsymbol{\Theta}_s \subset \mathbb{R}^{p_s},$$

with associated distribution functions $\boldsymbol{F}(.,\boldsymbol{\theta}_s) = (F_1(\boldsymbol{\theta}_s), \ldots, F_n(.,\boldsymbol{\theta}_s))$, where $\boldsymbol{\theta}_s$ is common for all density functions in model s. That is, each candidate model would represent a parametric family defined by a common parameter, which may contain different number of parameters. Based on the random sample $Y_1, \ldots, Y_n$, we need to select the best model from the collection $\{\boldsymbol{M}^{(s)}\}_{s\in\{1,\ldots,l\}}$ according to some suitable selection criterion. For such purpose, for each assumed model $\boldsymbol{M}^{(s)}$, we should first determine the best parameter $\boldsymbol{\theta}_s$ fitting the sample and subsequently select the best fitted model from the

collection. Then, given a set of observations, the model selection is performed in two steps: we first fit all the candidates models to the data, and then select the model with best trade-off between goodness of fit and complexity in terms of RP.

We next describe the first step of the model selection algorithm. Let consider a fixed parametric model $\boldsymbol{M}^{(s)}$ modeling the true underlying distribution. If the true distribution was known, the parameter that best fits the model $\boldsymbol{M}^{(s)}$, denoted by $\boldsymbol{\theta}^s_{\boldsymbol{g},\alpha}$, can be obtained by maximizing the theoretical averaged objective function $H_\alpha(\boldsymbol{\theta})$ defined in Eq. (8) under the s-model.

Following the discussion in Section 2, if the true underlying distribution is unknown but we have a random sample $Y_1, \ldots, Y_n$, the best estimate of the true parameter based on the sample from the RP approach is the MRPE defined in (15).

Once all candidate models are fitted to the observed data (or to the true distribution, if it is known), we should select the model with the best trade-off between fitness and complexity. Therefore, we need a measure of fairness between the best candidate for each model and the true distribution. The goodness of fit of a certain model $\boldsymbol{M}^{(s)}$ with associated densities $\boldsymbol{f}(\cdot, \boldsymbol{\theta}^s_g)$ and the best-fitting parameter $\boldsymbol{\theta}^s_g$ based on the RP can be quantified by the averaged objective function $H_\alpha(\boldsymbol{\theta}^s_g)$ given in Eq. (8).

As the true distribution is generally unknown, $\boldsymbol{\theta}^s_g$ is estimated by $\widehat{\boldsymbol{\theta}}^s_\alpha$. Hence, we can estimate $H_\alpha(\boldsymbol{\theta}^s_g)$ by $H_\alpha(\widehat{\boldsymbol{\theta}}^s_\alpha)$. But again H_α needs to be estimated, and the natural estimator is $H_{n,\alpha}(\widehat{\boldsymbol{\theta}}^s_\alpha)$. However, as the sample is used both for estimating the parameter and for estimating H_α, it does not hold that

$$E_{Y_1,\ldots,Y_n}\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\right] \neq E_{Y_1,\ldots,Y_n}\left[H_\alpha\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\right].$$

Moreover, the estimation bias would depend on the model and consequently, we need to add a term correcting the bias caused by the model assumption.

The AIC criterion selects the model that minimizes

$$-2\sum_{i=1}^{n}\log f_i(y_i, \boldsymbol{\theta}) + 2p = 2H_{n,0}(\boldsymbol{\theta}) + 2p,$$

where $2p$ is the term correcting the bias. Following the same idea, we define the $RP_{NH}-$Criterion as follows:

Definition 7. Let $\left\{\left(M_1^{(s)}, \ldots, M_n^{(s)}\right)\right\}_{s\in\{1,\ldots,l\}}$ be l candidate models for the i.n.i.d.o. $Y_1, \ldots, Y_n$. The selected model $\left(M_1^*, \ldots, M_n^*\right)$ according the $RP_{NH}-$**Criterion** is the one satisfying

$$\left(M_1^*, \ldots, M_n^*\right) = \min_{s\in\{1,\ldots,l\}} RP_{NH}\left(M_1^{(s)}, \ldots, M_n^{(s)}, \widehat{\boldsymbol{\theta}}^s_\alpha\right),$$

where

$$RP_{NH}\left(M_1^{(s)}, \ldots, M_n^{(s)}, \widehat{\boldsymbol{\theta}}^s_\alpha\right) = H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right) + \frac{1}{n}trace\left(\boldsymbol{\Omega_n}\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\boldsymbol{\Psi}_n^{-1}\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\right). \tag{24}$$

We can observe that

$$\lim_{\alpha\to 0} RP_{NH}\left(M_1^{(s)}, \ldots, M_n^{(s)}, \widehat{\boldsymbol{\theta}}^s_\alpha\right) = -\frac{1}{n}\sum_{i=1}^{n}\log f_i(Y_i, \boldsymbol{\theta}) + \frac{p}{n},$$

and hence we recover AIC criterion up to the multiplicative constant $2n$.

The tuning parameter α controls the trade-off between efficiency and robustness. Hence, for small values of α (in the limit $\alpha = 0$), the corresponding results will be more efficient while less robust. On the other hand, for large values of α, the results will lead to robustness but with a loss of efficiency.

In order to justify the $RP_{NH}-$Criterion, we shall establish that the estimated function $RP_{NH}\left(M_1^{(s)}, \ldots, M_n^{(s)}\right)$ quantifying the loss of choosing a model is an unbiased estimator of it theoretical version, $E_{Y_1,\ldots,Y_n}\left[H_\alpha\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\right]$.

Theorem 8. *Assume that conditions* **C1–C8** *hold. Then,*

$$E_{Y_1,\ldots,Y_n}\left[RP_{NH}\left(M_1^{(s)}, \ldots, M_n^{(s)}, \widehat{\boldsymbol{\theta}}^s_\alpha\right)\right] = E_{Y_1,\ldots,Y_n}\left[H_\alpha\left(\widehat{\boldsymbol{\theta}}^s_\alpha\right)\right], \forall s = 1, \ldots, l.$$

Proof. See Appendix. ■

Remark 9. A nice robust model selection criterion in the context of penalized regression called robust Akaike information criterion (robust AIC) is proposed in [29]. Analyzing Theorem 3 in that paper, it can be seen that when the sample size tends to infinity, the robust AIC obtains the same expression as the RP_{NH} criterion proposed in this section from a structural point of view. The differences between both approaches rely on the expressions for the estimators. Consequently, it could

be interesting in a future research to compare the behavior of these two model selection criteria in the context of penalized regression. This similarity appears only in the case of penalized regression and cannot be generalized for other situations.

We next develop explicit expressions for the RP_{NH}-criterion under the MLRM.

3.1. Example: The RP-based model selection under the multiple linear regression model

We consider the MLRM defined in Section 2.1.

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n. \tag{25}$$

We consider several models $\{(M_1^{(s)}, \dots, M_n^{(s)})\}_{s=1,\dots,l}$ where each model differs on the parameter $\boldsymbol{\beta}$ considered. For example, consider four explanatory variables (X_1, X_2, X_3, X_4) and four different models given by

$$\begin{aligned}
&(M_1^{(1)}, \dots, M_n^{(1)}) \equiv Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i,\\
&(M_1^{(2)}, \dots, M_n^{(2)}) \equiv Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \epsilon_i\\
&(M_1^{(3)}, \dots, M_n^{(3)}) \equiv Y_i = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i,\\
&(M_1^{(4)}, \dots, M_n^{(4)}) \equiv Y_i = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i.
\end{aligned}$$

Each of the models has five parameters that need to be estimated. Let us then determine the corresponding values of $RP_{NH}\left(M_1^{(s)}, \dots, M_n^{(s)}, \widehat{\boldsymbol{\theta}}_\alpha^s\right)$ for $s = 1, 2, 3, 4$.

As stated in Section 2.1, for each $s = 1, 2, 3, 4$, the estimators of $\widehat{\boldsymbol{\beta}}_\alpha^s$ and $\widehat{\sigma}_\alpha^s$ are the solutions of the system

$$\begin{aligned}
&\sum_{i=1}^n \exp\left(-\tfrac{\alpha}{2}\left(\tfrac{Y_i - \boldsymbol{X}_{s,i}^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left(\tfrac{Y_i - \boldsymbol{X}_{s,i}^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{X}_{s,i} = \boldsymbol{0}_4\\
&\sum_{i=1}^n \exp\left(-\tfrac{\alpha}{2}\left(\tfrac{Y_i - \boldsymbol{X}_{s,i}^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left\{\left(\tfrac{Y_i - \boldsymbol{X}_{s,i}^T\boldsymbol{\beta}}{\sigma}\right)^2 - \tfrac{1}{1+\alpha}\right\} = 0
\end{aligned}, \tag{26}$$

where $X_{s,i}$ corresponds to the values of observation i restricted to the variables appearing in model s. Note that, although $\boldsymbol{\beta}$ has a different meaning for the different models, this is not the case of σ. However, the estimation of σ is different for the different models and so this estimation is denoted for by $\widehat{\sigma}_\alpha^s$ for the sth model.

At $\alpha = 0$, we have that the model parameters can be explicitly obtained as

$$\widehat{\boldsymbol{\beta}}_0^s = (\mathbb{X}_s^T\mathbb{X}_s)^{-1}\mathbb{X}_s^T\mathbf{Y} \text{ and } (\widehat{\sigma}_0^s)^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \boldsymbol{X}_{s,i}^T\widehat{\boldsymbol{\beta}}_0\right)^2.$$

Thus, according to Eq. (13),

$$H_{n,\alpha}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \frac{1}{\alpha}\frac{1}{n}\sum_{i=1}^n -k\widehat{\sigma}^{-\frac{\alpha}{\alpha+1}} \exp\left(-\frac{\alpha}{2}\left(\frac{Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right)^2\right) + \frac{1}{\alpha},$$

with k as defined in (22).

Next, let us obtain expressions of $\boldsymbol{\Psi}_{s,n}\left(\boldsymbol{\beta}^s, \sigma\right)$ and $\boldsymbol{\Omega}_{s,n}\left(\boldsymbol{\beta}^s, \sigma\right)$. Note that these matrices also depend on the model s. Applying again the results of the previous section, we obtain

$$\begin{aligned}
\boldsymbol{\Psi}_{s,n}\left(\boldsymbol{\beta}^s, \sigma\right) &= K_1\,(\alpha+1)^{-\frac{3}{2}}\begin{bmatrix}\frac{1}{n}\mathbb{X}_s^T\mathbb{X}_s & 0\\ 0 & \frac{2}{\alpha+1}\end{bmatrix},\\
\boldsymbol{\Omega}_{s,n}\left(\boldsymbol{\beta}^s, \sigma\right) &= K_1^2\sigma^2\frac{1}{(2\alpha+1)^{3/2}}\begin{bmatrix}\frac{1}{n}\mathbb{X}_s^T\mathbb{X}_s & 0\\ 0 & \frac{(3\alpha^2+4\alpha+2)}{2(\alpha+1)(2\alpha+1)}\end{bmatrix},
\end{aligned} \tag{27}$$

where K_1 was defined in (22). Note that these matrices have dimension $(p+1)\times(p+1)$ where p is the dimension of vector $\boldsymbol{\beta}$ for each model. In our example, $p = 4$ and therefore,

$$\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\beta}}_\alpha^s, \widehat{\sigma}_\alpha^s\right)\boldsymbol{\Psi}_{s,n}^{-1}\left(\widehat{\boldsymbol{\beta}}_\alpha^s, \widehat{\sigma}_\alpha^s\right) = (\widehat{\sigma}_\alpha^s)^2 K_1\frac{(\alpha+1)^{\frac{3}{2}}}{(2\alpha+1)^{\frac{3}{2}}}\begin{bmatrix}\boldsymbol{I}_{p\times p} & \boldsymbol{0}\\ \boldsymbol{0}^T & \frac{3\alpha^2+4\alpha+2}{(\alpha+1)^2(2\alpha+1)}\end{bmatrix},$$

and hence,

$$trace\left(\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\beta}}_\alpha^s, \widehat{\sigma}_\alpha^s\right)\boldsymbol{\Psi}_{s,n}^{-1}\left(\widehat{\boldsymbol{\beta}}_\alpha^s, \widehat{\sigma}_\alpha^s\right)\right) = (\widehat{\sigma}_\alpha^s)^2 K_1\left(p\frac{(\alpha+1)^{\frac{3}{2}}}{(2\alpha+1)^{\frac{3}{2}}} + \frac{(\alpha+1)^{\frac{1}{2}}\left(3\alpha^2+4\alpha+2\right)}{2\,(2\alpha+1)^{5/2}}\right).$$

Therefore, applying the RP_{NH}−Criterion given in Definition 7

$$RP_{NH}(M_1^{(s)},\dots,M_n^{(s)},\widehat{\boldsymbol{\beta}}_\alpha^s,\widehat{\sigma}_\alpha^2) = -\frac{1}{\alpha}\left(\frac{1+\alpha}{2\pi}\right)^{\frac{\alpha}{2(\alpha+1)}}\frac{1}{n}\sum_{i=1}^{n}(\widehat{\sigma}_\alpha^s)^{-\frac{\alpha}{\alpha+1}}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_{s,i}^T\widehat{\boldsymbol{\beta}}_\alpha^s}{\widehat{\sigma}_\alpha^s}\right)^2\right) + \frac{1}{\alpha}+\frac{1}{n}(\widehat{\sigma}_\alpha^s)^2K_1\left(p\frac{(\alpha+1)^{\frac{3}{2}}}{(2\alpha+1)^{\frac{3}{2}}}+\frac{(\alpha+1)^{\frac{1}{2}}\left(3\alpha^2+4\alpha+2\right)}{2\,(2\alpha+1)^{5/2}}\right). \tag{28}$$

Finally, we select the model with minimum, in s, $RP_{NH}(M_1^{(s)},\dots,M_n^{(s)},\widehat{\boldsymbol{\beta}}_\alpha^s,\widehat{\sigma}_\alpha^s)$ as the most appropriate model among the four candidates.

4. The restricted model

Let us consider a particular case of the model selection problem. In some situations it is interesting to compare a full model based on $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^p$, with p parameters with other restricted models where the parameter has to satisfy additionally linear constraints of the form

$$\{\boldsymbol{\theta}\in\boldsymbol{\Theta}/\ \boldsymbol{m}(\boldsymbol{\theta})=\mathbf{0}_r\}, \tag{29}$$

where $\mathbf{0}_r$ denotes the null vector of dimension r with $r<p$ and $\boldsymbol{m}:\mathbb{R}^p\to\mathbb{R}^r$ is a vector-valued function such that the $p\times r$ matrix

$$\mathbf{M}(\boldsymbol{\theta})=\frac{\partial\boldsymbol{m}^T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{30}$$

exists and is continuous in $\boldsymbol{\theta}$, and $\operatorname{rank}(\mathbf{M}(\boldsymbol{\theta}))=r,\ \forall\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Related to the divergence-based restricted estimation, in [30] the restricted minimum DPD estimator was defined. Later, in [31] the restricted MRPE for general populations was given.

Given a candidate model, we have already established that the best fitting parameter for this model based on the RP is defined by

$$\boldsymbol{\theta}_{\boldsymbol{g},\alpha}=\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^p}H_\alpha(\boldsymbol{\theta}),$$

where $H_\alpha(\boldsymbol{\theta})$ was defined in Eq. (8). On the other hand, applying the same criterion for the restricted model, we obtain that the best-fitting parameter for the restricted model is given by

$$\boldsymbol{\theta}_{\boldsymbol{g},\alpha}^R=\arg\min_{\boldsymbol{\theta}\in\Theta/\ \boldsymbol{m}(\boldsymbol{\theta})=\mathbf{0}_r}H_\alpha(\boldsymbol{\theta}).$$

Following similar arguments than in Section 2, we defined the restricted MRPE as follows.

Definition 10. Given $Y_1,\dots,Y_n$ be i.n.i.d.o., the **restricted MRPE** (RMRPE), $\widetilde{\boldsymbol{\theta}}_\alpha$, is given by

$$\widetilde{\boldsymbol{\theta}}_\alpha=\arg\min_{\boldsymbol{\theta}\in\Theta/\boldsymbol{m}(\boldsymbol{\theta})=\mathbf{0}_r}H_{n,\alpha}(\boldsymbol{\theta}), \tag{31}$$

with $H_{n,\alpha}(\boldsymbol{\theta})$ defined in (13) for $\alpha>0$ and in (14) for $\alpha=0$.

Note that

$$H_{n,\alpha}(\widehat{\boldsymbol{\theta}}_\alpha)\le H_{n,\alpha}(\widetilde{\boldsymbol{\theta}}_\alpha).$$

The following theorem presents a representation of the RMPRE.

Theorem 11. *Assume conditions* **C1–C8** *and suppose that* $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$ *satisfies the conditions of the restricted model. Then,*

$$n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha-\boldsymbol{\theta}_{\boldsymbol{g},\alpha})=\boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}+o_p(1),$$

being

$$\boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})=\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}-\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}, \tag{32}$$

with

$$\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})=\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\left[\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\right]^{-1}. \tag{33}$$

Proof. See Appendix. ■

In the following lemma we establish a property about matrix $\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ that will be required for the next theorem.

Lemma 12. *Given $\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ and $\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$, it follows*

$$\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = -\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}).$$

Proof. Applying the definitions and denoting

$$\boldsymbol{A}^{-1}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \left[\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\right]^{-1},$$

we obtain

$$\begin{aligned}
&\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\\
&= \left[\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{A}^{-1}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} - \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\right]\\
&\quad\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\\
&= \left[\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{A}^{-1}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T - Id\right]\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\\
&= \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{A}^{-1}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{A}^{-1}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\\
&\quad - \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{A}^{-1}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} - \boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\\
&= -\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}).
\end{aligned}$$

Hence, the result holds. ■

Suppose now that we have chosen a model as the best fitting model and we wonder if this model overfits the data and a restricted model is more accurate. Then, we can pose this problem as a model selection problem with two models, the big one and a restricted model, and apply the results of the previous section. Hence, it suffices to compute $RP_{NH}(M_1^{(s)},\ldots,M_n^{(s)},\boldsymbol{\theta})$ for both models and select the one attaining the minimum. Assuming the restricted model is correct, in the following theorem we shall establish the asymptotic distribution of

$$2n\left[RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha\right) - RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha\right)\right],$$

where $RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha\right)$ was given in (24) and

$$RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha\right) = H_{n,\alpha}\left(\widetilde{\boldsymbol{\theta}}_\alpha\right) + \frac{1}{n}trace\left(\boldsymbol{\Omega}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)\,\boldsymbol{\Psi}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)^{-1}\right),$$

being $\boldsymbol{\Psi}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)$ and $\boldsymbol{\Omega}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)$ the matrices defined in (16) and (17) but for the restricted model.

Note that the probability of selecting the restricted model is

$$\Pr\left(RP_{NH}\left(M_1^{(k)},\ldots,M_n^{(k)},\widehat{\boldsymbol{\theta}}_\alpha\right) - RP_{NH}\left(M_1^{(k)},\ldots,M_n^{(k)},\widetilde{\boldsymbol{\theta}}_\alpha\right) > 0\right).$$

Theorem 13. *Assume conditions* **C1–C8** *hold and suppose that the fitting parameter, $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}$, belongs to the restricted model. Then, the asymptotic distribution of*

$$2n\left(RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha\right) - RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha\right)\right)$$

coincides with the distribution of the random variable

$$\sum_{j=1}^{r}\lambda_j(\boldsymbol{\theta}_{g,\alpha})Z_j^2 + 2trace\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\,\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right) - 2trace\left(\boldsymbol{\Omega}_n^R\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)(\,\boldsymbol{\Psi}_n^R)^{-1}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right),$$

where $Z_1,\ldots,Z_k$ are independent standard normal variables, $\lambda_1(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}),\ldots,\lambda_r(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ are the nonzero eigenvalues of $-\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\,\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and

$$r = rank\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\,\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right).$$

Proof. See Appendix. ■

The above result provides a way to asymptotically compute the probability of over-fitting, which is of great interest in model selection theory.

4.1. Example: The RP-based model selection under the multiple linear regression model and restricted parameter spaces.

We shall consider the MLRM as defined in Section 3.1 and we are interested in comparing a full model with a restricted model under the restrictions

$$\beta_{p-r+1} = \cdots = \beta_p = 0.$$

In this case the model parameter is $\boldsymbol{\theta} = \left(\beta_0, \ldots, \beta_p, \sigma\right)$ and the function $\boldsymbol{m}(\boldsymbol{\theta})$ defining the restrictions is

$$\boldsymbol{m}(\boldsymbol{\theta}) = \boldsymbol{m}\left(\beta_0, \ldots, \beta_p, \sigma\right) = (\beta_{p-r+1}, \ldots, \beta_p).$$

Consequently, its derivative is given by

$$\boldsymbol{M}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \mathbf{0}_{(p-r+1)\times r} \\ \boldsymbol{I}_{r\times r} \\ \mathbf{0}_{1\times r} \end{pmatrix}.$$

Let us expressed the design matrix $\mathbb{X}$ as

$$\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2),$$

with $\mathbb{X}_1$ a $n \times (p-r+1)$ matrix and $\mathbb{X}_2$ a $n \times r$ matrix. It is clear that $\mathbb{X}_1$ is the design matrix for the restricted model and $\mathbb{X}_2$ corresponds to the design matrix for the full model whose parameters are not in the small model. The matrices $\boldsymbol{\Psi}_n(\boldsymbol{\beta}, \sigma)$ and $\boldsymbol{\Omega}_n(\boldsymbol{\beta}, \sigma)$ given in Eq. (27) can be rewritten, using the notation $\mathbb{X}_1$ and $\mathbb{X}_2$, as

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}, \sigma) = K_1 (\alpha+1)^{-\frac{3}{2}} \begin{bmatrix} \frac{1}{n}\mathbb{X}_1^T\mathbb{X}_1 & \frac{1}{n}\mathbb{X}_1^T\mathbb{X}_2 & 0 \\ \frac{1}{n}\mathbb{X}_2^T\mathbb{X}_1 & \frac{1}{n}\mathbb{X}_2^T\mathbb{X}_2 & 0 \\ 0 & 0 & \frac{2}{\alpha+1} \end{bmatrix},$$

being K_1 as defined in (22) and

$$\boldsymbol{\Omega}_n(\boldsymbol{\beta}, \sigma) = K_1^2\sigma^2 \frac{1}{(2\alpha+1)^{3/2}} \begin{bmatrix} \frac{1}{n}\mathbb{X}_1^T\mathbb{X}_1 & \frac{1}{n}\mathbb{X}_1^T\mathbb{X}_2 & 0 \\ \frac{1}{n}\mathbb{X}_2^T\mathbb{X}_1 & \frac{1}{n}\mathbb{X}_2^T\mathbb{X}_2 & 0 \\ 0 & 0 & \frac{(3\alpha^2+4\alpha+2)}{(\alpha+1)^2(2\alpha+1)} \end{bmatrix}.$$

Now, the inverse of the matrix $\boldsymbol{\Psi}_n(\boldsymbol{\beta}, \sigma)$ is given by

$$\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\beta}, \sigma) = K_1 (\alpha+1)^{3/2} \begin{bmatrix} n\boldsymbol{A}_{11} & n\boldsymbol{A}_{12} & 0 \\ n\boldsymbol{A}_{21} & n\boldsymbol{A}_{22} & 0 \\ 0 & 0 & \frac{\alpha+1}{2} \end{bmatrix},$$

with

$$\begin{aligned} \boldsymbol{A}_{11} &= \left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1} + \left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2\boldsymbol{D}^{-1}\mathbb{X}_2^T\mathbb{X}_1\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}, \\ \boldsymbol{A}_{12} &= -\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2\boldsymbol{D}^{-1}, \\ \boldsymbol{A}_{21} &= -\boldsymbol{D}^{-1}\mathbb{X}_2^T\mathbb{X}_1\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}, \\ \boldsymbol{A}_{22} &= \boldsymbol{D}^{-1}, \end{aligned}$$

being

$$\boldsymbol{D} = \mathbb{X}_2^T\mathbb{X}_2 - \mathbb{X}_2^T\mathbb{X}_1\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2.$$

Therefore, we have that the matrix $\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\beta}, \sigma)$ can be computed as

$$\begin{aligned} \boldsymbol{\Psi}_n^{-1}(\boldsymbol{\beta}, \sigma)\boldsymbol{M}(\boldsymbol{\beta}, \sigma) &= K_1^{-1}(\alpha+1)^{3/2} \begin{bmatrix} n\boldsymbol{A}_{11} & n\boldsymbol{A}_{12} & 0 \\ n\boldsymbol{A}_{21} & n\boldsymbol{A}_{22} & 0 \\ 0 & 0 & \frac{\alpha+1}{2} \end{bmatrix} \begin{pmatrix} \mathbf{0}_{(p-r)\times r} \\ \boldsymbol{I}_{r\times r} \\ \mathbf{0}_r \end{pmatrix} \\ &= K_1^{-1}(\alpha+1)^{3/2} n \begin{bmatrix} -\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2\boldsymbol{D}^{-1} \\ \boldsymbol{D}^{-1} \\ 0 \end{bmatrix}. \end{aligned}$$

On the other hand,

$$\left(\boldsymbol{M}(\boldsymbol{\beta}, \sigma)^T\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\beta}, \sigma)\boldsymbol{M}(\boldsymbol{\beta}, \sigma)\right)^{-1} = \frac{K_1(\alpha+1)^{-3/2}}{n}\boldsymbol{D},$$

and

$$\boldsymbol{M}(\boldsymbol{\beta},\sigma)^T\,\boldsymbol{\Psi}_n^{-1}\,(\boldsymbol{\beta},\sigma)\,\boldsymbol{\Omega}_n\,(\boldsymbol{\beta},\sigma)=(\alpha+1)^{3/2}\,\frac{K_1\sigma^2}{(2\alpha+1)^{3/2}}\,(\mathbf{0},\boldsymbol{I}_{\boldsymbol{r}\times\boldsymbol{r}},\mathbf{0}),$$

and so, multiplying the above expressions we obtain that

$$\boldsymbol{Q}_\alpha(\boldsymbol{\beta},\sigma)=\boldsymbol{\Psi}_n^{-1}\,(\boldsymbol{\beta},\sigma)\,\boldsymbol{M}(\boldsymbol{\beta},\sigma)\left[\boldsymbol{M}(\boldsymbol{\beta},\sigma)^T\,\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\beta},\sigma)\boldsymbol{M}(\boldsymbol{\beta},\sigma)\right]^{-1}=\begin{bmatrix}-\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2\\ \boldsymbol{I}_{r\times r}\\ 0\end{bmatrix},$$

and

$$\boldsymbol{Q}_\alpha(\boldsymbol{\beta},\sigma)\boldsymbol{M}(\boldsymbol{\beta},\sigma)^T\,\boldsymbol{\Psi}_n\,(\boldsymbol{\beta},\sigma)^{-1}\,\boldsymbol{\Omega}_n\,(\boldsymbol{\beta},\sigma)=(\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}}\begin{pmatrix}\mathbf{0}&\left(\mathbb{X}_1^T\mathbb{X}_1\right)^{-1}\mathbb{X}_1^T\mathbb{X}_2&\mathbf{0}\\ \mathbf{0}&\boldsymbol{I}_{r\times r}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}\end{pmatrix}.$$

Consequently, in this case we can compute the $r-$first eigenvalues as

$$\lambda_1(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})=\cdots=\lambda_r(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})=(\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}},$$

and hence,

$$\sum_{i=1}^{r}\lambda_i(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})Z_i^2=-\,(\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}}\chi_r^2.$$

On the other hand, we have

$$\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)\,\boldsymbol{\Psi}_n^{-1}\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)=K_1\sigma_{\boldsymbol{g},\alpha}^2\frac{(\alpha+1)^{3/2}}{(2\alpha+1)^{3/2}}\begin{bmatrix}\boldsymbol{I}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\boldsymbol{I}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\frac{(3\alpha^2+4\alpha+2)}{2(2\alpha+1)(\alpha+1)}\end{bmatrix},$$

and hence the trace of the above matrix is given by

$$trace\left(\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)\,\boldsymbol{\Psi}_n^{-1}\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)\right)\to\sigma_{\boldsymbol{g},\alpha}^2K_1\frac{(\alpha+1)^{3/2}}{(2\alpha+1)^{3/2}}\left((p+1)+\frac{\left(3\alpha^2+4\alpha+2\right)}{2\,(2\alpha+1)\,(\alpha+1)}\right),$$

and

$$trace\left(\boldsymbol{\Omega}_n^R\left(\widetilde{\boldsymbol{\beta}}_\alpha,\widetilde{\sigma}_\alpha\right)\right)\left(\boldsymbol{\Psi}_n^R\right)^{-1}\left(\widetilde{\boldsymbol{\beta}}_\alpha,\widetilde{\sigma}_\alpha\right)\to\sigma_{\boldsymbol{g},\alpha}^2K_1\frac{(\alpha+1)^{3/2}}{(2\alpha+1)^{3/2}}\left((p-r+1)+\frac{\left(3\alpha^2+4\alpha+2\right)}{2\,(2\alpha+1)\,(\alpha+1)}\right).$$

Therefore,

$$trace\left(\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)\,\boldsymbol{\Psi}_n^{-1}\left(\widehat{\boldsymbol{\beta}}_\alpha,\widehat{\sigma}_\alpha\right)\right)-trace\left(\boldsymbol{\Omega}_n^R\left(\widetilde{\boldsymbol{\beta}}_\alpha,\widetilde{\sigma}_\alpha\right)\right)\left(\boldsymbol{\Psi}_n^R\right)^{-1}\left(\widetilde{\boldsymbol{\beta}}_\alpha,\widetilde{\sigma}_\alpha\right)\to\sigma_{\boldsymbol{g},\alpha}^2K_1\frac{(\alpha+1)^{3/2}}{(2\alpha+1)^{3/2}}r.$$

Finally, the asymptotic probability of selecting the restricted model when this model is correct is

$$\Pr\left(2n\left(RP_{NH}(M_1^{(s)},\ldots,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha)-RP_{NH}(M_1^{(s)},\ldots,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha)\right)>0\right)\to$$

$$\Pr\left((-\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}}\chi_r^2+2\,(\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}}r>0\right)=$$

$$\Pr\left((\alpha+1)^{3/2}\,\frac{K_1\sigma_{\boldsymbol{g},\alpha}^2}{(2\alpha+1)^{3/2}}\left(2r-\chi_r^2\right)>0\right)=\Pr\left(\chi_r^2<2r\right).$$

5. Influence function analysis

In this section we obtain the Influence Function for RP_{NH} criterion. The results presented in this section are based on Toma et al. [15] and Kurata and Hamada [20]. The influence function (IF), introduced by Hampel in [32,33] is the main tool to discriminate whether an estimator or test statistic is robust. The IF describes the effect of an infinitesimal contamination of the model on the estimator or test statistic. The desired property associated to a robust estimator is the IF to be bounded. When this is the case, the estimator is said to be B-robust. In [26], the IF of the minimum RP functional

for the non-homogeneous case was obtained. We shall introduce some of the notation considered in that paper in order to clarify the new concepts introduced in this section.

We shall denote by G_i the true distribution function associated to the observation Y_i and by $\boldsymbol{T}_\alpha(G_1,\ldots,G_n)=\boldsymbol{T}_\alpha(\boldsymbol{G})$ the minimum RP functional with $\boldsymbol{G}=(G_1,\ldots,G_n)$. Let $\triangle_{t_i}$ be the distribution that takes the value t_i with probability one, and let us consider

$$G_{i,\varepsilon}=(1-\varepsilon)G_i+\varepsilon\triangle_{t_i},$$

for some small $\varepsilon>0$. and we also denote $\boldsymbol{G}_{\boldsymbol{t},\boldsymbol{\varepsilon}}=\left(G_{1,\varepsilon},\ldots,G_{n,\varepsilon}\right)$ with $\boldsymbol{t}=(t_1,\ldots,t_n)$. The IF of $\boldsymbol{T}_\alpha(\boldsymbol{G})$ in the i_0-th direction is defined by

$$IF\left(t_{i_0},\boldsymbol{T}_\alpha,G_1,\ldots,G_n\right)=\lim_{\varepsilon\to 0}\frac{\boldsymbol{T}_\alpha(G_1,\ldots,G_{i_0,\varepsilon},\ldots,G_n)-\boldsymbol{T}_\alpha(G_1,\ldots,G_n)}{\varepsilon}$$

and the influence function in all directions is given by

$$IF\left(\boldsymbol{t},\boldsymbol{T}_\alpha,G_1,\ldots,G_n\right)=\lim_{\varepsilon\to 0}\frac{\boldsymbol{T}_\alpha\left(\boldsymbol{G}_{\boldsymbol{t},\boldsymbol{\varepsilon}}\right)-\boldsymbol{T}_\alpha(\boldsymbol{G})}{\varepsilon},$$

Their expressions can be seen in Castilla et al. [26]. In the particular but important case that the true distribution belongs to the model, i.e.,

$$G_i(y)=F_i\left(y,\boldsymbol{\theta}\right)\ \text{ for } i=1,\ldots,n$$

we get,

$$IF\left(t_{i_0},\boldsymbol{T}_\alpha,F_1\left(.,\boldsymbol{\theta}\right),\ldots,F_n\left(.,\boldsymbol{\theta}\right)\right)=\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}\right)D_{i_0,\alpha}(\boldsymbol{\theta})$$

and

$$IF\left(t,\boldsymbol{T}_\alpha,F_1\left(.,\boldsymbol{\theta}\right),\ldots,F_n\left(.,\boldsymbol{\theta}\right)\right)=\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}\right)\sum_{i=1}^{n}D_{i,\alpha}(\boldsymbol{\theta})$$

with

$$D_{i,\alpha}(\boldsymbol{\theta})=-\frac{l_{i,\alpha}(\boldsymbol{\theta})}{\left(\int f_i(y,\boldsymbol{\theta})^{\alpha+1}dy\right)^2}$$

and

$$l_{i,\alpha}(\boldsymbol{\theta})=f_i\left(t_i,\boldsymbol{\theta}\right)\int f_i\left(y,\boldsymbol{\theta}\right)^{\alpha+1}\boldsymbol{u}_i\left(y,\boldsymbol{\theta}\right)dy-f_i\left(t_i,\boldsymbol{\theta}\right)^{\alpha+1}\boldsymbol{u}_i\left(y,\boldsymbol{\theta}\right)\int f_i\left(y,\boldsymbol{\theta}\right)^{\alpha+1}dy.$$

The best fitting parameter $\boldsymbol{\theta}_{g,\alpha}$ can be expressed as $\boldsymbol{T}_\alpha(\boldsymbol{G})$. Considering $RP_{NH}(M_1^{(s)},\ldots,M_1^{(s)},\widehat{\boldsymbol{\theta}}_\alpha^s)$, the term for correcting the bias, i.e. $\frac{1}{n}trace\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}\right)\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}\right)\right)$, is a function of the parameter and it is a continuous function by condition **C8**. Consequently, the behavior of the bias depends on $\boldsymbol{T}_\alpha$.

We can observe that other term in $RP_{NH}(M_1^{(s)},\ldots,M_1^{(s)},\widehat{\boldsymbol{\theta}}_\alpha^s)$, i.e. $H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)$ depends on $\boldsymbol{T}_\alpha$ and on the data at the same time. Therefore, the influence function of $H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)$ cannot be bounded, even if the estimator is robust. The functional form of $H_{n,\alpha}$, ${}^*R_\alpha(\boldsymbol{G})$, is in accordance with (8) given by

$${}^*R_0(\boldsymbol{G})=-\frac{1}{n}\sum_{i=1}^{n}\int\log f_i(y,\boldsymbol{T}_0(\boldsymbol{G}))dG_i(y)$$

for $\alpha=0$ and

$${}^*R_\alpha(\boldsymbol{G})=\frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{1}{\alpha\ L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G})\right)}\int f_i(y,\boldsymbol{T}_\alpha(\boldsymbol{G}))^\alpha dG_i(y)+\frac{1}{\alpha}\right\}$$

with

$$L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G})\right)=\left(\int f_i(y,\boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha+1}dy\right)^{\frac{\alpha}{\alpha+1}},$$

for $\alpha>0$. Then, the IF associated to the functional ${}^*R_\alpha(\boldsymbol{G})$ is defined by

$$IF\left(\boldsymbol{t},{}^*R_\alpha,G_1,\ldots,G_n\right)=\lim_{\varepsilon\to 0}\frac{{}^*R_\alpha(\boldsymbol{G}_{\boldsymbol{t},\boldsymbol{\varepsilon}})-{}^*R_\alpha(\boldsymbol{G})}{\varepsilon}$$

Therefore, for $\alpha = 0$, we have

$$IF\left(\boldsymbol{t}, {}^{*}R_0, G_1, \ldots, G_n\right) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int \boldsymbol{u}_i\left(y, \boldsymbol{T}_0(\boldsymbol{G})\right)^T dG_i(y) IF(\boldsymbol{t}, \boldsymbol{T}_0, G_1, \ldots, G_n) - \int \log \boldsymbol{f}_i\left(y, \boldsymbol{T}_0(\boldsymbol{G})\right) dG_i(y) + \log \boldsymbol{f}_i\left(\mathbf{t}_i, \boldsymbol{T}_0(\boldsymbol{G})\right)\right\},$$

with

$$u_i\left(y, \boldsymbol{T}_0(\boldsymbol{G})\right) = \frac{\partial \log f_i(y, \boldsymbol{T}_0(\boldsymbol{G}))}{\partial \boldsymbol{T}_0(\boldsymbol{G})}.$$

In the particular but important case that the true distribution belongs to the model, i.e.,

$$G_i(y) = F_i\left(y, \boldsymbol{\theta}\right) \text{ for } i = 1, \ldots, n,$$

we have,

$$\begin{aligned} IF\left(\boldsymbol{t}, {}^{*}R_0, F_{1,\boldsymbol{\theta}}, \ldots, F_{n,\boldsymbol{\theta}}\right) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\int \boldsymbol{u}_i\left(y, \boldsymbol{\theta}\right)^T \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right) dy IF\left(t_{i_0}, \boldsymbol{T}_\alpha, F_1\left(., \boldsymbol{\theta}\right), \ldots, F_n\left(., \boldsymbol{\theta}\right)\right)\right. \\ &\quad \left. - \int \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right) \log \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right) dy - \log \boldsymbol{f}_i\left(\mathbf{t}_i, \boldsymbol{\theta}\right)\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\int \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right) \log \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right) dy - \log \boldsymbol{f}_i\left(\mathbf{t}_i, \boldsymbol{\theta}\right)\right\}. \end{aligned}$$

In order to get $IF\left(\boldsymbol{t}, {}^{*}R_\alpha, G_1, \ldots, G_n\right)$, for $\alpha > 0$, we need to calculate

$$\left(\frac{\partial L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G}_\varepsilon)\right)}{\partial \varepsilon}\right)_{\varepsilon=0}.$$

We have,

$$\begin{aligned} \left(\frac{\partial L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G}_\varepsilon)\right)}{\partial \varepsilon}\right)_{\varepsilon=0} &= \alpha\left(\int f_i(y, \boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha+1} dy\right)^{-\frac{1}{\alpha+1}} \\ &\quad \times \int f_i(y, \boldsymbol{T}_\alpha(\boldsymbol{G}))^{\alpha+1} \boldsymbol{u}_i\left(y, \boldsymbol{T}_\alpha(\boldsymbol{G})\right)^T dy IF(\boldsymbol{t}, \boldsymbol{T}_\alpha, G_1, \ldots, G_n). \end{aligned}$$

Therefore we have,

$$\begin{aligned} \left(\frac{\partial {}^{*}R_\alpha(\boldsymbol{G}_\varepsilon)}{\partial \varepsilon}\right)_{\varepsilon=0} &= \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{1}{\alpha L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G})\right)}\left(\int \alpha f_i(y, \boldsymbol{T}_\alpha(\boldsymbol{G}))^\alpha \boldsymbol{u}_i\left(y, \boldsymbol{T}_\alpha(\boldsymbol{G})\right)^T dG_i(y)\right.\right. \\ &\quad \times IF(\boldsymbol{t}, \boldsymbol{T}_\alpha, G_1, \ldots, G_n) \\ &\quad \left. - \int f_i(y, \boldsymbol{T}_\alpha(\boldsymbol{G}))^\alpha dG_i(y) + \boldsymbol{f}_i\left(\mathbf{t}_i, \boldsymbol{\theta}\right)^\alpha\right) \\ &\quad \left. - \frac{1}{L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G})\right)^2}\left(\frac{\partial L_\alpha^i\left(\boldsymbol{T}_\alpha(\boldsymbol{G}_\varepsilon)\right)}{\partial \varepsilon}\right)_{\varepsilon=0} \int f_i(y, \boldsymbol{T}_\alpha(\boldsymbol{G}))^\alpha dG_i(y)\right\}. \end{aligned}$$

In the particular case that the true distribution belong to the model, i.e.,

$$G_i(y) = F_i\left(y, \boldsymbol{\theta}\right) \text{ for } i = 1, \ldots, n,$$

we have,

$$IF\left(\boldsymbol{t}, {}^{*}R_\alpha, F_{1,\boldsymbol{\theta}}, \ldots, F_{n,\boldsymbol{\theta}}\right) = \frac{1}{\alpha L_\alpha^i\left(\boldsymbol{\theta}\right) n}\sum_{i=1}^{n}\left\{\int \boldsymbol{f}_i\left(y, \boldsymbol{\theta}\right)^{\alpha+1} - \boldsymbol{f}_i\left(\mathbf{t}_i, \boldsymbol{\theta}\right)^\alpha\right\}.$$

It is immediate to see that

$$\lim_{\alpha \to 0} IF\left(\boldsymbol{t}, {}^{*}R_\alpha, F_{1,\boldsymbol{\theta}}, \ldots, F_{n,\boldsymbol{\theta}}\right) = IF\left(\boldsymbol{t}, {}^{*}R_0, F_{1,\boldsymbol{\theta}}, \ldots, F_{n,\boldsymbol{\theta}}\right).$$

Finally, note that in the case of independent and identically distributed random variables we have,

$$IF\left(\boldsymbol{t}, {}^{*}R_0, F_{\boldsymbol{\theta}}\right) = \int \boldsymbol{f}\left(y, \boldsymbol{\theta}\right) \log \boldsymbol{f}\left(y, \boldsymbol{\theta}\right) dy - \log \boldsymbol{f}\left(\mathbf{t}_i, \boldsymbol{\theta}\right)$$

Table 1
Results for uncontaminated data.

p	0	1	2	3	4	5
AIC	0	0	0	0	840	160
BIC	0	0	0	0	975	25
AIC_c	0	0	0	0	881	119
$RPNH_{0.01}$	0	0	0	0	822	178
$RPNH_{0.02}$	0	0	0	0	822	178
$RPNH_{0.04}$	0	0	0	0	826	174
$RPNH_{0.1}$	0	0	0	0	826	174
$RPNH_{0.2}$	0	0	0	0	827	173
$RPNH_{0.4}$	0	0	0	0	823	177
$RPNH_{0.5}$	0	0	0	0	824	176
$RPNH_{0.7}$	0	0	0	0	819	181
$RPNH_{1.0}$	0	0	0	0	828	172

for $\alpha = 0$ and

$$IF\left(t, {}^{*}R_{\alpha}, F_{\boldsymbol{\theta}}\right) = \frac{1}{\alpha L_{\alpha}\left(\boldsymbol{\theta}\right)} \left\{ \int \boldsymbol{f}\left(y, \boldsymbol{\theta}\right)^{\alpha+1} - \boldsymbol{f}\left(t, \boldsymbol{\theta}\right)^{\alpha} \right\}$$

for $\alpha > 0$.

6. Simulation study

To evaluate the performance of the RP_{NH}-criterion introduced in this paper, we consider the situation of a polynomial regression model. We take the model

$$Y_i = X_i + 2X_i^2 - X_i^3 + X_i^4 + \epsilon_i, i = 1, \ldots, n,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$ and the variables X_i are fixed in a way such that the interval $[-2,2]$ is divided in $n + 1$ intervals of the same length. Next, we take $n = 100$, so that

$$X_i = -2 + \frac{4}{102}(i + 1), i = 1, \ldots, 100.$$

We consider several theoretical models aiming to fit this data. These models are given by the degree of the polynomial defining the model. Note that the regression coefficients adopt the same expression as in MLRM, just taking X^i as X_i, and thus we can use the formulas developed in the previous sections. In our case, we have considered six different models, varying from constants (degree 0) to polynomials of degree 5. Thus defined, each model is characterized by the degree, denoted by p.

We take 1000 different sample data $(Y^s, X^s), s = 1, \ldots, 1000$ and for each sample, we select the best fitting model according to several criteria. We have considered AIC, BIC, AIC_c and the RP_{NH}-criterion for different values of the tuning parameter, namely $\alpha = 0.01, 0.02, 0.04, 0.1, 0.2, 0.4, 0.5, 0.7$ and 1.

In Table 1 we have written the number of times that each model is selected for each model selection criterion. From these results, it can be seen that BIC seems to be the best fitting selection criterion, the other model selection criteria having a similar performance.

As it was explained throughout the paper, we expect RP_{NH} to be a robust selection criterion. To check this hypothesis, we have considered a situation of contamination. Thus, we consider the previous model but we introduce contamination in some of the data. More concretely, we define

$$\epsilon_i \sim \mathcal{U}(-r, r + 30),$$

for some of the data chosen at random. Here, r is a constant measuring the strength of contamination, in the sense that the bigger r, the strongest the contamination. We have considered three values $r = 1, 5, 10$. We have also chosen an asymmetric contamination model, so that no compensations due to symmetry might appear. The choice of 30 is given by approximating the range of values of the original model in the interval $[-2, 2]$. Moreover, we have varied the proportion of data affected by contamination. In this study, we have chosen the proportion of contamination as 0.05, 0.10, 0.20, 0.30.

Again, we have obtained the best fitting model according different model selection criteria, and we have conducted this experiment 1000 times. The number of times that each model is selected for each combination of contamination and strength of contamination r is given in Tables 2–5. The left part of each table corresponds to $r = 1$, the center part for $r = 5$ and the right part for $r = 10$.

Finally, in Tables 6–8 we show the quadratic error between the estimated parameters and the real values of the parameters for each method and each degree of contamination. This error is obtained by

$$\frac{\sum_{i=0}^{5}(\hat{\beta}_i - \beta_i)^2}{6},$$

Table 2
Results when a 5% of data selected randomly are contaminated.

p	$r=1$						$r=5$						$r=10$					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	4	21	651	324	0	0	8	31	648	313	0	0	19	50	612	319
BIC	0	0	25	78	784	113	0	0	42	97	748	113	0	0	74	122	689	115
AIC_c	0	0	6	28	681	285	0	0	9	39	661	291	0	0	24	61	634	281
$RPNH_{0.01}$	0	0	4	17	836	143	0	0	5	17	829	149	0	0	13	28	799	160
$RPNH_{0.02}$	0	0	2	29	851	118	0	0	6	62	823	109	0	0	9	70	798	123
$RPNH_{0.04}$	0	0	2	25	842	131	0	0	2	18	832	148	0	0	1	14	834	151
$RPNH_{0.1}$	0	0	0	0	825	175	0	0	0	0	827	173	0	0	0	0	824	176
$RPNH_{0.2}$	0	0	0	0	826	174	0	0	0	0	826	174	0	0	0	0	826	174
$RPNH_{0.4}$	0	0	0	0	822	178	0	0	0	0	820	180	0	0	0	0	817	183
$RPNH_{0.5}$	0	0	0	0	827	173	0	0	0	0	828	172	0	0	0	0	816	184
$RPNH_{0.7}$	0	0	0	0	825	175	0	0	0	0	820	180	0	0	0	0	818	182
$RPNH_{1.0}$	0	0	0	0	826	174	0	0	0	0	823	177	0	0	0	0	820	180

Table 3
Results when a 10% of data selected randomly are contaminated.

p	$r=1$						$r=5$						$r=10$					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	32	80	575	313	0	0	46	94	546	314	0	0	68	116	513	303
BIC	0	0	115	165	608	112	0	0	149	176	562	113	0	0	207	192	491	110
AIC_c	0	0	38	88	587	287	0	0	56	100	565	279	0	0	82	128	529	261
$RPNH_{0.01}$	0	0	28	47	750	165	0	0	32	54	724	190	0	0	40	64	685	211
$RPNH_{0.02}$	0	0	26	37	777	160	0	0	25	54	762	159	0	0	31	69	731	169
$RPNH_{0.04}$	0	0	20	78	784	118	0	0	27	115	737	121	0	0	33	109	740	118
$RPNH_{0.1}$	0	0	0	2	840	158	0	0	0	0	835	165	0	0	0	0	852	148
$RPNH_{0.2}$	0	0	0	0	817	183	0	0	0	0	822	178	0	0	0	0	832	168
$RPNH_{0.4}$	0	0	0	0	825	175	0	0	0	0	835	165	0	0	0	0	835	165
$RPNH_{0.5}$	0	0	0	0	833	167	0	0	0	0	835	164	0	0	0	0	834	166
$RPNH_{0.7}$	0	0	0	0	837	163	0	0	0	0	831	169	0	0	0	0	834	166
$RPNH_{1.0}$	0	0	0	0	846	154	0	0	0	0	828	172	0	0	0	0	835	165

Table 4
Results when a 20% of data selected randomly are contaminated.

p	$r=5$						$r=10$						$r=20$					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	74	159	457	310	0	0	94	171	428	307	0	0	121	194	391	294
BIC	0	0	248	236	403	113	0	0	290	239	369	102	0	0	350	243	313	94
AIC_c	0	0	84	171	474	271	0	0	107	185	435	273	0	0	144	207	386	283
$RPNH_{0.01}$	0	0	46	94	567	293	0	0	49	96	527	328	0	0	53	96	494	357
$RPNH_{0.02}$	0	0	43	86	587	284	0	0	44	93	566	297	0	0	49	91	546	314
$RPNH_{0.04}$	0	0	33	76	653	238	0	0	35	93	637	235	0	0	46	96	636	222
$RPNH_{0.1}$	0	0	44	211	622	123	0	0	22	118	711	149	0	0	16	78	759	147
$RPNH_{0.2}$	0	0	0	0	838	162	0	0	0	0	835	165	0	0	0	0	834	166
$RPNH_{0.4}$	0	0	0	0	825	175	0	0	0	0	834	166	0	0	0	0	824	176
$RPNH_{0.5}$	0	0	0	0	820	180	0	0	0	0	833	167	0	0	0	0	820	180
$RPNH_{0.7}$	0	0	0	0	823	177	0	0	0	0	818	182	0	0	0	0	816	184
$RPNH_{1.0}$	0	0	0	0	823	177	0	0	0	0	812	188	0	0	0	0	826	174

where $\hat{\beta}_1$ is the estimated parameter for the coefficient of degree i for the selected model and β_i is the corresponding real value, i.e. $(0, 1, 2, -1, 1, 0)$.

From the results in these tables, it can be seen that the performance of AIC, BIC and AIC_c dramatically decreases, in the sense that the proportion of times obtaining the true degree $p = 4$ decreases if contamination is present. As expected, the bigger the rate of contaminated data, the poorer the performance. Note however that they are not very affected for different values of r.

On the other hand, the results are quite similar to the uncontaminated case for RP_{NH} and big values of the tuning parameter. This was the expected result and it follows the same behavior as other situations where RP has been considered. The best behavior appears for $\alpha = 0.4$ and $\alpha = 0.5$, where the efficiency is good and the performance in terms of robustness is very good.

Finally, it can be seen that the quadratic errors follow the same behavior. Hence, the estimations are much better in the case of contamination when the RP_{NH} criterion is applied. As expected, the quadratic errors increase as the degree of contamination increases. Note however that for the RP_{NH} criterion, the corresponding errors for different degrees of

Table 5
Results when a 30% of data selected randomly are contaminated.

p	$r=1$						$r=5$						$r=10$					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	117	162	424	297	0	0	152	170	396	282	0	0	187	192	350	271
BIC	0	0	354	232	317	97	0	0	411	231	272	86	0	0	480	223	222	75
AIC_c	0	0	139	182	423	256	0	0	170	192	401	237	0	0	218	201	345	236
$RPNH_{0.01}$	0	0	57	83	493	367	0	0	54	94	468	384	0	0	55	92	409	444
$RPNH_{0.02}$	0	0	53	80	510	357	0	0	54	84	497	365	0	0	56	87	458	399
$RPNH_{0.04}$	0	0	47	80	553	320	0	0	49	85	537	329	0	0	53	79	562	312
$RPNH_{0.1}$	0	0	76	200	527	197	0	0	88	204	544	164	0	0	58	145	633	164
$RPNH_{0.2}$	0	0	11	48	779	162	0	0	7	15	817	161	0	0	2	11	824	163
$RPNH_{0.4}$	0	0	0	0	826	174	0	0	0	0	806	194	0	0	0	0	823	177
$RPNH_{0.5}$	0	0	0	0	827	173	0	0	0	0	797	203	0	0	0	0	818	182
$RPNH_{0.7}$	0	0	0	0	833	167	0	0	0	0	809	191	0	0	0	0	820	180
$RPNH_{1.0}$	0	0	0	0	828	172	0	0	0	1	810	189	0	0	0	0	827	173

Table 6
Quadratic error between the vector of estimated parameters and the realvector of parameters for different methods and degrees of contamination for $r=1$.

Method	0%	5%	10%	20%	30%
AIC	0.3232	6.9323	14.4651	29.8924	50.0073
BIC	0.2296	6.2863	13.4257	27.5233	43.2754
AIC_c	0.3039	6.7781	14.3572	29.5171	48.9798
$RPNH_{0.01}$	0.3333	4.4628	10.6302	23.1182	35.0075
$RPNH_{0.02}$	0.3336	3.1688	9.1788	21.4306	33.7234
$RPNH_{0.04}$	0.3320	1.1511	6.0465	17.8433	30.7373
$RPNH_{0.1}$	0.3340	0.3604	0.4596	6.2461	21.1560
$RPNH_{0.2}$	0.3416	0.3623	0.4126	0.4572	1.8985
$RPNH_{0.4}$	0.3733	0.3985	0.4330	0.4849	0.5891
$RPNH_{0.5}$	0.3949	0.4211	0.4496	0.5115	0.6042
$RPNH_{0.7}$	0.4548	0.4862	0.5013	0.5869	0.6867
$RPNH_{1.0}$	0.5554	0.6059	0.6305	0.7512	0.8627

Table 7
Quadratic error between the vector of estimated parameters and the realvector of parameters for different methods and degrees of contamination for $r=5$.

Method	0%	5%	10%	20%	30%
AIC	0.3232	7.9160	16.3224	33.6331	54.9498
BIC	0.2296	7.4007	15.0274	29.7602	45.6349
AIC_c	0.3039	7.8843	16.0214	33.1256	53.7613
$RPNH_{0.01}$	0.3333	4.8847	11.5048	24.5893	36.2259
$RPNH_{0.02}$	0.3336	3.3245	9.4987	22.1142	34.1580
$RPNH_{0.04}$	0.3320	0.9610	5.6644	17.2625	30.0149
$RPNH_{0.1}$	0.3340	0.3683	0.4346	3.6863	16.6361
$RPNH_{0.2}$	0.3416	0.3698	0.4062	0.5003	1.1497
$RPNH_{0.4}$	0.3733	0.3996	0.4278	0.5002	0.6641
$RPNH_{0.5}$	0.3949	0.4181	0.4525	0.5252	0.6865
$RPNH_{0.7}$	0.4548	0.4867	0.5124	0.6106	0.7693
$RPNH_{1.0}$	0.5554	0.6046	0.6520	0.7844	0.9405

contamination are more stable, and indeed, they are almost the same when the tuning parameter is greater than 0.2. Note also that the results are very similar for different values of r.

In order to test if this is the usual behavior of these methods, we have repeated this study for different values of the polynomial regression, each coefficient varying in $\{-2,-1,0,1,2\}$. This leads to 3125 different models for each value of $r=1,5,10$, so that we have 9375 different situations. And for all of them we can extract the same conclusions.

6.1. Choice of optimal tuning parameter

A practical concern for the implementation of the RP_{NH} criterion is determining the optimal tuning parameter. Because the best trade-off between robustness and efficiency would depend on the amount of contamination in the data (which is unknown in practice), selecting a suitable value of α is a challenging issue in real-data applications. Several studies on inferential methods based on divergences suggest moderate values of the tuning parameter for a suitable compromise between loss in efficiency and gain in robustness. Indeed, from our results, moderate values around $\alpha=0.4$ and $\alpha=0.5$ have shown a competitive performance with the classical AIC, BIC and AIC_c methods under uncontaminated data, but with

Table 8
Quadratic error between the vector of estimated parameters and the realvector of parameters for different methods and degrees of contamination for $r = 10$.

Method	0%	5%	10%	20%	30%
AIC	0.3232	9.7049	18.9952	39.1522	62.6121
BIC	0.2296	9.0393	17.4135	33.0922	49.2856
AIC_c	0.3039	9.5909	18.6002	38.5421	61.3608
$RPNH_{0.01}$	0.3333	5.7905	12.8467	26.2409	38.0303
$RPNH_{0.02}$	0.3336	3.5522	10.0976	22.8845	34.9776
$RPNH_{0.04}$	0.3320	0.8402	5.1664	16.4898	29.1456
$RPNH_{0.1}$	0.3340	0.3769	0.4206	2.9669	13.6491
$RPNH_{0.2}$	0.3416	0.3698	0.4035	0.5040	1.0845
$RPNH_{0.4}$	0.3733	0.4043	0.4298	0.5134	0.6376
$RPNH_{0.5}$	0.3949	0.4296	0.4514	0.5430	0.6703
$RPNH_{0.7}$	0.4548	0.4953	0.5121	0.6168	0.7535
$RPNH_{1.0}$	0.5554	0.6106	0.6388	0.7671	0.9396

Table 9
The Hald cement data.

X_1	X_2	X_3	X_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

a clear gain in robustness highly desirable in contaminated scenarios. Moreover, values exceeding $\alpha = 1$ are discouraged because of their remarkable loss of efficiency. Thus, the choice of the optimal α typically falls within the interval [0, 1].

Some works has been made for developing a data-driven criteria for selecting the best value of the tuning parameter based on the specific data under study. For example, Warwick and Jones [34] proposed selecting the optimal α that minimizes the estimated mean squared error, an approach that has to deal with the drawback of requiring a pilot estimation of the model parameters. Hence, this approach was somewhat pilot-dependent in some statistical models and so Basak et al. [35] improved upon this method by introducing iterative updates to the pilot estimator, reducing the dependency of the criterion on the initial pilot choice. Their proposed optimal choice of the tuning parameter could be adopted for selecting the best RP_{NH} model selection criterion. However, from our results, moderate values of α yield comparable performance under the different contamination scenarios, making the implementation of an optimality algorithm less critical in this model. As a result, any moderate value around $\alpha = 0.4$ would be appropriate for robust model selection.

7. Real data example

In this section we analyze two sets of real data at the light of this new model selection tool based on RP. We have considered two examples and look for a MLRM assuming that the explanatory variables are fixed, so that the dependent variables are not identically distributed and we can apply the results in the previous sections.

7.1. Hald cement data

We consider the problem proposed in [36] and later studied in [15]. The dependent variable Y measures the heat evolved in calories per gram as a function of four ingredients: tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium alumino-ferrite (X_3) and dicalcium silicate (X_4). The data are given in Table 9. It is assumed that Y can be written in terms of X_1, X_2, X_3, X_4 as a MLRM. We have considered the RP_{NH} procedure to select the best model for different values of the tuning parameter.

Considering different subsets of independent variables, we obtain 15 different multiple linear models and the goal is to select the best one. However, it is known that at least two independent variables are needed because cement needs a combination of at least two reactants. Hence, we can remove the four simple linear regression models and we finally consider 11 possible models.

Table 10
Results for the Hald cement data.

	$RPNH_{0.01}$	$RPNH_{0.02}$	$RPNH_{0.04}$	$RPNH_{0.05}$	$RPNH_{0.07}$
X_1, X_2	2.4179	2.3655	2.2642	2.2170	2.1280
X_1, X_3	3.8824	4.3871	4.1321	4.0138	3.7933
X_1, X_4	2.5408	2.4827	2.3738	2.3226	2.2261
X_2, X_3	3.3638	3.2836	3.1140	3.0349	2.8868
X_2, X_4	3.7164	3.8865	3.6579	3.5523	3.3565
X_3, X_4	2.9519	2.8785	2.7413	2.6771	2.5564
X_1, X_2, X_3	2.4024	2.3493	2.2495	2.2026	2.1141
X_1, X_2, X_4	2.4013	2.3484	2.2490	2.2023	2.1142
X_1, X_3, X_4	2.4295	2.3759	2.2752	2.2279	2.1386
X_2, X_3, X_4	2.6091	2.5490	2.4363	2.3834	2.2837
X_1, X_2, X_3, X_4	2.4747	2.4197	2.3164	2.2679	2.1765
Best model	(X_1, X_2, X_4)	(X_1, X_2, X_4)	(X_1, X_2, X_4)	(X_1, X_2, X_4)	(X_1, X_2, X_3)
	$RPNH_{0.1}$	$RPNH_{0.2}$	$RPNH_{0.4}$	$RPNH_{0.5}$	$RPNH_{0.7}$
X_1, X_2	2.0064	1.6822	1.2656	1.1249	0.9197
X_1, X_3	3.4984	2.7476	1.8439	1.5662	1.2019
X_1, X_4	2.0946	1.7454	1.3004	1.1517	0.9411
X_2, X_3	2.6873	2.1710	1.5474	1.3482	1.0702
X_2, X_4	3.0967	2.4472	1.7055	1.4771	1.1616
X_3, X_4	2.3930	1.9647	1.4322	1.2638	1.0347
X_1, X_2, X_3	1.9933	1.6716	1.2598	1.1264	0.9173
X_1, X_2, X_4	1.9939	1.6735	1.2623	1.1240	0.9228
X_1, X_3, X_4	2.0167	1.6921	1.2756	1.1352	0.9985
X_2, X_3, X_4	2.1482	1.7891	1.3340	1.1824	1.0089
X_1, X_2, X_3, X_4	2.0521	1.7221	1.3014	1.1601	0.9548
Best model	(X_1, X_2, X_3)	(X_1, X_2, X_3)	(X_1, X_2, X_3)	(X_1, X_2, X_4)	(X_1, X_2, X_3)

Table 11
Sum of error predictions and selected model for several values of α.

α	Sum of prediction errors	Parameters of the best model
0.01	47.972729	$71.648307 + 1.451938X_1 + 0.416110X_2 - 0.236540X_4$
0.02	47.974591	$71.714100 + 1.450493X_1 + 0.415189X_2 - 0.237022X_4$
0.04	47.980177	$71.785102 + 1.449055X_1 + 0.414200X_2 - 0.237562X_4$
0.05	47.984374	$71.822679 + 1.448337X_1 + 0.413678X_2 - 0.237854X_4$
0.07	48.146040	$48.204309 + 1.694829X_1 + 0.655214X_2 - 0.255725X_3$
0.10	48.185386	$48.207083 + 1.694688X_1 + 0.654400X_2 - 0.258594X_3$
0.20	48.463301	$48.207252 + 1.695774X_1 + 0.651180X_2 - 0.270680X_3$
0.40	51.506012	$48.132841 + 1.712755X_1 + 0.637983X_2 - 0.326003X_3$
0.50	50.229813	$76.333477 + 1.406477X_1 + 0.353684X_2 - 0.280523X_4$
0.70	64.749860	$48.377648 + 1.758178X_1 + 0.605115X_2 + 0.415870X_3$

We have applied the RP_{NH}-criterion defined in (28) for different values of the tuning parameter to select the most appropriate model. The solution is given in Table 10. As it can be seen in this table, the combinations of X_1, X_2, X_3 and X_1, X_2, X_4 seem to be the best candidates, with tiny differences between them. These results are similar to the conclusions obtained in [15]. Remark also the good performance of model X_1, X_2.

The corresponding sum of prediction errors and the values of the estimations for the best model for each value of α are given in Table 11. In this case, the possible outliers do not affect the predictions, so that the best values for α are the smaller ones.

7.2. Studying body fat

For this example, we study the body fat percentage in terms of various (more concretely 14) body measurements such as height, weight, age, chest circumference, abdomen circumference, hip circumference and thigh circumference. We look for a multiple regression model and we consider the data corresponding to 252 adult men. This data can be found in [37] an studied e.g. in [38]. It is assumed that Y can be written in terms of $X_1, \ldots, X_{14}$ as a MLRM and we have not used any transformation of the data. Exploring the data, it seems that there are two outliers.

Considering different subsets of independent variables, we obtain 2^{14} different multiple linear models and the goal is to select the best one. We have considered the RP_{NH} procedure to select the best model for different values of the tuning parameter. The results are shown in Table 12.

In these models, X_1 corresponds to variable "Density determined from underwater weighing", X_2 corresponds to "Age", X_4 corresponds to "Height" and X_7 to "Abdomen circumference". Note that we obtain almost the same result for any value of α and that the corresponding coefficients are almost the same. These similarities come from the fact that there are not

Table 12
Explaining body fat for different values of α.

α	Best model	Sum of errors
0	$445.30998 - 408.76351X_1 + 0.01197X_2 + 0.05168X_7$	1.2573046
0.01	$481.81620 - 437.85347X_1 - 0.00760X_4$	1.3061157
0.02	$482.64988 - 439.16676X_1$	1.3068753
0.04	$482.64988 - 439.17177X_1$	1.3071916
0.05	$482.64988 - 439.17338X_1$	1.3072978
0.07	$482.64988 - 439.17629X_1$	1.3074954
0.1	$482.64988 - 439.18037X_1$	1.3077842
0.2	$482.64988 - 439.19357X_1$	1.3088156
0.4	$482.64988 - 439.21993X_1$	1.3113155
0.5	$482.64988 - 439.22947X_1$	1.3123650
0.7	$482.64988 - 439.24154X_1$	1.3138016

many outliers in the data. Remark however that RP_{NH} proposes a model simpler than ML, where the outliers lead to include two new variables.

8. Conclusions

In this paper we have developed a new procedure for model selection for independent but not identically distributed observations aiming to compete with other methods based on maximum likelihood in terms of efficiency but being more robust against outlying data. For this purpose, we have considered RP, a tool that has proved itself to provide robust estimations in many statistical problems. We have developed a model selection criterion, the RP_{NH}-criterion, extending the well-known AIC. Besides, we have shown that the sample estimator is an unbiased estimator. Next, we have considered the case of having a restricted model and we have developed a procedure to decide whether the large model is more appropriate for modeling the available data. As an example of application, we have developed the MLRM when we aim to find the best model fitting a set of data. We have conducted a simulation study that shows that this new procedure works very well under contamination, i.e. simulations suggest that the procedure is robust. Besides, it seems that the cost in terms of efficiency is reduced. Finally, we have applied this new procedure in two situations with real data.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Spanish Grant PID2021-124933NB-I00. We would like to thank the referees and editors for their helpful comments that have improved a lot the final version of the paper.

Appendix. Proofs of the results

Assumptions

We consider the following regularity conditions:

C1. The support, $\mathcal{X}$, of the density functions $f_i(y, \boldsymbol{\theta})$ is the same for all i and it does not depend on $\boldsymbol{\theta}$. Besides, the true probability density functions $g_1, \ldots, g_n$ have the same support $\mathcal{X}$.

C2. For almost all $y \in \mathcal{X}$ the density $f_i(y, \boldsymbol{\theta})$ admits all third derivatives with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $i = 1, \ldots, n$.

C3. For $i = 1, 2, \ldots, n$ the integrals

$$\int f_i(y, \boldsymbol{\theta})^{1+\alpha} dy$$

can be differentiated thrice with respect to $\boldsymbol{\theta}$ and we can interchange integration and differentiation. As a consequence of this condition, it follows that

$$\frac{\partial V_{i,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = E_{Y_i}\left[\frac{\partial \widehat{V}_{i,\alpha}(Y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \quad \frac{\partial^2 V_{i,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = E_{Y_i}\left[\frac{\partial^2 \widehat{V}_{i,\alpha}(Y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] = \boldsymbol{J}_\alpha^{(i)}(\boldsymbol{\theta}).$$

C4. For $i = 1, 2, \ldots, n$ the matrices $\boldsymbol{J}_\alpha^{(i)}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ are positive definite.

C5. There exist functions $M_{jkl}^{(i)}$ and constants m_{jkl} such that

$$\left|\frac{\partial^3 \widehat{V}_{i,\alpha}(y;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| \leq M_{jkl}^{(i)}(y), \qquad \forall\boldsymbol{\theta}\in\boldsymbol{\Theta},\ \forall j,k,l$$

and

$$E_Y\left[M_{jkl}^{(i)}(Y)\right] = m_{jkl} < \infty, \qquad \forall\boldsymbol{\theta}\in\boldsymbol{\Theta},\ \forall j,k,l.$$

C6. For all j,k,l and $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, the sequences $\left\{\frac{\partial\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j}\right\}_{j=1,\ldots,p}$, $\left\{\frac{\partial^2\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right\}_{j,k=1,\ldots,p}$ and $\left\{\frac{\partial^3\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial l}\right\}_{j,k,l=1,\ldots,p}$ are uniformly integrable in the Cesàro sense, i.e.

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n}E_{Y_i}\left[\left|\frac{\partial\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j}\right| I_{\left\{\frac{\partial V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j}>n\right\}}(Y_i)\right]\right)=0,$$

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n}E_{Y_i}\left[\left|\frac{\partial^2\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right| I_{\left\{\frac{\partial^2 V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}>n\right\}}(Y_i)\right]\right)=0,$$

$$\lim_{n\to\infty}\left(\sup_{n>1}\frac{1}{n}\sum_{i=1}^{n}E_{Y_i}\left[\left|\frac{\partial^3\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| I_{\left\{\frac{\partial^3 V_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}>n\right\}}(Y_i)\right]\right)=0.$$

C7. For all $\varepsilon>0$

$$\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}E_{Y_i}\left[\left\|\boldsymbol{\Omega}_n^{-\frac{1}{2}}(\boldsymbol{\theta})\frac{\partial\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\|_2^2 I_{\left\{\left\|\boldsymbol{\Omega}_n^{-\frac{1}{2}}(\boldsymbol{\theta})\frac{\partial\widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\|_2^2\right\}}(Y_i)\right] > \varepsilon\sqrt{n}\right\}=0.$$

C8. The matrices $\boldsymbol{\Psi}_n^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}_n(\boldsymbol{\theta})$ are continuous for arbitrary $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

Proof of Theorem 8. Consider a fixed $s=1,\ldots,l$. A Taylor expansion of $V_{i,\alpha}(\boldsymbol{\theta})$ defined in Eq. (6) around $\boldsymbol{\theta}_{\mathbf{g},\alpha}^s$ and evaluated at $\widehat{\boldsymbol{\theta}}_\alpha^s$ gives

$$\begin{aligned}V_{i,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right) &= V_{i,\alpha}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+\left(\frac{\partial V_{i,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{g},\alpha}^s}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\\ &\quad+\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\left(\frac{\partial^2 V_{i,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{g},\alpha}^s}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+o\left(\left\|\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right\|^2\right)\\ &= V_{i,\alpha}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+\left(\frac{\partial V_{i,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{g},\alpha}^s}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\\ &\quad+\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{J}_\tau^{(i)}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+o\left(\left\|\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right\|^2\right).\end{aligned}$$

Summing over i and dividing by n, taking into account that $\boldsymbol{\theta}_{\mathbf{g},\alpha}^s$ maximizes $H_\alpha(\boldsymbol{\theta})$, we get

$$H_\alpha(\widehat{\boldsymbol{\theta}}_\alpha^s)=H_\alpha\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+o\left(\left\|\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right\|^2\right)$$

and hence,

$$E_{Y_1,\ldots,Y_n}\left[nH_\alpha\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]=nH_\alpha\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+\frac{1}{2}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right]+o_p(1). \tag{34}$$

But by Eq. (18), and applying Corollary 2.1 in [39], we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\xrightarrow[n\longrightarrow\infty]{\mathcal{L}}\sum_{i=1}^{k}\lambda_i(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s)Z_i^2,$$

where $\lambda_1(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s),\ldots,\lambda_n(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s)$ are the eigenvalues of the matrix

$$\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^{-1}=\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^{-1}$$

and $Z_1, \ldots, Z_k$ are independent normal random variables with mean zero and variance 1. Therefore,

$$\begin{aligned} & E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right] \\ = & \sum_{i=1}^k \lambda_i(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s)+o_P(1) \\ = & \ trace\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^{-1}\right)+o_P(1). \end{aligned}$$

On the other hand, taking into account that $\widehat{\boldsymbol{\theta}}_\alpha^s$ maximizes $H_{n,\alpha}(\boldsymbol{\theta})$, a Taylor expansion of $H_{n,\alpha}(\boldsymbol{\theta})$ at $\widehat{\boldsymbol{\theta}}_\alpha^s$ and evaluated at $\boldsymbol{\theta}_{\mathbf{g},\alpha}^s$ gives

$$H_{n,\alpha}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)=H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)+\frac{1}{2}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)^T\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha^s}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)+o\left(\left\|\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right\|^2\right).$$

But then, multiplying by n and considering the expected values,

$$\begin{aligned} nH_\alpha\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right) = & E_{Y_1,\ldots,Y_n}\left[nH_{n,\alpha}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right]=E_{Y_1,\ldots,Y_n}\left[nH_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] \\ & +\frac{1}{2}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)^T\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha^s}\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+o_p(1). \end{aligned}$$

Besides,

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha^s}\xrightarrow[n\longrightarrow\infty]{\mathcal{P}}\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right). \tag{35}$$

by the continuity of $\boldsymbol{\Psi}_n$. Hence, substituting in (34)

$$\begin{aligned} & E_{Y_1,\ldots,Y_n}\left[nH_\alpha\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] \\ & =nH_\alpha\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)+\frac{1}{2}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right]+o_p(1) \\ & =E_{Y_1,\ldots,Y_n}\left[nH_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+\frac{1}{2}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right]+o_p(1) \\ & \quad +\frac{1}{2}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+o_p(1) \\ & =E_{Y_1,\ldots,Y_n}\left[nH_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s-\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right]+o_p(1), \end{aligned}$$

and thus,

$$\begin{aligned} E_{Y_1,\ldots,Y_n}\left[H_\alpha\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] = & E_{Y_1,\ldots,Y_n}\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] \\ & +\frac{1}{n}E_{Y_1,\ldots,Y_n}\left[\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\sqrt{n}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s-\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+o_p(1) \\ = & E_{Y_1,\ldots,Y_n}\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right]+\frac{1}{n}trace\left(\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\theta}_{\mathbf{g},\alpha}^s\right)\right). \end{aligned}$$

Hence, the result holds.

Proof of Theorem 11.

The RMRPE estimator of $\boldsymbol{\theta}$, $\widetilde{\boldsymbol{\theta}}_\alpha$, must satisfy

$$\left\{\begin{array}{r}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_\alpha}+\boldsymbol{M}(\widetilde{\boldsymbol{\theta}}_\alpha)\boldsymbol{\lambda}_n=\mathbf{0}_p,\\ \boldsymbol{m}(\widetilde{\boldsymbol{\theta}}_\alpha)=\mathbf{0}_r,\end{array}\right\}\Leftrightarrow\left\{\begin{array}{r}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_\alpha}=-\boldsymbol{M}(\widetilde{\boldsymbol{\theta}}_\alpha)\boldsymbol{\lambda}_n\\ \boldsymbol{m}(\widetilde{\boldsymbol{\theta}}_\alpha)=\mathbf{0}_r\end{array}\right., \tag{36}$$

where $\boldsymbol{\lambda}_n$ is a vector of Lagrangian multipliers. Now, applying Eq. (18), we can write $\widetilde{\boldsymbol{\theta}}_\alpha=\boldsymbol{\theta}_{\mathbf{g},\alpha}+\boldsymbol{t}n^{-1/2}$, where $\|\boldsymbol{t}\|<c$, for some $0<c<\infty$. We have, applying Taylor,

$$\begin{aligned} \left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_\alpha} = & \left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{g},\alpha}}+\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{g},\alpha}}(\widetilde{\boldsymbol{\theta}}_\alpha-\boldsymbol{\theta}_{\mathbf{g},\alpha}) \\ & +o(\|\widetilde{\boldsymbol{\theta}}_\alpha-\boldsymbol{\theta}_{\mathbf{g},\alpha}\|^2), \end{aligned}$$

and hence

$$n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_\alpha} = n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + \left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) + o(n^{1/2}\|\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\|^2).$$

However,

$$o(n^{1/2}\|\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\|^2) = o(n^{1/2}\|\boldsymbol{t}\|^2/n) = o(n^{-1/2}\|\boldsymbol{t}\|^2) = o(O_p(1)) = o_p(1).$$

Now,

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \hat{V}_i(Y_i;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}$$

$$\xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n} E_{Y_i}\left[\left(\frac{\partial^2 \hat{V}_i(Y;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\right] = \left(\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right)_{jk}.$$

Therefore,

$$n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_\alpha} = n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) + o_p(1). \tag{37}$$

As the RMRPE $\widetilde{\boldsymbol{\theta}}_\alpha$ must satisfy the conditions in (36), and in view of (37) we have

$$n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} = -\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) - \boldsymbol{M}(\widetilde{\boldsymbol{\theta}}_\alpha) n^{1/2}\boldsymbol{\lambda}_n + o_p(1).$$

And applying the continuity of $\boldsymbol{M}$, this can be written as

$$-\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) - \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) n^{1/2}\boldsymbol{\lambda}_n = n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1). \tag{38}$$

On the other hand, applying Taylor to $\boldsymbol{m}$, we obtain

$$n^{1/2}\boldsymbol{m}(\widetilde{\boldsymbol{\theta}}_\alpha) = n^{1/2}\boldsymbol{m}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) + \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) + o_p(1). \tag{39}$$

From (39) and applying that $\boldsymbol{m}(\widetilde{\boldsymbol{\theta}}_\alpha) = \boldsymbol{0}_r$, $\boldsymbol{m}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{0}_r$, it follows that

$$\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) + o_p(1) = \boldsymbol{0}_r. \tag{40}$$

Now, we can express Eqs. (38) and (40) in matrix form as

$$\begin{pmatrix} -\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) & -\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T & \boldsymbol{0}_{r\times r} \end{pmatrix}\begin{pmatrix} n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ n^{1/2}\boldsymbol{\lambda}_n \end{pmatrix} = \begin{pmatrix} n^{1/2}\left(\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \\ \boldsymbol{0}_r \end{pmatrix} + o_p(1).$$

Therefore,

$$\begin{pmatrix} n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ n^{1/2}\boldsymbol{\lambda}_n \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) & -\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T & \boldsymbol{0}_{r\times r} \end{pmatrix}^{-1}\begin{pmatrix} n^{1/2}\left(\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \\ \boldsymbol{0}_r \end{pmatrix} + o_p(1).$$

But

$$\begin{pmatrix} -\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) & -\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T & \boldsymbol{0} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\boldsymbol{\alpha}}) & -\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\ -\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T & \boldsymbol{R}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \end{pmatrix},$$

where $\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ and $\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ are given in (32) and (33), respectively. The matrix $\boldsymbol{R}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ is the matrix needed to make the right hand side of the above equation equal to the indicated inverse. Then,

$$n^{1/2}(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) n^{1/2}\left(\frac{\partial H_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1) \tag{41}$$

and the result holds.

Proof of Theorem 13. Let us denote

$$L = 2n\left[RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widehat{\boldsymbol{\theta}}_\alpha\right) - RP_{NH}\left(M_1^{(s)},\ldots,M_n^{(s)},\widetilde{\boldsymbol{\theta}}_\alpha\right)\right].$$

Then,

$$L = 2n\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)\right] + 2trace\left[\boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\theta}}_\alpha\right)\boldsymbol{\Psi}_n\left(\widehat{\boldsymbol{\theta}}_\alpha\right)^{-1}\right] - 2trace\left[\boldsymbol{\Omega}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)\boldsymbol{\Psi}_n^R\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)^{-1}\right].$$

First, note that

$$H_n\left(\widetilde{\boldsymbol{\theta}}_\alpha\right) = H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + \left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + \frac{1}{2}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^T\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + o(\|\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\|^2).$$

Hence,

$$2n\left[H_n\left(\widetilde{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right] = 2\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^T\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + o_p(1).$$

Now, taking into account that

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) = \boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1),$$

and

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \to \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right),$$

by Eq. (35), we conclude that

$$\begin{aligned}&2n\left[H_n\left(\widetilde{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right]\\ &=2\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\\ &\quad+\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1).\end{aligned}$$

Now, applying Lemma 12, we know that

$$\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = -\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}),$$

and thus,

$$2n\left[H_n\left(\widetilde{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right] = \sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}}\boldsymbol{P}_\alpha^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1).$$

On the other hand,

$$H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) = H_n\left(\widehat{\boldsymbol{\theta}}_\alpha\right) + \left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha} - \widehat{\boldsymbol{\theta}}_\alpha\right) + \frac{1}{2}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha} - \widehat{\boldsymbol{\theta}}_\alpha\right)^T\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha} - \widehat{\boldsymbol{\theta}}_\alpha\right) + o(\|\boldsymbol{\theta}_{\boldsymbol{g},\alpha} - \widehat{\boldsymbol{\theta}}_\alpha\|^2).$$

Now, taking into account that

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha} \longrightarrow \left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \longrightarrow \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right),$$

and

$$\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha} = 0,$$

we conclude that

$$2n\left[H_n\left(\widehat{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right] = -\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + o_p(1).$$

Applying $\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_\alpha} = \mathbf{0}$, we have by Taylor

$$\mathbf{0} = n^{1/2}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) n^{1/2}\left(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) + o_p(1),$$

so that

$$n^{1/2}\left(\widehat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right) = -n^{1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^{-1}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1).$$

Hence,

$$2n\left[H_n\left(\widehat{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right] = -\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1).$$

But as $\boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} - \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}$, we obtain

$$\begin{aligned}2n\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)\right] = &-\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\\ &\times\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} + o_p(1).\end{aligned}$$

Finally we have,

$$\sqrt{n}\left(\frac{\partial H_{n,\alpha}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\boldsymbol{g},\alpha}} \xrightarrow[n\to\infty]{L} N(\mathbf{0}_k, \boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)),$$

and thus the asymptotic distribution of $2n\left[H_{n,\alpha}\left(\widehat{\boldsymbol{\theta}}_\alpha\right) - H_{n,\alpha}\left(\widetilde{\boldsymbol{\theta}}_\alpha\right)\right]$ coincides with the distribution of the random variable

$$\sum_{i=1}^{r}\lambda_i(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})Z_i^2,$$

where $Z_1,\ldots,Z_r$ are independent standard normal variables, $\lambda_1(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}),\ldots,\lambda_r(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})$ are the nonzero eigenvalues of $-\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and

$$r = \operatorname{rank}\left(\mathbf{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\right).$$

For more details see Corollary 2.1 in [39]. This finishes the proof.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki (Eds.), 2nd International Symposium on Information Theory, Akadémia Kiadó, Budapest, Hungary, 1973, pp. 267–281.
[2] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control AC-19 (1974) 716–723.
[3] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[4] G.E. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.
[5] C.M. Hurvich, C.L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297–307.
[6] C.M. Hurvich, C.L. Tsai, A corrected akaike information criterion for vector autoregressive model selection, J. Time Series Anal. 14 (1993) 271–279.
[7] C.M. Hurvich, C.L. Tsai, Model selection for extended quasi–likelihood models in small samples, Biometrics 51 (1995) 1077–1084.
[8] S. Konishi, G. Kitagawa, Generalised information criteria in model selection, Biometrika 83 (1996) 875–890.
[9] K. Takeuchi, Distribution of information statistics and criteria for adequacy of models, Math. Sci. 153 (1976) 12–18, (In Japanese).
[10] H. Bozdogan, Model selection and akaike's information criterion (AIC): The general theory and its analytical extensions, Psychometrika 52 (1987) 345–370.
[11] C.R. Rao, Y. Wu, On Model Selection, in: IMS Lectures Notes. Monograph Series, vol. 312, 2001, pp. 1–57.
[12] J.E. Cavanaugh, A.A. Neath, Akaike's information criterion: Background, derivation, properties, and refinements, Int. Ency. Stat. Sci. (2011) 26–29, http://dx.doi.org/10.1007/978-3-642-04898-2_111.
[13] K. Mattheou, S. Lee, A. Karagrigoriou, A model selection criterion based on the BHHJ measure of divergence, J. Stat. Plann. Inference 139 (2009) 228–235.
[14] A. Basu, I.R. Harris, N.L. Hjort, M.C. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (3) (1998) 549–559.
[15] A. Toma, A. Karagrigoriou, P. Trentou, Robust model selection criteria based on pseudodistances, Entropy 22 (3) (2020) 304.
[16] M.C. Jones, N.L. Hjort, I.R. Harris, A. Basu, A comparison of related density-based minimum divergence estimators, Biometrika 88 (2001) 865–873.
[17] T. Kawashima, H. Fujisawa, Robust and sparse regression via γ-divergence, Entropy 19 (11) (2017) 608.
[18] A. Ghosh, S. Majumdar, Ultrahigh-dimensional robust and efficient sparse regression using non-concave penalized density power divergence, IEEE Trans. Inform. Theory 66 (12) (2020) 7812–7827.

[19] A. Mandal, S. Ghosh, Robust LASSO and Its Applications in Healthcare Data. Trends in Mathematical, Information and Data Sciences, Studies in Systems, Decision and Control, Vol. 445, Springer, 2023.
[20] S. Kurata, E. Hamada, A robust generalization and asymptotic properties of the model selection criterion family, Commun. Stat. (Theor. Methods), 47 3 (2018) 532–547.
[21] L. Pardo, Statistical Inference Based on Divergence Measures, Chapman and Hall CRC., Boca Raton (USA), 2006.
[22] H. Fujisawa, S. Eguchi, Robust parameter estimation with a small bias agains theavy contamination, J. Multivariate Anal. 99 (2008) 2053–2081.
[23] A. Toma, S. Leoni-Aubin, Robust tests based on dual divergence estimators and saddle points approximation, J. Multivariate Anal. 101 (2010) 1143–1155.
[24] E. Castilla, N. Martín, S. Muñoz, L. Pardo, Robust wald-type tests based on minimum rényi pseudodistance estimators for the multiple regression model, J. Stat. Comput. Simul. 14 (2020) 2592–2613.
[25] M. Jaenada, L. Pardo, Robust statistical inference in generalized linear models based on minimum rényi pseudodistance estimators, Entropy 24 (123) (2022).
[26] E. Castilla, M. Jaenada, L. Pardo, Estimation and testing on independent not identically distributed observations based on rényi's pseudodistances, IEEE Trans. Inform. Theory 68 (7) (2022) 4588–4609.
[27] M. Jaenada, P. Miranda, L. Pardo, Robust tests statistics based on restricted minimum rényi pseudodistance estimators, Entropy 24 (616) (2022).
[28] M. Broniatowski, A. Toma, I. Vajda, Decomposable pseudodistances and applications in statistical estimation, J. Statist. Plann. Inference 142 (2012) 2574–2585.
[29] A. Mandal, S. Ghosh, Robust variable selection criteria for the penalized regression, 2019, arXiv preprint arXiv:1912.12550.
[30] A. Basu, A. Mandal, N. Martín, L. Pardo, Testing composite hypothesis based on density power divergence, Sankhya 80 (13) (2018) 222–262.
[31] E. Castilla, M. Jaenada, N. Martín, L. Pardo, Robust approach for comparing two dependent normal populations through waldtype tests based on rényi's pseudodistance estimators, Stat. Comput. 32 (2023) 100, http://dx.doi.org/10.1007/s11222-022-10162-7.
[32] F.R. Hampel, Contributions To the Theory of Robust Estimation, University of California, Berkeley, 1968.
[33] F.R. Hampel, The influence curve and its role in robust estimation, J. Amer. Statist. Assoc. 69 (346) (1974) 383–393.
[34] J. Warwick, M.C. Jones, Choosing a robustness tuning parameter, J. Stat. Comput. Simul. 75 (7) (2005) 581–588.
[35] S. Basak, A. Basu, M.C. Jones, On the optimal density power divergence tuning parameter, J. Appl. Stat. 48 (3) (2021) 536–556.
[36] N.R. Draper, H. Smith, Applied Regression Analysis, second ed., Wiley Blackwell, Hoboken, NJ (USA), 1981.
[37] A.R. Behnke, J.H. Wilmore, Evaluation and Regulation of Body Build and Composition, Prentice-Hall, Englewood Cliffs, N.J, 1974.
[38] F. Katch, W. McArdle, Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston, 1977.
[39] J.J. Dik, M.C.M. Gunst, The distribution of general quadratic forms in normal variables, Stat. Neerl. 39 (1985) 14–26.