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

Angel Felipe[1], Maria Jaenada[1], Pedro Miranda[1] and Leandro Pardo[1]
[1]Department of Statistics and O.R., Complutense University of Madrid, Spain

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 measure 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.

Keywords: Rényi's pseudodistance, robustness, restricted model, multiple linear regression model.

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 [20] 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.

Following similar ideas than the AIC, several other model selection criteria have been proposed in the literature. For example, Schwarz in [24] developed the "Bayesian information criterion" (BIC), which imposes a stronger penalty for model complexity than AIC. Also derived from AIC, Hurvich and Tsai [13, 14, 15] 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 [19] 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 [25] when the fitting method is maximum likelihood. Finally, Bozdogon [5] proposed another variant of AIC, called CAIC, that corrected its lack of consistency. Interesting surveys about model selection criteria can be found in [23, 9].

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, [22] considered the density power divergence (DPD) [3] to define a robust model selection criterion. Similarly, Toma et al. [26] introduced another robust criterion for model selection based on the Rényi pseudodistance (RP) [18].

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 [21] considered a criterion based on DPD, extending the theory of [22]. 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 [26].

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 a simulation study illustrates the robustness of the proposed criterion and compare it with other model selection criteria. Section 6 deals with a real data example. Some final conclusions are presented in Section 7.

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

Let $Y_1, ..., Y_n$ be i.n.i.d.o. observations, where each Y_i has true probability distribution function $G_i, i = 1, ..., n$, and probability density function $g_i, i = 1, ..., n$, respectively. For inferential purposes, it is assumed that the true density function g_i could belong to a parametric family of densities, $f_i(y, \boldsymbol{\theta}), i = 1, ..., n$, 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 $f_i(y, \boldsymbol{\theta}), i = 1, ..., n$.

The value of $\boldsymbol{\theta}$ that best fits the original distributions $g_1, ..., g_n$, would naturally minimize some kind of distance between the true and assumed densities, $(g_1(y), ..., g_n(y))$ and $(f_1(y, \boldsymbol{\theta}), ..., f_n(y, \boldsymbol{\theta}))$. Here, we will use the family of RP divergence measures defined in [18] 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*

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

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.

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

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

Hence, $R_0\left(f(\cdot, \boldsymbol{\theta}), g(\cdot)\right)$ coincides with the Kullback-Leibler divergence measure between g and f. 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, [12] considered the RP divergence under the name of γ-cross entropy. Additionally, Toma and Leoni-Auban [27] defined new robust and efficient measures based on RP. In [7], Wald-type tests based on RP were developed in the context of MLRM, and were extended later in [17] for the generalized multiple regression model. Moreover, in [17] a robust approach for comparing two dependent normal populations via a Wald-type test based on RP was carried out. In [16] 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

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

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

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

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

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

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

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

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*

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

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

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

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 $(f_i(y,\boldsymbol{\theta}),g_i(y)), i=1,...,n,$ 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 weighting to all functions $V_{i,\alpha}(\boldsymbol{\theta})$. Hence, we consider

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

Definition 3 *Consider $(g_1(y),...,g_n(y))$ and $(f_1(y,\boldsymbol{\theta}),...,f_n(y,\boldsymbol{\theta}))$, n pairs of true and assumed densities for i.n.i.d.o. random variables $Y_i, i=1,...,n$. 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 $i=1,...,n$, 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

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

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

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

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

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

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 for $\alpha > 0$ as

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

with

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

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

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

Definition 4 *Given $Y_1, \dots, Y_n$ 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 $H_{n,\alpha}(\boldsymbol{\theta})$

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

That is, the estimation equations of the MRPE are

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

with

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

and

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

It is interesting to observe that if $Y_1,...,Y_n$ 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 [6].

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

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

with

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

and

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

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

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 $g_1,...,g_n$ 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 $i=1,...,n$.

C3. For $i=1,2,...,n$ 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 $i = 1, 2, ..., n$ 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

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

and

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

C6. For all j,k,l and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the sequences $\left\{\frac{\partial \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j}\right\}_{j=1,\ldots,p}$,$\left\{\frac{\partial^2 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right\}_{j,k=1,..,p}$ and $\left\{\frac{\partial^3 \widehat{V}_{i,\alpha}(Y_i,\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial l}\right\}_{j,k,l=1,..,p}$ 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 $\varepsilon > 0$

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

Now, the following result, whose proof can be seen in [7], holds.

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

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

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

2.1 Example: The MPRE under the MLRM

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

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

where the errors $\varepsilon_i' s$ are i.i.d. normal random variables with mean zero and variance σ^2, $\boldsymbol{X}_i^T = (X_{i1}, ..., X_{ip})$ is the vector of independent variables corresponding to the i-th condition and $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)^T$ 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 $f_i\left(y, \boldsymbol{\beta}, \sigma\right) \equiv \mathcal{N}(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma^2)$ 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

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

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

$$\begin{array}{c} \sum\limits_{i=1}^n \exp\left(-\frac{\alpha}{2}\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{X}_i = \boldsymbol{0}_p \\ \sum\limits_{i=1}^n \exp\left(-\frac{\alpha}{2}\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right)\left\{\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}}{\sigma}\right)^2 - \frac{1}{1+\alpha}\right\} = 0 \end{array}, \tag{21}$$

which is exactly the same system as the one obtained in [7]. For $\alpha = 0$, if we denote $\mathbb{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T_{n\times p}$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$, 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 [7], it can be seen that matrices $\boldsymbol{\Psi}_{n,\alpha}(\boldsymbol{\beta}, \sigma)$ and $\boldsymbol{\Omega}_{n,\alpha}(\boldsymbol{\beta}, \sigma)$ are given by

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

and

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

with

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

Therefore, for $\alpha = 0$ we get the Fisher information matrix for $(\boldsymbol{\beta}, \sigma)$ in both matrices, 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

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

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

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

with associated distribution functions $\boldsymbol{F}(.,\boldsymbol{\theta}_s) = (F_1(\boldsymbol{\theta}_s), ..., F_n(.,\boldsymbol{\theta}_s))$, 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 $Y_1, ..., Y_n$, we need to select the best model from the collection $\{\boldsymbol{M}^{(s)}\}_{s\in\{1,...,l\}}$ 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 distribution underlying. 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 distribution underlying is unknown but we have a random sample $Y_1, ..., Y_n$, 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

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

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

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

where

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

We can observe that

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

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

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

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

Theorem 7 *Assume that conditions* **C1-C8** *hold. Then,*

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

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

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

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

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

and hence,

$$E_{Y_1,\dots,Y_n}\left[nH_\alpha\left(\widehat{\boldsymbol{\theta}}_\alpha^s\right)\right] = nH_\alpha\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}^s\right) - \frac{1}{2}E_{Y_1,\dots,Y_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}^s\right)^T \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}^s\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_\alpha^s - \boldsymbol{\theta}_{\boldsymbol{g},\alpha}^s\right)\right] + o_p(1). \tag{25}$$

But by Eq. (18), and applying Corollary 2.1 in [10], we have

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

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

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

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

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

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 $\boldsymbol{\theta}_{\boldsymbol{g},\alpha}^s$ gives

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

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

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

Besides,

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\ \partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^s_\alpha} \underset{n\longrightarrow\infty}{\overset{\mathcal{P}}{\longrightarrow}} -\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}^s_{\boldsymbol{g},\alpha}\right). \tag{26}$$

by the continuity of $\boldsymbol{\Psi}_n$. Hence, substituting in (25)

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

and thus,

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

Hence, the result holds. ■

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.

$$Y_i=\boldsymbol{X}_i^T\boldsymbol{\beta}+\varepsilon_i,\quad i=1,\ldots,n. \tag{27}$$

We consider several models $\{(M_1^{(s)},...,M_n^{(s)})\}_{s=1,\ldots,l}$ 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

$$\begin{aligned}
(M_1^{(1)},...,M_n^{(1)}) &\equiv Y_i=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\epsilon_i,\\
(M_1^{(2)},...,M_n^{(2)}) &\equiv Y_i=\beta_0+\beta_1X_1+\beta_2X_2+\beta_4X_4+\epsilon_i\\
(M_1^{(3)},...,M_n^{(3)}) &\equiv Y_i=\beta_0+\beta_1X_1+\beta_3X_3+\beta_4X_4+\epsilon_i,\\
(M_1^{(4)},...,M_n^{(4)}) &\equiv Y_i=\beta_0+\beta_2X_2+\beta_3X_3+\beta_4X_4+\epsilon_i.
\end{aligned}$$

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

$$\left.\begin{array}{l}\sum\limits_{i=1}^{n}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_{s,i}^{T}\boldsymbol{\beta}}{\sigma}\right)^{2}\right)\left(\frac{Y_i-\boldsymbol{X}_{s,i}^{T}\boldsymbol{\beta}}{\sigma}\right)\boldsymbol{X}_{s,i}=\mathbf{0}_4\\ \sum\limits_{i=1}^{n}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_{s,i}^{T}\boldsymbol{\beta}}{\sigma}\right)^{2}\right)\left\{\left(\frac{Y_i-\boldsymbol{X}_{s,i}^{T}\boldsymbol{\beta}}{\sigma}\right)^{2}-\frac{1}{1+\alpha}\right\}=0\end{array}\right., \tag{28}$$

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 s-th 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

$$\begin{aligned}\boldsymbol{\Psi}_{s,n}\left(\boldsymbol{\beta}^s,\sigma\right)&=K_1\left(\alpha+1\right)^{-\frac{3}{2}}\left[\begin{array}{cc}\frac{1}{n}\mathbb{X}_s^T\mathbb{X}_s & 0\\ 0 & \frac{2}{\alpha+1}\end{array}\right],\\ \boldsymbol{\Omega}_{s,n}\left(\boldsymbol{\beta}^s,\sigma\right)&=K_1^2\sigma^2\frac{1}{\left(2\alpha+1\right)^{3/2}}\left[\begin{array}{cc}\frac{1}{n}\mathbb{X}_s^T\mathbb{X}_s & 0\\ 0 & \frac{(3\alpha^2+4\alpha+2)}{2(\alpha+1)(2\alpha+1)}\end{array}\right],\end{aligned} \tag{29}$$

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,

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

and hence,

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

Therefore, applying the $RP_{NH}-$Criterion defined in (6)

$$\begin{aligned}RP_{NH}(M_1^{(s)},...,M_n^{(s)},\widehat{\boldsymbol{\beta}}_\alpha^s,\widehat{\sigma}_\alpha^2)\\ =& -\frac{1}{\alpha}\left(\frac{1+\alpha}{2\pi}\right)^{\frac{\alpha}{2(\alpha+1)}}\frac{1}{n}\sum_{i=1}^n(\widehat{\sigma}_\alpha^s)^{-\frac{\alpha}{\alpha+1}}\exp\left(-\frac{\alpha}{2}\left(\frac{Y_i-\boldsymbol{X}_{s,i}^T\widehat{\boldsymbol{\beta}}_\alpha^s}{\widehat{\sigma}_\alpha^s}\right)^2\right)\\ &+\frac{1}{\alpha}+\frac{1}{n}(\widehat{\sigma}_\alpha^s)^2K_1\left(p\frac{(\alpha+1)^{\frac{3}{2}}}{(2\alpha+1)^{\frac{3}{2}}}+\frac{(\alpha+1)^{\frac{1}{2}}\left(3\alpha^2+4\alpha+2\right)}{2\left(2\alpha+1\right)^{5/2}}\right).\end{aligned} \tag{30}$$

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

$$\left\{\boldsymbol{\theta}\in\boldsymbol{\Theta}/\ \boldsymbol{m}(\boldsymbol{\theta})=\mathbf{0}_r\right\}, \tag{31}$$

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

$$\mathbf{M}\left(\boldsymbol{\theta}\right)=\frac{\partial\boldsymbol{m}^T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{32}$$

exists and is continuous in $\boldsymbol{\theta}$, and rank$(\mathbf{M}\left(\boldsymbol{\theta}\right))=r,\forall\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Related to the divergence-based restricted estimation, in [4] the restricted minimum density power divergence estimator was defined. Later, in [8] 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 8 *Given $Y_1, ..., Y_n$ be i.n.i.d.o., the* **restricted MRPE** *(RMRPE), $\widetilde{\boldsymbol{\theta}}_\alpha$, is given by*

$$\widetilde{\boldsymbol{\theta}}_\alpha = \arg \min_{\boldsymbol{\theta}\in\Theta/\boldsymbol{m}(\boldsymbol{\theta})=\mathbf{0}_r} H_{n,\alpha}(\boldsymbol{\theta}), \tag{33}$$

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

Note that

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

The following theorem presents a representation of the RMPRE.

Theorem 9 *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

$$\boldsymbol{P}^*(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} - \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}, \tag{34}$$

with

$$\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) = \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \left[\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1} \boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\right]^{-1}. \tag{35}$$

Proof. 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 $\widetilde{\boldsymbol{\theta}}_\alpha = \boldsymbol{\theta}_{\boldsymbol{g},\alpha} + \boldsymbol{t}n^{-1/2}$, where $||\boldsymbol{t}|| < c$, for some $0 < c < \infty$. We have, applying Taylor,

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

and hence

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

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,

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

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

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

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

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

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 (34) and (35), 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. ■

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 10 *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}) \\
&\quad= \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] \\
&\qquad \boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)\boldsymbol{P}^*_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}) \\
&\quad= \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}) \\
&\quad= \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} \\
&\qquad - \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}) \\
&\quad= -\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 $RP_{NH}((M_1^{(s)},...,M_n^{(s)},\boldsymbol{\theta})$ 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

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

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

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

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

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

Theorem 11 *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*

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

coincides with the distribution of the random variable

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

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 $-\boldsymbol{Q}_\alpha(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and

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

Proof. Let us denote

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

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

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

Hence,

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

Now, taking into account that

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

and

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

by Eq. (26), we conclude that

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

Now, applying the previous lemma, 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,

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

On the other hand,

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

Now, taking into account that

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

and

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

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

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

so that

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

Hence,

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

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

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

Finally we have,

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

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 $-\boldsymbol{Q}_{\alpha}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})\boldsymbol{M}(\boldsymbol{\theta}_{\boldsymbol{g},\alpha})^T\boldsymbol{\Psi}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)^{-1}\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}_{\boldsymbol{g},\alpha}\right)$ and

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

For more details see Corollary 2.1 in [10]. This finishes the proof. ■

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

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

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

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

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 $\boldsymbol{\Psi}_n\left(\boldsymbol{\beta},\sigma\right)$ and $\boldsymbol{\Omega}_n\left(\boldsymbol{\beta},\sigma\right)$ given in Eq. (29) can be rewritten, using the notation $\mathbb{X}_1$ and $\mathbb{X}_2$, as

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

being K_1 as defined in (22) and

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

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

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

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 $\boldsymbol{\Psi}_n^{-1}\left(\boldsymbol{\beta},\sigma\right)$ can be computed as

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

On the other hand,

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

and

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

and so, multiplying the above expressions we obtain that

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

and

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

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

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

and hence,

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

On the other hand, we have

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

and hence the trace of the above matrix is given by

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

and

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

Therefore,

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

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

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

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

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

5 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

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

where $\epsilon_i \sim \mathcal{N}(0,1)$ and the variables X_i are fixed and chosen uniformly in the interval [-2, 2]. Next, we take $n = 100$, so that

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

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 $(Y^s, X^s), s = 1, ..., 1000$ 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 $\alpha = 0.01, 0.02, 0.04, 0.07, 0.1, 0.2, 0.4, 0.5, 0.7$ 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.

p	0	1	2	3	4	5
AIC	0	0	0	0	836	164
BIC	0	0	0	0	967	33
AIC_c	0	0	0	0	864	136
$RPNH_{0.01}$	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	822	178
$RPNH_{0.2}$	0	0	0	0	834	166
$RPNH_{0.4}$	0	0	0	0	842	158
$RPNH_{0.5}$	0	0	0	0	841	159
$RPNH_{0.7}$	0	0	0	0	838	162
$RPNH_{1.0}$	0	0	0	0	837	163

Table 1: Results for uncontaminated data.

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

$$\epsilon_i \sim \mathcal{U}(\min_i(X_i + 2X_i^2 - X_i^3 + X_i^4) - r, \max_i(X_i + 2X_i^2 - X_i^3 + X_i^4) + r),$$

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 valus $r = 1, 5, 10$. 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, 3, 4 and 5. The left part of each table corresponds to $r = 1$, the center part for $r = 5$ and the right part for $r = 10$.

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	1	19	659	321	0	0	10	27	622	341	0	0	9	38	616	337
BIC	0	0	16	64	802	118	0	0	39	84	751	126	0	0	57	96	715	132
AIC_c	0	0	1	22	694	283	0	0	12	33	651	304	0	0	12	43	634	311
$RPNH_{0.01}$	0	0	3	14	844	139	0	0	5	18	830	147	0	0	9	22	812	157
$RPNH_{0.02}$	0	0	0	13	866	121	0	0	2	47	833	118	0	0	6	62	810	122
$RPNH_{0.04}$	0	0	2	20	844	134	0	0	1	18	833	148	0	0	1	18	833	148
$RPNH_{0.1}$	0	0	0	0	835	165	0	0	0	0	836	164	0	0	0	0	830	170
$RPNH_{0.2}$	0	0	0	0	829	171	0	0	0	0	833	167	0	0	0	0	833	167
$RPNH_{0.4}$	0	0	0	0	837	163	0	0	0	0	835	165	0	0	0	0	839	161
$RPNH_{0.5}$	0	0	0	0	837	163	0	0	0	0	848	152	0	0	0	0	834	166
$RPNH_{0.7}$	0	0	0	0	842	158	0	0	0	0	836	164	0	0	0	0	831	169
$RPNH_{1.0}$	0	0	0	0	838	162	0	0	0	0	836	164	0	0	0	0	830	170

Table 2: Results for contamination degree of 5%

	$r=1$						$r=5$						$r=10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	17	64	591	328	0	0	18	82	575	325	0	0	47	106	538	309
BIC	0	0	94	155	629	122	0	0	95	180	611	114	0	0	153	178	558	111
AIC_c	0	0	21	75	621	283	0	0	24	93	601	282	0	0	55	123	556	266
$RPNH_{0.01}$	0	0	24	37	770	169	0	0	26	48	750	176	0	0	37	61	705	197
$RPNH_{0.02}$	0	0	19	30	800	151	0	0	24	40	780	156	0	0	30	60	747	163
$RPNH_{0.04}$	0	0	16	60	809	115	0	0	19	100	764	117	0	0	23	94	770	113
$RPNH_{0.1}$	0	0	0	5	845	150	0	0	0	1	839	160	0	0	0	1	851	148
$RPNH_{0.2}$	0	0	0	0	829	171	0	0	0	0	835	165	0	0	0	0	844	156
$RPNH_{0.4}$	0	0	0	0	829	171	0	0	0	0	840	160	0	0	0	0	850	150
$RPNH_{0.5}$	0	0	0	0	824	176	0	0	0	0	845	155	0	0	0	0	841	159
$RPNH_{0.7}$	0	0	0	0	831	169	0	0	0	0	833	167	0	0	0	0	834	166
$RPNH_{1.0}$	0	0	0	0	841	159	0	0	0	0	835	165	0	0	0	0	833	167

Table 3: Results for a contamination degree of 10%.

	$r = 1$						$r = 5$						$r = 10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	52	134	509	305	0	0	76	176	464	284	0	0	148	169	397	286
BIC	0	0	238	210	440	112	0	0	278	234	398	90	0	0	367	223	325	85
AIC_c	0	0	64	154	512	270	0	0	91	191	476	242	0	0	179	180	400	241
$RPNH_{0.01}$	0	0	41	92	596	271	0	0	43	93	561	303	0	0	52	95	514	339
$RPNH_{0.02}$	0	0	37	85	625	253	0	0	39	92	589	280	0	0	47	90	561	302
$RPNH_{0.04}$	0	0	29	75	676	220	0	0	32	88	661	219	0	0	43	93	648	216
$RPNH_{0.1}$	0	0	42	214	631	113	0	0	41	138	693	128	0	0	20	64	810	106
$RPNH_{0.2}$	0	0	0	0	836	164	0	0	0	0	831	169	0	0	0	0	849	151
$RPNH_{0.4}$	0	0	0	0	837	163	0	0	0	0	833	167	0	0	0	0	840	160
$RPNH_{0.5}$	0	0	0	0	836	164	0	0	0	0	829	171	0	0	0	0	840	160
$RPNH_{0.7}$	0	0	0	0	845	155	0	0	0	0	827	173	0	0	0	0	849	151
$RPNH_{1.0}$	0	0	0	0	834	166	0	0	0	0	823	177	0	0	0	0	836	164

Table 4: Results for a contamination degree of 20%.

	$r = 1$						$r = 5$						$r = 10$					
p	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
AIC	0	0	112	178	433	277	0	0	114	209	373	304	0	0	192	183	374	251
BIC	0	0	327	256	327	90	0	0	385	240	276	99	0	0	457	212	253	78
AIC_c	0	0	136	191	436	237	0	0	137	233	368	262	0	0	219	189	371	221
$RPNH_{0.01}$	0	0	51	79	519	351	0	0	55	90	488	367	0	0	58	90	428	424
$RPNH_{0.02}$	0	0	46	77	540	337	0	0	48	84	520	348	0	0	52	87	472	389
$RPNH_{0.04}$	0	0	44	81	573	302	0	0	46	80	555	319	0	0	53	78	533	336
$RPNH_{0.1}$	0	0	70	187	550	193	0	0	63	221	537	179	0	0	55	139	628	178
$RPNH_{0.2}$	0	0	17	68	774	141	0	0	11	13	817	159	0	0	2	8	854	136
$RPNH_{0.4}$	0	0	0	0	856	144	0	0	0	0	833	167	0	0	0	0	841	159
$RPNH_{0.5}$	0	0	0	0	845	155	0	0	0	0	830	170	0	0	0	0	832	168
$RPNH_{0.7}$	0	0	0	0	834	166	0	0	0	0	815	185	0	0	0	0	826	174
$RPNH_{1.0}$	0	0	0	0	828	172	0	0	1	1	813	185	0	0	0	0	841	159

Table 5: Results for a contamination degree of 30%.

From the results in these tables, it can be seen that the performance of AIC, BIC and AIC_c dramatically decrease, 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, 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 9 375 different situations. And for all of them we can extract the same conclusions.

6 Real data example

In this section we analyze a set of real data at the light of this new model selection tool based on RP. We consider the problem proposed in [11] and later studied in [26]. 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 6. 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.

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

Table 6: The Hald cement data.

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.

We have applied the RP_{NH}-criterion defined in (30) for different values of the tuning parameter to select the most appropriate model. The solution is given in Table 7. 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 [26]. Remark also the good performance of model X_1, X_2.

7 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

	$RPNH_{0.01}$	$RPNH_{0.02}$	$RPNH_{0.04}$	$RPNH_{0.05}$	$RPNH_{0.07}$
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 7: Results for the Hald cement data.

more robust agaisnt 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 a situation with real data.

Acknowledgements

This work was supported by the Spanish Grant PID2021-124933NB-I00.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), 2nd international symposium on information theory (pp. 267–281). Budapest, Hungary: Akadémia Kiadó.

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

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

[4] Basu, A., Mandal, A., Martín, N. and Pardo, L. (2018). Testing composite hypothesis based on density power divergence. *Sankhya*, **80 (13)**, 222–262.

[5] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

[6] Broniatowski, M., Toma, A. and Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, **142**, 2574–2585.

[7] Castilla, E., Jaenada, M. and Pardo, L. (2022). Estimation and testing on independent not identically distributed observations based on Rényi's pseudodistances. *IEEE Transactions on Information Theory*, **68, 7**, 4588–4609.

[8] Castilla, E., Jaenada, M., Martín, N. and Pardo, L. (2023). Robust approach for comparing two dependent normal populations through Waldtype tests based on rényi's pseudodistance estimators. *Statistics and Computing*, DOI: 10.1007/s11222-022-10162-7.

[9] Cavanaugh, J. E. and Neath, A. A. (2011). Akaike's Information Criterion: Background, Derivation, Properties, and Refinements. *International Encyclopedia of Statistical Science*, 26–29. doi:10.1007/978-3-642-04898-2_111.

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

[11] Draper, N.R. and Smith, H. (1981). Applied Regression Analysis, 2nd ed. Wiley Blackwell. Hoboken, NJ (USA).

[12] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias agains theavy contamination. *Journal of Multivariate Analysis*, **99**, 2053–2081.

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

[14] Hurvich, C. M. and Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, **14**, 271–279.

[15] Hurvich, C. M. and Tsai, C. L. (1995). Model selection for extended quasi–likelihood models in small samples. *Biometrics*, **51**, 1077–1084.

[16] Jaenada, M., Miranda, P. and Pardo, L. (2022). Robust tests Statistics based on restricted minimum Rényi Pseudodistance estimators. *Entropy*, **24**, 616.

[17] Jaenada, M. and Pardo, L. (2022). Robust Statistical Inference in Generalized Linear Models based on minimum Rényi Pseudodistance estimators. *Entropy*, **24**, 123.

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

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

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

[21] Kurata, S. and Hamada, E. (2018). A robust generalization and asymptotic properties of the model selection criterion family. *Communication In Statistics (Theory and Methods)*, **47**, 3, 532-547.

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

[23] Rao, C. R. and Wu, Y. (2001). On model Selection. I*MS Lectures Notes. Monograph Series*, **312**, 1-57.

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

[25] Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Math. Sci.*, **153**, 12-18 (In Japanese).

[26] Toma, A., Karagrigoriou, A., and Trentou, P. (2020). Robust model selection criteria based on pseudodistances. *Entropy*, **22(3)**, 304.

[27] Toma, A. and Leoni-Auban, S. (2010). Robust tests based on dual divergence estimators and saddle points approximation. *Journal of Multivariate Analysis*, **101**, 1143–1155.