Universal and Automatic Elbow Detection for Learning the Effective Number of Components in Model Selection Problems

Eduardo Morgado*, Luca Martino*, Roberto San Millán-Castillo*,
* Universidad Rey Juan Carlos (URJC), Madrid, Spain.

2022

Abstract

We design a Universal Automatic Elbow Detector (UAED) for deciding the effective number of components in model selection problems. The relationship with the information criteria widely employed in the literature is also discussed. The proposed UAED does not require the knowledge of a likelihood function and can be easily applied in diverse applications, such as regression and classification, feature and/or order selection, clustering, and dimension reduction. Several experiments involving synthetic and real data show the advantages of the proposed scheme with benchmark techniques in the literature.
Keywords: model selection, order selection, automatic elbow detection, variable selection, clustering.

1 Introduction

Model selection is vast and one of the most relevant tasks in signal processing, statistics and machine learning [1, 2, 3, 4]. It is the process of selecting a statistical model from a set of candidates. Model selection includes as special cases the following well-known sub-tasks: order selection (e.g., in polynomial functions or auto-regressive models [5]), variable selection [6], dimension reduction [7], and clustering [8], to name a few.
More specifically, in a large amount of research works from the most diverse fields, researchers and practitioners face a trade-off between the number of components/variables to consider in their analyses and the performance of the obtained results. Note that we use the term "variables" as a general concept that can equivalently represent variables, features, or the number of clusters, depending on the nature of the considered problem. This trade-off occurs because increasing the number of variables taken into account in the analysis allows for better results, at the expense of obtaining a more complex model. In other words, the model performance and the model complexity generate the so-called bias-variance trade-off. Therefore, in many applications, researchers must obtain the optimal number of components/variables addressing the

aforementioned trade-off [1].

Related works. The solutions in the literature belong to different families and approaches. A first class of methods is formed by the *resampling techniques*, such as cross-validation or bootstrap, where the dataset is split into training and test sets [9, 10, 11]. However, the proportion of data to include in the training and test sets is a crucial parameter that affects critically the results. Another important family is the class of the *information criteria* [12], such as the Bayesian Information Criterion (BIC) [13], the Akaike Information Criterion (AIC) [14], or the Hannan-Quinn Information Criterion (HQIC) [15], to name a few [2, 16]. The information criteria consider a linear penalization of the model complexity, and they differ for the choice of the slope of this penalization. These choices are motivated by theoretical probabilistic derivations which involve several assumptions and approximations. Hence, the good performance of an information criterion is often restricted to very specific scenarios. Moreover, the computation of the information criteria often requires the knowledge of the maximum of a likelihood function. More recent and advanced works related to information criteria can be found in [17, 18, 19]. Other probabilistic strategies related to the information criteria are the so-called minimum description length principle, Mallows's Cp coefficient and the structural risk minimization [20, 21]. In the Bayesian framework, the use of marginal likelihood and posterior predictive approaches are usually employed [2, 22, 23]. The connection between the marginal likelihood and information criteria is discussed in the appendices of [16]. The posterior predictive approach is related to the cross-validation idea. Furthermore, standard frequentist approaches based on p-values have a vast use in some specific applications and deserve to be cited [24, 25]. Finally, some authors apply a visual inspection of an error curve looking for an "elbow", specially in the clustering literature.

In this work, we design a Universal Automatic Elbow Detector (UAED) based on a geometric approach. The proposed scheme is inspired by the concept of the maximum "area under the curve" in receiver operator characteristic curves [1, 26], which is well-known and vastly employed in signal processing and machine learning. The resulting UAED technique also induces a linear penalization of the model complexity. We discuss the connections, differences and the advantages of UAED with respect to the information criteria already presented in the literature. It is important to remark that the range of applicability of UAED is much wider than other techniques in the literature, since no likelihood function is needed. The application of UAED only requires the knowledge of an error curve, that can be defined in different ways according to the user's desire. Moreover, we describe several appealing behaviors of UAED and test it in different numerical examples, three of them involving real datasets. The results show the benefits of UAED with respect to other benchmark techniques in the literature. Therefore, the main contributions of the work are the following:

- We introduce a Universal Automatic Elbow Detector (UAED) based on a geometric approach.
- Four equivalent derivations of UAED are presented, two of them in Section 3 and other two derivations in Appendices A-B.
- The behavior, invariance and other properties of UAED are discussed. See Section 3.4 and Appendix C.

- UAED is tested and compared with other benchmark methods in six numerical experiments, regarding different applications (clustering, order selection and variable selection). Three of them involve applications with real datasets. See Section 4.
- We provide a related Matlab code of UAED.[1]

The rest of the article is organized as follows. Section 2 describes the framework and the notation employed in the derivation of UAED. Section 3 presents and discusses UAED in detail, whereas Section 4 shows six numerical experiments and practical applications (three experiments involve real datasets). Finally, in Section 5, some conclusions are given. In the appendices, we present (a) two alternative derivations, (b) an additional property of UAED, and (c) a possible extension of UAED (which provides more flexibility).

2 Framework and main notation

In many applications, we desire to infer a vector of parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$ of dimension k given a data vector $\mathbf{y} = [y_1, ..., y_N]^\top$. A likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is usually available, often induced by a related physical model. Furthermore, in different types of real-world application problems (clustering, variable selection, or dimension reduction) and specially in model selection problems, an *error function* (i.e., a fitting measure) is obtained. Here, we denote it as

$$V(k) : \mathbb{N} \to \mathbb{R}, \qquad k = 0, 1, 2, ..., K,$$

where k denotes the number of components (e.g., variables, clusters, or order of the polynomial function), i.e., k defines the complexity of the model. In the literature, we often have

$$V(k) = -2\log(\ell_{\texttt{max}}), \quad \text{where} \quad \ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k),$$

as in [12]. However, in this work, $V(k)$ could be directly the mean square error (MSE), or the mean absolute error (MAE). For instance, $V(k)$ can represent the prediction error in regression problems with a polynomial function, where k is the order of the polynomial, or the sum of the inner variances within clusters where k is the number of clusters. We assume that k starts in 0 and grows with step 1 for simplicity, but more general cases can easily be addressed. Namely, a different incremental step could be also considered.
Generally, $V(k)$ is a *non-increasing* error curve, i.e., for any pair of non-negative integers n_1, n_2 such that $n_2 > n_1$, then we have $V(n_2) \leq V(n_1)$.[2] Indeed, $V(k)$ is a fitting term that decreases as the complexity of the model (given by the number k of parameters) grows. Therefore, we have $V(0) \geq V(k), \forall k$. Hence, in this work, we assume $V(k)$ to be a non-increasing function. See Figure 1 for some graphical examples.
Observe that $V(0)$ represents the value of the error function corresponding, for instance, to a constant model in a regression problem, or a single cluster (for all the data) in a clustering

[1] A related Matlab code is given at `http://www.lucamartino.altervista.org/PUBLIC_UAED_CODE.zip`.
[2] This condition could be also relaxed. We keep it, for the sake of simplicity.

problem. In some applications, the score function $V(k)$ should be also convex, i.e., the differences $V(k+1)-V(k)$ will decrease as k increases. This is the case of a variable selection problem, if the variables have been ranked correctly. However, this work does not require conditions regarding the concavity of $V(k)$.

Additional assumptions. Just for the sake of simplicity and without loss of generality, we assume that $\min V(k) = V(K) = 0$. Note that this condition can be always obtained with a simple subtraction, defining a new curve

$$V'(k) = V(k) - \min V(k) = V(k) - V(K). \tag{1}$$

Figure 1(a) depicts a graphical example of the curve $V(k)$. Moreover, above we have assumed $k = 0, 1, ..., K$ but, if there exists a value $k_{\texttt{max}} \leq K$ such that $V(k)$ has not an additional drop for $k \geq k_{\texttt{max}}$, i.e., $k_{\texttt{max}} = \min\left[\arg\min_k V(k)\right]$, so that

$$V(k_{\texttt{max}}) = V(k_{\texttt{max}} + 1) = V(k_{\texttt{max}} + 2) = ... = V(K) = 0. \tag{2}$$

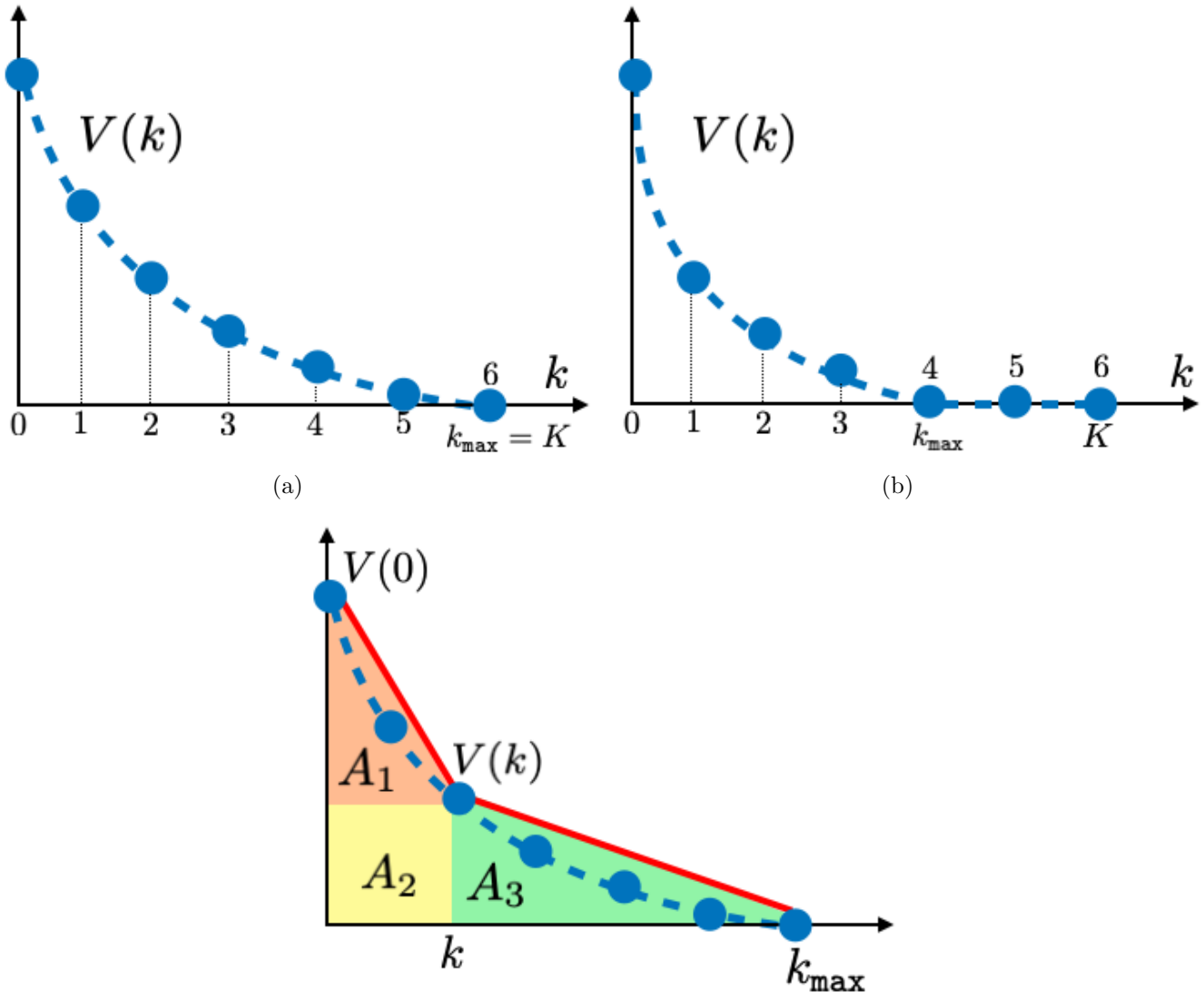
In this scenario, we can consider $k = 0, 1, ..., k_{\texttt{max}}$, since the rest of the components, from $k_{\texttt{max}} + 1$ to K, must be discarded because they do not cause a drop in the error function. See Figures 1(a)-1(b) for two graphical examples. Clearly, if the minimum value of k is different from 0, let us say $k_{\texttt{min}}$, we can always set $k' = k - k_{\texttt{min}}$.

3 The Universal Automatic Elbow Detector (UAED)

In this section, we provide two equivalent geometric derivations of the proposed method, and discuss the similarities, differences, and connections with other methods in the literature. The behavior of the proposed technique is described and some interesting considerations are also highlighted.

3.1 First derivation

Considering the decay $V(k)$ described in the previous section, the underlying idea is "inspired" by the concept of the maximum AUC in ROC curves [1, 26]. Namely, we desire to extract geometric information from the curve $V(k)$ looking for an "elbow" in order to determine the optimal number of components, denoted $k^* \in \{0, 1..., k_{\texttt{max}}\}$, to include in our model (i.e., in the vector $\boldsymbol{\theta}_{k^*}$).
We consider the construction of two straight lines passing through the points $(0, V(0))$, $(k, V(k))$ and $(k, V(k))$, $(k_{\texttt{max}}, 0)$ as shown in Figure 1(c) (where $k \in \{0, 1, ..., k_{\texttt{max}}\}$). These two straight lines form a piece-wise linear approximation of the curve $V(k)$. The goal is to minimize the area under this approximation. More specifically, as we can see in Figure 1(c), the area to minimize consist of three sub-areas: the two areas of two triangles (A_1 and A_3) and the area of a rectangle



(a) (b)

(c)

Figure 1: **(a)-(b)** Example of error curve $V(k)$ where **(a)** $k_{\texttt{max}} = K = 6$, **(b)** $k_{\texttt{max}} = 4$ and $K = 6$. **(c)** Construction with two straight lines and the areas A_1, A_2 and A_3.

in the middle (A_2). Namely, we have

$$A_1 = \frac{k \cdot (V(0) - V(k))}{2}, \tag{3}$$

$$A_2 = k \cdot V(k), \tag{4}$$

$$A_3 = \frac{(k_{\texttt{max}} - k) \cdot V(k)}{2}, \tag{5}$$

hence the definition of k^* is

$$\begin{aligned} k^* = \arg\min_k \{A_1 + A_2 + A_3\} &= \arg\min_k \left\{ \frac{k(V(0) - V(k))}{2} + kV(k) + \frac{(k_{\max} - k)V(k)}{2} \right\}, \\ &= \arg\min_k \left\{ \frac{kV(0)}{2} + \frac{k_{\max}V(k)}{2} \right\}, \\ &= \arg\min_k \left\{ \frac{V(k)}{V(0)} + \frac{k}{k_{\max}} \right\}, \quad \text{for } k = 1, ..., k_{\max}, \end{aligned} \tag{6}$$

where clearly we are assuming $V(0) \neq 0$ and $k_{\max} \neq 0$. Note that in the last equation we have multiplied by the constant factor $\frac{2}{V(0)k_{\max}}$. Now, multiplying the last expression by the constant value $V(0)$, we can equivalently write

$$k^* = \arg\min_k \left\{ V(k) + \frac{V(0)}{k_{\max}} k \right\}, \quad \text{for } k = 1, ..., k_{\max}. \tag{7}$$

It is important to remark that, since k belongs to a discrete and finite set, solving the optimization above is straightforward. In the case of multiple minima, e.g., having M different minima, $k_1^*, k_2^*, ..., k_M^*$, the user can choose the best solution (within the M possible one) according to some specific requirement depending on the specific application. Here, we suggest the most conservative choice, i.e.,

$$k^* = \max k_j^*, \quad \text{for } j = 1, ..., M. \tag{8}$$

3.2 Second equivalent derivation

The solution offered by the expressions (6)-(7) is equivalent to finding the k^* such that the difference between $V(k^*)$ and the value of the straight line (evaluated at k^*, as well) connecting the extreme points $(0, V(0))$ and $(k_{\max}, 0)$ is maximized, as depicted in Figure 2. More specifically, this straight line has an equation

$$v(k) = -\frac{V(0)}{k_{\max}} \cdot k + V(0), \tag{9}$$

hence, the difference that we maximize is the following:

$$\begin{aligned} d(k) &= v(k) - V(k), \\ &= -\frac{V(0)}{k_{\max}} \cdot k + V(0) - V(k), \\ &= V(0) - \left(\frac{V(0)}{k_{\max}} \cdot k + V(k) \right). \end{aligned} \tag{10}$$

Since $V(0)$ does not depend on k (i.e., it is a constant value), we can write

$$\begin{aligned} k^* = \arg\max_k d(k) &= \arg\max_k \left[V(0) - \left(\frac{V(0)}{k_{\max}} \cdot k + V(k) \right) \right], \\ &= \arg\max_k \left[- \left(\frac{V(0)}{k_{\max}} \cdot k + V(k) \right) \right], \\ &= \arg\min_k \left[\frac{V(0)}{k_{\max}} \cdot k + V(k) \right], \end{aligned} \tag{11}$$

which is exactly the expression in Eq. (7). Two additional and equivalent derivations are given in Appendix A and Appendix B. They are also represented graphically in Figures 5(a) and 5(b), respectively.

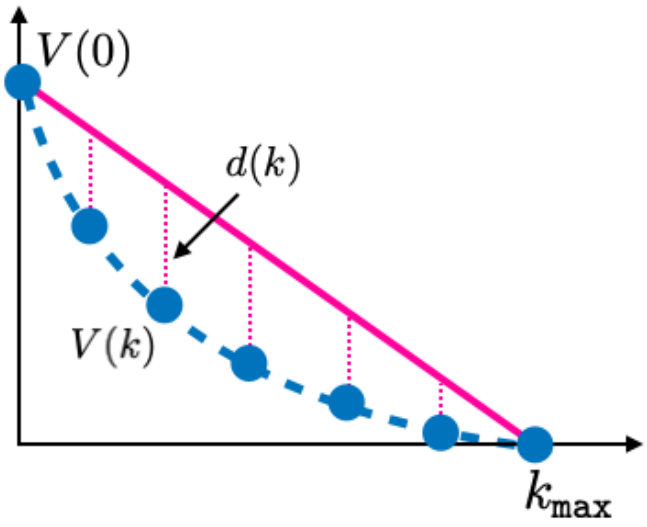


Figure 2: Graphical representation of alternative derivation in Section 3.2.

3.3 Relation with the information criteria

Recalling the expression in (7), i.e.,

$$k^* = \arg\min_k \left\{ V(k) + \frac{V(0)}{k_{\max}} k \right\},$$

here we show that this cost function can be interpreted in the same form of other information criteria, i.e., with a linear penalization of the model complexity,

$$C(k) = V(k) + \frac{V(0)}{k_{\max}} k, \tag{12}$$

$$= V(k) + \lambda k, \tag{13}$$

where we set $\lambda = \frac{V(0)}{k_{\max}}$. Note that Eq. (13) has exactly the same form of the cost function used in the information criteria like BIC and AIC, for instance, when $V(k)$ is defined as

$$V(k) = -2\log \ell_{\max}, \quad \text{with} \quad \ell_{\max} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k).$$

BIC corresponds to the choice $\lambda = \log(N)$ where N is the number of data in $\mathbf{y}$, and AIC corresponds to the choice $\lambda = 2$. Therefore, when $V(k) = -2\log \ell_{\max}$, UAED can be interpreted as an information criterion with the particular choice of $\lambda = \frac{V(0)}{k_{\max}}$. Table 1 summarizes this information.

Table 1: Different information criteria and the proposed UAED.

Criterion	**Choice of** λ
Bayesian-Schwarz Information Criterion (BIC) [13]	$\log N$
Akaike Information Criterion (AIC) [14]	2
Hannan-Quinn Information Criterion (HQIC) [15]	$\log(\log(N))$
Universal Automatic Elbow Detector (UAED)	$\frac{V(0)}{k_{\max}}$

3.4 Behavior of the proposed solution

Analyzing the involved parameters in the expression (7) or (12), we can highlight the following considerations about the behavior of the UAED method. We list some important points below:

• Observing Eq. (12), the penalization of the complexity of the model depends on $V(0)$ and $k_{\max}$: since $\lambda = \frac{V(0)}{k_{\max}}$ increasing $V(0)$ or decreasing $k_{\max}$, intensifies the penalty. This is a reasonable and desirable behavior. Indeed, increasing the value of $V(0)$ also increases the differences $V(0) - V(k)$, which means that the first components/variables have more impact in the fitting - the decay of $V(k)$ - so that fewer components/variables can form a reasonable model. Otherwise, decreasing the value of $V(0)$ means more variables have a similar impact in the decay of $V(k)$. Therefore, we should consider more components, in fact the slope of the penalization, $\lambda = \frac{V(0)}{k_{\max}}$, decreases in this case.
• Regarding $k_{\max}$, we can notice that a decrease of $k_{\max}$ means that fewer components/variables produces a drop in the curve $V(k)$. On the other hand, an increase in $k_{\max}$ means that the use of more variables causes a drop $V(k)$, so we should consider more components, indeed, the slope of the penalization, $\lambda = \frac{V(0)}{k_{\max}}$, decreases.
• **Scaling of the axes.** Looking the expression (7) or (12), it is possible to show that the solution

does not depend on different possible re-normalization of the axes, i.e., after a scaling of the axes (one of them, or both, even with different scales) the solution remains invariant, or is a scaled version of the previous one (with the same scaling factor). Indeed, considering a scaling on the vertical axis, i.e., assuming $V(k)' = aV(k)$ with $a > 0$, we have

$$k^* = \arg\min_k \left[aV(k) + \frac{aV(0)}{k_{\max}} k \right] = \arg\min_k \left[a\left(V(k) + \frac{V(0)}{k_{\max}} k \right) \right], \\ = \arg\min_k \left[V(k) + \frac{V(0)}{k_{\max}} k \right]. \tag{14}$$

Let now consider the case of scaling the horizontal axis, for instance, instead of having $k = 0, 1, 2..., k_{\max}$, we have $k' = 0, b, 2b, ..., bk_{\max}$ (i.e., $k' = bk$), and another error curve defined as $\widetilde{V}(k') = V(k'/b)$, where b is a positive integer. Hence we can write

$$(k')^* = \arg\min_{k'} \left[\widetilde{V}(k') + \frac{V(0)}{k'_{\max}} \cdot k' \right] = \arg\min_{k'} \left[V(k'/b) + \frac{V(0)}{k'_{\max}} \cdot k' \right], \\ = b \arg\min_k \left[V(bk/b) + \frac{V(0)}{bk_{\max}} \cdot bk \right], \\ = b \arg\min_k \left[V(k) + \frac{V(0)}{k_{\max}} \cdot k \right] = bk^*. \tag{15}$$

Namely, the new solution $(k')^* = bk^*$ is just a scaled version of the previous one, taking into account the same scaling factor b.

• **Shift of the axes.** Let us consider a $V(k)$ which fulfills that assumptions provided above (to be non-increasing and $\min V(k) = V(K) = 0$). A shift of the vertical axis, i.e., $\widetilde{V}(k) = V(k) + c$ where $c \in \mathbb{R}$ does not affect the results since, by assumption, we have always to consider an error curve such that

$$V'(k) = \widetilde{V}(k) - \min \widetilde{V}(k) = V(k) + c - c = V(k).$$

then $V'(k)$ satisfies again $\min V'(k) = 0$. A shift on the horizontal axis only produces the same shift in the solution k^*. Furthermore, given the considerations in Appendices A and B, we can see that the solution k^* is invariant even if the axes are exchanged.

• Here, we describe two ideal scenarios and discuss the behavior of UAED. For clarity in the exposition, let us consider as an example a variable selection problem. First of all, we consider the case that all the input variables are equally important for predicting the output variable. Then, we have $k_{\max} = K$, and the error curve $V(k)$ is a straight line connecting the points $(0, V(0))$ and $(k_{\max}, V(k_{\max}))$ (i.e., each variable has the same impact to the error decay). In this scenario, we have $k_{\max} = K$, and UAED provides $M = k_{\max} + 1$ different minima $k_1^* = 0, k_2^* = 1, k_3^* = 2, ..., k_M^* = k_{\max}$ (i.e., M possible candidates to be an elbow). Thus, the UAED solution is given by Eq. (8), i.e., $k^* = \max k_j^* = k_{\max}$. Namely, UAED suggests to select all the variables, that is the correct solution.

On the other hand, let us consider now a scenario where all the input variables are independent from the output variable. In this case, $V(k)$ is a constant function, i.e., $V(k) = V(0)$ for all k and, as a consequence, $k_{\max} = 0$. Hence, since $k_{\max} = 0$, UAED gives $k^* = 0$, which is the correct

solution (i.e., no variables should be selected). Thus, in both scenarios, UAED provides the correct results.

4 Experiments with synthetic and real data

In this section, we test the UAED in six real-world applications. In each experiment, we consider a different function $V(k)$, in order to show the vast range of applicability of UAED. Sections 4.1, 4.2, 4.3 consider synthetic data in a clustering example and two order selection problems. In Sections 4.4, 4.5 and 4.6, the experiments involve the analysis of real datasets: the first one is a variable selection in a regression problem with soundscape emotion data, whereas the second and third ones involve classification problems with real biomedical datasets. We compare the performance of UAED with BIC, AIC, and other information criteria described in the literature, in those examples where these schemes can be also applied.

4.1 Clustering

We consider 2500 simulated data from a mixture of 5 bidimensional Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_1 = [3, 0]$, $\boldsymbol{\Sigma}_1 = [0.3, 0; 0, 2]$, $\boldsymbol{\mu}_2 = [14, 5]$, $\boldsymbol{\Sigma}_2 = [1.5, 0.7; 0.7, 1.5]$;, $\boldsymbol{\mu}_3 = [-5, -10]$, $\boldsymbol{\Sigma}_3 = [1.5, 0.7; 0.7, 1.5]$, $\boldsymbol{\mu}_4 = [10, -10]$, $\boldsymbol{\Sigma}_4 = [1.5, 0; 0, 1.5]$;, and $\boldsymbol{\mu}_5 = [-5, 5]$, $\boldsymbol{\Sigma}_5 = [1, -0.8; -0.8, 1]$. Figure 3(a) shows these data points.
We assume $V(k) = \log\left[\sum_{j=1}^{k+1} \text{var}(j)\right]$, where $\text{var}(j)$ represents the inner variance of the j-th cluster, as shown in Figure 3(b). Each value of $\text{var}(j)$ has been computed and averaged over 200 runs, applying a k-means algorithm. In this setting, the total number of clusters is given by $k+1$ (i.e., $k = 0$ corresponds to a unique, single cluster). We assume $K = 50$ as the maximum number of possible clusters.
It is important to remark that, with this choice of $V(k)$, the other information criteria cannot be directly applied[3]. We apply UAED and obtain $k^* + 1 = 5$ as the chosen number of clusters, which is the correct solution.

4.2 Order selection of a polynomial function in a regression problem

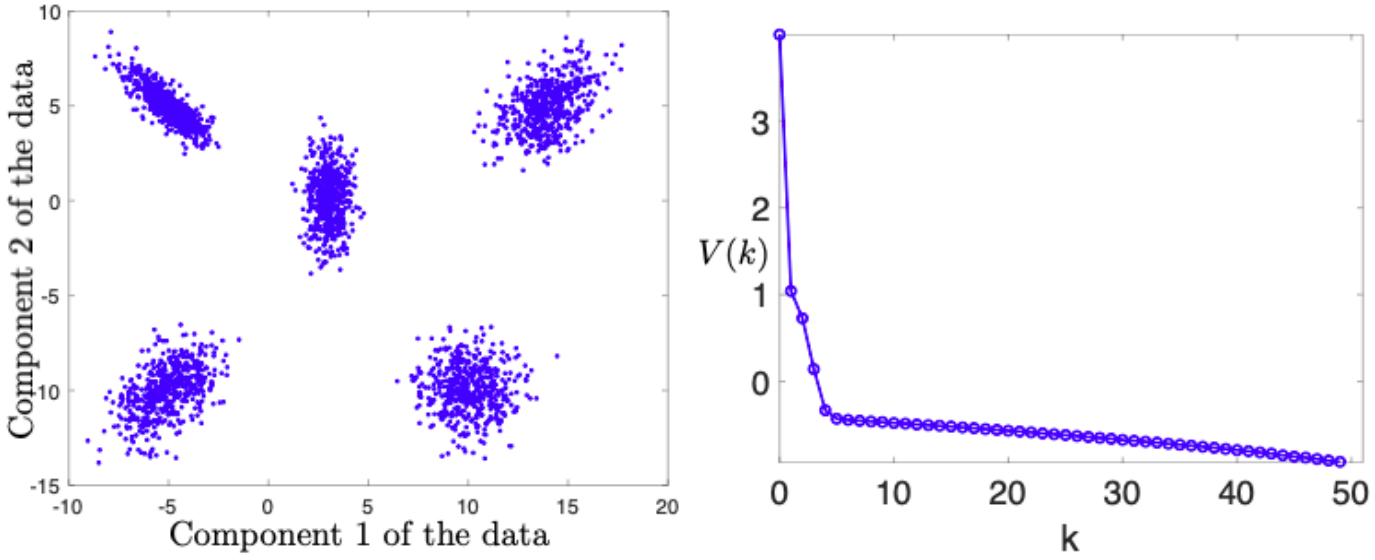
We generate a dataset of $N = 100$ pairs $\{x_n, y_n\}_{n=1}^N$, where both inputs x_n's and outputs y_n's are scalar values, considering the following observation model,

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + ...\theta_k x_n^k + \epsilon_n, \tag{16}$$

where $\boldsymbol{\theta}_k = [\theta_0, \theta_1, ..., \theta_k]^\top$, ϵ_n is Gaussian noise with zero mean and variance $\sigma_\epsilon^2 = 1$. The dataset has been generated with a polynomial function of order $k^* = 4$, and with the coefficients

$$\theta_0 = 4.05,\ \theta_1 = -2.025,\ \theta_2 = -2.225,\ \theta_3 = 0.1,\ \theta_4 = 0.1.$$

[3]The information criteria require the choice of the error curve of type $V(k) = -2\log \ell_{\texttt{max}}$ where $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ and, as a consequence, a definition of a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$.



(a) (b)

Figure 3: **(a)** Artificial Data of the clustering experiment. **(b)** The function $V(k) = \log\left[\sum_{j=1}^{k+1} \text{var}(j)\right]$ where $\text{var}(j)$ represents the inner variance in the j-th cluster. Note that $k = 0$ corresponds to a unique, single cluster.

In this experiment, we consider $V(k) = -2\log(\ell_{\texttt{max}})$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, where $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by Eq. (16), in order to allow the comparison with other schemes in the literature, as shown in Table 1. The corresponding function $V(k)$ is shown in Figure 4(a).

Applying BIC, AIC and HQIC we obtain the suggested order of polynomial is 4, 6, and 10, respectively. With the proposed UAED method, we obtain the suggested order is 4, which is the correct order of the underlying polynomial function. Therefore, in this experiment, BIC and UAED provide the correct answer.

4.3 Order selection in an auto-regressive model

We generate a dataset of T pairs $\{t, y_t\}_{t=1}^{T}$, where t is an integer temporal index and the signal y_t is a scalar value for each t. We consider the following auto-regressive model,

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + ...\theta_k y_{t-k} + \epsilon_t, \quad \text{for } t = 1, ..., T, \tag{17}$$

where $\boldsymbol{\theta}_k = [\theta_1, \theta_2, ..., \theta_k]^\top$, ϵ_t is Gaussian noise with zero mean and variance σ_ϵ^2. The order of the model is k.
In this example, we consider two possible values of the order of the model, $k^* \in \{3, 5\}$ and we have set $k_{\max} = K = 100$. We generate the data $\mathbf{y} = [y_1, ..., y_T]$ according to the model (17) considering the following coefficients

$$\begin{aligned} &\theta_1 = 1,\ \theta_2 = -0.7408,\ \theta_2 = 0.5488,\ \theta_3 = 0.1, && \text{for } k = 3, \\ &\theta_1 = 1,\ \theta_2 = -0.7408,\ \theta_2 = 0.5488,\ \theta_3 = 0.1,\ \theta_4 = -0.4066,\ \theta_5 = 0.3012, && \text{for } k = 5, \end{aligned}$$

where we have used the formula $\theta_i = (-1)^{i-1}\exp\{-0.3(i-1)\}$, which ensures that the system in Eq. (17) is stable. We test different number of data, $T \in \{200, 2000\}$, and different levels of noise, with standard deviation $\sigma_\epsilon \in \{0.5, 1, 2\}$. In all scenarios, we average the results with 10^3 independent runs where, for each simulation, we generate a new time series of T data according to Eq. (17).
Moreover, in all the simulations, we consider $V(k) = -2\log(\ell_{\texttt{max}})$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, where $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by Eq. (17), in order to allow the comparison with other schemes in the literature, as shown in Table 1. The value of $V(0)$ is obtained considering the log of the power of the signal y_t (for further details, see the alternative BIC computation with Gaussian noise in [27, page 375]).

Remark. Note this example satisfies all the assumptions in the derivation of BIC. Therefore, we expect excellent results of BIC. We desire to test the performance of UAED, even in this scenario where BIC is clearly favored.

The decided orders of the model given by AIC, BIC, HQIC, and UAED in the different scenarios and simulations are given from Figure 7 to Figure 18. All these figures show the histograms of the decided orders given in each run by the different methods. Namely, the bars represent the percentages of times that an index is chosen as order of the model. Each figure corresponds to a specific scenario with a true order of the model $k^* \in \{3, 5\}$, a certain number of data $T \in \{200, 2000\}$, and a noise level $\sigma_\epsilon \in \{0.5, 1, 2\}$. Table 2 summarizes the results, showing the correct-decision rate $p_A \in [0, 1]$, and using the following ranking for the performance:

- **Best:** for the method that provides the highest correct-decision rate p_A.
- **Excellent:** for the methods with $p_A \geq 0.95$.
- **Good:** for the methods with $0.80 \leq p_A < 0.95$.
- **Fair:** for the methods with $0.50 \leq p_A < 0.80$.
- **Poor:** for the methods with $0.20 \leq p_A < 0.50$.
- **Bad:** for the methods with $0.10 \leq p_A < 0.20$.
- **Very bad:** for the methods with $p_A < 0.10$.

We can observe that BIC and UAED provide the best performance. As remarked above, this example is particularly favorable for BIC but, even in this numerical experiment, UAED provides very good results and, in some scenarios, even the best results. Indeed, as we can observe from Table 2, UAED gives the best results with $T = 2000$ (i.e., when we have more data), regardless of the noise level and the true order of the model. It is important to highlight that well-known methods, such as AIC and HQIC, provide very poor performance. Clearly, additional information regarding the dispersion of the wrong decisions can be observed in Figures from 7 to 18. This experiment shows clearly that UAED is a competitive and robust methodology.

Table 2: Summary of the results in the example in Section 4.3.

Scenario			Method				
k	σ_ϵ	T	**UAED**	**BIC**	**AIC**	**HQIC**	**Fig.**
3	0.5	200	Good, $p_A \approx 0.82$	**Best**, $p_A \approx 0.97$	Bad, $p_A \approx 0.13$	Very bad, $p_A \approx 0.01$	7
		2000	**Best**, $p_A = 1$	**Best**, $p_A = 1$	Bad, $p_A \approx 0.1$	Bad, $p_A \approx 0.1$	8
	1	200	Good, $p_A \approx 0.82$	**Best**, $p_A \approx 0.97$	Bad, $p_A \approx 0.13$	Very bad, $p_A \approx 0.01$	9
		2000	**Best**, $p_A = 1$	Excellent, $p_A \approx 0.98$	Bad, $p_A \approx 0.13$	Bad, $p_A \approx 0.17$	10
	2	200	Good, $p_A \approx 0.81$	**Best**, $p_A \approx 0.96$	Bad, $p_A \approx 0.11$	Very bad, $p_A \approx 0.01$	11
		2000	**Best**, $p_A = 1$	Excellent, $p_A \approx 0.97$	Very bad, $p_A \approx 0.05$	Very bad, $p_A \approx 0.06$	12
5	0.5	200	Good, $p_A \approx 0.80$	**Best**, $p_A \approx 0.90$	Bad, $p_A \approx 0.14$	Very bad, $p_A \approx 0.01$	13
		2000	**Best**, $p_A = 1$	**Best**, $p_A = 1$	Very bad, $p_A \approx 0.07$	Bad, $p_A \approx 0.1$	14
	1	200	Good, $p_A \approx 0.80$	**Best**, $p_A \approx 0.90$	Bad, $p_A \approx 0.17$	Very bad, $p_A \approx 0.03$	15
		2000	**Best**, $p_A = 1$	Excellent, $p_A \approx 0.98$	Bad, $p_A \approx 0.10$	Bad, $p_A \approx 0.13$	16
	2	200	Good, $p_A \approx 0.80$	**Best**, $p_A \approx 0.90$	Bad, $p_A \approx 0.13$	Very bad, $p_A \approx 0.01$	17
		2000	**Best**, $p_A = 1$	Excellent, $p_A \approx 0.98$	Very bad, $p_A \approx 0.09$	Bad, $p_A \approx 0.11$	18

4.4 Variable selection in a regression problem with real data

In this section, we present a feature selection problem for regression. Moreover, we consider real data. More specifically, a dataset of N pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$ is given, where each input vector $\mathbf{x}_n = [x_{n,1}, ..., x_{n,K}]$ is formed by K variables, and the outputs y_n's are scalar values. We assume $K \leq N$ and a linear observation model,

$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + ...\theta_K x_{n,K} + \epsilon_n, \tag{18}$$

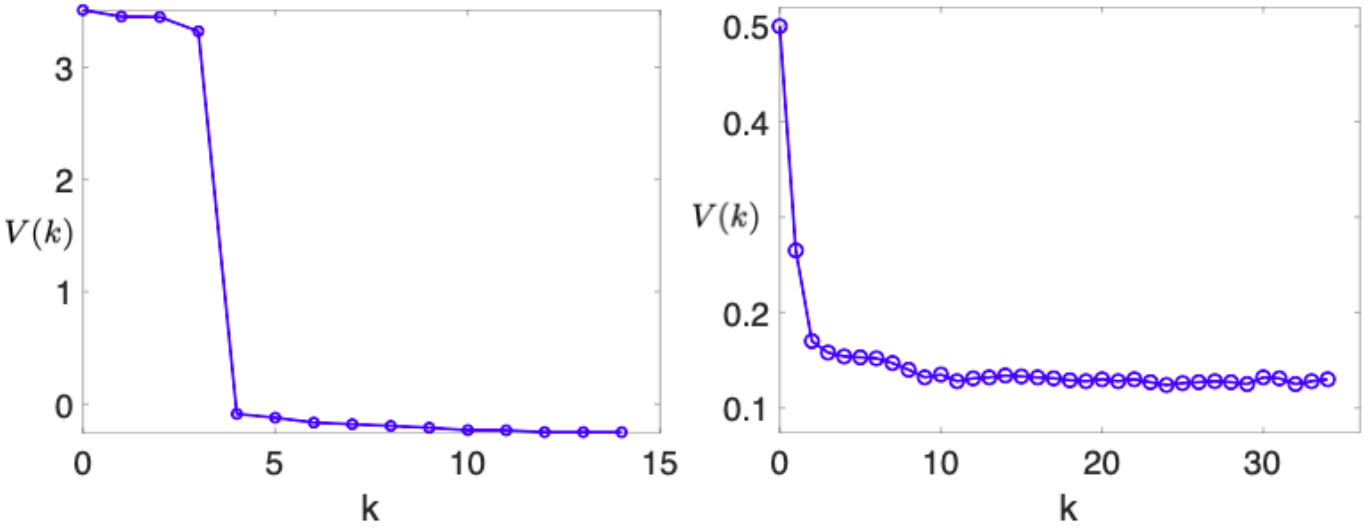
where ϵ_n is Gaussian noise with zero mean and variance σ_ϵ^2, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. In the real dataset studied in [28], there are $K = 122$ features and $N = 1214$ number of data points. The output represents the variable defined as "arousal" in [28].

In order to allow the comparison with other schemes in the literature, here we can set $V(k) = -2\log(\ell_{\texttt{max}})$ where $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, after ranking the 122 variables as in [28]. Clearly, The likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by Eq. (18). Therefore, in this experiment, we can compare UAED again with other information criterion measures in the literature, some of them are given in Table 1. BIC suggests a model with 17 variables, AIC chooses 44 variables, and HQIC selects 41 variables. The proposed UAED suggests considering only 11 variables. Therefore, the UAED suggestions is closer to the results given in other previous studies and to experts' recommendations in the literature, e.g., [28].

4.5 Variable selection in a classification problem with a nonalcoholic fatty liver disease real dataset

Biomedical applications are nowadays extremely relevant [29, 30]. The authors in [31] analyze the most important features for predicting patients at risk of developing nonalcoholic fatty liver disease. The authors collected data from 1525 patients who attended the Cardiovascular Risk Unit of Mostoles University Hospital (Madrid, Spain) from 2005 to 2021, and use a random forest (RF)

algorithm to classify patients and rank the input features, in order to select the most important one. They found that 4 features were the most relevant according to the ranking and the experts' opinions: (a) insulin resistance, (b) ferritin, (c) serum levels of insulin, and (d) triglycerides.
In this experiment, we set $V(k) = 1 - \text{accuracy}(k)$ that is given in Figure 4(b), after ranking the 35 features [31]. Note that $V(0) = 0.5$ representing a completely random binary classification. It is important to remark that, with this choice of $V(k)$, the other information criteria cannot be employed. The application of UAED suggests to select 4 variables which is exactly the result of the paper [31], obtained using a cross-validation approach, and supported by the experts' opinions.



(a) (b)

Figure 4: **(a)** The corresponding curve $V(k) = -2\log \ell_{\texttt{max}}$ (with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$) in Section 4.2; **(b)** The curve $V(k) = 1 - \text{accuracy}(k)$ of the experiment in Section 4.5.

4.6 Variable selection in a real dataset of ventricular fibrillation

The authors in [32] addressed the early recognition of ventricular fibrillation (VF). In their study, they considered a feature selection based on the percentage of the balanced error rate (BER), which plays the role of $V(k)$. See Equation 5 at page 7 of [32] and Figure 4-(a)-right (green curve) at page 10 of [32] for further information regarding this $V(k)$ curve. We consider as $V(0)$ the first point provided in this curve.
It is important to remark that again, with this choice of $V(k)$ based on the balanced error rate, the other information criteria cannot be employed. The application of UAED suggests to choose 6 features which is very close to the suggestion of the authors in [32] that was 5 features. The authors in [32] reached this conclusion after using much more complex analyses and comparing with expert's opinions. In this sense, UAED is also a simpler approach to employ.

5 Conclusions

A novel automatic elbow detector for model selection purposes has been introduced. The proposed UAED scheme is inspired by the concept of the maximum "area under the curve" (AUC) in receiver operator characteristic (ROC) curves. The contributions of the work can be divided into four main parts: (a) motivation and derivations of the proposed method, (b) analysis of the behavior in ideal scenarios and its properties, (c) test and comparison with several numerical simulations, and (d) a related Matlab code.

Four different geometrical derivations of UAED have been provided. The first three derivations are based on the vertical, horizontal and Euclidean distance of the $V(k)$ curve (with $k \in \mathbb{N}$) with respect to a decreasing straight line, which corresponds to the ideal decay when each component has equal importance. The last derivation shows an additional property of the proposed solution, in a generalized framework with k is a continuous variable instead of an integer (as in the rest of the work). This property can play a relevant role in future works regarding this research line.

Furthermore, we have analyzed other features and properties of UAED, as the invariance on scaling the axes and the behavior in ideal scenarios. The relationships and differences with several information criteria (already given in the literature) have been described and highlighted. More specifically, UAED can be also considered as an information criterion with the choice of the slope of the model penalty as $\lambda = \frac{V(0)}{k_{\texttt{max}}}$. However, the $V(k)$ curve in UAED can be also chosen differently from the usual definition $V(k) = -2 \log \ell_{\texttt{max}}$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$, which is required in the other information criteria. Hence, it is important to remark that the proposed procedure has a much wider range of application with respect to the other schemes in the literature, as also clarified by the numerical experiments.

Six experiments and comparisons show the benefits of the proposed UAED scheme. We have considered different examples in clustering, order selection, and variable selection within regression and classification problems. Moreover, three of these six experiments involve real datasets. We have compared with the most relevant information criteria in the literature (AIC, BIC and HQIC). Even in scenarios that are much favorable for one of them (i.e., AIC, BIC or HQIC), the proposed UAED method provides very competitive results.

Acknowledgement

The work was partially supported by the Young Researchers R&D Project, ref. num. F861 (AUTO-BA-GRAPH) funded by Community of Madrid and Rey Juan Carlos University, and by Agencia Estatal de Investigación AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00).

References

[1] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, pp. 1–58, 2006.

[2] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *SIAM Review (SIREV)*, vol. 65, no. 1, pp. 3–58, 2023.

[3] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.

[4] P. Stoica, X. Shang, and Y. Cheng, "The Monte-Carlo sampling approach to model selection: A primer [lecture notes]," *IEEE Signal Processing Magazine*, vol. 39, no. 5, pp. 85–92, 2022.

[5] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[6] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[7] Y. Ma and L. Zhu, "A review on dimension reduction," *International Statistical Review*, vol. 81, no. 1, pp. 134–150, 2013.

[8] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[9] P. Stoica and Y. Selén, "Cross-validation rules for order estimation," *Digital Signal Processing*, vol. 14, pp. 355–371, 2004.

[10] E. Fong and C. Holmes, "On the marginal likelihood and cross-validation," *Biometrika*, vol. 107, no. 2, pp. 489–496, 2020.

[11] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[12] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[13] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. B*, vol. 64, pp. 583–616, 2002.

[15] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[16] F. Llorente, L. Martino, E. Curbelo, J. Lopez-Santiago, and D. Delgado, "On the safe use of prior densities for bayesian model selection," *WIREs Computational Statistics*, p. e1595, 2022.

[17] M. Drton and M. Plummer, "A Bayesian information criterion for singular models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 79, no. 2, pp. 323–380, 2017.

[18] A. Mariani, A. Giorgetti, and M. Chiani, "Model order selection based on information theoretic criteria: Design of the penalty," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2779–2789, 2015.

[19] L. Martino, R. S. Millan-Castillo, and E. Morgado, "Spectral information criterion for automatic elbow detection," *viXra:2209.0123*, pp. 1–20, 2022.

[20] M. Kobayashi and S. Sakata, "Mallows' cp criterion and unbiasedness of model selection," *Journal of Econometrics*, vol. 45, no. 3, pp. 385–395, 1990.

[21] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE transactions on Information Theory*, vol. 44, no. 5, pp. 1926–1940, 1998.

[22] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.

[23] C. M. Pooley and G. Marion, "Bayesian model evidence as a practical alternative to deviance information criterion," *Royal Society Open Science*, vol. 5, no. 3, pp. 1–16, 2018.

[24] M. Efroymson, "Multiple regression analysis," *Mathematical methods for digital computers*, pp. 191–203, 1960.

[25] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, pp. 1–49, 1976.

[26] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[27] M. Priestley, *Spectral Analysis and Time Series*, A. Press, Ed. Wiley Series in Computational Statistics, 1981.

[28] R. San Millán-Castillo, L. Martino, E. Morgado, and F. Llorente, "An exhaustive variable selection study for linear models of soundscape emotions: Rankings and Gibbs analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2460–2474, 2022.

[29] K. Ali, Z. A. Shaikh, and A. A. Khan, "Multiclass skin cancer classification using efficientnets-a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[30] V. Laghari, A. A.and Estrela and S. Yin, "How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging," *Curr Med Imaging*, pp. 1–20, 2022.

[31] R. Gárcia-Carretero, R. Holgado-Cuadrado, and O. Barquero-Pérez, "Assessment of classification models and relevant features on nonalcoholic steatohepatitis using random forest," *Entropy*, vol. 23, no. 6, 2021.

[32] C. Figuera, U. Irusta, E. Morgado, E. Aramendi, U. Ayala, L. Wik, J. Kramer-Johansen, T. Eftestol, and F. Alonso-Atienza, "Machine learning techniques for the detection of shockable rhythms in automated external defibrillators," *PLOS ONE*, vol. 11, no. 7, pp. 1–17, 2016.

A Third alternative derivation

Let us consider Figure 5(a). First of all, we must find the value k' such that the straight line, connecting the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, which reaches the value $V(k)$ (where $k \neq k'$, and more precisely $k \leq k'$). Namely, we desire to obtain k' such that

$$V(k) = -\frac{V(0)}{k_{\texttt{max}}} \cdot k' + V(0), \tag{19}$$

hence

$$k' = -\frac{k_{\texttt{max}}}{V(0)} \left[V(k) - V(0)\right]. \tag{20}$$

Now, we could also consider to maximize the following difference

$$r(k) = k' - k = -\frac{k_{\texttt{max}}}{V(0)} \left[V(k) - V(0)\right] - k, \tag{21}$$

and the elbow is defined as

$$\begin{aligned} k^* = \arg\max_k \; r(k) &= \arg\max_k \left[-\frac{k_{\texttt{max}}}{V(0)} V(k) - k\right], \\ &= \arg\min_k \left[\frac{k_{\texttt{max}}}{V(0)} V(k) + k\right], \\ &= \arg\min_k \left[V(k) + \frac{V(0)}{k_{\texttt{max}}} k\right], \end{aligned} \tag{22}$$

where in the last we have multiplied by the constant $V(0)$. Note that Eq. (22) is exactly the same optimization problem (i.e., with the same cost function) in Sections 3.1-3.2.

B Fourth alternative derivation

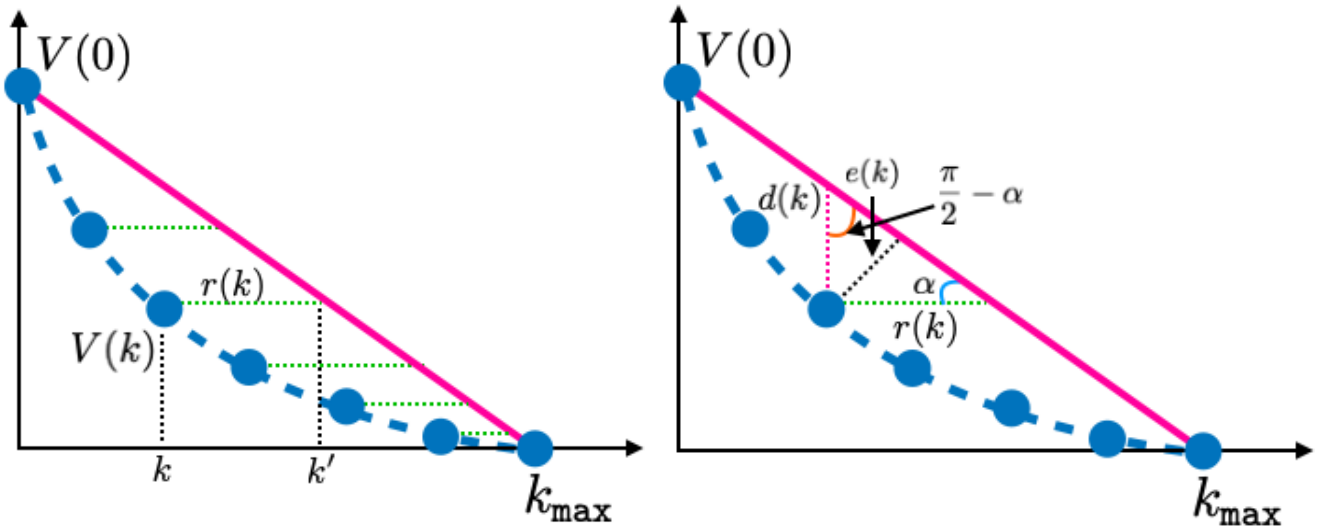
One could also consider the Euclidean distance $e(k)$ between the points in the curve $V(k)$ and the straight line connecting the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, as depicted in Figure 5(b). Observing this figure, we can notice that

$$e(k) = d(k) \sin(\pi/2 - \alpha) = d(k) \cos\alpha = r(k) \sin\alpha, \tag{23}$$

where α is the angle shown in Figure 5(b). Since the angle α is constant, then we can write

$$\begin{aligned} k^* = \arg\max_k [e(k)] &= \arg\max_k [d(k)\cos\alpha] = \arg\max_k [d(k)]\,, \\ &= \arg\max_k [r(k)\sin\alpha] = \arg\max_k [r(k)]\,. \end{aligned} \tag{24}$$

Therefore, maximizing $e(k)$ is equivalent to maximize $d(k)$ or $r(k)$.



(a) (b)

Figure 5: **(a)** Graphical representation of the other alternative derivation in Appendix A. **(b)** Graphical representation of derivation based on the Euclidean distance $e(k)$.

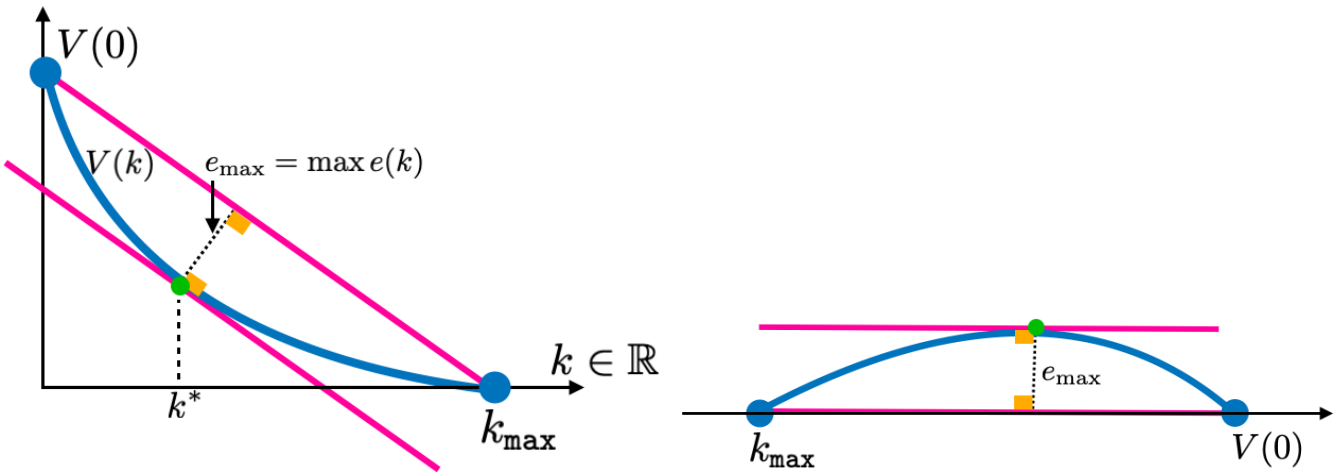
C An additional property

So far, we have considered that k is a discrete variable. If we assume that k can be a continuous parameter, i.e., $k \in \mathbb{R}$, we can obtain an additional property of the UAED solution. For the sake of simplicity, we also assume that $V(k)$ is convex. This condition can be relaxed, and the following discussion can be generalized for more general curves $V(k)$.
With these assumptions, it is possible to show that the derivative $\dot{V}(k) = \frac{dV}{dk}$ evaluated at the optimal value k^* (namely, the solution obtained by UAED) is equal to the slope of the straight line passing through the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, i.e.,

$$\dot{V}(k^*) = -\frac{V(0)}{k_{\texttt{max}}}.$$

Indeed, we know from App. B that $k^* = \arg\max_k [e(k)]$ where $e(k)$ is the Euclidean distance between the points $(k, V(k))$ and the straight line passing through $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, as shown again in Figure 6(a). Applying a rotation to the plot in Figure 6(a) and obtaining a new

horizontal axis such that it coincides with the straight line passing through $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, we can observe that the optimal point must be a stationary point (i.e., with null derivative) in this new coordinate system (by construction). This is depicted in Figure 6(b). Therefore, the tangent straight line at the green point in Figure 6(b) is parallel to the new horizontal axis. Inverting the rotation (namely, coming back the k-axis of Figure 6(a)), the previous consideration is equivalent to say that the derivative $\dot{V}(k) = \frac{dV}{dk}$ evaluated at the optimal value k^* must be $\dot{V}(k^*) = -\frac{V(0)}{k_{\texttt{max}}}$, i.e., the tangent straight line at the green point is parallel to the straight line passing through the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$.



(a) (b)

Figure 6: An additional property of UAED, when we consider k as a continuous parameter, i.e., $k \in \mathbb{R}$, instead of a discrete variable. The derivative $\dot{V}(k) = \frac{dV}{dk}$ evaluated at the optimal value k^* (obtained by UAED) is equal to the slope of the straight line passing through the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, i.e., $\dot{V}(k^*) = -\frac{V(0)}{k_{\texttt{max}}}$.

D Possible extension

We have already shown that the resulting expression in Eq. (6) provides good performance and is endowed with valuable behaviors.

However, we can add more flexibility that can be useful in the scenarios in which the researchers and/or practitioners determine that the benefit of reducing the error is greater than the benefit of reducing the number of considered variables or vice versa. We define an additional parameter $\alpha \in [0, 1]$, and consider the modified definition of the optimal k as

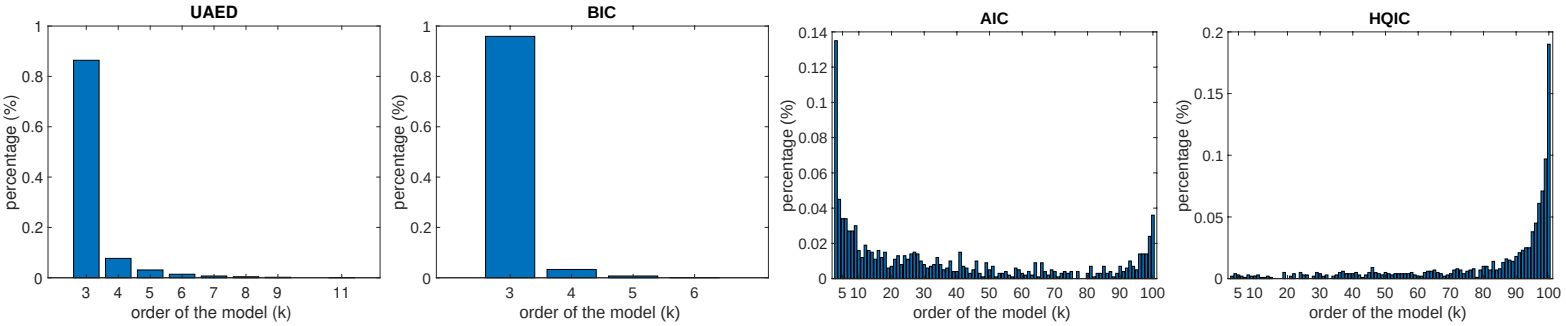
$$k^* = \arg\min_k \left[\alpha \cdot \frac{V(k)}{V(0)} + (1 - \alpha) \cdot \frac{k}{k_{\texttt{max}}} \right]. \tag{25}$$

Note that $\alpha = 0$ implies that all priority is to reduce the number of considered variables ($k^* = 0$), that $\alpha = 1$ implies that all priority is to reduce the resulting error (so that $k^* = k_{\texttt{max}}$). For $\alpha = 0.5$,

we come back to the definition in Eq. (7). As we have previously done in Section 3, can rewrite Eq. (25) as

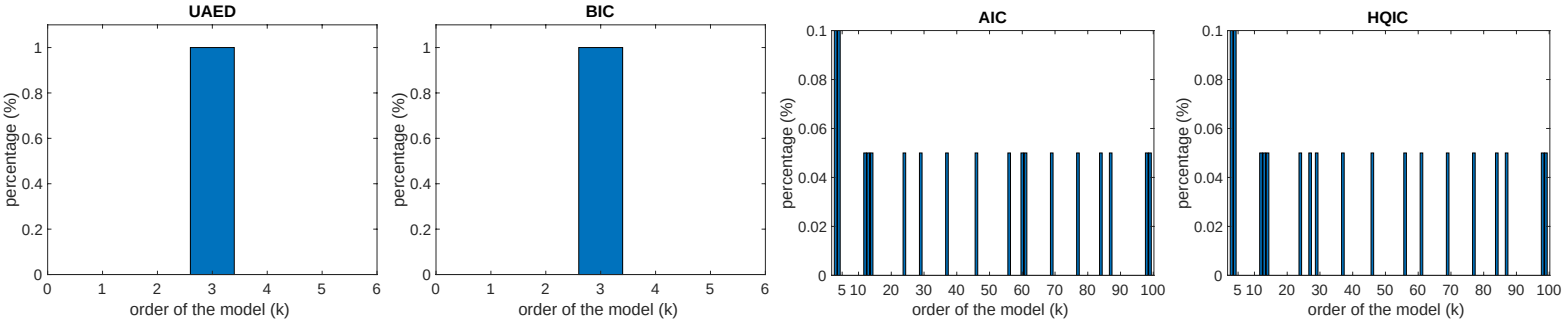
$$k^* = \arg\min_k \left[V(k) + \underbrace{\left(\frac{1-\alpha}{\alpha} \frac{V(0)}{k_{\texttt{max}}} \right)}_{\lambda} \cdot k \right], \\ = \arg\min_k \left[V(k) + \lambda k \right], \tag{26}$$

having the form of an information criterion with a different choice of λ which involves now the parameter α, as well.



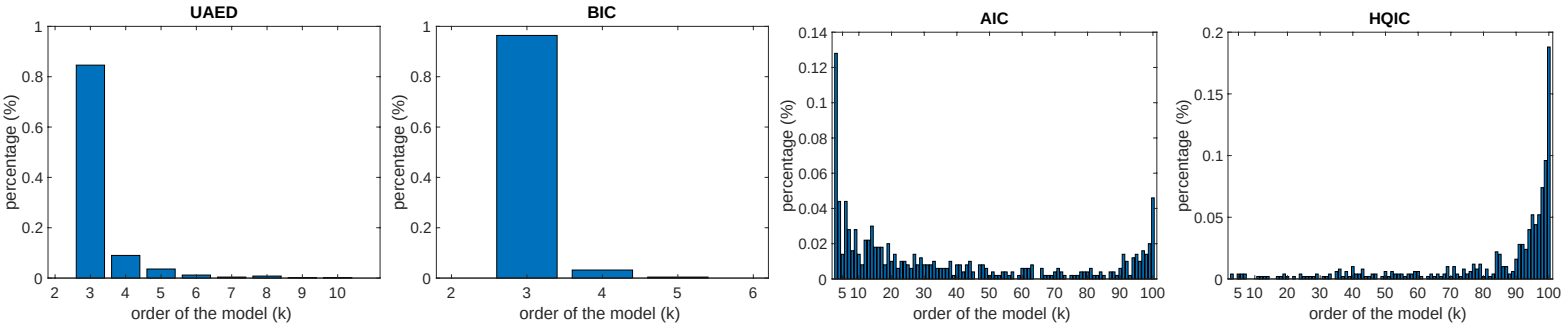
(a) (b) (c) (d)

Figure 7: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 0.5$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



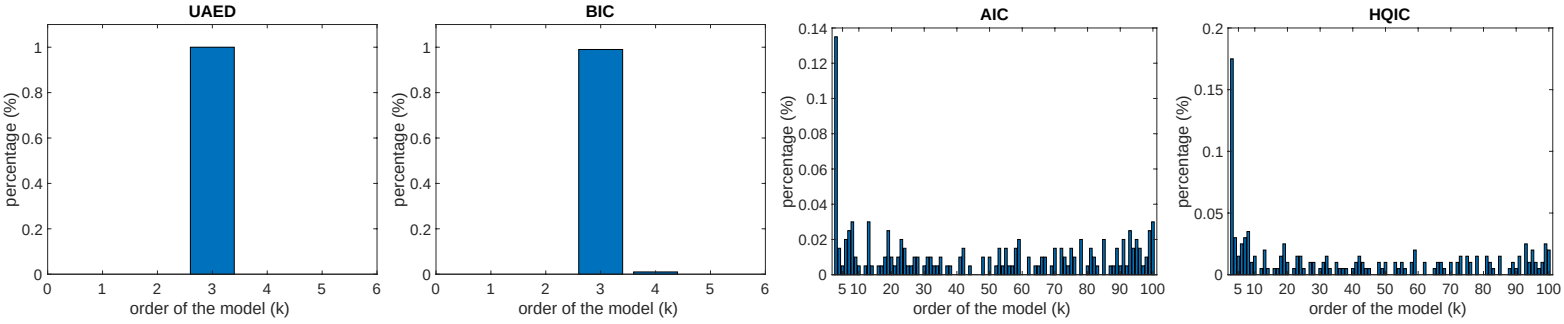
(a) (b) (c) (d)

Figure 8: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 0.5$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



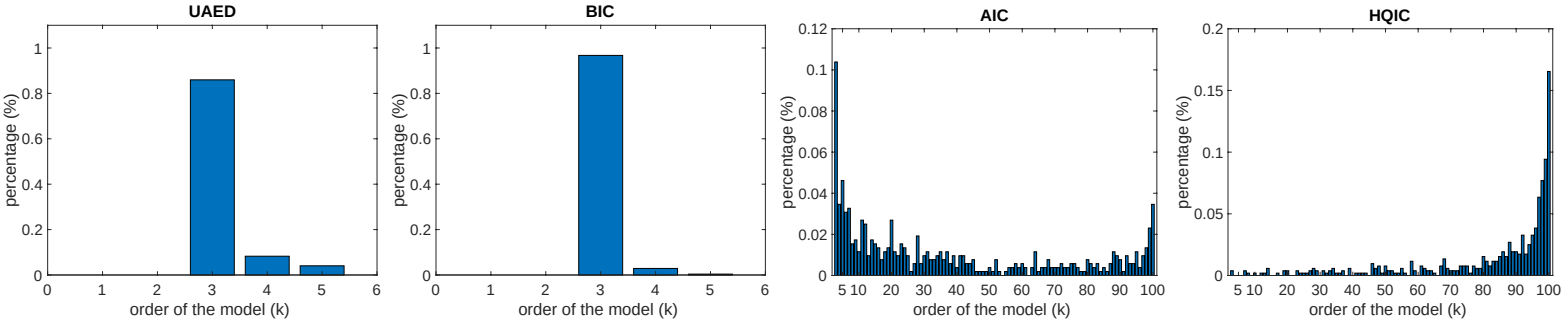
(a) (b) (c) (d)

Figure 9: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 1$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



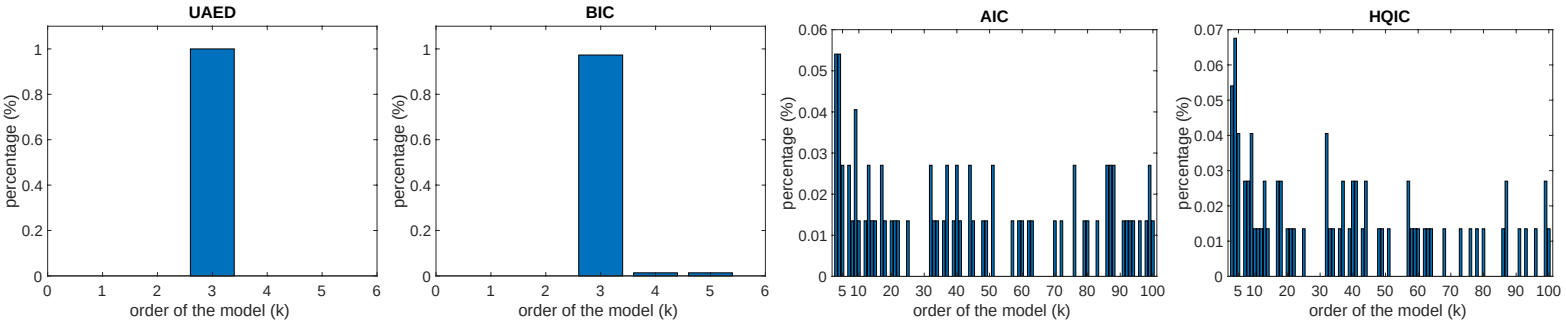
(a) (b) (c) (d)

Figure 10: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 1$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



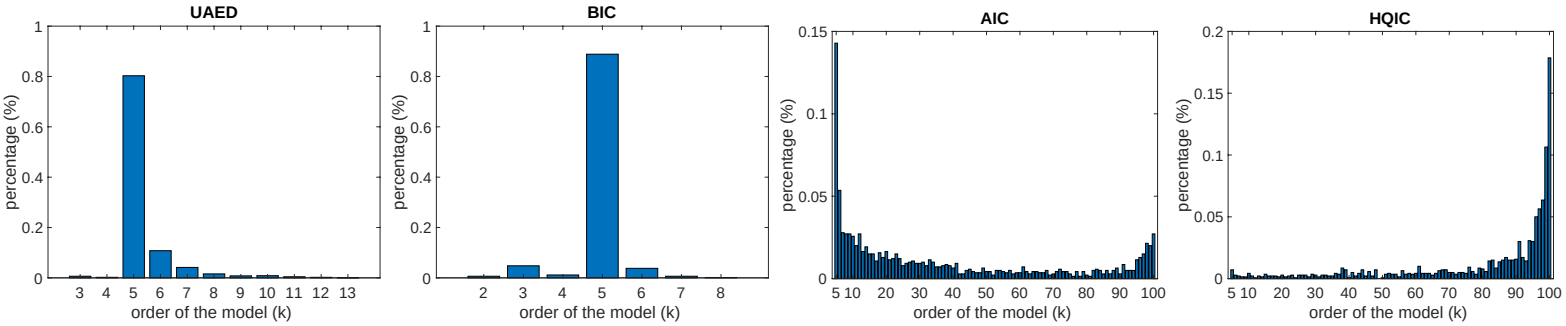
(a) (b) (c) (d)

Figure 11: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 2$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



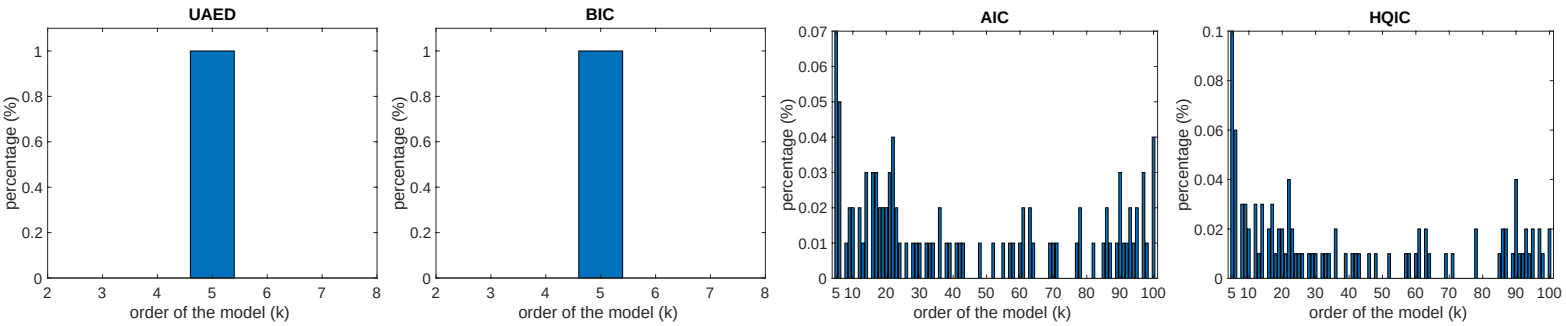
(a) (b) (c) (d)

Figure 12: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 3$, the standard deviation of the noise is $\sigma_\epsilon = 2$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



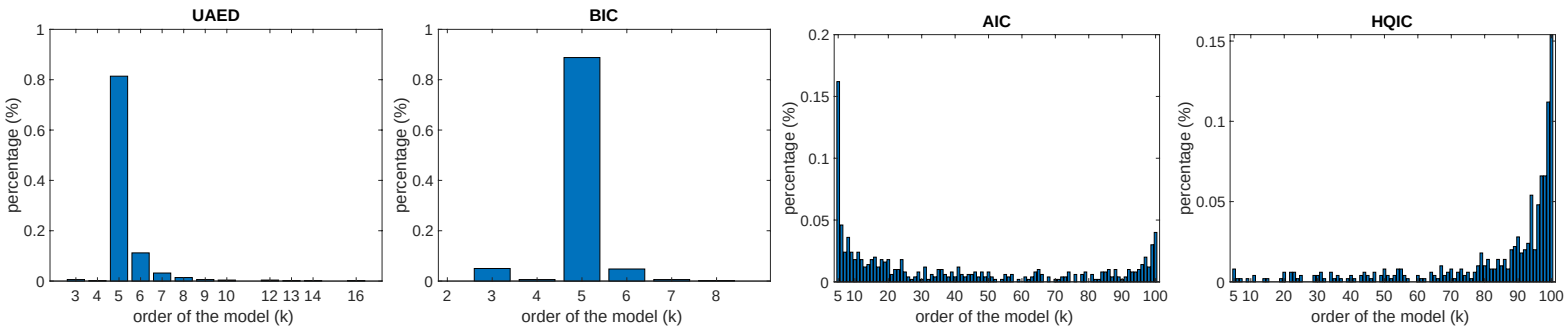
(a) (b) (c) (d)

Figure 13: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 0.5$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



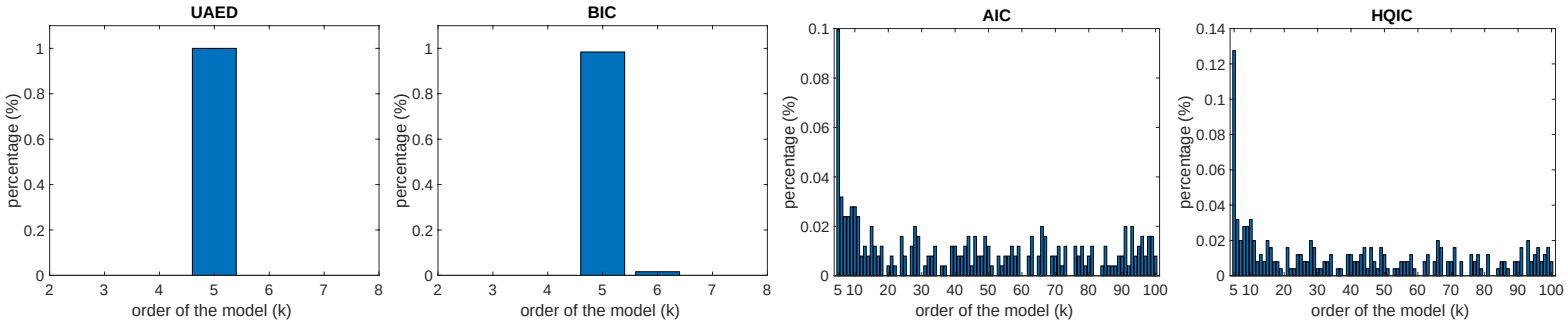
(a) (b) (c) (d)

Figure 14: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 0.5$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



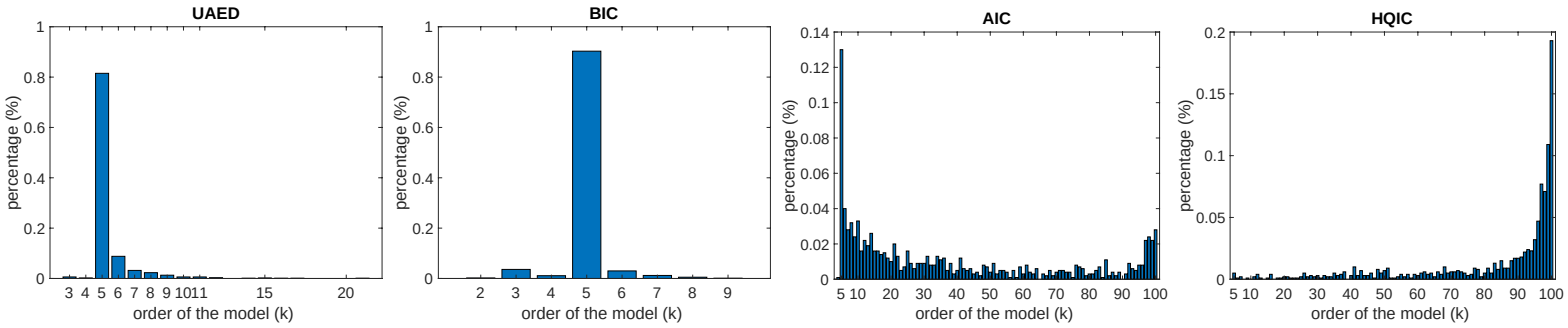
(a) (b) (c) (d)

Figure 15: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 1$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



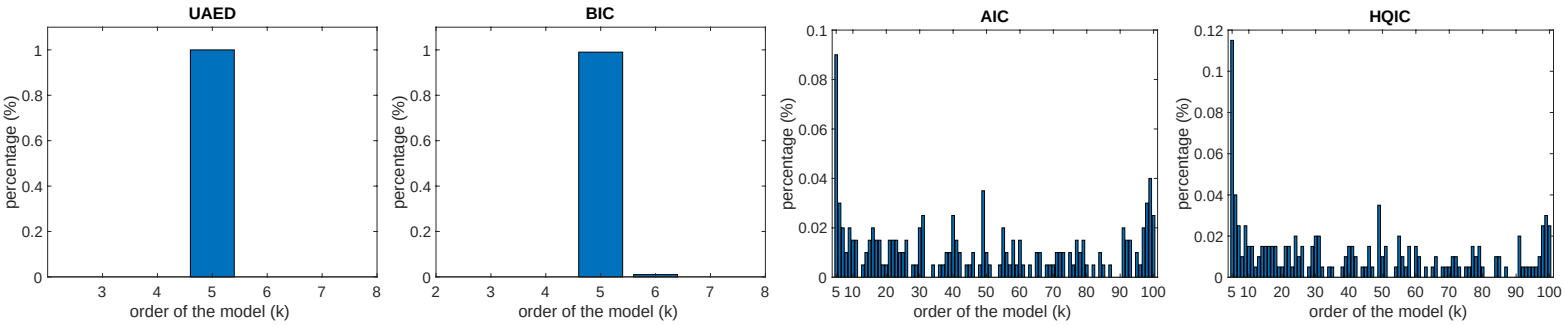
(a) (b) (c) (d)

Figure 16: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 1$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



(a) (b) (c) (d)

Figure 17: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 2$, and the number of data $T = 200$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.



(a) (b) (c) (d)

Figure 18: Percentages of model order decision by the different methods, in the scenario where the true order is $k = 5$, the standard deviation of the noise is $\sigma_\epsilon = 2$, and the number of data $T = 2000$; **(a)** results of UAED; **(b)** results of BIC; **(c)** results of AIC; **(d)** results of HQIC.