Spectral information criterion for automatic elbow detection

Luca Martino[†], Roberto San Millán-Castillo[†], Eduardo Morgado[†]

[†] Universidad Rey Juan Carlos, Fuenlabrada, Madrid.

Abstract

We introduce a generalized information criterion which contains other well-known information criteria, such as BIC and AIC, as special cases. Furthermore, the proposed spectral information criterion (SIC) is also more general than the other information criteria, e.g., since the knowledge of a likelihood function is not strictly required. SIC extracts geometric features of the error curve and, as a consequence, it can be considered an automatic elbow detector. SIC provides a subset of all possible models, with a cardinality that often is much smaller than the total number of possible models. The elements of this subset are "elbows" of the error curve. A practical rule for selecting a unique model within the sets of elbows is suggested as well. Several experiments involving ideal scenarios, synthetic data and real data show the benefits of the proposed scheme. Matlab code related to the experiments is available.

Keywords: Model selection, automatic elbow detection, information criterion, BIC, AIC, marginal likelihood

1 Introduction

Model selection is undoubtedly one of the most important task in signal processing, statistics and machine learning. It can be considered one of the fundamental tasks of the scientific inquiry. Indeed, the majority of the problems in statistical inference can be interpreted in some way as a statistical modeling problem [1, 2, 3, 4].

Model selection is the process of selecting one model among many candidate models given some data. We can distinguish three main scenarios. The first one (denoted as S1) is when completely different models are compared. The second setting S2 is when several models of the same parametric family are evaluated, i.e., the parameters or hyper-parameters of the model are tuned. The third scenario S3 is related to the previous one but, in this case, the family contains models of different complexity since the number of parameters can grow (building more complex models). This last case is also referred as *nested models*. Examples of model selection in nested models are the order selection in an autoregressive predictive method, variable or feature selection, clustering, and even dimension reduction, to name a few [5, 6, 7, 8].

The main competing concerns are (a) the model performance and (b) the model complexity, which generate the so-called bias-variance trade off. Namely, practitioners and researchers try to overcome the two extreme conditions in prediction, underfitting (high bias and low variance) and overfitting (low bias and high variance). The fitting of the current data usually requires more complex models, whereas the ability of good predictions with new unseen data demands for simpler models [9]. More generally, simpler models (e.g., with fewer parameters) are to be preferred for a principle of parsimony (a.k.a. Occam's razor). Therefore, the concept of selecting the best model is in some sense related to the idea of choosing a model that is "good enough". The issue is to define mathematically what "good enough" means exactly [9].

In the literature, there are two main classes of methods for addressing also the scenarios S1 and S3: they are *resampling methods* and *probabilistic statistical measures*. Examples of well-known resampling methods are the *bootstrap* and *cross-validation* (CV) techniques [10, 11, 8]. They are based on the splitting of the data in training and test sets into fitting a model on the training set, and evaluating it on the test set. This process may then be repeated several times and the performance can be averaged over the runs. Resampling methods can be also used to tune the constant value λ in the regularization term in the scenario S2. However, the proportion of data to use in training and in test is a crucial parameter to be chosen by the user, that affects critically the results in terms of required computational time and model complexity penalization. The *leave-one-out CV* approach is one of the faster CV strategies (if N is the number of data, the number of CV repetitions is exactly N) but tends to select more complex models (closer to the overfitting). More generally, in a CV scheme decreasing the percentage of data in the training set (and, as a consequence, increasing the data in the test set) yields to obtain simpler models (tending to the underfitting), whereas increasing the percentage of data in the training set yields to obtain more complex models (tending to the overfitting).

Alternatively, the probabilistic measures employ score rules for evaluating the different models, considering both their performance on the entire dataset and the model complexity. This family is mainly formed by the so-called *information criteria* [12, 2, 13, 14], such the Bayesian information criterion (BIC) which is an approximation of the marginal likelihood [15], and the Akaike information criterion (AIC), which is based on entropy maximization principle [16]. Other examples are the risk inflation criterion [17], the Mallows's C_p coefficient [18], minimum description length (MDL) [19]. The MDL is quite related to BIC and the Mallows's C_p coefficient is related to AIC in the context of Gaussian linear regression (and variable selection). Denoting as k the dimension of the problem (e.g., the number of parameters to infer), all the information criterion (IC) measures use the maximum log-likelihood as a fitting term (which is an error decay denoted as $V(k)$), and a linear penalization of the model complexity λk, where λ is a positive constant. They differ for the *slope* λ of this linear penalization term (see Table 2 and the appendices in [20]). The choice of this slope, i.e., coefficient multiplying the penalization term, is justified by different theoretical derivations, each one with several assumptions and approximations. In Bayesian inference, the marginal likelihood is used for model selection purposes. The marginal likelihood is strictly related to the BIC [21] and, more generally, it can be expressed similarly as an IC measure (see the appendices in [20]). The model penalization in the marginal likelihood is induced by the choice of the prior densities [4, 20]. Again in the Bayesian framework, the posterior predictive is another approach similar to CV [20]. Other approaches based on geometric considerations deserve to be mentioned. Some methods are based on visual inspection of an error curve looking for an "elbow" or "knee". Some automatic procedures for elbow detection, or similar

goals, have been proposed in the literature [22, 23]. Finally, some classical schemes based on p-values (the so-called stepwise regression) are designed for specific applications [24, 25].
In this work, we extend the IC approach extracting geometric information from the error curve. The proposed *spectral information criterion* (SIC) generalizes and contains several IC schemes in the literature as special cases. The underlying idea is to remove the dependence on a particular choice of the slope λ in the IC approach, varying this value and studying the corresponding distribution of minima of a suitable cost function. Namely, since the IC schemes given in the literature are good or even optimal *but only* in specific scenarios and under certain assumptions, the idea in this work can be interpreted as follow: to consider theoretically all the possible IC approaches (within of the BIC and AIC type of information criteria) and then analyze the obtained results. SIC has a wider range of application, since it can be employed even in scenarios where a likelihood function is not provided. Indeed, the SIC scheme can be applied to any error curve obtaining geometric information from it. In this sense, SIC can be considered an *automatic elbow detector*. Firstly, the proposed technique is able to reduce drastically the possible number of models, providing a subset of suggested models. Secondly, a final criterion for selecting a unique model is also provided. Several numerical experiments show the good performance obtained by the SIC scheme. We also provide Matlab code related to the experiments.[1]

2 Proposed framework

Let be $\theta_1, ..., \theta_k, ..., \theta_K$ the K possible components of a complete vector $\boldsymbol{\theta}_K = [\theta_1, ..., \theta_K]^\top$ to infer, which is related to some observed vector of data $\mathbf{y}$, i.e., $\boldsymbol{\theta}_K \rightarrow \mathbf{y}$. In many applications, the goal is to study all the possible models within a parametric family with parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$ with $k \leq K$. Note that k represents the actual dimension of the problem, e.g., the order of a polynomial function or the number of feature in a regression problem, or the number of clusters etc.. The maximum number of components/variables/clusters (depending on the specific application) is denoted as K. In this work, we focus on the task of selecting the optimal number of components $k^* \leq K$. We also refer to k^* as a possible "elbow" of the problem. In several parts of the work, for clarity in the exposition we refer specifically to the nomenclature and notation of a variable selection problem, without loss of generality.
In this work, we employ a generalized IC approach. Below, we introduce the cost function that we desire to minimize,

$$C(k, \lambda) = V(k) + \lambda k, \quad k = 0, ..., K, \quad \lambda \in [0, \lambda_{\mathtt{max}}], \tag{1}$$

where $V(k)$ is a generic fitting term, λk is a penalization term of model complexity, where λ is a constant and k represents the dimension of the model. We consider all the possible values of $\lambda \in [0, \lambda_{\mathtt{max}}]$ where $\lambda_{\mathtt{max}}$ is defined below in Section 3.1.
Linear penalization. It is important to remark that we employ in Eq. (1) a *linear* penalization of the complexity, since this linear term appears in different theoretical derivations in the literature [15, 16, 27]. Moreover, it appears not just in several IC formulations but also in other more general approaches, e.g., involving marginal likelihood with uniform priors [20, App. A and B]

[1] The Matlab code is given at `http://www.lucamartino.altervista.org/PUBLIC_SIC_CODE.zip`.

and alternative geometric solutions [22]. Therefore, choosing a linear penalty for the complexity seems to have a strong theoretical support by different points of view.

2.1 About the fitting term $V(k)$

The function $V(k)$ represents a generic non-increasing function[2] with a finite value in zero, i.e., $V(0) < \infty$ (hence $V(k)$ takes always finite values). Examples of function $V(k)$ are:

- Given the vector of parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$ of dimension k and denoting a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$, we can have $V(k) = -2\log(\ell_{\texttt{max}})$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$, exactly as in a standard IC approach. Thus, the SIC scheme is a more general approach and contains the standard IC strategies, that employ a cost function as Eq. (1), as a special cases;

- the root mean square error (MSE) or the mean absolute error (MAE) in a regression problems, i.e., $V(k) = \text{MSE}(k)$, as function of a integer k, where k can represent the order of a polynomial or the number of variables involved in the regression;

- $V(k) = 1-\text{Accuracy}(k)$ in a classification problem using the first k most important features;

- $V(k)$ can present the k-th eigenvalue of the covariance matrix of the data in a principal component analysis (PCA), where the eigenvalues are ordered in a decreasing order;

- $V(k)$ could be the sum of the inner variances in each cluster (or the log of this sum), in a clustering application.

The list above just contains some examples, but it is important to remark that the proposed method only required that $V(k)$ be non-increasing (condition that can be also relaxed).

3 Spectral information criterion (SIC) method

In this work, the underlying approach is inspired by the idea of "integrating out" λ as usually done Bayesian analysis, i.e., we would like to remove the dependence of λ in our problem. Namely, we would like to avoid to pick a specific value of λ, unlike the other IC schemes in the literature. In the next subsections, we first define properly $\lambda_{\texttt{max}}$ and a piecewise linear function $k^*(\lambda)$ of minima of $C(k, \lambda)$. Finally, in the last two subsections, we introduce the spectral information criterion (SIC).

3.1 Defining and computing $\lambda_{\texttt{max}}$

The value of $\lambda_{\texttt{max}}$ is defined as

$$\lambda_{\texttt{max}} = \{\min \lambda : \quad \arg\min_k C(k, \lambda) = 0\}, \tag{2}$$

[2] This condition can be even relaxed, as also shown in Figure 8(c).

so that

$$\arg\min_k C(k, \lambda_{\texttt{max}}) = \arg\min_k \left[V(k) + \lambda_{\texttt{max}} k\right] = 0, \tag{3}$$

and we have

$$\arg\min_k C(k, \lambda') = 0, \qquad \text{for any } \lambda' \geq \lambda_{\texttt{max}}. \tag{4}$$

Note that, as an example, $k = 0$ corresponds to a constant model in a regression problem, when the case of "no variables" are used (in a variable selection example), i.e., $V(0) = \text{var}(\mathbf{y})$ which is the variance of the data. The value of $\lambda_{\texttt{max}}$ can be analytically obtained as

$$\lambda_{\texttt{max}} = \max_k \left[\frac{V(0) - V(k)}{k}\right], \quad \text{for } k = 1, ..., K. \tag{5}$$

Since above we consider $k = 1, 2, ..., K$, we can perform an exhaustive search and, considering Eq. (5), obtain $\lambda_{\texttt{max}}$. If the value K is huge and/or for some reason the exhaustive search cannot be performed, classical numerical methods can be successfully implemented, such as *bisection method* [26, Chapter 3].

3.2 The function $k^*(\lambda)$

For the sake of simplicity, let assume in this section that $V(k)$ is a decreasing function, with $V(0) < \infty$. With this assumption, it can be proved that $C(k, \lambda)$ has a unique minimum. See a graphical example in Figure 1(a). Now, we study the function $k^*(\lambda) : [0, \lambda_{\texttt{max}}] \subset \mathbb{R} \to \{0, 1, 2..., K\}$, defined as

$$k^*(\lambda) = \arg\min_k C(k, \lambda), \tag{6}$$

which takes real values in the interval $[0, \lambda_{\texttt{max}}]$ and covert them in discrete values within the set $\{0, 1, 2..., K\}$. It is a non-increasing, piecewise constant function where $k^*(0) = K$ and $k^*(\lambda) = 0$ for $\lambda \geq \lambda_{\texttt{max}}$, i.e., more specifically,

$$\begin{cases} k^*(0) = K, \\ k^*(\lambda_{\texttt{max}}) = 0, \end{cases} \tag{7}$$

as shown in Figure 1(b). A relevant consideration is that some values of $k \in \{0, 1, 2..., K\}$ could not represent an output of the function $k^*(\lambda)$, i.e., they could not have a corresponding λ associated. For instance, this is the case of $k = 1$ in Figure 2(a).

An example of piecewise constant function $k^*(\lambda)$ is given in Figure 1(b). Several values of λ can be associated to the same minimum k^*, as shown in Figure 2(a). On the other hand, some value k' could not have any λ associated, that means that the value k' cannot be a minimum of $C(k, \lambda)$. More generally, to each k, we can associate an interval of lambda values, $\mathcal{S}_k \subset [0, \lambda_{\texttt{max}}]$. Observe that $\mathcal{S}_0 =$ by definition since we consider $\lambda \in [0, \lambda_{\texttt{max}}]$, so that $|\mathcal{S}_0| = 0$. These intervals, for $k = 1, ..., K$, form a partition of $[0, \lambda_{\texttt{max}}]$, i.e.,

$$\mathcal{S}_1 \cup \mathcal{S}_2 ... \cup \mathcal{S}_K = [0, \lambda_{\texttt{max}}],$$

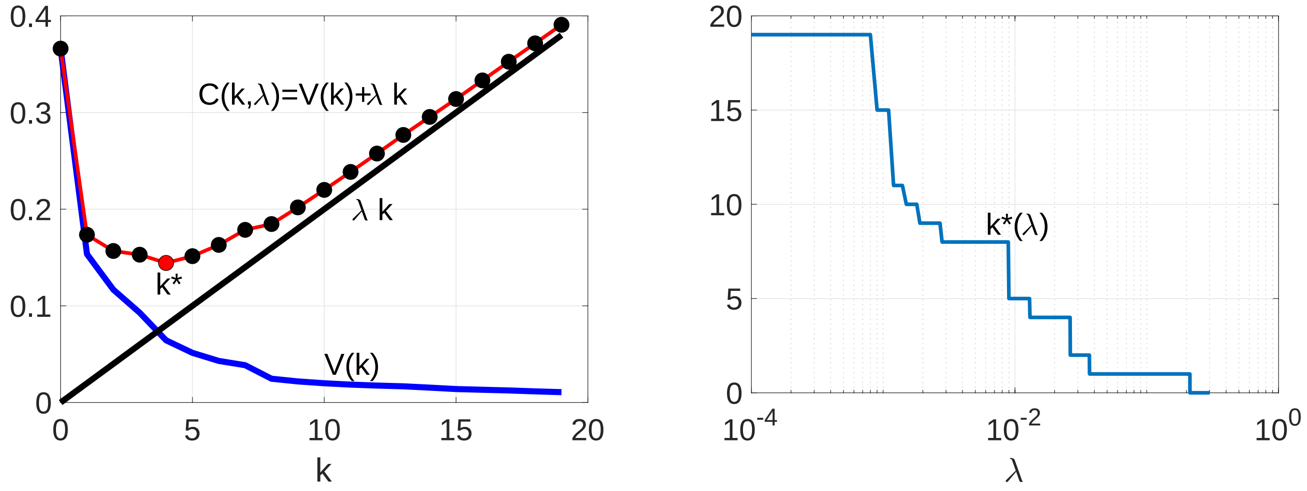
and $\mathcal{S}_k \cap \mathcal{S}_j = 0$, for all $k \neq j$. Figure 2(b) provides a graphical representation. As stated above, some value $k' \neq 0$ could be never a minimum, so that $|\mathcal{S}_{k'}| = 0$.

3.3 Description of the SIC method

As previously stated, we would like to remove the dependence of λ in our problem. Namely, we would like to avoid to pick a specific value of λ. Here, the idea is to use the information provided by the measures $|\mathcal{S}_k|$. With this goal, we define the weights $\bar{w}_k \propto |\mathcal{S}_k|$, i.e.,

$$\bar{w}_k = \frac{|\mathcal{S}_k|}{\sum_{j=0}^{K} |\mathcal{S}_j|} = \frac{|\mathcal{S}_k|}{\sum_{j=1}^{K} |\mathcal{S}_j|}, \tag{8}$$

where we have used $|\mathcal{S}_0| = 0$. Note that $\bar{w}_k$, for $k = 1, ..., K$, defines a probability mass function (pmf), $\sum_{k=1}^{K} \bar{w}_k = 1$. The main part of the SIC method is to compute (approximately) the probabilities $\bar{w}_k$. This approximation can be obtained with a quasi-Monte Carlo strategy (i.e., with a simple grid) or with a standard Monte Carlo approach using M number of samples. The latter is given in Table 1. An example of the weights $\bar{w}_k$ is given in Figure 3(a), which correspond to the $V(k)$ curve in Figure 1(a). The algorithm in Table 1 is generally fast even with choices of M such as $M = 10^6$, $M = 10^7$ or greater.[3]



(a)

(b)

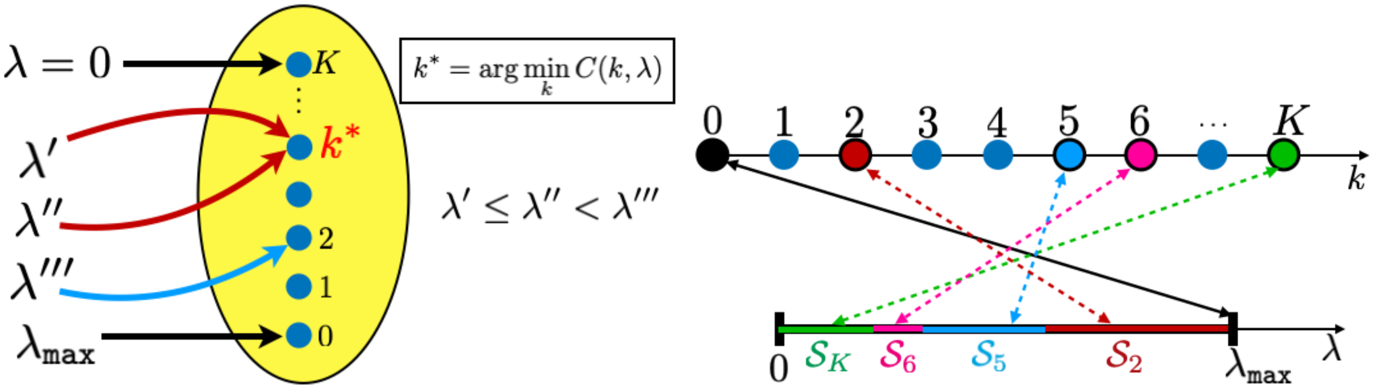
Figure 1: **(a)** Example of function $V(k)$, a penalization term λk and the corresponding cost function $C(k, \lambda)$ (shown with dots). **(b)** Example of piecewise constant function $k^*(\lambda)$.

3.4 Interpretation and model selection

We will show in the next sections that the set $\mathcal{E}$ of indices k such that the corresponding weight is non-zero, $\bar{w}_k > 0$,

$$\mathcal{E} = \{\text{all } k : \ \bar{w}_k > 0\} = \{k_E^{(1)}, k_E^{(2)}, ..., k_E^{(J)}\},$$

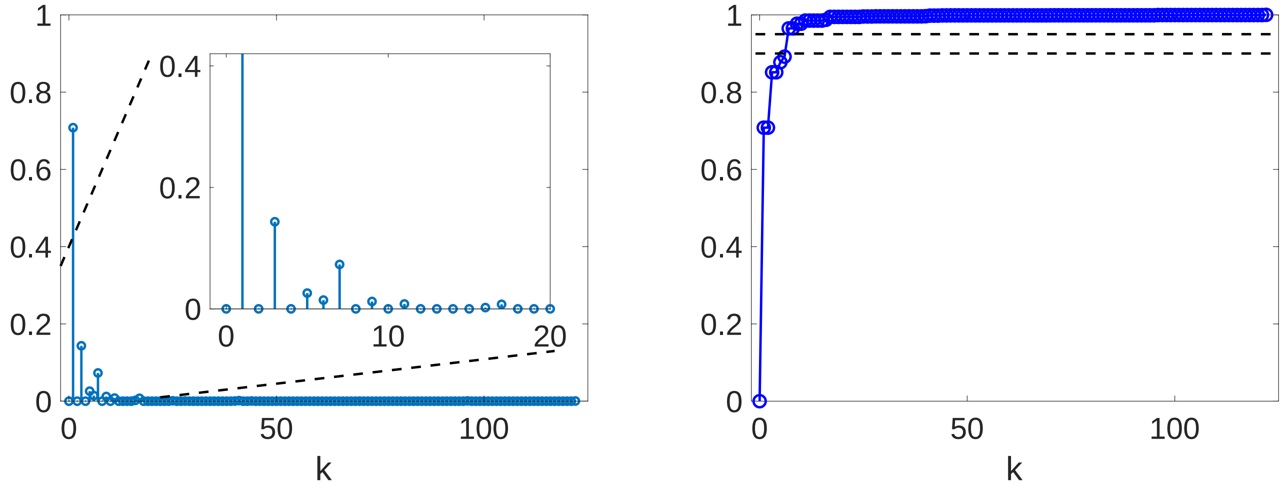
[3] See the (non-optimized) Matlab implementation at `http://www.lucamartino.altervista.org/PUBLIC_SIC_CODE.zip`.



(a)

(b)

Figure 2: **(a)** Correspondence between choice of $\lambda \in [0, \lambda_{\max}]$ and the corresponding minimum k^*. Different lambda can give the same minimum k^*. Each minimum has associate an interval $\mathcal{S}_{k^*}$ of values of λ's. **(b)** A graphical representation of the intervals $\mathcal{S}_k$ (and the measures $|\mathcal{S}_k|$) for all k. Note that, for some k', $|\mathcal{S}_{k'}| = 0$, i.e., the discrete value k' could be never a minimum, considering all the possible values of $\lambda \in [0, \lambda_{\max}]$.



(a)

(b)

Figure 3: **(a)** The weights $\bar{w}_k$ obtained by the SIC method applied in the experiment in Section 5.4. **(b)** The cumulative function W_k corresponding to the probability mass $\bar{w}_k$, with $k = 0, ..., K$. The dashed lines show the confidence level $\ell = 0.9$ and $\ell = 0.95$, respectively.

can be interpreted as a a possible "elbow" of the curve, i.e., a possible chosen models represented by the indices $k_E^{(j)}$. We have denoted $J = |\mathcal{E}|$. Note that $J \leq K$ and, in some cases, $J << K$. Therefore, we can have a sensible reduction of the number of possible models to choose. In order to select just one model, the more conservative solution is $k_E = \max k_E^{(j)}$ choosing the more complex model, whereas the simplest possible model is given by the choice $k_E = \min k_E^{(j)}$ with

$j = 1, ..., J$. Any intermediate solution can be motivated by the specific application. However, more considerations can be done. For this purpose, let us define the cumulative sum of the first m weights, i.e.,

$$W_m = \sum_{i=1}^{m} \bar{w}_i,$$

with $1 < m \leq K$. Figure 3(b) provides an example. In absence of any other user consideration to select a specific model within the set $\mathcal{E}$, we give here a possible suggestion, obtained by empirical studies. We suggest to choose as "elbow" the index defined as

$$k_E = \min\{k : \ W_k \geq \ell\}, \qquad \text{with} \qquad \ell \geq 0.9,$$

where ℓ is a confidence level. A conservative, robust choice (selecting a more complex model) can be get setting $\ell = 0.95$.

Table 1: Computation of the weights in the SIC method by Monte Carlo.

- For $i = 1, ..., M$:
  1. Draw $\lambda_i \sim \mathcal{U}([0, \lambda_{\texttt{max}}])$.
  2. Compute

$$k_i^* = \arg\min_k C(k, \lambda_i) = \arg\min_k \left[V(k) + \lambda_i k\right]. \qquad (9)$$

- Return the number of occurrences of the event $\{k_i^* = j\}$ for $j = 1, ..., K$, or equivalently return the weights

$$\bar{w}_j = \frac{\#\{k_i^* = j\}}{M}, \qquad j = 1, ..., K. \qquad (10)$$

4 Analysis of SIC performance and behavior

In this section, we analyze the results provided by SIC in ideal scenarios (Section 4.1), and its behavior (a) under variation of K, and (b) under translation and scale of the axes (Section 4.2).

4.1 SIC performance with piecewise linear decays $V(k)$

In this section, we consider ideal scenarios to check the performance of the proposed method in these settings. We describe the 4 different scenarios denoted as **I1**-**I2**-**I3** and **I4**. We also discuss the expected results in each case, and we check the performance of SIC.

• **I1.** The first ideal scenario is when $V(k)$ is constant, i.e., $V(k) = V(0)$ for all k. This means all the components $\theta_1, ..., \theta_K$ of the vector to infer $\boldsymbol{\theta}_K$ are independent from the output variable y, so that the correct solution is $k_E = 0$. Since $V(k)$ is constant in this scenario, we have $V(k) = V(0)$, and we would obtain $\lambda_{\texttt{max}} = 0$ by Eq. (5). Namely, we get $k^*(\lambda) = 0$ for any possible $\lambda > \lambda_{\texttt{max}} = 0$, by definition of $\lambda_{\texttt{max}}$. Hence, having $k^*(\lambda) = 0$, finally we have $k_E = 0$. Thus, the SIC method obtains the correct result.

• **I2** The second ideal scenario is when $V(k)$ is a linear straight line connection the points $(0, V(0))$ and $(K, V(K))$, as shown in Figure 7(e). In this situation, all the variables contribute in the same way to the decay of $V(k)$ (i.e., each variable has the same influence to the error decrease), so that the correct solution is $k_E = K$. In Figure 7(e), we can see that the SIC scheme selects $k_E = K$ having a unique non-zero weight $\bar{w}_K = 1$ (i.e., at $k = K$), which is the correct result.

• **I3** Another ideal scenario is when $V(k)$ is formed by two pieces of straight lines, as depicted in Figures 7. In this case, if $V(k)$ is convex (i.e., when the second slope is smaller than the first slope) the solution k_E (i.e., a possible "elbow") is given by the intersection of the two straight lines, as illustrated in Figures 7. If $V(k)$ is concave (i.e., when the second slope is greater than the first slope), as for instance in Figure 7(f), the intersection is not a possible solution, so that the correct solution is $k_E = K$ in this case.
As we can observe in Figures 7, SIC selects the right k_E in any of these cases. Generally, there is a main weight $\bar{w}_k$ close to 1, and in Figures 7(a), 7(e) and 7(f) we have even a unique non-zero weight. Note that as the value $V(k_E)$ grows the weight at at $k = K$ becomes bigger and bigger. This is a desirable behavior since, $V(k_E)$ grows, the scenario becomes more similar to **I2**, i.e., more similar to Fig. 7(e), where all the components/variables have the same impact to the results, i.e., they generate the same drop in $V(k)$. Indeed, if the value $V(k_E)$ is such that we have only one straight line connecting the points $(0, V(0))$ and $(K, V(K))$ as in Fig. 7(e), we have $\bar{w}_K = 1$ since we come back **I2**. As the value $V(k_E)$ grows more, $V(k)$ becomes concave and the SIC scheme correctly keeps $\bar{w}_K = 1$ at $k_E = K$. Therefore, in all settings, the SIC method provides the expected and desirable behavior.

• **I4** More generally, we can consider a piecewise linear decay $V(k)$, formed by several pieces of straight lines, as given in Figures 8. If $V(k)$ is convex, all the intersection points are possible candidates to be an "elbow". Let us denote the intersection points as

$$\mathcal{E} = \{k_E^{(1)}, k_E^{(2)}, ..., k_E^{(J)}\},$$

where J is the number of the intersections. In this framework, different users can have different opinions regarding the correct "elbow" to pick, i.e., the model to choose. These opinions can depend to the different context and application, as well as the computational budget etc. Note that this setting I4 is the more general scenario and contains the other ones, I1-I2 and I3, as special cases.
Also in this scenario, the SIC method provides desirable results, considering the criterion in Section

3.4, with both $\ell = 0.9$ or $\ell = 0.95$. As we can observe in Figures 8(a)-8(b)-8(c) and Figures 8(d)-8(e)-8(f) the only non-zero weights $\bar{w}_k$ correspond to the possible elbows $\{k_E^{(1)}, k_E^{(2)}, ..., k_E^{(J)}\}$. Moreover, in Figure 8(c), we consider an increase piece in $V(k)$ creating a concave part, so that a possible elbow point should be discarded as the SIC scheme does. An equivalent situation is given in Figure 8(b).

4.2 Additional considerations about the SIC behavior

We have seen that the SIC scheme provides the desirable results in all the ideal scenarios above, where a piecewise linear curve $V(k)$ is given. Moreover, it is important to remark that SIC presents a small dependence on possible changes of the value K (i.e., on an increase or a decrease of K), if there is not a significant drop variation in $V(k)$ associated to the variation of K. Indeed, we can also observe in Eq. (5) that $\lambda_{\texttt{max}}$ is also virtually insensible to variations to the value of K. This is clearly another desirable behavior.

Finally, it is important to remark that the results of SIC does not depend on a shift and/or a scale of the axes. Regarding a shift and scale of the horizontal axis $k' = \alpha k + \beta$, it is easy to show that the solutions are just shifted and scaled in the same way, i.e., $k'_E = \alpha k_E + b$. Regarding a shift of $V(k)$, we can see the solutions are completely invariant. For instance, defining $V'(k) = V(k) + b$, since the constant b does not depend on k, we can write the following sequences of equalities,

$$\begin{aligned} k^*(\lambda) &= \arg\min_k \left[V(k) + \lambda k\right], \qquad \lambda \in [0, \lambda_{\texttt{max}}], \\ &= \arg\min_k \left[V(k) + \lambda k + b\right], \\ &= \arg\min_k \left[V'(k) + \lambda k\right]. \end{aligned}$$

Hence the function $k^*(\lambda)$ does not change for any possible value of b. Considering now a scale factor, i.e., defining $V'(k) = aV(k)$, we have to observe that $\lambda_{\texttt{max}}$ is also scaled in the same way. Namely, we have

$$\begin{aligned} \lambda'_{\texttt{max}} &= \max_k \left[\frac{aV(0) - aV(k)}{k}\right], \quad \text{for } k = 1, ..., K, \\ &= a \max_k \left[\frac{V(0) - V(k)}{k}\right] = a\lambda_{\texttt{max}}. \end{aligned}$$

Thus, we have also that $\lambda' \in [0, \lambda'_{\texttt{max}}] = [0, a\lambda_{\texttt{max}}]$. Observe that we can write $\lambda' = a\lambda$ where $\lambda \in [0, \lambda_{\texttt{max}}]$, so that we can write

$$\begin{aligned} k^*(\lambda') &= \arg\min_k \left[V'(k) + \lambda' k\right], \quad \text{with} \quad \lambda' \in [0, \lambda'_{\texttt{max}}] = [0, a\lambda_{\texttt{max}}], \\ &= \arg\min_k \left[aV(k) + a\lambda k\right], \quad \text{with} \quad \lambda \in [0, \lambda_{\texttt{max}}], \\ &= \arg\min_k \left[a(V(k) + \lambda k)\right] \\ &= \arg\min_k \left[V(k) + \lambda k\right] = k^*(\lambda). \end{aligned}$$

Namely, again the function $k^*(\lambda)$ does not change. In the next section, we will consider experiments with real-word applications, and with real data in two of them.

5 Real-world applications and experiments

In this section, we test SIC in different real-world applications, considering different functions $V(k)$, in order to show the vast range of applicability of the proposed scheme. In Sections 5.4-5.5, the experiments involve real data problems: variable selection in a regression problem with soundscape emotion data, and in a classification problem with biomedical data. In Sections 5.3-5.5, a probabilistic model is involved so that the fitting term can be defined as $V(k) = -2\log(\ell_{\texttt{max}})$. Hence, in these two sections, this allows a comparison with BIC, AIC and other information criteria described in the literature.

5.1 Clustering

We generate 2500 artificial data from 5 different 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 4(a) depicts these data points.
We consider $V(k) = \log\left[\sum_{j=1}^{k+1} \text{var}(j)\right]$, where var$(j)$ is the internal variance in the j-th cluster, as shown in Figures 4(b). Each value of var(j) is compute and averaged after 200 runs of a k-means algorithm. Note that the total number of clusters is $k+1$ (e.g., $k = 0$ corresponds to a single cluster). We consider $K = 50$ as maximum number of possible clusters. Figures 4(c)-4(d) show the results obtained by SIC. Recalling that the number of clusters is $k+1$, we have the subset of possible clusters,

$$\mathcal{E} = \{2, 5, 6, 50\},$$

and the final SIC suggestion is $k_E = 5$ for both $\ell = 90$ and $\ell = 95$, which is the correct number of clusters in the synthetic data.

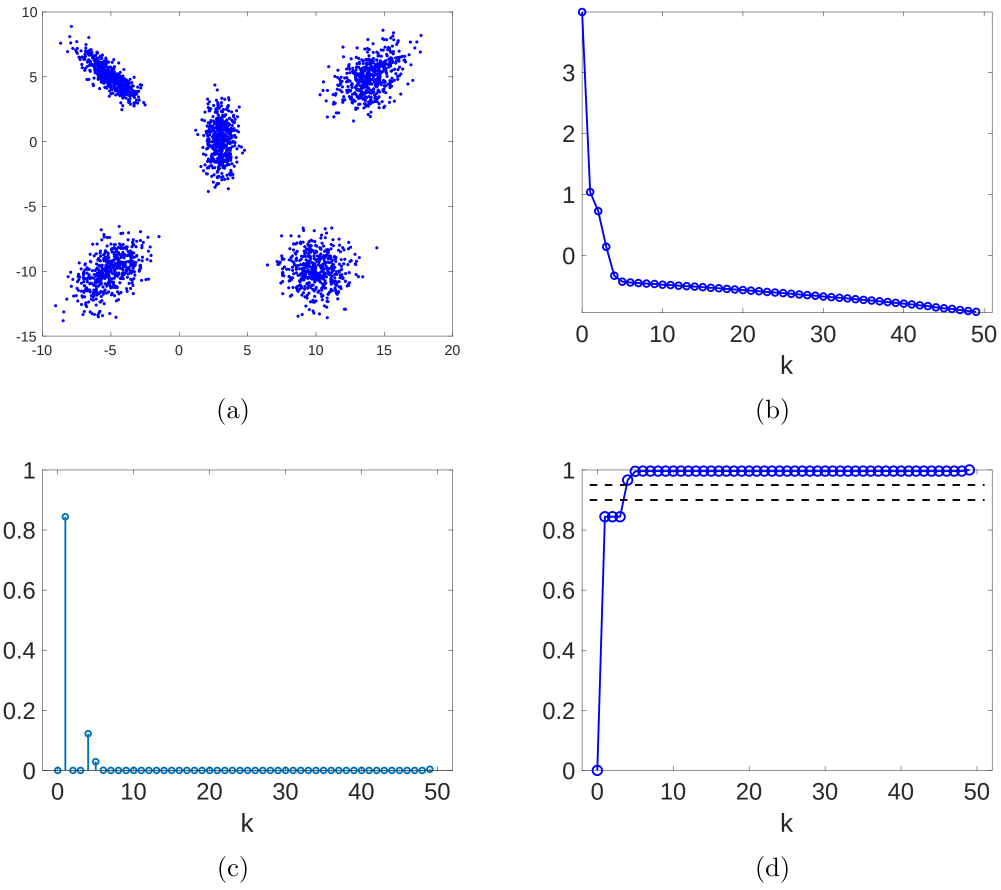
5.2 Dimension reduction

In this experiment, we generate 10^4 Gaussian data in $\mathbb{R}^5$ with a zero vector mean and the following covariance matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0.7 & 0 \\ 0 & 0 & 0.7 & 2 & 0.7 \\ 0 & 0 & 0 & 0.7 & 2 \end{bmatrix}. \tag{11}$$

where 2 dimensions are completely uncorrelated to the remaining ones. Three dimensions are correlated (i.e., they could be summarized by one of them). We consider as $V(k)$ the eigenvalues of $\boldsymbol{\Sigma}$ (in decreasing order) and the trace of the matrix $\boldsymbol{\Sigma}$ as $V(0)$, i.e.,

$$V(0) = 8,\ V(1) = 3.00, V(2) = 2.01,\ V(3) = 1.01\ V(4) = 1.00,\ V(5) = 0.98.$$



(a) (b)

(c) (d)

Figure 4: **(a)** Data points of the clustering experiment. **(b)** The curve $V(k) = \log\left[\sum_{j=1}^{k+1} \text{var}(j)\right]$ where var(j) is the internal variance in the j-th cluster. **(c)** The weights $\bar{w}_k$ obtained by the SIC method. **(d)** The cumulative function W_k. Recall that $k+1$ is the number of cluster (hence, $k = 0$ corresponds to a single cluster).

The function $V(k)$ and the results of SIC are depicted in Figure 6(a). Looking the weights in Figure 6(a), we can observe that SIC is mainly focus on the possible elbows $k_E^{(1)} = 1$ and $k_E^{(2)} = 3$. Applying the final SIC suggestion, we obtain $k_E = 3$ for both $\ell = 90$ and $\ell = 95$, that is the expected result for this dimension reduction problem. The AED method in [22] can be also applied in this example but suggests the use of $k_E = 1$, unlike SIC (that provides the correct answer in this example).

5.3 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,

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + ...\theta_k x_n^k + \epsilon_n, \tag{12}$$

where $\boldsymbol{\theta}_k = [\theta_0, \theta_1, ..., \theta_k]^\top$, ϵ_n is a 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

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

Figure 5(a) depicts the generated data points and the underlying polynomial function of order 4 in solid line. 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 the Eq. (12). The function $V(k)$ is shown in Figure 5(b). With this choice of $V(k)$, we can compare with other information criteria in the literature, as shown in Table 2. After applying SIC, we obtain the results in Figures 5(c)-5(d) and the following set of possible models,

$$\mathcal{E} = \{\underbrace{4,}_{\text{AED,BIC}}\ \underbrace{6,}_{\text{AIC}}\ \underbrace{10,}_{\text{HQIC}}\ 12, 13\}.$$

Above, we have also highlighted the suggested models by BIC (i.e., 4), AIC (i.e., 6), Hannan-Quinn IC (i.e., 10), and the AED method in [22] (i.e., 4), which are all contained in $\mathcal{E}$ (as expected by the design of SIC). The final SIC suggestion is $k_E = 4$ for both $\ell = 90$ and $\ell = 95$, which is the correct order of the underlying polynomial function. Therefore, in this experiment, BIC, AED and SIC provide the correct answer.

5.4 Variable selection in a regression problem with real data

In several real-world applications, we observe a dataset of N pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$, 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 consider the case that $K \leq N$ and assume a linear observation model,

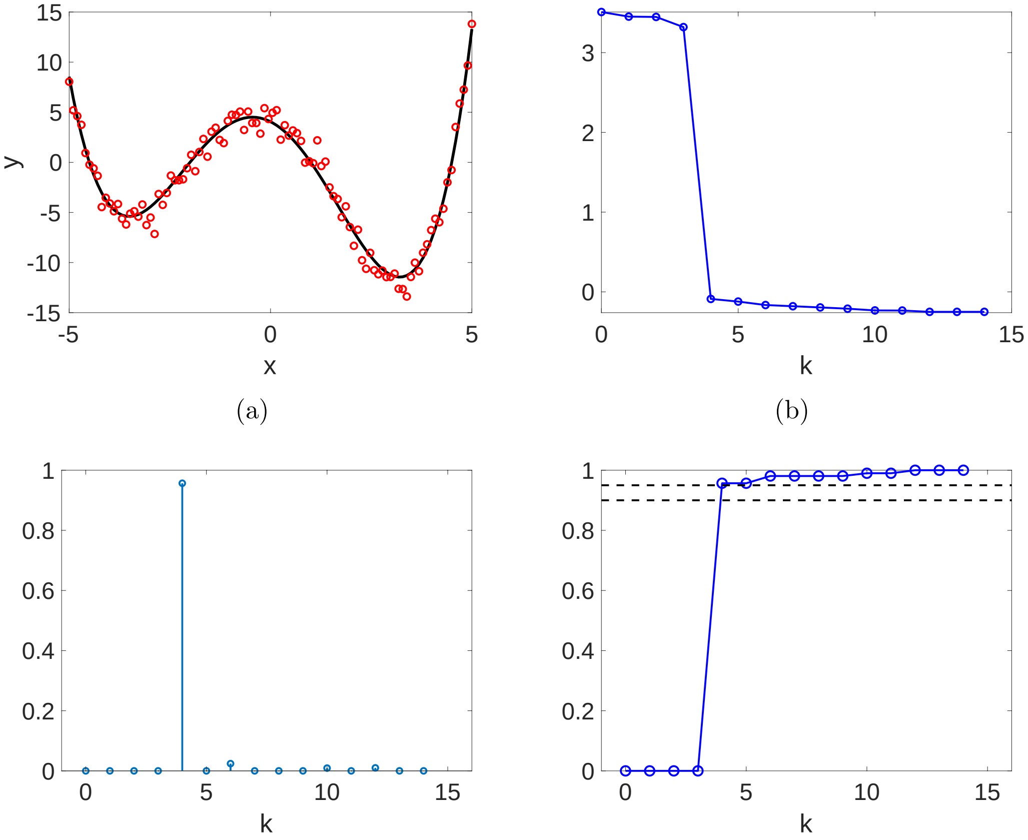
$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + ...\theta_K x_{n,K} + \epsilon_n, \tag{13}$$

where ϵ_n is a Gaussian noise with zero mean and variance σ_ϵ^2, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. More specifically, in [5], in the real dataset there are $K = 122$ features and $N = 1214$ number of data. The dataset studied in [5] has two outputs: "arousal" and "valence". Here, we focus on the "arousal".

In this experiment, we set $V(k) = -2\log(\ell_{\texttt{max}})$ with $\ell_{\texttt{max}} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$, after ranking the 122 variables (see [5]). The likelihood $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by the Eq. (13). Hence, in this experiment, we can compare again with other IC measures in the literature (see Table 2). We use $M = 10^6$ samples for SIC. The set $\mathcal{E}$ is formed by $J = |\mathcal{E}| = 19 << 122$ suggested models, more specifically,

$$\mathcal{E} = \{1, 3, 5, 6, 7, 9,\ \underbrace{11,}_{\text{AED}}\ 16\underbrace{, 17,}_{\text{BIC}} 25, 28, 40,\ \underbrace{41,}_{\text{HQIC}}\ \underbrace{44,}_{\text{AIC}} 46, 70, 71, 96, 122\}.$$

Above, we have also remarked the suggested models by BIC (i.e., 17), AIC (i.e., 44), Hannan-Quinn IC (i.e., 41), and AED (i.e., 11), which are all contained in $\mathcal{E}$ (as expected by the design of SIC). Figures 3(a)-3(b) shows the SIC weights and the cumulative function for this experiment.



(a) (b)

(c) (d)

Figure 5: **(a)** Data points ($N = 100$) of the experiment of the order selection of a polynomial function (shown with a solid line) in a regression problem. **(b)** The corresponding curve $V(k) = -2 \log \ell_{\texttt{max}}$. **(c)** The weights $\bar{w}_k$ obtained by the SIC method. **(d)** The cumulative function W_k. Recall that k represents the order of the polynomial function.

The final SIC suggestion is $k_E = 7$ for both $\ell = 90$ and $\ell = 95$. Therefore, SIC confirms the results given in other previous studies and experts have suggested in the literature. Hence, unlike in the previous experiment, here only SIC provides the correct result.

5.5 Variable selection in a biomedical classification problem with real data

In [28], the authors study 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) method to classify patients and rank the input variables. 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 depicted in Figure 6(b), after ranking

Table 2: Different information criteria contained in SIC as special cases; N denotes the number of observed data.

Criterion	**Choice of λ**
Bayesian-Schwarz information criterion (BIC) [15]	$\log N$
Akaike information criterion (AIC) [16]	2
Hannan-Quinn information criterion (HQIC) [27]	$\log(\log(N))$
Automatic Elbow Detector (AED) [22]	$\frac{V(0)}{\min[\arg\min V(k)]}$

the 35 features [28]. Note that $V(0) = 0.5$ representing a completely random binary classification. The set of possible elbows obtained by SIC is

$$\mathcal{E} = \{1, 2, 3, 9, 11, 24\},$$

where $J = |\mathcal{E}| = 6 << 35$. The final SIC suggestion is $K_E = 2$ features ($\ell = 0.9$), $K_E = 3$ features ($\ell = 0.95$), which is close to the result of the paper [28]. SIC suggests a model without the triglycerides.

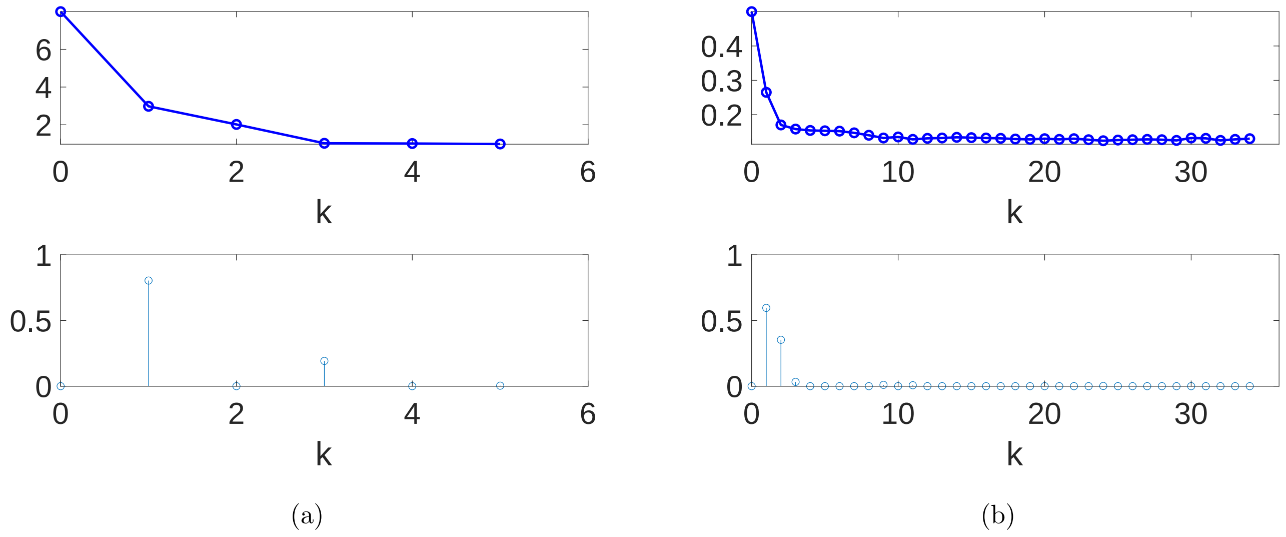


Figure 6: $V(k)$ curves and SIC results for the experiments **(a)** in Section 5.2 and **(b)** in Section 5.5.

6 Conclusions

In this work, we have introduced a generalized information criterion which contains, as special cases, several other information criteria introduced in the literature. First of all, we have introduced the novel approach based on the idea of considering all the possible slopes associated to the linear penalization of the model complexity. SIC returns two main products. The first one is the set of possible "elbows", which contains also the results of other well-known IC schemes in the literature. The second one is the suggestion of the choice of a unique elbow, i.e., a chosen model, within the set of possible ones.

We have tested the SIC technique in different ideal scenarios. These tests have proven that SIC can be considered as an automatic elbow detector, extracting geometrical information from the error curve $V(k)$. Additionally, several real-world experiments (two of them involving real data) have shown that SIC provides better results than the other existing IC measures, exactly coincident (or much closer) to the groundtruths or the experts opinions. Finally, it is important to remark that SIC does not require to assume the knowledge of a likelihood function, unlike other IC schemes in the literature, so that its range of application is much wider, as shown in the numerical experiments.

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] K. Aho, D. Derryberry, and T. Peterson, "Model selection for ecologists: the worldviews of AIC and BIC," *Ecology*, vol. 95, no. 3, pp. 631–636, 2014.

[2] T. Ando, "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models," *Biometrika*, vol. 94, no. 2, pp. 443–458, 2007.

[3] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[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] 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.

[6] A. O'Hagan, "Discussion on posterior Bayes factors (by M. Aitkin)," *Journal of the Royal Statistical Society Series B*, vol. 53, p. 136, 1991.

[7] P. Mukherjee, D. Parkinson, and A. R. Liddle, "A nested sampling algorithm for cosmological model selection," *The Astrophysical Journal Letters*, vol. 638, no. 2, p. L51, 2006.

[8] 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.

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

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

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

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

[13] T. Ando, "Predictive Bayesian model selection," *American Journal of Mathematical and Management Sciences*, vol. 31, no. 1-2, pp. 13–38, 2011.

[14] A. van der Linde, "DIC in variable selection," *Statistica Neerlandica*, vol. 59, no. 1, pp. 45–56, 2005.

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

[16] 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.

[17] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *The Annals of Statistics*, vol. 22, no. 4, pp. 1947–1975, 1994.

[18] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[19] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[20] 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.

[21] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *(to appear) SIAM Review - arXiv:2005.08334*, 2022.

[22] E. Morgado, L. Martino, and R. San Millán-Castillo, “Universal and automatic elbow detection for learning the effective number of components in model selection problems,” *preprint*, pp. 1–12, 2022.

[23] 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.

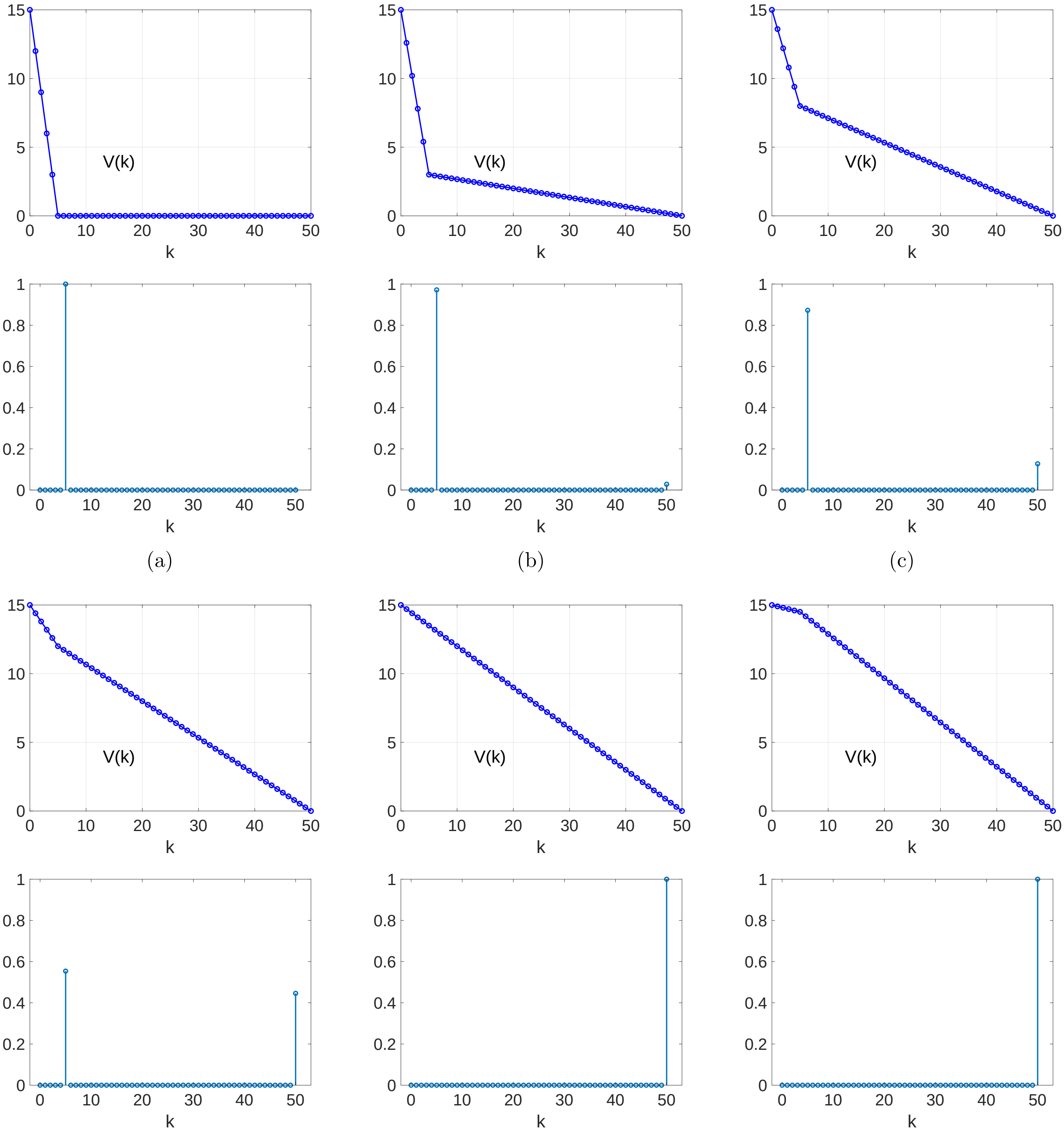
[24] M. Efroymson, “Multiple regression analysis,” *Mathematical methods for digital computers*, pp. 191–203, 1960.

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

[26] J. F. Epperson, *An Introduction to Numerical Methods and Analysis.* Wiley, 2007.

[27] 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.

[28] 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.



(a)

(b)

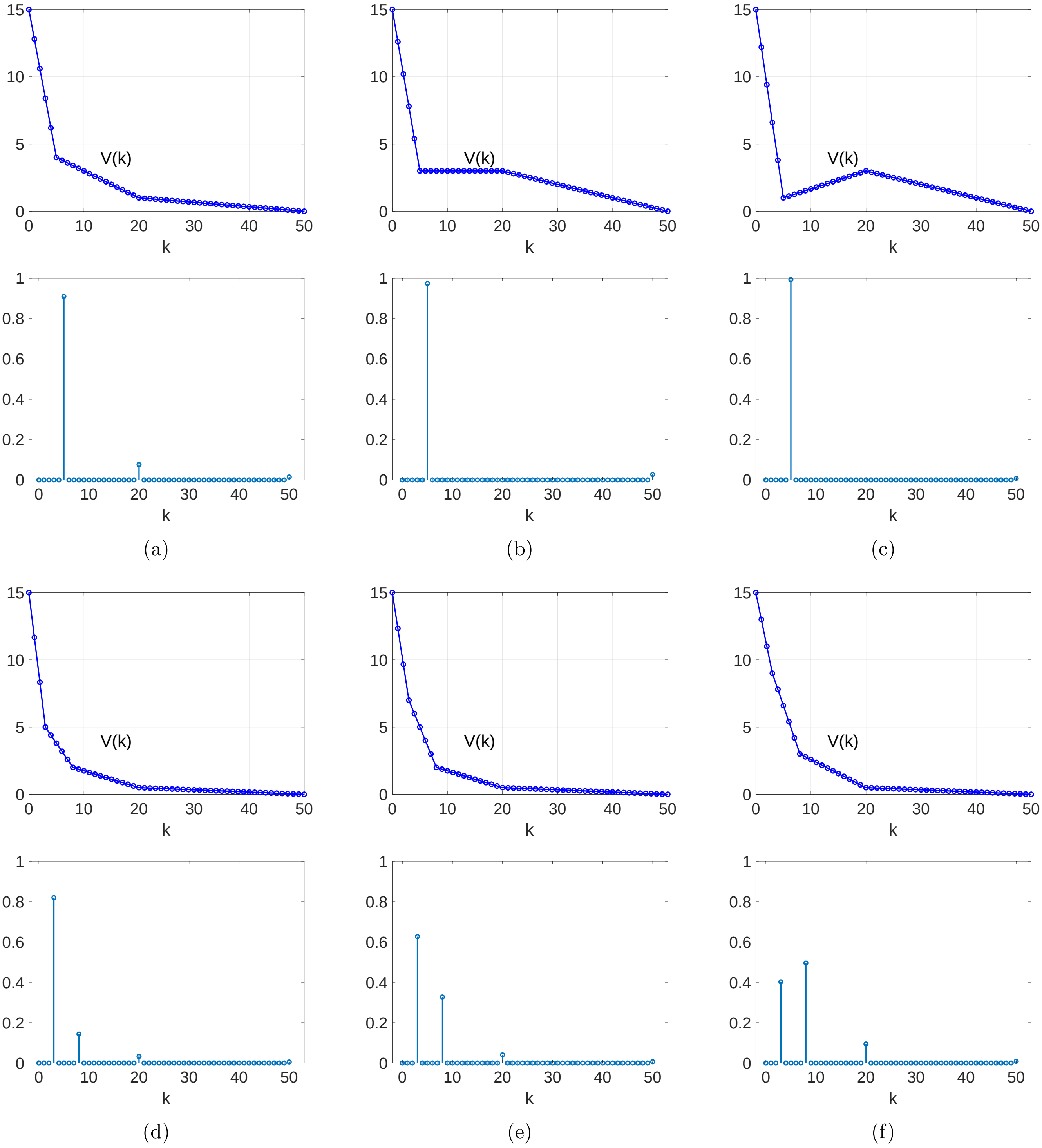
(c)

(d)

(e)

(f)

Figure 7: **(a)-(b)-(c)** Application of SIC to ideal cases where $V(k)$ has one unique elbow at $k = 5$. The elbow is clearer in 7(a) than in 7(b) and 7(c). We can observe that, as the value of $V(5)$ grows, SIC starts to suggest also to use all the 50 components. Clearly, it is a desirable behavior. **(d)-(e)-(f)** Application of SIC to ideal cases where $V(k)$ has a very "slight" elbow at $k = 5$ in 7(d), and there is not elbow in 7(e) and 7(f). SIC again provides desirable results.



(a) (b) (c)

(d) (e) (f)

Figure 8: **(a)-(b)-(c)** Application of SIC to ideal cases where $V(k)$ has two elbows in 8(a), and one elbow in 8(b) and 8(c). SIC again provides the desirable results. **(d)-(e)-(f)** Application of SIC to ideal cases where $V(k)$ has several elbows, that SIC is able to detect.