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 an automatic elbow detector (UAED) for deciding 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, dimension reduction etc. Several experiments involving synthetic and real data show the advantages of the proposed scheme with benchmark techniques in the literature.

1 Introduction

Model selection is vast and one of the most relevant task in signal processing, statistics and machine learning [1, 2]. It is the process of of selecting a statistical model from a set of candidate ones. Model selection includes as special cases very famous sub-tasks: order selection (e.g.,in polynomial functions or ARMA models), variable selection, dimension reduction, clustering [3, 4] etc.

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 analyzes and the goodness of the obtained results. Note that we use the term "variables" as a general concept that can equivalently represent variables, features or number of clusters, as an example, 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 to take into account the aforementioned trade-off [1].

The solution in the literature belongs to different families and approaches. A first class of methods is formed by the *resampling techniques*, such as cross-validation (CV) or bootstrap, where the dataset is split into a training and test sets [5, 6, 7]. 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* [8], such as the Bayesian information criterion (BIC) [9], the Akaike information criterion (AIC) [10], or the Hannan-Quinn information criterion (HQIC) [11], to name a few [2, 12]. 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 involves the knowledge of the maximum of a likelihood function is required. 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. In the Bayesian framework, the use of marginal likelihood and posterior predictive approaches are usually employed [2, 13, 14]. The connection between the marginal likelihood and information criteria is discussed in the appendices of [12]. The posterior predictive approach is related to the CV idea. Furthermore, standard frequentist approaches based on p-values have a vast use in some specific applications and deserve to be cited [15, 16]. Finally, specially in the clustering literature, some authors apply a visual inspection of an error curve looking for an "elbow".

In this work, we consider a geometric approach to design an *universal automatic elbow detector* (UAED). Our approach is inspired by the concept of the maximum "area under the curve" (AUC) in receiver operator characteristic (ROC) curves [1, 17], which is well-known and vastly employed in signal processing in machine learning. The resulting UAED technique also induces a linear penalization of the model complexity. We discuss the relationship and the advantages of the UAED with the information criteria. It is important to remark that range of applicability of the UAED is much wider than other techniques in the literature, since no likelihood function is required. The application of UAED only requires the knowledge of an error curve. Moreover, we describe several appealing features and behaviour of the UAED and test it in different numerical examples, two of them involving a real dataset. The results show the benefit of UAED with respect to other benchmark techniques in the literature.

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 applications problems (some of them are mentioned in the introduction) and specially in model selection problems, an *error function* (i.e., a fitting measure) is obtained, that we denote 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, order of the polynomial etc.), 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 [18, 19] but, in this work, $V(k)$ could directly the mean square error (MSE), or the mean absolute error (MAE). For instance, $V(k)$ can represent the prediction error in regression problem 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 addressed.
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 $V(n_2) \leq V(n_1)$.[1] 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

$$V(0) \geq V(k), \qquad \forall k.$$

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 problem. See Figure 1(a) for a graphical example of the curve $V(k)$. In some application, the score function $V(k)$ should be also convex, i.e., the differences $V(n+1) - V(n)$ will decrease as n increases. This is the case of a variable selection problem, if the variables have been ranked correctly. However, conditions regarding the concavity of $V(k)$ are not required in this work.

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 $V'(k) = V(k) - \min V(k) = V(k) - V(K)$. 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 drop for $k \geq k_{\texttt{max}}$, i.e.,

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

in this scenario we can consider $k = 0, 1, ..., k_{\texttt{max}}$, since the rest of components must be discarded due to 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 different from 0, let 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 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 concept of the maximum AUC in ROC curves [1, 17]. Namely, we desire to extract geometric information from the curve $V(k)$ looking for an "elbow" in order to determine the optimal number

[1] This condition could be also relaxed. We keep it, for the sake of simplicity.

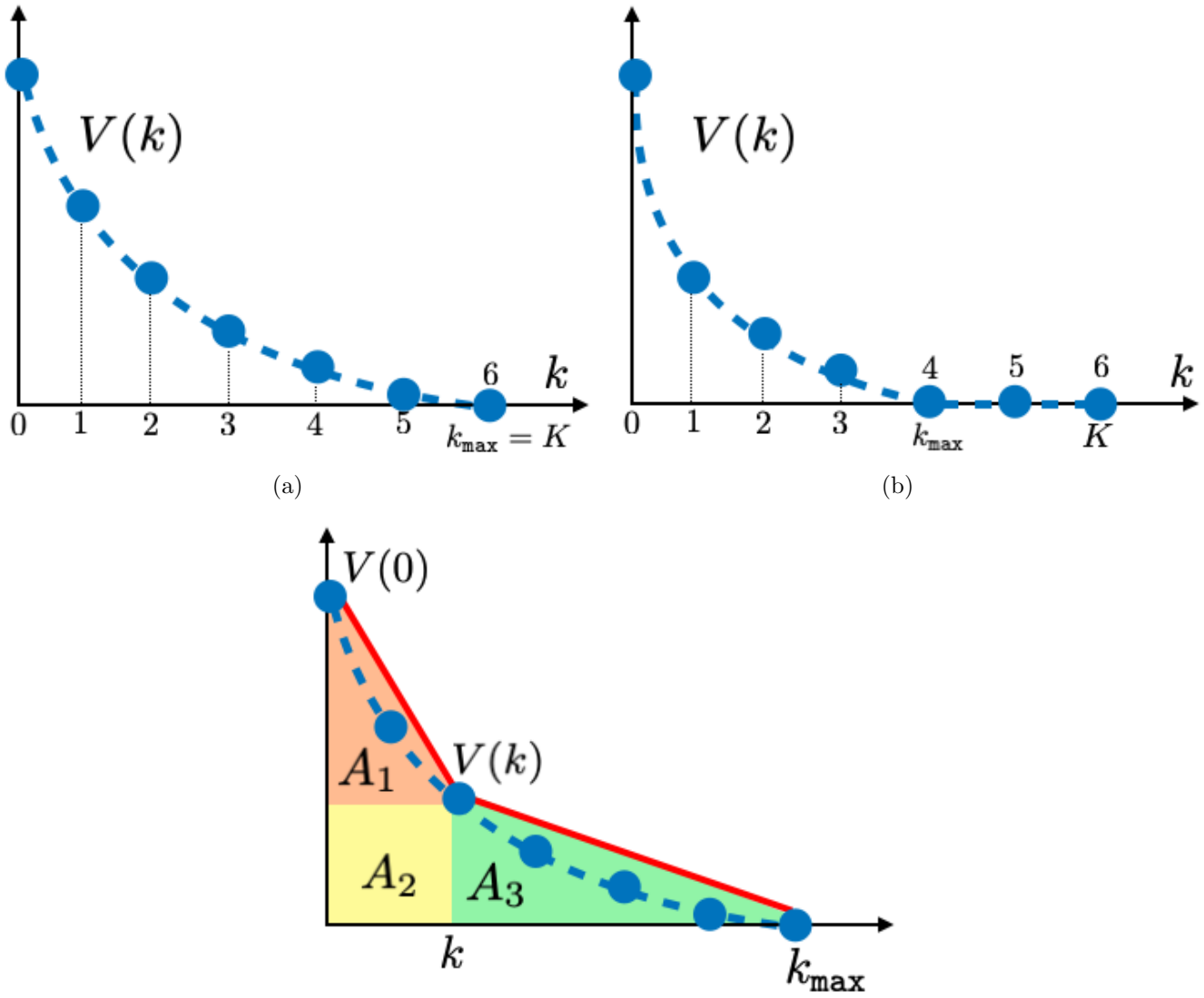


Figure 1: **(a)-(b)** Example of error curve $V(k)$ where **(a)** $k_{\max} = K = 6$, **(b)** $k_{\max} = 4$ and $K = 6$. **(c)** Construction with two straight lines and the areas A_1, A_2 and A_3.

of components, denoted $k^* \in \{0, 1..., k_{\max}\}$, to consider 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 $(k, V(k))$, $(k_{\max}, 0)$ as shown in Figure 1(c). 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 is composed by three sub-areas: two areas of two triangles (A_1 and A_3) and the area of rectangle in the middle (A_2). Namely,

we have

$$\begin{aligned} A_1 &= \frac{k \cdot (V(0) - V(k))}{2}, \\ A_2 &= k \cdot V(k), \\ A_3 &= \frac{(k_{\texttt{max}} - k) \cdot V(k)}{2}, \end{aligned} \tag{2}$$

hence the definition of k^* is

$$k^* = \arg\min_k \{A_1 + A_2 + A_3\}, \tag{3}$$

$$= \arg\min_k \left\{ \frac{V(k)}{V(0)} + \frac{k}{k_{\texttt{max}}} \right\}. \tag{4}$$

Multiplying by the constant value $V(0)$, we can equivalently write

$$k^* = \arg\min_k \left\{ V(k) + \frac{V(0)}{k_{\texttt{max}}} k \right\}. \tag{5}$$

It is important to remark that, since k belongs to discrete and finite set, solving the optimization above is straightforward (if K, or $k_{\texttt{max}}$, is not a huge value).

3.2 Second equivalent derivation

The solution offered by the expression (20) is equivalent to finding the k^* such that the difference between $V(k^*)$ and value of the straight line connecting the extreme points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$ is maximized, as depicted in Figure 2(a). More specifically, this straight line has equation

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

hence the difference that to maximize is

$$d(k) = v(k) - V(k), \tag{6}$$

$$= -\frac{V(0)}{k_{\texttt{max}}} \cdot k + V(0) - V(k), \tag{7}$$

$$= V(0) - \left(\frac{V(0)}{k_{\texttt{max}}} \cdot k + V(k) \right). \tag{8}$$

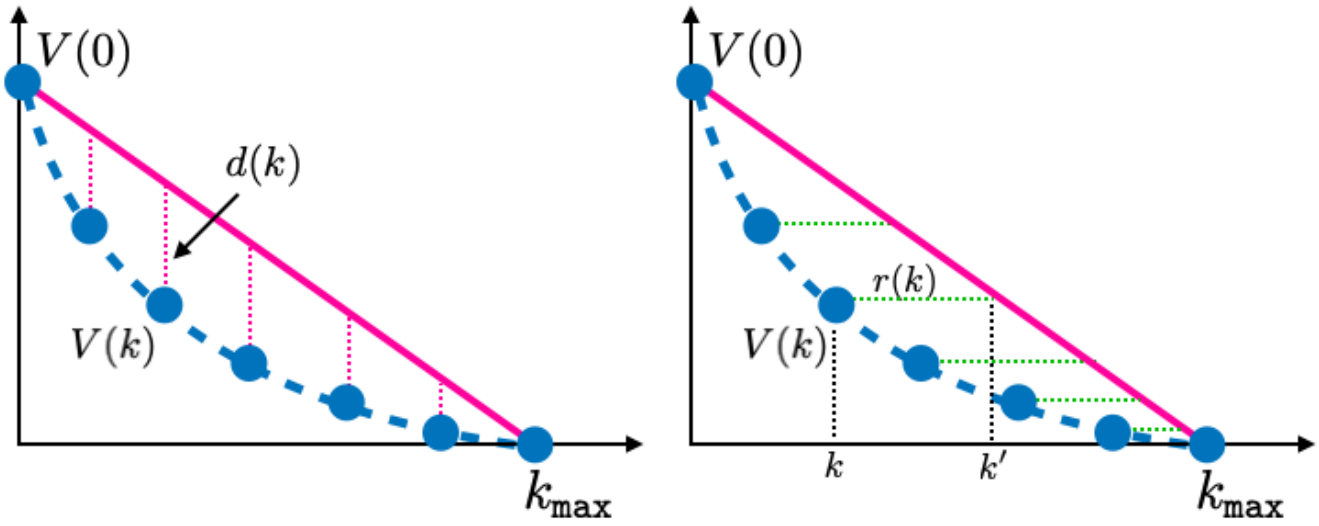
Since $V(0)$ does not depend on k (i.e., it is a constant value), we can write

$$k^* = \arg\max_k d(k) = \arg\max_k \left[V(0) - \left(\frac{V(0)}{k_{\texttt{max}}} \cdot k + V(k) \right) \right], \tag{9}$$

$$= \arg\max_k \left[- \left(\frac{V(0)}{k_{\texttt{max}}} \cdot k + V(k) \right) \right], \tag{10}$$

$$= \arg\min_k \left[\frac{V(0)}{k_{\texttt{max}}} \cdot k + V(k) \right], \tag{11}$$

which is exactly the expression in Eq. (5). Another alternative derivation is given in Appendix A and represented graphically in Figure 2(b).



(a) (b)

Figure 2: **(a)** Graphical representation of alternative derivation in Section 3.2. **(b)** Graphical representation of the other alternative derivation in Appendix A.

3.3 Relation with the information criteria

Recalling the expression in (5), i.e.,

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

here we show that this cost function can 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

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

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 a information criterion with the particular choice of $\lambda = \frac{V(0)}{k_{\max}}$. Table 1 summarizes this information.

3.4 Behaviour of the proposed solution

Analyzing the parameters involved in the expression (3) or (12), we can highlight the following considerations:

Table 1: Different information criteria and the proposed UAED.

Criterion	**Choice of λ**
Bayesian-Schwarz information criterion (BIC) [9]	$\log N$
Akaike information criterion (AIC) [10]	2
Hannan-Quinn information criterion (HQIC) [11]	$\log(\log(N))$
Universal Automatic Elbow Detector (UAED)	$\frac{V(0)}{k_{\texttt{max}}}$

• Observing Eq. (12), the penalization of the complexity of the model depends on $V(0)$ and $k_{\texttt{max}}$: since $\lambda = \frac{V(0)}{k_{\texttt{max}}}$ increasing $V(0)$ or decreasing $k_{\texttt{max}}$, intensifies the penalty. This is a reasonable and desirable behaviour. 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 less number of 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 he slope of the penalization, $\lambda = \frac{V(0)}{k_{\texttt{max}}}$, decreases in this case.
• Regarding $k_{\texttt{max}}$, we can notice that a decreasing of $k_{\texttt{max}}$ means that less components/variables produces a drop in the curve $V(k)$. On the other hand, increasing $k_{\texttt{max}}$ means that more variables causes a drop $V(k)$, so that we should consider more components, indeed, the slope of the penalization, $\lambda = \frac{V(0)}{k_{\texttt{max}}}$, decreases.

4 Experiments

4.1 Variable selection in a regression problem with real data

In several real-world applications, we observe a dataset of D pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$, where each input vector $\boldsymbol{\theta}_n = [\theta_{n,1}, ..., \theta_{n,k}]$ is formed by R variables, and the outputs y_n's are scalar values. We consider the case that $R \leq D$ and assume a linear observation model,

$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + ...\theta_R x_{n,K} + \epsilon_n, \tag{14}$$

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 [3], in the real dataset there are $K = 122$ features and $N = 1214$ number of data. The considered output represents the"arousal".
In this experiment, we set $V(k) = -2\log p(\mathbf{y}|\boldsymbol{\theta})$ after ranking the 122 variables (see [3]). Hence, we can compare also with other IC measures. We use $N = 10^6$ samples for SIC. The set $\mathcal{E}$ is

formed by $J = |\mathcal{E}| = 19 << 122$ suggested models, more specifically,

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

The number of components suggested by BIC is 17, by AIC is 44 and by Hannan-Quinn IC is 41. The UAED suggests to use 11 variables, which is a solution closer to the experts suggestions that was 7 variables [3].

5 Conclusions

A novel universal automatic elbow detector (UAED) has been introduced. Several experiments and comparisons show the benefits of the proposed UAED scheme. The relationships and differences with other information criteria given in the literature have been described and highlighted. Furthermore, the proposed procedure has a much wider range of application with respect to the other schemes in the literature.

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," *(to appear) SIAM Review - arXiv:2005.08334*, 2022.

[3] 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.

[4] A. Gupta and S. Das, "On efficient model selection for sparse hard and fuzzy center-based clustering algorithms," *Information Sciences*, vol. 590, pp. 29–44, 2022.

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

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

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] P. Stoica and Y. Selén, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, pp. 36–47, 2004.

[19] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, "The deviance information criterion: 12 years on," *J. R. Stat. Soc. B*, vol. 76, pp. 485–493, 2014.

A Other alternative derivation

Let us consider Figure 2(b). First of all, we must find the value k' such that the straight line, connecting the points $(0, V(0))$ and $(k_{\texttt{max}}, 0)$, reaches the value $V(k)$ (where $k \neq k'$, and more

precisely $k \leq k'$). Namely, we desire to obtain k' such that

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

hence

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

Now, we could also consider to maximize the following difference

$$r(k) = k' - k, \tag{15}$$

$$= -\frac{k_{\texttt{max}}}{V(0)} \left[V(k) - V(0) \right] - k, \tag{16}$$

and the elbow is defined as

$$k^* = \arg\max \; r(k) = \arg\max \left[-\frac{k_{\texttt{max}}}{V(0)} V(k) - k \right], \tag{17}$$

$$= \arg\min \left[\frac{k_{\texttt{max}}}{V(0)} V(k) + k \right], \tag{18}$$

$$= \arg\min \left[V(k) + \frac{V(0)}{k_{\texttt{max}}} k \right], \tag{19}$$

where in the last we have multiplied by the constant $V(0)$. Note that Eq. (19) is exactly the same optimization problem (i.e., with the same cost function) in Sections 3.1-3.2.

A.1 Possible extension

We have already shown that the resulting expression in Eq. (3) 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

$$k^* = \arg\min_k \left[\alpha \cdot \frac{V(k)}{V(0)} + (1 - \alpha) \cdot \frac{k}{k_{\texttt{max}}} \right]. \tag{20}$$

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 Eq. (3).