

Article

A Metric Based on the Efficient Determination Criterion

Jesús E. García , Verónica A. González-López * and Johsac I. Gomez Sanchez

Department of Statistics, University of Campinas, Campinas 13083-859, São Paulo, Brazil;
jg@ime.unicamp.br (J.E.G.); johsacgomez16@gmail.com or j216401@dac.unicamp.br (J.I.G.S.)

* Correspondence: veronica@ime.unicamp.br

Abstract: This paper extends the concept of metrics based on the Bayesian information criterion (BIC), to achieve strongly consistent estimation of partition Markov models (PMMs). We introduce a set of metrics drawn from the family of model selection criteria known as efficient determination criteria (EDC). This generalization extends the range of options available in BIC for penalizing the number of model parameters. We formally specify the relationship that determines how EDC works when selecting a model based on a threshold associated with the metric. Furthermore, we improve the penalty options within EDC, identifying the penalty $\ln(\ln(n))$ as a viable choice that maintains the strongly consistent estimation of a PMM. To demonstrate the utility of these new metrics, we apply them to the modeling of three DNA sequences of dengue virus type 3, endemic in Brazil in 2023.

Keywords: partition Markov models; Bayesian information criterion; entropy

1. Introduction

This article embarks on an exploration of the efficient determination criterion (EDC), as introduced in [1], with a particular emphasis on formulating an EDC-based metric. Our endeavor is bolstered by the presence of a Bayesian information criterion (BIC) metric proposed in [2], designed to provide consistent estimations of partition Markov models [2]. Our aim is to extend the scope of the BIC-based metric, thereby broadening the array of algorithms available for identifying partition Markov models.

To achieve our goal, we furnish a theoretical framework delineating the operational principles underlying the BIC/EDC, the BIC-based metric, and additionally, we conduct a brief survey of the current research landscape within this domain to provide context for our approach.

Let (X_t) be a discrete-time order o Markov chain on a finite and discrete alphabet Δ , with $o < \infty$; let us call $\Omega = \Delta^o$ the state space. Denote the string $a_k a_{k+1} \dots a_m$ by a_k^m , where $a_i \in \Delta$, $k \leq i \leq m$. For each $a \in \Delta$ and $s \in \Omega$, the transition probability from the state s to a is

$$P(a|s) = \text{Prob}(X_t = a | X_{t-o}^{t-1} = s). \quad (1)$$

Let $\mathcal{P} = \{\Gamma_1, \Gamma_2, \dots, \Gamma_{|\mathcal{P}|}\}$ be a partition of Ω , then for each pair of parts Γ_i and Γ_j , $i \neq j$, $i, j \in \{1, \dots, |\mathcal{P}|\}$, $\Gamma_i \cap \Gamma_j = \emptyset$, and $\Omega = \cup_{i=1}^{|\mathcal{P}|} \Gamma_i$.

Note that a part Γ of the partition is constituted by a collection of states coming from Ω ; we reformulate the notion introduced by Equation (1) as follows, for $a \in \Delta$, $\Gamma \in \mathcal{P}$:

$$\begin{aligned} P(\Gamma, a) &= \sum_{s \in \Gamma} \text{Prob}(X_{t-o}^{t-1} = s, X_t = a), \\ P(\Gamma) &= \sum_{s \in \Gamma} \text{Prob}(X_{t-o}^{t-1} = s), \\ P(a|\Gamma) &= \frac{P(\Gamma, a)}{P(\Gamma)} \text{ if } P(\Gamma) > 0. \end{aligned} \quad (2)$$



Citation: García, J.E.; González-López, V.A.; Gomez Sanchez, J.I. A Metric Based on the Efficient Determination Criterion. *Entropy* **2024**, *26*, 526. <https://doi.org/10.3390/e26060526>

Academic Editors: Carlos Alberto De Bragança Pereira, Paulo Canas Rodrigues and Mark Andrew Gannon

Received: 16 March 2024

Revised: 9 June 2024

Accepted: 17 June 2024

Published: 19 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Given the previous notation, we appeal to a model in (X_t) which allows a more efficient estimation of the transition probabilities, introduced by Equation (1); see [2].

Definition 1. Let (X_t) be a discrete-time order o Markov chain on a finite and discrete alphabet $\Delta, o < \infty$. Two states $s, r \in \Omega = \Delta^o$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r) \forall a \in \Delta$. For any $s \in \Omega$, the equivalence class of s is given by the set of states $\{r \in \Omega : r \sim_p s\}$.

The previous notion allows the definition of a Markov chain with minimal partition $\mathcal{P}$, that is, one which follows the equivalence relationship.

Definition 2. Let (X_t) be a discrete-time order o Markov chain on a finite and discrete alphabet $\Delta, o < \infty$, and let $\mathcal{P} = \{\Gamma_1, \Gamma_2, \dots, \Gamma_{|\mathcal{P}|}\}$ be a partition of $\Omega = \Delta^o$; (X_t) is a Markov chain with minimal partition $\mathcal{P}$ if $\mathcal{P}$ is defined by the relationship $\sim_p$ introduced by Definition 1.

As previously indicated, the objective of this model is to allow a more efficient estimation of the probabilities introduced by Equation (1), which occurs in the most efficient way possible by identifying the parts of the minimal partition (Definition 2), and thus, being able to use all the states inserted in each part to estimate a single probability per part. To identify the partition $\mathcal{P}$ introduced in Definition 2, a strategy must be implemented as shown below.

In a given sample x_1^n , of size n , coming from the stochastic process (X_t) under the assumptions of Definition 2, given the state $s \in \Omega$ and the element of the alphabet $a \in \Delta$, we denote the number of occurrences of s followed by a in the sample x_1^n by $N_n(s, a) = |\{t : 0 < t \leq n, x_{t-o}^{t-1} = s, x_t = a\}|$ and $N_n(s) = \sum_{a \in \Delta} N_n(s, a)$ is the number of occurrences of s in the sample x_1^n . Also, given a partition $\mathcal{P}$ of Ω , denote the number of occurrences of elements into Γ (part of $\mathcal{P}$) followed by a as

$$N_n(\Gamma, a) = \sum_{s \in \Gamma} N_n(s, a), \quad a \in \Delta, \quad (3)$$

the accumulated number of values $N_n(s)$ for $s \in \Gamma$ is denoted by

$$N_n(\Gamma) = \sum_{s \in \Gamma} N_n(s). \quad (4)$$

Note that $N_n(\Gamma, a)$ and $N_n(\Gamma)$ can be computed for any partition $\mathcal{P}$ of Ω , not only for the partition introduced by Definition 2.

The counts of occurrences, in this case $N_n(\Gamma, a)$ and $N_n(\Gamma)$, allow the estimation of probabilities (Equation (2)) subject to a modification of the likelihood function of the sample. The likelihood of the sample is

$$P(x_1^n) = P(x_1^0) \prod_{a \in \Delta, \Gamma \in \mathcal{P}} P(a|\Gamma)^{N_n(\Gamma, a)},$$

then, the maximum of the modified log-likelihood is

$$\sum_{a \in \Delta, \Gamma \in \mathcal{P}} N_n(\Gamma, a) \ln \left(\frac{N_n(\Gamma, a)}{N_n(\Gamma)} \right), \quad \text{with } N_n(\Gamma) > 0, \forall \Gamma.$$

And,

$$\frac{N_n(\Gamma, a)}{N_n(\Gamma)}, \quad \text{with } N_n(\Gamma) > 0, \quad (5)$$

is the maximum likelihood estimator of $P(a|\Gamma)$ given in Equation (2).

As shown in [2], under the assumptions of Definition 2, the partition $\mathcal{P}$ can be consistently (strong consistency) retrieved using the Bayesian information criterion (BIC), defined as

$$\text{BIC}(x_1^n, \mathcal{P}) = \sum_{a \in \Delta, \Gamma \in \mathcal{P}} N_n(\Gamma, a) \ln \left(\frac{N_n(\Gamma, a)}{N_n(\Gamma)} \right) - \frac{(|\Delta| - 1)|\mathcal{P}|}{\alpha} \ln(n), \quad (6)$$

with $\alpha > 0$, a constant value. Then, the BIC takes into consideration the maximum of the modified log-likelihood term penalized by $\frac{(|\Delta| - 1)|\mathcal{P}|}{\alpha} \ln(n)$, where $(|\Delta| - 1)|\mathcal{P}|$ is the number of probabilities to be estimated.

In practice, candidates to be the partition according to Definition 2 are compared, and the partition with the higher BIC value is considered more suitable. Also, in [2] a metric is introduced based on the BIC criterion, along with clustering algorithms, which are used to obtain $\mathcal{P}$; the metric is defined below. To achieve consistent estimation, such a metric operates on partitions of the state space that follow certain rules. The metric is then able to refine partitions until it identifies the one cited by Definition 2. The partitions in which we will apply the metric, are made up of members (parts) formed by states sharing all the transition probabilities. The definition to follow formalizes the concept.

Definition 3. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet $\Delta, o < \infty$, and state space $\Omega = \Delta^o$. Set a partition of Ω , $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$,

- i. given a part Γ of $\mathcal{P}$, Γ is a good part if $\forall a, a \in \Delta, P(a|s) = P(a|r), \forall r, s \in \Gamma, r \neq s$.
- ii. $\mathcal{P}$ is a good partition of Ω if Γ satisfies i. $\forall \Gamma \in \mathcal{P}$.

Under the validity of Definition 3-i, the probabilities introduced by Equation (2) are

$$P(a|\Gamma) = \text{Prob}(X_t = a | X_{t-o}^{t-1} = s) \quad \forall a \in \Delta, \forall s \in \Gamma, \quad (7)$$

since all the elements of the good part Γ of $\mathcal{P}$ share the transition probabilities. Note that the partition identified by Definition 2 verifies Definition 3-ii, but the reciprocal is naturally not valid. A straightforward example of a good partition is one composed of all the states being isolated.

The following introduces a notion used to estimate the minimal partition (Definition 2). This criterion operates on good parts (Definition 3-i).

Definition 4. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet $\Delta, o < \infty$, and state space $\Omega = \Delta^o$; x_1^n is a sample of the process and let $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ be a good partition of Ω ,

$$d_{\mathcal{P}}(i, j) = \frac{\alpha}{(|\Delta| - 1) \ln(n)} \sum_{a \in \Delta} \left\{ \sum_{k \in \{i, j\}} N_n(\Gamma_k, a) \ln \left(\frac{N_n(\Gamma_k, a)}{N_n(\Gamma_k)} \right) - N_n(\Gamma_{ij}, a) \ln \left(\frac{N_n(\Gamma_{ij}, a)}{N_n(\Gamma_{ij})} \right) \right\}.$$

where α is a constant and positive value, $N_n(\Gamma_{ij}) = N_n(\Gamma_i) + N_n(\Gamma_j)$, $N_n(\Gamma_{ij}, a) = N_n(\Gamma_i, a) + N_n(\Gamma_j, a)$, $\forall a \in \Delta$.

In [2], it is proved that $d_{\mathcal{P}}$ of Definition 4 is a metric, meaning that, if $\Gamma_l \in \mathcal{P}, l \in \{i, j, k\}$,

- i. $d_{\mathcal{P}}(i, j) \geq 0$, with equality, if and only if $\frac{N_n(\Gamma_i, a)}{N_n(\Gamma_i)} = \frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} \quad \forall a \in \Delta$;
- ii. $d_{\mathcal{P}}(i, j) = d_{\mathcal{P}}(j, i)$;
- iii. $d_{\mathcal{P}}(i, k) \leq d_{\mathcal{P}}(i, j) + d_{\mathcal{P}}(j, k)$.

As a consequence of the property i of a metric, the ability of $d_{\mathcal{P}}$ to operate adequately depends on the accuracy of the maximum likelihood estimation of the transition probabilities $P(a|\Gamma_i)$ and $P(a|\Gamma_j)$, $\forall a \in \Delta$. That is, when the estimators of those probabilities

$\frac{N_n(\Gamma_i, a)}{N_n(\Gamma_i)}, \frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} \forall a \in \Delta$ are near and the sample size n is large enough, we have evidence of proximity between $P(a|\Gamma_i)$ and $P(a|\Gamma_j)$, $\forall a \in \Delta$. And, such a finding indicates that the elements of both parts must be together.

Partition Markov models, commonly referred to as those delineated by Definition 2, have found application in diverse realms. For instance, they have been employed in data compression in conjunction with Huffman coding, as exemplified in [3]. Across these investigations, the utilization of the BIC-based metric $d_{\mathcal{P}}$ has proven indispensable. Also, in [2], this metric has been pivotal for modeling the behavior of internet users. The partition Markov model allows identifying the chances of a user visiting a certain internet site in their next step, based on their history, and identifies equivalent histories in the sense introduced by Definition 2.

Since the support of $d_{\mathcal{P}}$ is the BIC criterion, the question arises whether there is a broader criterion than BIC that is capable of maintaining strong consistency in the estimation of $\mathcal{P}$. The next section shows that such a criterion exists (a generalization of BIC) and was proved by [4]. Then, the next question that we propose to answer is whether such a generalization of the BIC allows the creation of a metric that generalizes the one introduced in [2].

The next section (Section 2) addresses the problem by introducing the efficient determination criterion, and then, presenting how this criterion is linked to a metric, also introducing a cut-off point that enables the practical use of the metric based on the efficient determination criterion, for sufficiently large values of n . Section 3 shows an application in which different fits of model—Definition 2—are compared, inferred by variants of the efficient determination criterion, indicated as recommended in Section 2. This article ends with the Conclusions—Section 4—in which we highlight the main contributions, and the Bibliography section.

2. Efficient Determination Criterion

Ref. [1] proposes a criterion generalizing the BIC criterion, the efficient determination criterion (EDC). In that paper, the proposal is to introduce a sequence $\{w_n\}_{n \geq 1}$ in the place of $\{\ln(n)\}_{n \geq 1}$; see Equation (6). The generalization also offers more options in the penalty term of Equation (6), instead of the number of parameters a function $\gamma(\cdot)$ is introduced acting over the number of parameters; this function is strictly increasing in the number of parameters. Under the assumptions of Definition 2, the criterion is formulated as follows:

$$\text{EDC}(x_1^n, \mathcal{P}) = \sum_{a \in \Delta, \Gamma \in \mathcal{P}} N_n(\Gamma, a) \ln \left(\frac{N_n(\Gamma, a)}{N_n(\Gamma)} \right) - \gamma \left(\frac{(|\Delta| - 1)|\mathcal{P}|}{\alpha} \right) w_n. \quad (8)$$

With $\alpha > 0$ a constant value, $\gamma(\cdot)$ being a strictly increasing function, and $\{w_n\}$ a sequence of positive numbers depending on n . As well as BIC, candidates to be the partition according to Definition 2 are compared, and the higher the EDC, the more indicated the partition is. Note that if we choose $\gamma(\cdot)$ as being the identity function, $\gamma \left(\frac{(|\Delta| - 1)|\mathcal{P}|}{\alpha} \right) = \frac{(|\Delta| - 1)|\mathcal{P}|}{\alpha}$, and $w_n = \ln(n)$, then Equation (6) is recovered. Then, clearly the EDC criterion is a generalization of the BIC criterion.

Ref. [4] proves that the EDC criterion provides a strongly consistent way to estimate the partition $\mathcal{P}$ of Definition 2 if

$$\lim_{n \rightarrow \infty} \frac{w_n}{n} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{w_n}{\ln(\ln(n))} = \infty. \quad (9)$$

Note that if we take $w_n = n^a$ for $a \in (0, 1)$, the conditions given in Equation (9) are valid. Also, we can use $w_n = a \ln(n)$ for $a > 0$. Another option is to use $w_n = n^a \ln(n)$ for $a \in (0, 1)$. Figure 1 shows penalty functions w_n verifying Equation (9). We see in the figure that functions w_n are positioned between the functions n and $\ln(\ln(n))$. And between n and $\ln(\ln(n))$ is also the w_n related to the BIC criterion ($w_n = \ln(n)$).

Clearly, the penalty $\ln(\ln(n))$ does not verify the second statement of Equation (9), but according to [5] it is an optimal penalty term for estimating the order of a Markov chain. With such inspiration in mind, the following proposition guarantees that $\ln(\ln(n))$ can also be used to obtain a consistent estimate of $\mathcal{P}$. To state the proposition we introduce the notion of relative entropy.

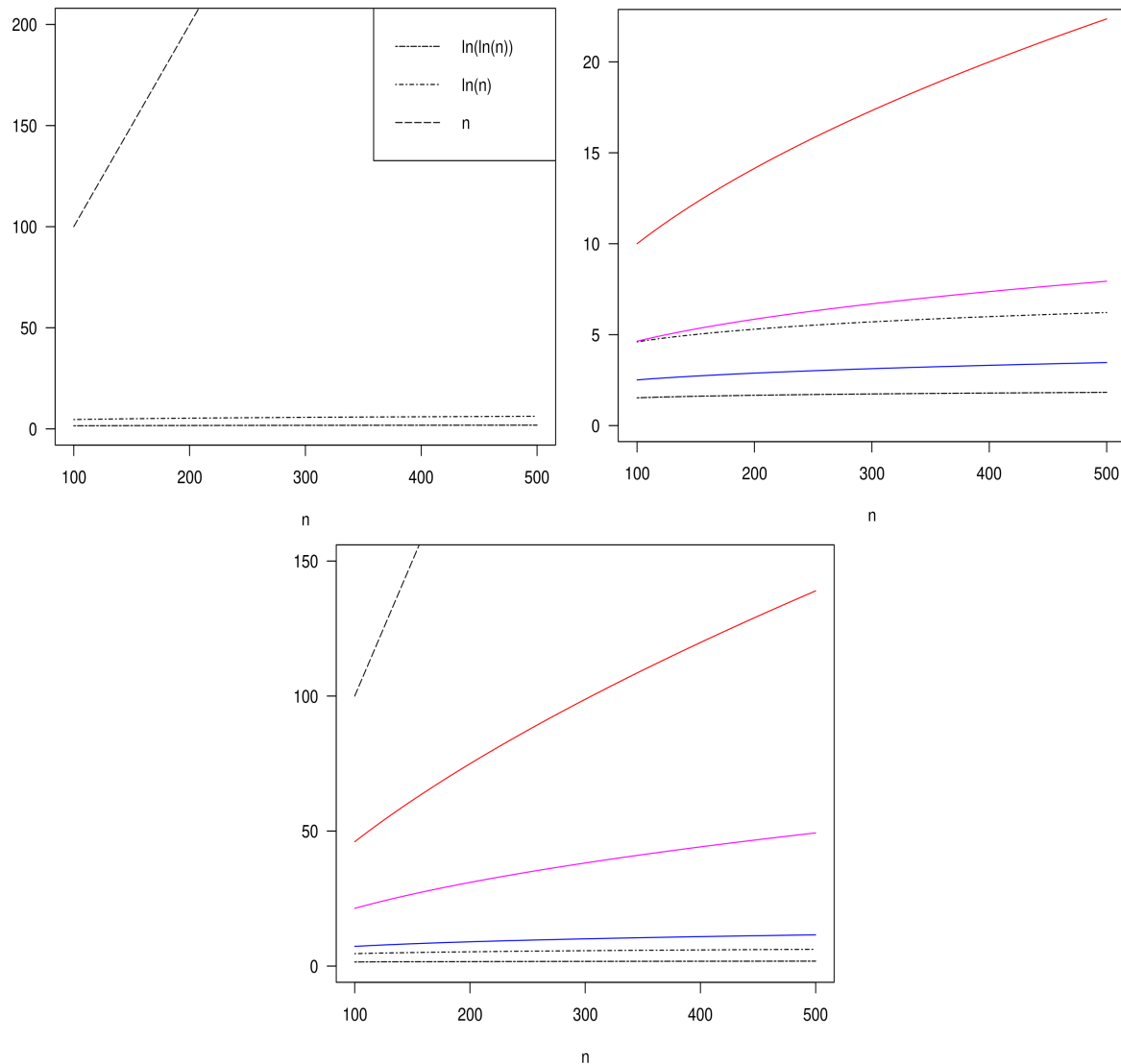


Figure 1. **Top left:** Double-dashed line, $\ln(\ln(n))$; dotted-dashed line, $\ln(n)$; long-dashed line, n . **Top right:** $w_n = n^{1/2}$ (in red); $w_n = n^{1/3}$ (in magenta); $w_n = n^{1/5}$ (in blue); double-dashed line, $\ln(\ln(n))$; dotted-dashed line, $\ln(n)$. **Bottom:** $w_n = n^{1/2} \ln(n)$ (in red); $w_n = n^{1/3} \ln(n)$ (in magenta); $w_n = n^{1/10} \ln(n)$ (in blue); double-dashed line, $\ln(\ln(n))$; dotted-dashed line $\ln(n)$; long-dashed line n .

Definition 5. Let P and Q be probability distributions on Δ . The relative entropy between P and Q is given by $D(P(\cdot)||Q(\cdot)) = \sum_{a \in \Delta} P(a) \ln\left(\frac{P(a)}{Q(a)}\right)$, with $Q(a) \neq 0, \forall a \in \Delta$.

Proposition 1. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet $\Delta, o < \infty$, and state space $\Omega = \Delta^o$; x_1^n is a sample of the process and let $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ be a partition of Ω , and $P(\cdot|\Gamma)$ be the probability given by Equation (2) related to a good part Γ (Definition 3-i). To any $\delta > 0$ there exists $\kappa > 0$ (depending on $P(\cdot|\cdot)$) such that, eventually, almost surely as $n \rightarrow \infty$

$$\left| \frac{N_n(\Gamma, a)}{N_n(\Gamma)} - P(a|\Gamma) \right| < \sqrt{\frac{\delta \ln(\ln(n))}{N_n(\Gamma)}},$$

for all Γ , good part, with $N_n(\Gamma) \geq 1$ and $o < \kappa \ln(\ln(n))$.

Proof. From the proof of Corollary 2 of [6] (on page 1621), we obtain that for any $\epsilon > 0$ there is $\kappa > 0$ (depending on $P(\cdot|\cdot)$) such that, eventually, almost surely as $n \rightarrow \infty$

$$\left| \frac{N_n(s, a)}{N_n(s)} - P(a|s) \right| < \sqrt{\frac{\epsilon \ln(\ln(n))}{N_n(s)}}, \quad (10)$$

for all $s \in \Omega$ with $N_n(s) \geq 1$ and $o < \kappa \ln(\ln(n))$.

Consider $\delta > 0$ and set $\epsilon = \frac{\delta}{|\Delta|^{2o}}$, in Equation (10), then

$$\begin{aligned} \left| \frac{N_n(s, a)}{N_n(s)} - P(a|s) \right| &\leq \sqrt{\frac{\delta \ln(\ln(n))}{|\Delta|^{2o} N_n(s)}} \\ N_n(s, a) - N_n(s)P(a|s) &\leq \sqrt{\frac{\delta \ln(\ln(n))}{|\Delta|^{2o}}} N_n(s). \end{aligned}$$

Because Γ is a good part of $\mathcal{P}$, $s \in \Gamma$, we obtain

$$\sum_{s \in \Gamma} N_n(s, a) - P(a|\Gamma) \sum_{s \in \Gamma} N_n(s) \leq \sum_{s \in \Gamma} \sqrt{\frac{\delta \ln(\ln(n))}{|\Delta|^{2o}}} N_n(s).$$

Following Equations (3), (4) and (7), we have

$$N_n(\Gamma, a) - P(a|\Gamma)N_n(\Gamma) \leq \frac{\sqrt{\delta \ln(\ln(n))}}{|\Delta|^o} \sum_{s \in \Gamma} \sqrt{N_n(s)},$$

then,

$$\begin{aligned} \left| \frac{N_n(\Gamma, a)}{N_n(\Gamma)} - P(a|\Gamma) \right| &\leq \frac{\sqrt{\delta \ln(\ln(n))}}{|\Delta|^o N_n(\Gamma)} |\Gamma| \sqrt{\max_{s \in \Gamma} N_n(s)} \\ &\leq \frac{\sqrt{\delta \ln(\ln(n))}}{|\Delta|^o N_n(\Gamma)} |\Delta|^o \sqrt{\sum_{s \in \Gamma} N_n(s)} \\ &= \frac{\sqrt{\delta \ln(\ln(n))}}{N_n(\Gamma)} \sqrt{N_n(\Gamma)} \\ &= \sqrt{\frac{\delta \ln(\ln(n))}{N_n(\Gamma)}}. \end{aligned}$$

□

The next results show that despite $\ln(\ln(n))$ violating the second condition imposed by Equation (9), the EDC (with $\ln(\ln(n))$) provides a consistent estimate of the minimal partition.

Theorem 1. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet Δ , $o < \infty$, and state space $\Omega = \Delta^o$; x_1^n is a sample of the process and let $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ be a partition of Ω , and suppose that i and j exist; $i \neq j$ such that Γ_i and Γ_j following Definition 3-i. Then, $P(a|\Gamma_i) = P(a|\Gamma_j)$, $\forall a \in \Delta$ if, and only if, eventually, almost surely as $n \rightarrow \infty$,

$$EDC(x_1^n, \mathcal{P}^{ij}) > EDC(x_1^n, \mathcal{P}).$$

where $EDC(x_1^n, \mathcal{P})$ is defined by Equation (8), with $w_n = \ln(\ln(n))$ and $EDC(x_1^n, \mathcal{P}^{ij})$ is given by Equation (8) (with $w_n = \ln(\ln(n))$) over the partition $\mathcal{P}^{ij} = \left\{ \{\mathcal{P} \setminus \{\Gamma_i\}\} \setminus \{\Gamma_j\} \right\} \cup \Gamma_{ij}$ and $\Gamma_{ij} = \{\Gamma_i \cup \Gamma_j\}$.

Proof. The proof is a variant of the one presented in [2], theorem 1. $\Leftarrow$ is direct from that proof, just considering (i) $\frac{\ln(\ln(n))}{n} \rightarrow 0$ instead of $\frac{\ln(n)}{n} \rightarrow 0$, when $n \rightarrow \infty$ and considering (ii) that $\gamma(\cdot)$ is an increasing function. For $\Rightarrow$, we have that $P(a|\Gamma_i) = P(a|\Gamma_j), \forall a \in \Delta$, and we want to prove that $EDC(x_1^n, \mathcal{P}) - EDC(x_1^n, \mathcal{P}^{ij}) < 0$. Again, following the steps in such a proof, we obtain that $EDC(x_1^n, \mathcal{P}) - EDC(x_1^n, \mathcal{P}^{ij})$ is bounded above by

$$N_n(\Gamma_i) D\left(\frac{N_n(\Gamma_i, \cdot)}{N_n(\Gamma_i)} \middle\| P(\cdot|\Gamma_i)\right) + N_n(\Gamma_j) D\left(\frac{N_n(\Gamma_j, \cdot)}{N_n(\Gamma_j)} \middle\| P(\cdot|\Gamma_j)\right) - \left(\gamma\left(\frac{(|\Delta|-1)|\mathcal{P}|}{\alpha}\right) - \gamma\left(\frac{(|\Delta|-1)(|\mathcal{P}|-1)}{\alpha}\right)\right) \ln(\ln(n)),$$

where $D(P(\cdot)|Q(\cdot))$ is the relative entropy, given by Definition 5.

For each $\Gamma \in \{\Gamma_i, \Gamma_j\}$, $\frac{N_n(\Gamma, \cdot)}{N_n(\Gamma)}$ and $P(\cdot|\Gamma)$ are probabilities on Δ ; then, Equation (11) follows from lemma 6.3 in [7]. On the other hand, since $\Gamma \in \{\Gamma_i, \Gamma_j\}$ is a good part, by hypothesis, from Proposition 1, for any $\delta > 0$ and large enough n , Equation (12) follows,

$$D\left(\frac{N_n(\Gamma, \cdot)}{N_n(\Gamma)} \middle\| P(\cdot|\Gamma)\right) \leq \sum_{a \in \Delta} \frac{\left(\frac{N_n(\Gamma, a)}{N_n(\Gamma)} - P(a|\Gamma)\right)^2}{P(a|\Gamma)} \quad (11)$$

$$\leq \sum_{a \in \Delta} \frac{\frac{\delta \ln(\ln(n))}{N_n(\Gamma)}}{P(a|\Gamma)}. \quad (12)$$

Then, set $c_0 = \left(\gamma\left(\frac{(|\Delta|-1)|\mathcal{P}|}{\alpha}\right) - \gamma\left(\frac{(|\Delta|-1)(|\mathcal{P}|-1)}{\alpha}\right)\right)$, which is >0 , since $\gamma(\cdot)$ is a strictly increasing function. For any $\delta > 0$ and large enough n ,

$$\begin{aligned} EDC(x_1^n, \mathcal{P}) - EDC(x_1^n, \mathcal{P}^{ij}) &\leq \frac{2\delta|\Delta|}{p} \ln(\ln(n)) - c_0 \ln(\ln(n)) \\ &= \ln(\ln(n)) \left(\frac{2\delta|\Delta|}{p} - c_0\right) \end{aligned}$$

where $p = \min\{P(a|\Gamma) : a \in \Delta, \Gamma \in \{\Gamma_i, \Gamma_j\}\}$. In particular, taking $\delta < \frac{pc_0}{2|\Delta|}$, for a large enough n , $EDC(x_1^n, \mathcal{P}) - EDC(x_1^n, \mathcal{P}^{ij}) < 0$. $\square$

As a result of the previous theorem, it turns out that it is possible to guarantee that the EDC with the penalty term $w_n = \ln(\ln(n))$ allows the consistent estimation of the minimal partition. As a consequence, we have:

Corollary 1. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet $\Delta, o < \infty$, and state space $\Omega = \Delta^o$; x_1^n is a sample of the process. Let Ψ be the set of all the partitions of Ω . Define

$$\mathcal{P}_n^* = \operatorname{argmax}_{\mathcal{P} \in \Psi} \{EDC(x_1^n, \mathcal{P})\}$$

where $EDC(x_1^n, \mathcal{P})$ is defined by Equation (8), with $w_n = \ln(\ln(n))$. Then, eventually, almost surely as $n \rightarrow \infty$, $\mathcal{P}^* = \mathcal{P}_n^*$, where $\mathcal{P}^*$ is the partition of Ω , following Definition 2.

Proof. Following the same steps as the proof of Theorem 3 of [2]. It is enough to replace the BIC criterion with the EDC criterion (Equation (8)) with $w_n = \ln(\ln(n))$ and apply Theorem 1 instead of Theorem 1 and Corollary 1 of [2]. $\square$

Corollary 1 complements the results of [4], showing that the minimal partition (Definition 2) is consistently recovered by the EDC (Equation (8)) when it is formulated by a strictly increasing function γ and w_n follows Equation (9), or when $w_n = \ln(\ln(n))$.

In order to generalize the BIC-based metric $d_{\mathcal{P}}$, given by Definition 4, the following notion is introduced.

Definition 6. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet Δ , $o < \infty$, and state space $\Omega = \Delta^o$; x_1^n is a sample of the process, let $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ be a good partition of Ω , and $1 \leq i, j \leq |\mathcal{P}|$, $i \neq j$:

$$\delta_{\mathcal{P}}(i, j) = v_n \sum_{a \in \Delta} \left\{ \sum_{k \in \{i, j\}} N_n(\Gamma_k, a) \ln \left(\frac{N_n(\Gamma_k, a)}{N_n(\Gamma_k)} \right) - N_n(\Gamma_{ij}, a) \ln \left(\frac{N_n(\Gamma_{ij}, a)}{N_n(\Gamma_{ij})} \right) \right\},$$

where $N_n(\Gamma_{ij}) = N_n(\Gamma_i) + N_n(\Gamma_j)$, $N_n(\Gamma_{ij}, a) = N_n(\Gamma_i, a) + N_n(\Gamma_j, a)$, $\forall a \in \Delta$. With $v_n^{-1} = w_n \left(\gamma \left(\frac{(|\Delta|-1)|\mathcal{P}|}{\alpha} \right) - \gamma \left(\frac{(|\Delta|-1)(|\mathcal{P}|-1)}{\alpha} \right) \right)$, α a constant and positive value, $\gamma(\cdot)$ being a strictly increasing function, and $\{w_n\}$ a sequence of positive numbers depending on n .

It is evident that if we take $\gamma \left(\frac{(|\Delta|-1)|\mathcal{P}|}{\alpha} \right) = \frac{(|\Delta|-1)|\mathcal{P}|}{\alpha}$ and $w_n = \ln(n)$, Definition 6 coincides with Definition 4.

The next result shows the relationship between the EDC criterion and the notion introduced in Definition 6.

Theorem 2. Let (X_t) be a Markov chain of order o , with finite and discrete alphabet Δ , $o < \infty$, and $\Omega = \Delta^o$; x_1^n is a sample of the process. Let $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ be a good partition of Ω , and $1 \leq i, j \leq |\mathcal{P}|$, $i \neq j$, then,

$$\text{EDC}(x_1^n, \mathcal{P}) < \text{EDC}(x_1^n, \mathcal{P}^{ij}) \Leftrightarrow \delta_{\mathcal{P}}(i, j) < 1.$$

where $\delta_{\mathcal{P}}(i, j)$ is given by Definition 6, $\text{EDC}(x_1^n, \mathcal{P})$ is defined by Equation (8), and $\text{EDC}(x_1^n, \mathcal{P}^{ij})$ is given by Equation (8) over the partition $\mathcal{P}^{ij} = \left\{ \{\mathcal{P} \setminus \{\Gamma_i\}\} \setminus \{\Gamma_j\} \right\} \cup \Gamma_{ij}$ and $\Gamma_{ij} = \{\Gamma_i \cup \Gamma_j\}$.

Proof.

$$\begin{aligned} \text{EDC}(x_1^n, \mathcal{P}) - \text{EDC}(x_1^n, \mathcal{P}^{ij}) &= \sum_{a \in \Delta} \sum_{k \in \{i, j\}} N_n(\Gamma_k, a) \ln \left(\frac{N_n(\Gamma_k, a)}{N_n(\Gamma_k)} \right) - \\ &\quad \sum_{a \in \Delta} N_n(\Gamma_{ij}, a) \ln \left(\frac{N_n(\Gamma_{ij}, a)}{N_n(\Gamma_{ij})} \right) - v_n^{-1} \end{aligned} \quad (13)$$

Note that $\text{EDC}(x_1^n, \mathcal{P}) < \text{EDC}(x_1^n, \mathcal{P}^{ij}) \Leftrightarrow \text{EDC}(x_1^n, \mathcal{P}) - \text{EDC}(x_1^n, \mathcal{P}^{ij}) < 0$, and $\text{EDC}(x_1^n, \mathcal{P}) - \text{EDC}(x_1^n, \mathcal{P}^{ij}) < 0 \Leftrightarrow \delta_{\mathcal{P}}(i, j) < 1$, applying Equation (13), since $v_n > 0$. $\square$

Remark 1. In order to guarantee the consistent estimation of the partition given by Definition 2, we note that Theorem 2 must be used for a large enough n and with weights w_n following Equation (9) or $w_n = \ln(\ln(n))$.

The following theorem characterizes the notion given by Definition 6 as being a metric.

Theorem 3. Let (X_t) be a Markov chain of order o over a finite and discrete alphabet Δ , $o < \infty$, $\Omega = \Delta^o$ the state space, and x_1^n a sample of the Markov process. If $\mathcal{P} = \{\Gamma_1, \dots, \Gamma_{|\mathcal{P}|}\}$ is a good partition of Ω , for each n , and for any $i, j, k \in \{1, 2, \dots, |\mathcal{P}|\}$, given $\delta_{\mathcal{P}}$ as Definition 6,

- $\delta_{\mathcal{P}}(i, j) \geq 0$, with equality, if and only if, $\frac{N_n(\Gamma_i, a)}{N_n(\Gamma_i)} = \frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} \quad \forall a \in \Delta$;
- $\delta_{\mathcal{P}}(i, j) = \delta_{\mathcal{P}}(j, i)$;

$$\text{iii. } \delta_{\mathcal{P}}(i, k) \leq \delta_{\mathcal{P}}(i, j) + \delta_{\mathcal{P}}(j, k).$$

Proof. Here, we only prove iii. since item i. is straightforward from Theorem 2 [2] and ii. follows from definition. Consider the relative entropy between two probabilities P and Q on the alphabet Δ , $D(P(\cdot) || Q(\cdot)) = \sum_{a \in \Delta} P(a) \ln(P(a)/Q(a))$. D is non-negative, furthermore D is zero if and only if $P(\cdot) = Q(\cdot)$. Returning to our goal, iii. occurs if and only if

$$0 \leq v_n^{-1}(\delta_{\mathcal{P}}(i, j) + \delta_{\mathcal{P}}(j, k) - \delta_{\mathcal{P}}(i, k)) = (*) \quad (14)$$

since $v_n > 0$. We inspect the right side of Equation (14),

$$\begin{aligned} (*) &= \sum_{s=i,k} \sum_{a \in \Delta} N_n(\Gamma_j, a) \left(\ln \left(\frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} \right) - \ln \left(\frac{N_n(\Gamma_{js}, a)}{N_n(\Gamma_{js})} \right) \right) + \\ &\quad \sum_{s=i,k} \sum_{a \in \Delta} N_n(\Gamma_s, a) \left(\ln \left(\frac{N_n(\Gamma_{ik}, a)}{N_n(\Gamma_{ik})} \right) - \ln \left(\frac{N_n(\Gamma_{sj}, a)}{N_n(\Gamma_{sj})} \right) \right) \\ &= \sum_{s=i,k} N_n(\Gamma_j) \sum_{a \in \Delta} \frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} \ln \left(\frac{N_n(\Gamma_j, a)}{N_n(\Gamma_j)} / \frac{N_n(\Gamma_{js}, a)}{N_n(\Gamma_{js})} \right) + \\ &\quad \sum_{s=i,k} \sum_{a \in \Delta} N_n(\Gamma_s, a) \ln \left(\frac{N_n(\Gamma_{ik}, a)}{N_n(\Gamma_{ik})} / \frac{N_n(\Gamma_{sj}, a)}{N_n(\Gamma_{sj})} \right) \\ &\stackrel{(1)}{=} N_n(\Gamma_j) \sum_{s=i,k} D \left(\frac{N_n(\Gamma_j, \cdot)}{N_n(\Gamma_j)} || \frac{N_n(\Gamma_{js}, \cdot)}{N_n(\Gamma_{js})} \right) + \\ &\quad \sum_{s=i,k} \sum_{a \in \Delta} N_n(\Gamma_s, a) \frac{N_n(\Gamma_{ik})}{N_n(\Gamma_{ik}, a)} \frac{N_n(\Gamma_{ik}, a)}{N_n(\Gamma_{ik})} \ln \left(\frac{N_n(\Gamma_{ik}, a)}{N_n(\Gamma_{ik})} / \frac{N_n(\Gamma_{sj}, a)}{N_n(\Gamma_{sj})} \right) \\ &\geq^{(2)} N_n(\Gamma_j) \sum_{s=i,k} D \left(\frac{N_n(\Gamma_j, \cdot)}{N_n(\Gamma_j)} || \frac{N_n(\Gamma_{js}, \cdot)}{N_n(\Gamma_{js})} \right) + \sum_{s=i,k} \frac{1}{n} D \left(\frac{N_n(\Gamma_{ik}, \cdot)}{N_n(\Gamma_{ik})} || \frac{N_n(\Gamma_{sj}, \cdot)}{N_n(\Gamma_{sj})} \right) \geq^{(3)} 0. \end{aligned}$$

where (1) follows from the definition of the relative entropy D , between two empirical laws $\frac{N_n(\Gamma_j, \cdot)}{N_n(\Gamma_j)}$ and $\frac{N_n(\Gamma_{js}, \cdot)}{N_n(\Gamma_{js})}$. (2) follows from $\frac{1}{N_n(\Gamma_{ik}, a)} \geq \frac{1}{n}$ and $N_n(\Gamma_s, a) N_n(\Gamma_{ik}) \geq 1$, and using the relative entropy D between the two empirical laws $\frac{N_n(\Gamma_{ik}, \cdot)}{N_n(\Gamma_{ik})}$ and $\frac{N_n(\Gamma_{sj}, \cdot)}{N_n(\Gamma_{sj})}$. The last inequality (3) is valid since D is non-negative. Then, Equation (14) is valid. $\square$

We saw by Theorem 2 that the improvement in the construction of the partition, when joining the parts Γ_i and Γ_j is detected by the metric $\delta_{\mathcal{P}}$ when it takes a value less than 1, a value that could be used as a reference. It is clear that this property should only be used for large enough values of n , and a penalization term following Equation (9), according to [4], or when $w_n = \ln(\ln(n))$ according to Theorem 1 and Corollary 1, which is when the EDC criterion is capable of consistently estimating the partition that follows Definition 2.

The following section shows an example of applying the metric—Definition 6—to real data. We seek to show how the model—Definition 2—varies by varying the penalty term w_n for cases where the consistency of the estimate is guaranteed, when n is large enough, that is, with terms w_n following Remark 1.

In the application we consider genetic sequences of dengue virus in FASTA format. In [8], a variant of the model specified in Definition 2 is designed specifically to model the first DNA sequence (FASTA format) of SARS-CoV-2 virus, Genbank number MN908947 (accessible at: <https://www.ncbi.nlm.nih.gov/nuccore/MN908947>, accessed on 10 March 2024). Furthermore, partition Markov models have been instrumental in modeling the DNA of the SARS-CoV-2 virus, exposing the evolution of the various variants during the pandemic period (see, for example, [9]).

3. Application

We examine and model three DNA sequences sourced from dengue virus type 3 (DENV-3), originating from Brazil. These sequences were sequenced and made publicly

available in early 2023 (<https://www.ncbi.nlm.nih.gov/>, accessed on 10 March 2024). We then proceed to compare the models derived from these sequences by applying the metric—Definition 6—and employing the agglomerative algorithm. Our analysis focuses on observing the variations in partition composition and probability magnitudes as we change the penalization term w_n .

According to [10], the genesis of the initial autochthonous case of DENV-3 (GIII-American-I lineage) in Brazil dates back to December 2000, specifically within Rio de Janeiro. Over the course of the 2000s multiple incursions of this lineage were documented from the Caribbean into Brazil. The northern and southeastern regions of Brazil swiftly emerged as the epicenters of dissemination. The advent of this lineage precipitated a significant dengue outbreak in Brazil, in Rio de Janeiro, in 2002, followed by subsequent outbreaks in diverse locales.

However, since 2010, publicly available data indicate a downward tendency in the prevalence of DENV-3; DENV-3 has represented a mere fraction ($<1\%$) of the total dengue cases in Brazil, with scant confirmed instances reported. Consequently, the transmission of DENV-3 has not been substantiated in recent years, pointing out a potential extinction of the DENV-3 (GIII-American-I lineage) within Brazil. The resurgence of DENV-3 is a real challenge in Brazil, since it is expected that the population will not have immunity, given the time that this virus has not been found in the region.

Table 1 shows the GenBank numbers, collection date, and origins of three sequences, introduced by [10]. The records correspond to three complete genetic sequences in FASTA format (alphabet $\Delta = \{a, c, g, t\}$), of DENV-3, which is already native to Brazil.

Table 1. Three autochthonous sequences of DENV-3 in FASTA format, $\Delta = \{a, c, g, t\}$.

Origin	Collection Date	GenBank	Sequence Nickname	Size
Roraima, Canta (Brazil)	4 March 2023	OQ706226	A	10,697
Roraima, Boa Vista (Brazil)	22 January 2023	OQ706227	B	10,511
Roraima, Boa Vista (Brazil)	3 January 2023	OQ706228	C	10,553

We assume that each of these sequences is a sample of a process that meets Definition 2. We proceed to fit the model—Definition 2—using the metric—Definition 6—and the agglomerative algorithm. For this, we take into account the alphabet $\Delta = \{a, c, g, t\}$, with cardinal $|\Delta| = 4$, where the sequences take their values. In Table 2, we show the frequencies for each element of the alphabet.

Table 2. Frequencies $N_n(a)$, $a \in \Delta = \{a, c, g, t\}$.

GenBank	Sequence Nickname	$N_n(a)$	$N_n(c)$	$N_n(g)$	$N_n(t)$
OQ706226	A	3435	2209	2782	2271
OQ706227	B	3378	2164	2736	2233
OQ706228	C	3395	2173	2743	2242

Considering that $\min\{10,697, 10,511, 10,553\} = 10,511$ and $\log_{|\Delta|}(10,511) = 6.68$, with integer part equal to 6, we adopt $o = 3$, since $3 < 6$ and the elements of the genetic alphabet Δ are organized in multiples of 3.

We fit four scenarios for each of the three sequences OQ706226, OQ706227, and OQ706228, with each scenario governed by a different penalty w_n . All of them are considered in Definition 6, $\Delta = \{a, c, g, t\}$, $o = 3$, $\alpha = 2$ (see [11]), and the γ function is the identity function. For each penalty, we identify, using the metric (Definition 6), the estimated partition of the partition given by Definition 2, and then, determine the transition probabilities of each part for each element of Δ . We denote by Γ_i^v the part i estimated for the sequence v , where v can be A, B, C, corresponding to OQ706226, OQ706227, and OQ706228, respectively. In Tables 3 and 4, we record the results for the three sequences with penalty $w_n = n^{1/2}$; Tables 5 and 6 report the results for the three sequences with penalty $w_n = n^{1/3}$. While Tables 7 and 8 show, for the three sequences, the results using the usual BIC penalty

($w_n = \ln(n)$). Finally, Tables 9–11 report the results with the penalty $w_n = \ln(\ln(n))$ (see Corollary 1): for the sequence OQ706226, Table 9; for the sequence OQ706227, Table 10; and for the sequence OQ706228, Table 11.

Table 3. Minimal partition—Definition 2—estimated by Definition 6. From top to bottom, for the sequences *A* (GenBank OQ706226), *B* (GenBank OQ706227), and *C* (GenBank OQ706228). $\Delta = \{a,c,g,t\}, o = 3, w_n = n^{1/2}, \alpha = 2, \gamma$ function given by the identity.

Sequence	Part	States
OQ706226	Γ_1^A	ccc, tgc, cgc, ggc, gac, gtc, acg, att, gct, gcc, atc, agc, ggt, gag, ata, tcc, agt, tct, ctc, ttc, cct, ctt, gtt, aaa, ttt, aag, ccg, gcg, tgg, gtg, tat, agg, ctg, gat, ttg, aat, tcg, cgg, cat, ggg, act, atg
	Γ_2^A	cag, cac, taa, cgt, aga, gga, tta, tac, cga, acc, tag, gta, cca, tga, tca, cta, gca, aac, tgt, gaa, caa, aca
OQ706227	Γ_1^B	ccc, ggc, cgc, gac, gtc, acg, gct, tag, gcc, atc, agc, ggt, gag, ata, att, tcc, agt, tgc, tct, gtt, ctc, ttc, cct, ctt, aaa, ttt, aag, ccg, gcg, tgg, gtg, tat, agg, ctg, ttg, aat, gat, tcg, cgg, cat, act, ggg, atg
	Γ_2^B	cag, cgt, aga, gga, tta, tac, cga, cac, acc, taa, tga, cta, cca, tca, gca, gta, aac, tgt, gaa, caa, aca
OQ706228	Γ_1^C	ccc, cgc, ggc, gac, tgc, gct, tcc, tct, agt, gtc, acg, att, tag, gcc, atc, agc, ggt, gag, ata, ctc, cct, ttc, aaa, ctt, gtt, ttt, aag, ccg, gcg, tgg, gtg, agg, ctg, ttg, aat, gat, tcg, cat, tat, cgg, ggg, atg, act
	Γ_2^C	cag, cgt, tta, cac, aga, gga, cga, tac, acc, taa, tga, cta, gca, cca, tca, gta, aac, tgt, gaa, caa, aca

Table 4. Transition probabilities—Equation (2)—estimated by Equation (5). From top to bottom, for the sequences *A* (GenBank OQ706226), *B* (GenBank OQ706227), and *C* (GenBank OQ706228). $\Delta = \{a,c,g,t\}, w_n = n^{1/2}$; full estimated partitions displayed in Table 3. In bold, the highest probability per part.

OQ706226	<i>i</i>	$P(a \Gamma_i^A)$	$P(c \Gamma_i^A)$	$P(g \Gamma_i^A)$	$P(t \Gamma_i^A)$
	1	0.37032	0.21336	0.21256	0.20376
	2	0.20595	0.19061	0.37152	0.23192
OQ706227	<i>i</i>	$P(a \Gamma_i^B)$	$P(c \Gamma_i^B)$	$P(g \Gamma_i^B)$	$P(t \Gamma_i^B)$
	1	0.36846	0.21441	0.21374	0.20339
	2	0.20723	0.18508	0.37341	0.23428
OQ706228	<i>i</i>	$P(a \Gamma_i^C)$	$P(c \Gamma_i^C)$	$P(g \Gamma_i^C)$	$P(t \Gamma_i^C)$
	1	0.36871	0.21373	0.21387	0.20369
	2	0.20760	0.18681	0.37167	0.23392

Table 5. Minimal partition—Definition 2—estimated by Definition 6. From top to bottom, for the sequences *A* (GenBank OQ706226), *B* (GenBank OQ706227), and *C* (GenBank OQ706228). $\Delta = \{a,c,g,t\}, o = 3, w_n = n^{1/3}, \alpha = 2, \gamma$ function given by the identity.

Sequence	Part	States
OQ706226	Γ_1^A	ccc, tgc, cgc, ggc, gac, gtc, acg, att, gct, gcc, atc, agc, ggt, gag, ata
	Γ_2^A	tcc, tct, agt, ctc, ttc, cct, ctt, gtt, aaa, ttt
	Γ_3^A	aag, ccg, gcg, tgg, gtg, agg, ctg, gat, ttg, aat, tcg, cgg, cat, ggg, tat, act, atg
	Γ_4^A	cag, cac, taa, cta, cgt, aga, gga, tta, tac, cga, acc, tag, gta, cca, tga, tca, aac, tgt, gaa, caa, aca, gca
OQ706227	Γ_1^B	ccc, tgc, tct, gtt, cgc, ggc, gac, gtc, gct, acg, att, tag, gcc, atc, agc, ggt, gag, ata
	Γ_2^B	tcc, agt, ctc, ttc, cct, ctt, aaa, ttt
	Γ_3^B	aag, ccg, gcg, tgg, gtg, agg, ctg, ttg, aat, gat, tcg, cgg, cat, tat, act, ggg, atg
	Γ_4^B	cag, cgt, aga, gga, tta, tac, cga, cac, acc, tgt, gaa, caa, taa, tga, cta, cca, tca, aca, gca, gta, aac
OQ706228	Γ_1^C	ccc, cgc, ggc, gac, tcc, tct, agt, tgc, gct, gtc, acg, att, gcc, atc, agc, ggt, gag, ata
	Γ_2^C	ctc, cct, ttc, aaa, ctt, gtt, ttt
	Γ_3^C	aag, ccg, gcg, tat, tgg, gtg, agg, ctg, ttg, aat, gat, tcg, cat, cgg, act, ggg, atg
	Γ_4^C	cag, taa, tta, cac, cgt, cga, tac, acc, aga, gga, tag, gta, cca, tca, tga, cta, gca, aac, tgt, gaa, caa, aca

Table 6. Transition probabilities—Equation (2)—estimated by Equation (5). From top to bottom, for the sequences A (GenBank OQ706226), B (GenBank OQ706227), and C (GenBank OQ706228). $\Delta = \{a, c, g, t\}$, $w_n = n^{1/3}$; full estimated partitions displayed in Table 5. In bold, the highest probability per part.

OQ706226	i	$P(a \Gamma_i^A)$	$P(c \Gamma_i^A)$	$P(g \Gamma_i^A)$	$P(t \Gamma_i^A)$
	1	0.29882	0.22627	0.25608	0.21882
	2	0.40166	0.18760	0.25987	0.15087
	3	0.41291	0.22673	0.11709	0.24328
	4	0.20595	0.19061	0.37152	0.23192
OQ706227	i	$P(a \Gamma_i^B)$	$P(c \Gamma_i^B)$	$P(g \Gamma_i^B)$	$P(t \Gamma_i^B)$
	1	0.30233	0.22596	0.25963	0.21208
	2	0.40612	0.18593	0.25902	0.14893
	3	0.41410	0.22668	0.11608	0.24314
	4	0.20723	0.18508	0.37341	0.23428
OQ706228	i	$P(a \Gamma_i^C)$	$P(c \Gamma_i^C)$	$P(g \Gamma_i^C)$	$P(t \Gamma_i^C)$
	1	0.31038	0.22205	0.26436	0.20321
	2	0.42182	0.17773	0.24578	0.15467
	3	0.41362	0.22573	0.11644	0.24422
	4	0.20761	0.19017	0.37147	0.23075

Table 7. Minimal partition—Definition 2—estimated by Definition 6. From top to bottom, for the sequences A (GenBank OQ706226), B (GenBank OQ706227), and C (GenBank OQ706228). $\Delta = \{a, c, g, t\}$, $o = 3$, $w_n = \ln(n)$, $\alpha = 2$, γ function given by the identity.

Sequence	Part	States
OQ706226	Γ_1^A	ccc, tgc, cgc, ggc, gac, gtc, acg, att, gct, gcc, atc, agc, ggt, gag, ata
	Γ_2^A	tcc, agt, ctc, tct, ttc, cct, ctt, gtt, aaa, ttt
	Γ_3^A	aag, ccg, gcg, tgg, gtg, act, agg, ctg, gat, ttg, aat, tcg, cgg, cat, ggg, tat, atg
	Γ_4^A	cag, cac, taa, cgt, aga, gga, tta, tac, cga, acc, tag, gta, cca, tca, tga, cta, aac
	Γ_5^A	tgt, gaa, caa, aca, gca
OQ706227	Γ_1^B	ccc, tgc, tct, gtt, cgc, ggc, gac, gtc, gct, acg, att, tag
	Γ_2^B	gcc, atc, agc, ggt, gag, ata
	Γ_3^B	tcc, agt, ctc, ttc, cct, ctt, aaa, ttt
	Γ_4^B	aag, ccg, gcg, tgg, gtg, agg, ctg, ttg, aat, gat, tcg, cgg, cat, tat, act, ggg, atg
	Γ_5^B	cag, cgt, aga, gga, tta, tac, cga, cac, acc
	Γ_6^B	tgt, gaa, caa, taa, tga, cta, cca, tca, aca, gca, gta, aac
OQ706228	Γ_1^C	ccc, cgc, ggc, gac, tcc, tct, agt, tgc, gct, gtc, acg, att
	Γ_2^C	gcc, atc, agc, ggt, gag, ata
	Γ_3^C	ctc, cct, ttc, aaa, ctt, gtt, ttt
	Γ_4^C	aag, ccg, gcg, tgg, gtg, agg, ctg, ttg, aat, gat, tcg, cgg, cat, tat, act, ggg, atg
	Γ_5^C	cag, taa, tta, cac, cgt, cga, tac, acc, aga, gga
	Γ_6^C	tag, gta, cca, tca, tga, cta, gca, aac, tgt, gaa, caa, aca

We observe from Tables 3, 5, 7 and 9 (right)–11 (right) that as the penalty w_n is reduced (that is, when w_n approaches the lower limit $\ln(\ln(n))$), the model is allowed to acquire more parameters, in this case, more parts.

Given a penalization w_n , the three sequences, A (GenBank OQ706226), B (GenBank OQ706227), and C (GenBank OQ706228), show a similar number of parts. More specifically, for the penalty $w_n = n^{1/2}$ the behavior of the three sequences is represented by two parts (Table 3), for $w_n = n^{1/3}$ the behavior of the three sequences is described by four parts (Table 5). For $w_n = \ln(n)$, OQ706226 is modeled by five parts while the other two are modeled by six parts (see Table 7). For penalty $w_n = \ln(\ln(n))$, OQ706226 is modeled by a partition with 13 parts (see Table 9, right) while the other two sequences are modeled by 14 parts, see Tables 10 (right) and 11 (right).

Table 8. Transition probabilities—Equation (2)—estimated by Equation (5). From top to bottom, for the sequences *A* (GenBank OQ706226), *B* (GenBank OQ706227), and *C* (GenBank OQ706228). $\Delta = \{a,c,g,t\}, w_n = \ln(n)$; full estimated partitions displayed in Table 7. In bold, the highest probability per part.

OQ706226	<i>i</i>	$P(a \Gamma_i^A)$	$P(c \Gamma_i^A)$	$P(g \Gamma_i^A)$	$P(t \Gamma_i^A)$
	1	0.29882	0.22627	0.25608	0.21882
	2	0.40166	0.18760	0.25987	0.15087
	3	0.41291	0.22673	0.11709	0.24328
	4	0.18351	0.20596	0.35967	0.25086
	5	0.26507	0.15017	0.40273	0.18203
OQ706227	<i>i</i>	$P(a \Gamma_i^B)$	$P(c \Gamma_i^B)$	$P(g \Gamma_i^B)$	$P(t \Gamma_i^B)$
	1	0.30945	0.21394	0.28167	0.19493
	2	0.28468	0.25573	0.20507	0.25452
	3	0.40612	0.18593	0.25902	0.14893
	4	0.41410	0.22668	0.11608	0.24314
	5	0.18868	0.20926	0.31990	0.28216
	6	0.21860	0.17026	0.40620	0.20494
OQ706228	<i>i</i>	$P(a \Gamma_i^C)$	$P(c \Gamma_i^C)$	$P(g \Gamma_i^C)$	$P(t \Gamma_i^C)$
	1	0.31933	0.21040	0.28482	0.18545
	2	0.28451	0.25570	0.20528	0.25450
	3	0.42182	0.17773	0.24578	0.15467
	4	0.41362	0.22573	0.11644	0.24422
	5	0.19202	0.20975	0.31905	0.27917
	6	0.21932	0.17546	0.41088	0.19434

Table 9. Sequence OQ706226. Right: Minimal partition—Definition 2—estimated by Definition 6. Left: Transition probabilities—Equation (2)—estimated by Equation (5); in bold, the highest probability per part. $\Delta = \{a,c,g,t\}, o = 3, \gamma$ function given by the identity, $w_n = \ln(\ln(n))$.

<i>i</i>	$P(a \Gamma_i^A)$	$P(c \Gamma_i^A)$	$P(g \Gamma_i^A)$	$P(t \Gamma_i^A)$	Part	States
1	0.31140	0.19342	0.28411	0.21107	Γ_1^A	ccc, tgc, cgc, ggc, gac
2	0.32218	0.25314	0.19456	0.23013	Γ_2^A	gcc, atc
3	0.38732	0.19249	0.27934	0.14085	Γ_3^A	tcc, tct, agt, ctc, ttc, cct
4	0.23757	0.25967	0.21823	0.28453	Γ_4^A	agc, ggt, gag, ata
5	0.28879	0.26078	0.27371	0.17672	Γ_5^A	gtc, acg, att, gct
6	0.40686	0.21316	0.14272	0.23726	Γ_6^A	aag, ccg, gcg, tgg, gtg, ttg, tat
7	0.19113	0.20930	0.32195	0.27762	Γ_7^A	cag, cac, taa, cgt, aga, gga, tta, tac, cga, acc
8	0.17164	0.29104	0.41791	0.11940	Γ_8^A	tag, gta
9	0.46675	0.21455	0.08532	0.23337	Γ_9^A	tcg, cgg, cat, ggg, act, atg
10	0.34566	0.27172	0.11275	0.26987	Γ_{10}^A	agg, ctg, gat, aat
11	0.26507	0.15017	0.40273	0.18203	Γ_{11}^A	tgt, gaa, caa, aca, gca
12	0.43116	0.17754	0.21981	0.17150	Γ_{12}^A	ctt, gtt, aaa, ttt
13	0.17246	0.18610	0.41439	0.22705	Γ_{13}^A	cca, tca, tga, cta, aac

The formal determination of whether the identified models, under each penalty, exhibit significant differences lies beyond the scope of this application. However, we acknowledge it as an open question worthy of further exploration.

The following observation applies to all three sequences. Observing the magnitudes of the transition probabilities, marked in bold, we note that, as reported in Tables 6, 8, and 9 (left)–11 (left), there is a predominant number of parts whose prevalence is the transition to element a of the alphabet $\{a,c,g,t\}$, and secondly, parts that indicate a prevalence for element g of the alphabet. As for Table 4, which reports the most penalized case ($w_n = n^{1/2}$), one part is recorded with prevalence for a and another with prevalence

for g, which is natural, since the model only has two parts. As Tables 9–11 show, based on the penalty $w_n = \ln(\ln(n))$, the three sequences show the same part {agc,ggg,gag,ata} with a prevalence for the element t of the alphabet Δ , of lower magnitude, than those previously mentioned.

Table 10. Sequence OQ706227. Right: Minimal partition—Definition 2—estimated by Definition 6. Left: Transition probabilities—Equation (2)—estimated by Equation (5); in bold, the highest probability per part. $\Delta = \{a,c,g,t\}$, $o = 3$, γ function given by the identity, $w_n = \ln(\ln(n))$.

i	$P(a \Gamma_i^B)$	$P(c \Gamma_i^B)$	$P(g \Gamma_i^B)$	$P(t \Gamma_i^B)$	Part	States
1	0.34479	0.19313	0.27725	0.18483	Γ_1^B	ccc, tgc, gtt, tct
2	0.31992	0.25636	0.19280	0.23093	Γ_2^B	gcc, atc
3	0.39127	0.19183	0.27978	0.13712	Γ_3^B	tcc, agt, ctc, ttc, cct
4	0.23810	0.25490	0.22129	0.28571	Γ_4^B	agc, ggt, gag, ata
5	0.29094	0.19737	0.28509	0.22661	Γ_5^B	cgc, ggc, gac
6	0.27672	0.26908	0.28435	0.16985	Γ_6^B	gtc, gct, acg, att, tag
7	0.40705	0.21628	0.14945	0.22722	Γ_7^B	aag, ccg, gcg, tgg, gtg
8	0.18868	0.20926	0.31990	0.28216	Γ_8^B	cag, cgt, aga, gga, tta, tac, cga, cac, acc
9	0.47018	0.20642	0.08601	0.23739	Γ_9^B	tcg, cgg, cat, tat, act, ggg, atg
10	0.35015	0.26558	0.11424	0.27003	Γ_{10}^B	agg, ctg, ttg, aat, gat
11	0.27992	0.13996	0.42191	0.15822	Γ_{11}^B	tgt, gaa, caa
12	0.43490	0.17450	0.21880	0.17181	Γ_{12}^B	ctt, aaa, ttt
13	0.21127	0.17647	0.39188	0.22038	Γ_{13}^B	taa, tga, cta, cca, tca, aca, gca
14	0.11330	0.20690	0.45321	0.22660	Γ_{14}^B	gta, aac

Table 11. Sequence OQ706228. Right: Minimal partition—Definition 2—estimated by Definition 6. Left: Transition probabilities—Equation (2)—estimated by Equation (5); in bold, the highest probability per part. $\Delta = \{a,c,g,t\}$, $o = 3$, γ function given by the identity, $w_n = \ln(\ln(n))$.

i	$P(a \Gamma_i^C)$	$P(c \Gamma_i^C)$	$P(g \Gamma_i^C)$	$P(t \Gamma_i^C)$	Part	States
1	0.30740	0.18975	0.28463	0.21822	Γ_1^C	ccc, cgc, ggc, gac
2	0.31924	0.25581	0.19450	0.23044	Γ_2^C	gcc, atc
3	0.35319	0.20914	0.29501	0.14266	Γ_3^C	tcc, tct, agt
4	0.23889	0.25556	0.21944	0.28611	Γ_4^C	agc, ggt, gag, ata
5	0.30048	0.24642	0.27345	0.17965	Γ_5^C	tgc, gct, gtc, acg, att
6	0.41354	0.17708	0.26875	0.14063	Γ_6^C	ctc, ttc, cct
7	0.41304	0.21196	0.14565	0.22935	Γ_7^C	aag, ccg, gcg, tat, tgg, gtg
8	0.19202	0.20975	0.31905	0.27917	Γ_8^C	cag, taa, tta, cac, cgt, cga, tac, acc, aga, gga
9	0.17778	0.28889	0.41481	0.11852	Γ_9^C	tag, gta
10	0.47315	0.20716	0.08440	0.23529	Γ_{10}^C	tcg, cat, cgg, act, ggg, atg
11	0.34564	0.26588	0.11374	0.27474	Γ_{11}^C	agg, ctg, ttg, aat, gat
12	0.27205	0.15097	0.41555	0.16144	Γ_{12}^C	tgt, gaa, caa, aca
13	0.43154	0.17848	0.21883	0.17115	Γ_{13}^C	ctt, gtt, aaa, ttt
14	0.18957	0.17653	0.40722	0.22668	Γ_{14}^C	cca, tca, tga, cta, gca, aac

4. Conclusions

The main objective of this paper, developed in Section 2, is to introduce a new notion based on the Equation (8), as given in Definition 6. This concept is used to identify the minimal partition of a Markov chain—Definition 2. Theorem 3 proves that the concept in Definition 6 constitutes a metric. Furthermore, Theorem 2 establishes the relationship between this new metric and the operation of the EDC criterion, showing that in an iterative process, selecting a partition with a higher EDC value is equivalent to using the value 1 as a

threshold in the metric. In this way, we achieve our main goal of proposing an EDC-based metric to estimate the minimal partition.

Our results add to those of [4], in the search to characterize penalty terms that can be used in the EDC criterion to obtain the consistent estimation of the minimal partition. Ref. [4] demonstrates that the EDC, under certain conditions on the term w_n (Equation (9)), provides a strongly consistent estimate of the minimal partition, as defined in Definition 2. Building on the results from [5], we conjectured that using $w_n = \ln(\ln(n))$ might preserve strong consistency, even though this term does not satisfy the second condition imposed by Equation (9). We confirm in Theorem 1 and Corollary 1 that strong consistency is indeed achieved using the EDC with the penalization term $w_n = \ln(\ln(n))$.

We conclude the article with an application demonstrating the effect of the metric introduced in Definition 6 on estimating the minimal partition—Definition 2, using various penalty terms discussed in Remark 1. For this purpose, we analyze three Dengue virus type 3 sequences, native to Brazil and collected in 2023, in FASTA format. The application shows that relaxing the penalty results in higher cardinalities for the estimated partition. We identify which parts (collections of states) of the Dengue sequences have a greater or lesser preference for transitioning to the next element (a, c, g, or t) in the alphabet $\Delta = \{a, c, g, t\}$. As expected, the models identified for each sequence exhibit similar features when the penalty is applied, which is natural given that the sequences share the same collection date and region of origin.

Author Contributions: Conceptualization, J.E.G. and V.A.G.-L.; Methodology, J.E.G. and V.A.G.-L.; Software, J.E.G. and V.A.G.-L.; Validation, J.E.G., V.A.G.-L. and J.I.G.S.; Formal analysis, J.E.G. and V.A.G.-L.; Investigation, J.E.G., V.A.G.-L. and J.I.G.S.; Resources, J.E.G. and V.A.G.-L.; Data curation, J.E.G. and V.A.G.-L.; Writing—original draft, J.E.G. and V.A.G.-L.; Writing—review & editing, J.E.G. and V.A.G.-L.; Visualization, J.E.G. and V.A.G.-L.; Supervision, J.E.G. and V.A.G.-L.; Project administration, J.E.G. and V.A.G.-L.; Funding acquisition, J.E.G. and V.A.G.-L. All authors have read and agreed to the published version of the manuscript.

Funding: Johsac I. Gomez Sanchez gratefully acknowledge the financial support provided by CAPES with fellowships from the Master Graduate Program in Statistics—University of Campinas.

Data Availability Statement: The National Center for Biotechnology Information Advances Science and Health, <https://www.ncbi.nlm.nih.gov/> (accessed on 10 March 2024).

Acknowledgments: The authors wish to express their gratitude to the three referees and the editors for their helpful comments on an earlier draft of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhao, L.C.; Dorea, C.C.Y.; Gonçalves, C.R. On determination of the order of a Markov chain. *Stat. Inference Stoch. Process.* **2001**, *4*, 273–282. [\[CrossRef\]](#)
2. García Jesús, E.; González-López, V.A. Consistent Estimation of Partition Markov Models. *Entropy* **2017**, *19*, 160. [\[CrossRef\]](#)
3. García, J.E.; González-López, V.A.; Tasca, G.H.; Yaginuma, K.Y. An Efficient Coding Technique for Stochastic Processes. *Entropy* **2022**, *24*, 65. [\[CrossRef\]](#)
4. Pereira, D.F.S. Critério de Determinação Eficiente Para Estimção de Cadeias de Markov de Partição Mínima. Master's Thesis, University of Brasilia, Brasilia, Federal District, Brazil, 2021. Available online: <http://repositorio2.unb.br/jspui/handle/10482/42891> (accessed on 10 March 2024).
5. Dorea, C.C.Y. Optimal penalty term for EDC Markov chain order estimator. *Ann. de l'ISUP* **2008**, *52*, 15–25. Available online: <https://hal.science/hal-03633210> (accessed on 10 March 2024).
6. Csiszár, I. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inf. Theory* **2002**, *48*, 1616–1628. [\[CrossRef\]](#)
7. Csiszár, I.; Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Theory* **2006**, *52*, 1007–1016. [\[CrossRef\]](#)
8. García, J.E.; González-López, V.A.; Tasca, G.H. Partition Markov Model for COVID-19 Virus. *4open* **2020**, *3*, 13. [\[CrossRef\]](#)
9. García, J.E.; González-López, V.A.; Tasca, G.H. Multiple partition Markov model for B.1.1.7, B.1.351, B.1.617.2, and P.1 variants of SARS-CoV 2 virus. *Comput. Stat.* **2022**. [\[CrossRef\]](#)

10. Naveca, F.G.; Santiago, G.A.; Maito, R.M.; Meneses, C.A.R.; do Nascimento, V.A.; de Souza, V.C.; do Nascimento, F.O.; Silva, D.; Mejía, M.; Gonçalves, L.; et al. Reemergence of dengue virus serotype 3, Brazil, 2023. *Emerg. Infect. Dis.* **2023**, *29*, 1482–1484. [[CrossRef](#)]
11. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.