Independent block identification in multivariate time series

Florencia Leonardi[*][†], Matías Lopez-Rosenfeld[‡], Daniela Rodriguez[‡], Magno T. F. Severino[†], Mariela Sued[‡]

July 10, 2020

Abstract

In this work we propose a model selection criterion to estimate the points of independence of a random vector, producing a decomposition of the vector distribution function into independent blocks. The method, based on a general estimator of the distribution function, can be applied for discrete or continuous random vectors, and for iid data or dependent time series. We prove the consistency of the approach under general conditions on the estimator of the distribution function and we show that the consistency holds for iid data and discrete time series with mixing conditions. We also propose an efficient algorithm to approximate the estimator and show the performance of the method on simulated data. We apply the method in a real dataset to estimate the distribution of the flow over several locations on a river, observed at different time points.

Keywords: Model selection, regularized estimator, structure estimation, dimensionality reduction.

MSC: 62M10; 62M15; 60G10

1 Introduction

The discharge of water flowing in a river or a channel is measured using stream gauges. Let X_u denote the flow recorded at the uth gauging station ($u = 1, \ldots, d$) and $\mathbf{X}$ the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ containing the d records. Let us suppose this random vector is observed on different days, and denote by $\mathbf{X}^{(i)} = (X_1^i, \ldots, X_d^i)$ the vector observed at the ith day. Time series are one of the most popular tools to model the process $\{\mathbf{X}^{(i)} : 1 \leq i \leq n\}$, where $\mathbf{X}^{(i)} \in \mathbb{R}^d$. In general, the number of parameters to be estimated is polynomial in the dimension d, and this could be large in comparison to the sample size n, leading to overfitting. In examples such as the water discharge presented above, the river dynamics may generate

[*] Corresponding author: florencia@usp.br
[†] Universidade de São Paulo
[‡] Universidad de Buenos Aires

independence in the behavior of some points of its course. In this case, a hydroelectric dam or a interbasin transfer can cause independence among observations taken before and after these man-made interventions. For instance, we can consider that $(X_1, \ldots, X_u)$ is independent of $(X_{u+1}, \ldots, X_d)$; therefore we can model the first u gauges and the rest separately. Thus, we reduce the dimension of the problem in such a way that we can work with independent processes in u and $d - u$ dimensions, respectively. It could also occur that $(X_1, \ldots, X_u)$, $(X_{u+1}, \ldots, X_v)$, and $(X_{v+1}, \ldots, X_d)$ are independent, giving rise to spaces with even smaller dimensions. The aim of this work is to propose a method to determine the greater possible decomposition of the vector $\mathbf{X}$ into independent subvectors.

Inference about independence has historically been addressed by means of hypothesis testing. Tests for independence between random variables have been extensively discussed in the literature by different authors. Contingency tables for categorical data and tests based on correlation coefficients, like Pearson's, Kendall's and Spearman's, are some of the most popular methods to deal with independence. Current approaches to general independence testing include distance based methods such as distance correlation, as presented in Székely and Rizzo (2009) or in Székely et al. (2007). Also kernel-methods have been proposed, including the Hilbert-Schmidt Information Criterion considered in Gretton and Gyorfi (2010) and Gretton et al. (2005). A different approach can be obtained by testing correlations on multiscale graphs, such as found in Shen et al. (2019). Copula functions have also been used to test for independence, as in Dugué (1975), Deheuvels (1981) or Genest and Rémillard (2004), among others.

These are some of the many existing references that use hypothesis testing to discover or study independence. However, to the best of our knowledge, the estimation of points of independence, as proposed in this work, has not received much attention, aside from the work presented in Castro et al. (2018). In the later, the authors consider this problem in order to detect recombination hotspots in Single Nucleotide Polymorphisms (SNPs) data, assuming that the random vector takes values in A^d, where A is a finite alphabet and the observations are independent. In this paper we consider a more general setting where the random vectors assume values in $\mathbb{R}^d$ and are not necessarily independent.

The paper is organized as follows. In Section 2 we define the estimator of the independent blocks and state the main theoretical results. In Section 3 we introduce an efficient binary splitting algorithm to approximate the estimator and state its convergence under the same conditions as the exact criterion. In Section 4 we show the results of the estimators on simulated data and in Section 5 we apply the method to a real dataset of water flow in the São Francisco River in Brazil.

2 Independent block estimator

Let $\mathbf{X} \sim F$ be a multivariate random vector taking values in $\mathbb{R}^d$. For u, v with $1 \leq u < v < d$, consider the following subvectors of $\mathbf{X}$

$$\mathbf{X}_{1:u} = (X_1, \ldots, X_u)\ , \quad \mathbf{X}_{u:v} = (X_{u+1}, \ldots, X_v)\ , \quad \mathbf{X}_{v:d} = (X_{v+1} \ldots, X_d),$$

and let $F_{1:u}$, $F_{u:v}$ and $F_{v:d}$ denote the cumulative distribution functions of $\mathbf{X}_{1:v}$, $\mathbf{X}_{u:v}$ and $\mathbf{X}_{v:d}$, respectively. We say that $U = \{u_1, \ldots, u_k\}$, with $1 \leq u_1 < \ldots < u_k < d$, is a set of independence for F if $\mathbf{X}_{1:u_1}$, $\mathbf{X}_{u_i:u_{i+1}}$ $(i = 1, \ldots, k-1)$, $\mathbf{X}_{u_k:d}$ are independent. Note that if U is a set of independence for F, any smaller set $\tilde{U} \subset U$ is also a set of independence for F. Moreover, if U and V are sets of independence for F, $U \cup V$ is a set of independence for F too. This suggests to define $U^*(F)$ as the biggest set of independence for F, in the sense that any other set of independence is included in $U^*(F)$. The aim of this work is to estimate $U^*(F)$ on the basis of $\{\mathbf{X}^{(i)} : 1 \leq i \leq n\}$, a stationary random process with $\mathbf{X}^{(i)} \sim F$. This is a model selection problem, a core topic in data science. As explained in Massart (2007), the main objective of model selection is to construct a data-driven criterion to select a model among a given list of candidates. Once a model is chosen, it can be used to produce accurate estimations of some parameters of interest. In the present setting, each model $\mathcal{M}_U$ postulate that U is a set of independence for F. These models are nested, in the sense that if $\tilde{U} \subset U$, then $\mathcal{M}_U \subset \mathcal{M}_{\tilde{U}}$. Through the estimation of $U^*(F)$ we can determine which is the smallest model that generates our data. Typically, the larger is the postulated model, the more flexible it is to describe the data, risking to lead to overfitting. To avoid this type of phenomena, a penalization term is added to a given empirical minimum contrast that can be used to choose a parsimonious model. To be more precise, given F and $U = \{u_1, \ldots, u_k\}$, define the U-product of F by

$$F_U(x_1, \ldots, x_d) = F_{1:u_1}(x_1, \ldots, x_{u_1}) \prod_{i=1}^{k-1} F_{u_i:u_{i+1}}(x_{u_i+1}, \ldots x_{u_{i+1}})\, F_{u_{k+1}:d}(x_{u_k+1}, \ldots, x_d). \quad (1)$$

For instance, if $U = \{1, 4\}$ and $d = 5$, we are considering the product of the marginal distribution of the subvectors (X_1), (X_2, X_3, X_4) and (X_5). For $U = \emptyset$, define $F_U = F$. We can measure the discrepancy between F and its U-product considering

$$\ell(U, F) = \sup_{\mathbf{x} \in \mathbb{R}^d} |F_U(\mathbf{x}) - F(\mathbf{x})|. \quad (2)$$

Note that U is a set of independence for F if and only if $F_U \equiv F$, which means that $\ell(U, F) = 0$. Since $U^* = U^*(F)$ is the maximal set of independence for F, there exists $\alpha > 0$ such that

$$\ell(U, F) = 0 \quad \text{if } U \subseteq U^*(F), \text{ while} \quad \ell(U, F) > \alpha \quad \text{if } U \not\subseteq U^*(F). \quad (3)$$

This characterization of U^* suggests that it can be estimated by looking at the *biggest* set that minimizes an empirical version of $\ell(U, F)$. In this work, the empirical version will be defined through a plug–in procedure while the penalization term will take care of choosing the biggest set, as indicated in what follows.

Given $\mathbf{X}^n = \{\mathbf{X}^{(i)} : 1 \leq i \leq n\}$, a sample of the process with stationary distribution F, let $\widehat{F}_{\mathbf{X}^n}$ denote any estimator of F. For instance, a distribution-free consistent estimator of F is given by the empirical distribution, defined by

$$\widehat{F}_{\mathbf{X}^n}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\mathbf{X}^{(i)} \leq \mathbf{x}\}}. \quad (4)$$

However, if a model is postulated for F, other estimators can be used; for instance, if F is assumed to be a Gaussian distribution with parameters μ and $\mathbf{\Sigma} = \{\sigma_{i,j}\}$, $F_{u:v}$ is also a normal distribution but in $\mathbb{R}^{v-u}$, with mean $\mu_{u:v} = \mathbb{E}(\mathbf{X}_{u:v}) = (\mu_u, \ldots, \mu_{v-1})^t$ and variance-covariance matrix $\mathbf{\Sigma}_{u:v} = \mathrm{cov}(\mathbf{X}_{u:v})$. In such a case, F_U is also a multivariate Gaussian distribution in $\mathbb{R}^d$ with parameters μ and $\mathbf{\Sigma}_U$, where $\mathbf{\Sigma}_U$ stands for the U-block matrix obtained by replacing the coefficients $\sigma_{i,j}$ in $\mathbf{\Sigma}$ with zero whenever $i \leq u < j$, for some $u \in U$. Thus, we can use a Gaussian distribution with estimated parameters in lieu of the empirical distribution, defined in (4).

Consider

$$\mathrm{PL}(U, \mathbf{X}^n) = \ell(U, \widehat{F}_{\mathbf{X}^n}) + \lambda_n \, (|U| + 1)^{-1}, \tag{5}$$

where $|U|$ denotes the cardinal of the set U. In this way, we have combined an empirical version of $\ell(U, F)$ with a penalization term, giving rise to the new objective function to be minimized. Define

$$\widehat{U}_n = \underset{U \subseteq \{1,\ldots,d-1\}}{\arg\min} \ \mathrm{PL}(U, \mathbf{X}^n). \tag{6}$$

That is, $\widehat{U}_n$ satisfies

$$\mathrm{PL}(\widehat{U}_n, \mathbf{X}^n) \ \leq \ \mathrm{PL}(U, \mathbf{X}^n) \ , \quad \text{for all } U \subseteq \{1, \ldots, d-1\}.$$

The following result establishes the consistency of $\widehat{U}_n$ as far as the penalization term and the convergence rate of $\widehat{F}_{\mathbf{X}^n}$ satisfy certain conditions.

Theorem 1 *Assume that*

$$sup_{\mathbf{x}\in\mathbb{R}^d} |\widehat{F}_{\mathbf{X}^n}(\mathbf{x}) - F(\mathbf{x})| \leq a_n \ , \qquad \textit{eventually almost surely as } n \to \infty. \tag{7}$$

If $\lambda_n \to 0$ and $a_n/\lambda_n \to 0$, then $\widehat{U}_n = U^*$ eventually almost surely when $n \to \infty$.

Remark 2 *The convergence of $\widehat{U}_n$ to U^* established in Theorem 1 does not require the process to be in a stationary regime. It holds as far as the empirical distribution $\widehat{F}_{\mathbf{X}^n}$ converges uniformly to the limit distribution F at a certain rate a_n, related to the penalization factor λ_n as indicated in this theorem.*

Adler and Brown (1986) studied the tail behavior of the suprema of the centered empirical distribution in the iid case, giving rise to the following result.

Corollary 3 *Assume that $\{\mathbf{X}^{(i)} : i \geq 1\}$ are iid and consider the empirical distribution $\widehat{F}_{\mathbf{X}^n}$ defined in (4) to estimate F. Take $\lambda_n = cn^{-\xi}$, with $\xi \in (0, 1/2)$. Then, $\widehat{U}_n = U^*$ eventually almost surely when $n \to \infty$.*

As discussed in Adams et al. (2010), even if the uniform consistency of the centered empirical distribution for the non iid case can be deduced for general ergodic sampling schemes, distribution-free probability bounds like those required in (7) cannot be obtained without further constrains. That is to say, besides the iid case, universal rates can not be established in general. However, specific rates can be deduced for particular cases. For instance, assume now that $\mathbf{X}$ is a discrete random vector, that is $\mathbf{X} \in A^d$, with A a finite alphabet and let $\{\mathbf{X}^{(i)} : i \geq 1\}$ be a stationary and ergodic mixing process with marginal distribution F. For $i \leq j$ denote by $\mathbf{X}^{(i:j)}$ the cylinder (projection) $\mathbf{X}^{(i:j)} = \{\mathbf{X}^{(k)} \colon i \leq k \leq j\}$. Denote also by $\mathbf{x}_1^k$, with $k \geq 1$, a sequence of length k of vectors in A^d. Then the process $\{\mathbf{X}^{(i)} \colon i \geq 1\}$ satisfies a mixing condition with rate $\{\psi(\ell)\} \downarrow 0$ as $\ell \to \infty$ if for each k, m and each $\mathbf{x}_1^k \in A^k$, $\mathbf{x}_1^m \in A^m$ with $\mathbb{P}(\mathbf{X}^{(1:m)} = \mathbf{x}_1^m) > 0$ we have

$$\left|\mathbb{P}(\mathbf{X}^{(n:(n+k-1))} = \mathbf{x}_1^k \,|\, \mathbf{X}^{(1:m)} = \mathbf{x}_1^m) - \mathbb{P}(\mathbf{X}^{(n:(n+k-1))} = \mathbf{x}_1^k)\right| \leq \psi(\ell)\mathbb{P}(\mathbf{X}^{(n:(n+k-1))} = \mathbf{x}_1^k), \quad (8)$$

for $n \geq m + \ell$. Csiszár (2002) obtained a result on the rate of convergence for the empirical probabilities in a stationary stochastic process with exponential mixing sequence. Based on this approach we can prove the following result.

Corollary 4 *Assume* $\{\mathbf{X}^{(i)} : i \geq 1\}$ *satisfies the mixing condition* (8) *with* $\psi(\ell) = \delta^\ell$ *for some* $0 < \delta < 1$. *Consider the empirical distribution function* $\widehat{F}_{\mathbf{X}^n}(\mathbf{x})$ *defined in (4) to estimate* $F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$. *Then* $\widehat{U}_n$ *defined in (6), with* $\lambda_n = cn^{-\xi}$, $\xi \in (0, 1/2)$, *satisfies* $\widehat{U}_n = U^*$ *eventually almost surely when* $n \to \infty$.

3 Efficient computation by binary splitting

To calculate the estimator in (6) we need to compute the function $\mathrm{PL}(U, \mathbf{X}^n)$ over all possible subsets $U \subseteq \{1, 2, \ldots, d-1\}$. The number of subsets is exponential in d so the complexity of the exhaustive search algorithm is $O(2^d T)$, where T is the time needed to compute $\mathrm{PL}(U, \mathbf{X}^n)$. Observe that T could also depend on d, but at most linearly. In any case, the problem becomes computationally infeasible even for moderate values of d. To overcome this computational problem, in this section we introduce a more efficient divide and conquer algorithm to approximate the estimator given by 6, with time complexity $O(d^2 T)$. At each step, we include an independence point in the estimation of $U^*(F)$, as far as it improves the behavior of the penalized discrepancy defined in (5). To be more precise, let

$$\mathrm{PL}(U, \mathbf{X}^n_{u:v}) = \ell(U, \widehat{F}_{\mathbf{X}^n_{u:v}}) + \lambda_n \, (|U| + 1)^{-1},$$

for all $1 \leq u \leq v \leq d$ and $U \subseteq \{u, \ldots, v-1\}$, where $|U|$ denotes the cardinal of the set U, as defined before. Consider

$$h(u\colon v, \mathbf{X}^n_{u:v}) = \underset{i \in u:v}{\arg\min} \{\, \mathrm{PL}(\{i\}, \mathbf{X}^n_{u:v})\}, \quad (9)$$

where, by convention, we set $\mathrm{PL}(\{v\}, \mathbf{X}^n_{u:v}) = \mathrm{PL}(\emptyset, \mathbf{X}^n_{u:v})$, with v the biggest element in $u\colon v$.

The binary splitting algorithm constructs a binary tree with nodes indexed by subintervals of $1 : d$, such that the set of terminal nodes of the tree is a partition of $1 : d$ and the end points of these intervals correspond to the estimated points of independence in $\widehat{U}_n^{\mathrm{bin}}$. The algorithm works as follows.

1. Initialize $\widehat{U}_n^{\mathrm{bin}} = \emptyset$ and $I = 1\colon d$ (the root of the tree).

2. Compute $h(I, \mathbf{X}^n_I)$. If $h(I, \mathbf{X}^n_I) < \max(I)$ add $h(I, \mathbf{X}^n_I)$ to $\widehat{U}_n^{\mathrm{bin}}$ and two leaves to node I in the tree, with labels $I_1 = I \cap \{i \colon i \leq h(I, \mathbf{X}^n_I)\}$ and $I_2 = I \cap \{i \colon i > h(I, \mathbf{X}^n_I)\}$.

3. Repeat step 2 for the new terminal nodes in the tree, until no more leaves are added.

The final estimated set of points of independence $\widehat{U}_n^{\mathrm{bin}}$ is the set of right extremes of the terminal nodes in the tree (excluding the root and the end d of the entire interval).

Even if this algorithm is an approximation to the minimum in (6), we can show that the estimated set $\widehat{U}_n^{\mathrm{bin}}$ converges to $U^*(F)$ eventually almost surely, under the same conditions of Theorem 1.

Proposition 5 *Under the same assumptions of Theorem 1,* $\widehat{U}_n^{\mathrm{bin}} = U^*$ *eventually almost surely when* $n \to \infty$*.*

To prove Proposition 5 we use the characterization of $U^*(F)$ given in (4). The proofs of the results presented in this section are given in Section 7.

4 Simulations

In this section we study the behavior of the proposed estimators through experiments on synthetic data, in which we can evaluate the ability of the proposed methods to recover the correct set of independence. In the sequel, we use exhaustive and binary to refer to the estimators defined by (6) and in Section 3, respectively.

In order to measure the difference between the estimates and the target set we use the Hausdorff distance. This distance is defined for two non-empty sets $A, B \subset \{1, \dots, d-1\}$ and is given by

$$\rho_H(A, B) = \max\{\rho(A||B), \rho(B||A)\},$$

where $\rho(B||A) = \sup_{b\in B} \inf_{a\in A} |a - b|$ and $\rho_H(\emptyset, A) = d - 1$.

In the sequel we show the results of the estimation procedures for the sets of independence in two scenarios, under independence and dependence of the time series, respectively. The data is generated in a stationary regime, as described in the following sub-sections.

4.1 Gaussian independent scenario

In this case, $\{\mathbf{X}^{(i)} : 1 \le i \le n\}$ are iid random vectors distributed as $\mathbf{X} = (X_1, \dots, X_5) \in \mathbb{R}^5$, with centered multivariate Gaussian distribution and the set of points of independence is $U^* = \{2, 3\}$; i.e., the subvectors (X_1, X_2), (X_3) and (X_4, X_5) are independent components of the vector $\mathbf{X}$. We consider the correlation structure $\mathbf{\Sigma}_\rho$ given by

$$\mathbf{\Sigma}_\rho = \begin{pmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3/\sqrt{10} \\ 0 & 0 & 0 & 3/\sqrt{10} & 1 \end{pmatrix},$$

where ρ stands for the correlation between X_1 and X_2. For any ρ, (X_3, X_4, X_5) has the same joint distribution. Moreover, the marginal distribution of X_1 and X_2 is also the same.

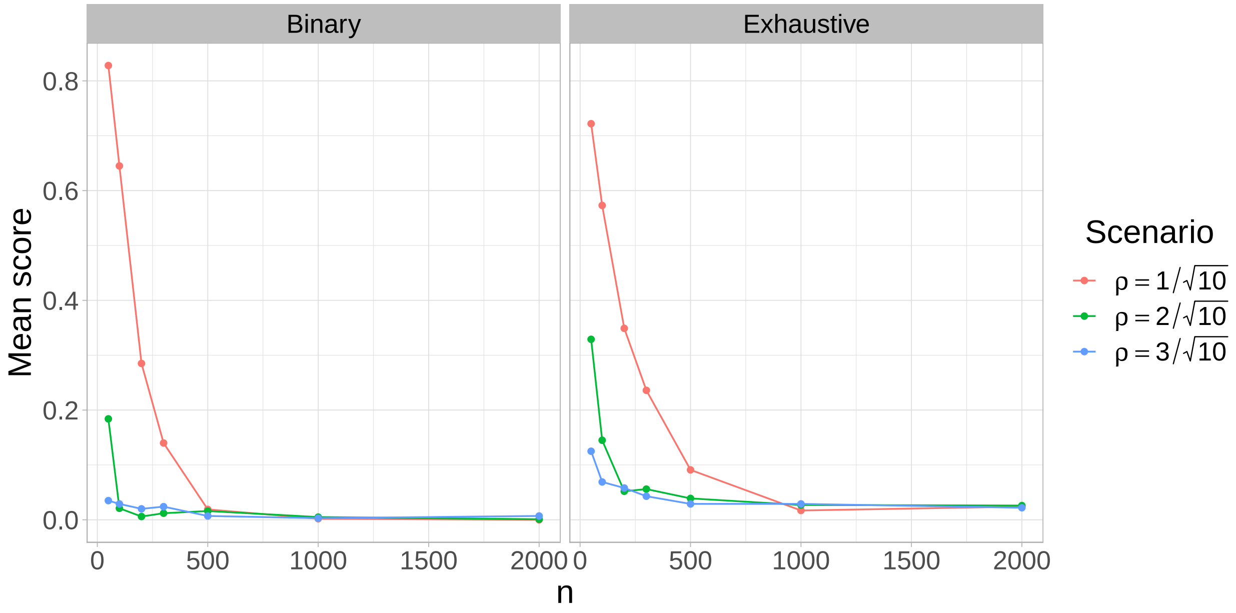


Figure 1: Performance of binary and exhaustive algorithms for the three scenarios considered.

However, in the simulation study we consider three different correlation values between X_1 and X_2: (i) $\rho = 1/\sqrt{10}$, (ii) $\rho = 2/\sqrt{10}$ and (iii) $\rho = 3/\sqrt{10}$. It is worth noticing that the criterion is location-scale invariant, in the sense that the estimators remain the same if the variables are linearly transformed. In particular, the variable can be standardized, thus in the simulations we use unit variance random variables and the matrices Σ_ρ are also the covariance matrices of the random vectors in each scenario.

For each Σ_ρ considered, we generate $Nrep = 1000$ data sets of different size (n=50, 100, 200, 300, 500, 1000, and 2000). Additionally, for each data set, using the empirical cumulative distribution function $\widehat{F}_{\mathbf{X}^n}$ defined in (4), we compute both $\widehat{U}_n$, the estimator defined in (6) and $\widehat{U}_n^{\text{bin}}$, obtained by means of the binary splitting algorithm presented in Section 3. The penalty term was chosen accordingly to Corollary 3 with $c = 1$ and $\xi = 0.4$, that is, we use the penalty $\lambda_n = n^{-0.4}$.

Figure 1 shows the mean value among the replications of the Hausdorff distance between each one of these estimators and the true set of independence points, as a function of the sample size n. The error rate of each procedure, computed as the proportions of replications where the estimated set differs from the true set of independence, is presented in Figure 2. It can be seen that both algorithms reach the true set of independence as far as n increases in the three scenarios. It is worth noticing that as far as the correlation between X_1 and X_2 increases, it becomes easier for the algorithms to discover the true set of independence. In general, we can expect the exhaustive algorithm to perform better than the approximate binary search algorithm. While this is not totally appreciated in Figures 1 and 2, we see that the exhaustive algorithm have a better performance when considering a large set of penalizing constants c and ξ, see for example Figure 1 in the supplementary material to this article.

The mean processing time (in seconds) for the three scenarios is presented in Figure 3.

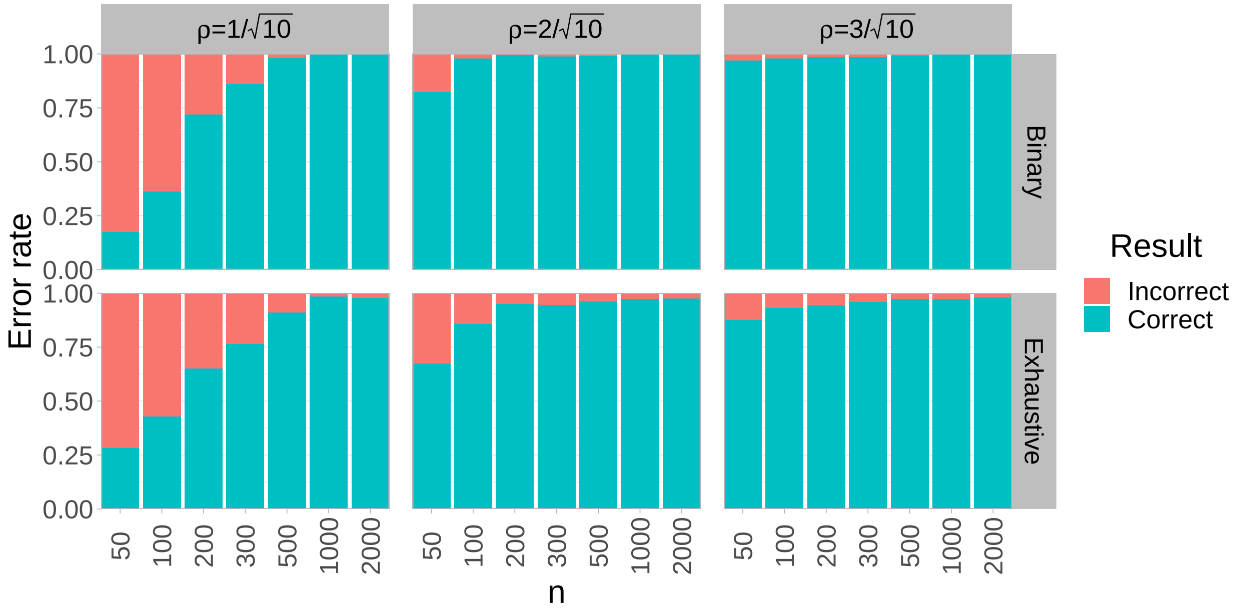


Figure 2: Error rate for binary and exhaustive algorithms in the three scenarios considered.

We observe that the mean speed up of the simulation when using the exhaustive algorithm is around 1.93 times that of the binary algorithm.

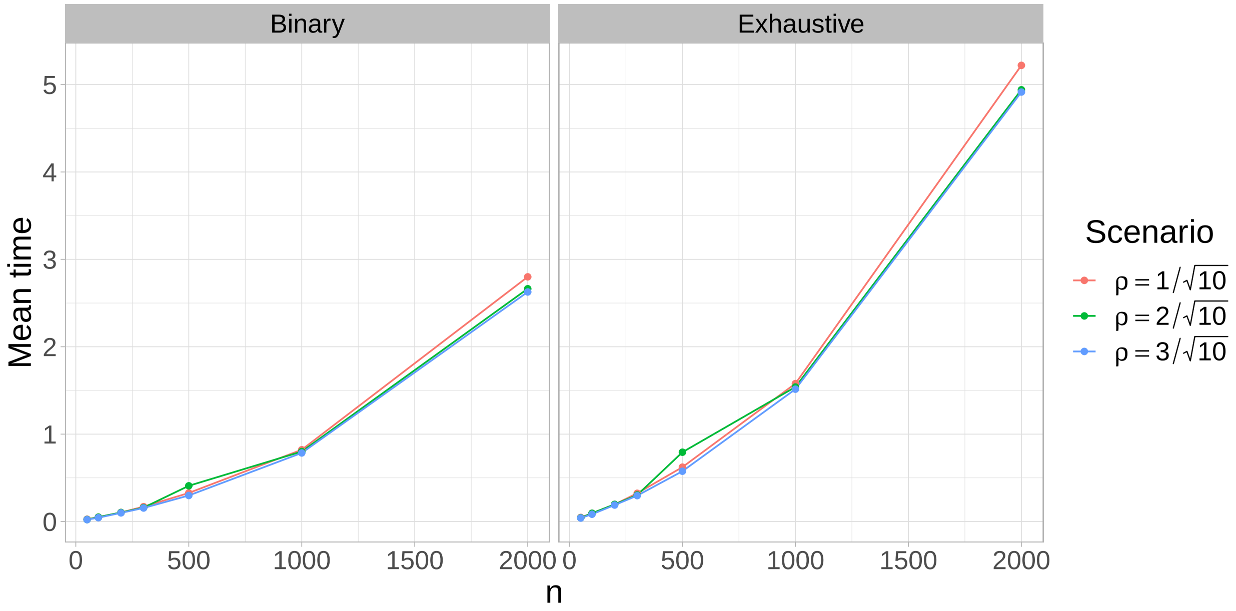


Figure 3: Mean processing time (in seconds) using binary and exhaustive algorithms in the three considered scenarios.

4.2 Time Series

Now, $\mathbf{X}^{(i)} = (X_1^i, \ldots, X_5^i)$ is generated combining different kinds of time series, giving rise to two data generating processes. Following the standard notation used in time series, we use $\mathbf{Y}_i$ to denote a 2-variate process satisfying

$$\mathbf{Y}_i = \boldsymbol{A}_1 \mathbf{Y}_{i-1} + \boldsymbol{A}_2 \mathbf{Y}_{i-2} + \mathbf{U}_i + \theta_1 \mathbf{U}_{i-1}, \tag{10}$$

where $\mathbf{U}_i$ is a two-dimensional sequence of uncorrelated and centered Gaussian process with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{U}}$, $\boldsymbol{A}_i \in \mathbb{R}^{2\times 2}$, for $i = 1, 2$, and $\theta_1 \in \mathbb{R}$. This is a unified representation of many of the most popular time series models. For instance, when $\boldsymbol{A}_2 = 0$, $\mathbf{Y}_i \sim$ VARMA(1,1); on the other hand, if $\boldsymbol{A}_2 = 0$ and $\theta_1 = 0$, $\mathbf{Y}_i \sim$ VAR(1) while $\mathbf{Y}_i \sim$ VAR(2) when $\theta_1 = 0$.

Finally, we consider the univariate AR(1) processes $T_i \in \mathbb{R}$, satisfying

$$T_i = \gamma_1 \, T_{i-1} + \varepsilon_i, \tag{11}$$

where now ε_i stands for an iid sequence of centered Gaussian random variables with variance σ^2. In the sequel we describe the data generating process considered for the non iid samples.

Model 1: In this case, $\mathbf{X}^{(i)} = (X_1^i, X_2^i, X_3^i, X_4^i, X_5^i)$, where

- $(X_1^i, X_2^i) \sim$ VAR(2), in the sense that satisfies (10), with $\theta_1 = 0$,

$$\boldsymbol{A}_1 = \begin{pmatrix} -0.3 & -0.4 \\ 0.6 & 0.5 \end{pmatrix}, \qquad \boldsymbol{A}_2 = \begin{pmatrix} -0.1 & 0.1 \\ -0.2 & 0.05 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{U}} = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

- $X_3^i \sim$ AR(1), in the sense that satisfies (11), with $\gamma = 0.5$ and $\sigma^2 = 9$.
- $(X_4^i, X_5^i) \sim$VAR(1), in the sense that satisfies (10), with $\boldsymbol{A}_2 = 0$, $\theta_1 = 0$,

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{U}} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

These processes are generated independently and, therefore, in the present case, we have that $U^* = \{2, 3\}$. To know the extent of dependence between the components of the subvectors (X_1^i, X_2^i) and (X_4^i, X_5^i) we performed a simulation with $n = 10{,}000$ time steps and computed the empirical correlation between the variables, obtaining

$$\mathrm{Cor}(X_1^i, X_2^i) = -0.25 \qquad \text{and} \qquad \mathrm{Cor}(X_4^i, X_5^i) = 0.90\,.$$

Model 2: Now, $\mathbf{X}^{(i)} = (X_1^i, X_2^i, X_3^i, X_4^i, X_5^i)$, with

- $(X_1^i, X_2^i) \sim$ VARMA(1,1), in the sense that satisfies (10), with $\boldsymbol{A}_2 = 0$, $\theta_1 = 0.9$,

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & -0.6 \\ 0.7 & 0.3 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{\mathbf{U}} = \begin{pmatrix} 1.3 & 0.91 \\ 0.91 & 1.3 \end{pmatrix}.$$

- $(X_3^i, X_4^i) \sim$ VAR(1), with the same parameters used in the last two coordinates of Model 1.

- $X_5^i \sim$AR(1), with the same parameters used in third coordinate of Model 1.

As in the previous model, these processes are generated independently and, therefore, we have now that $U^* = \{2, 4\}$. As before, we also computed the empirical correlation between the variables in the sub-vectors, obtaining

$$\mathrm{Cor}(X_1^i, X_2^i) = 0.32 \qquad \text{and} \qquad \mathrm{Cor}(X_3^i, X_4^i) = 0.90\,.$$

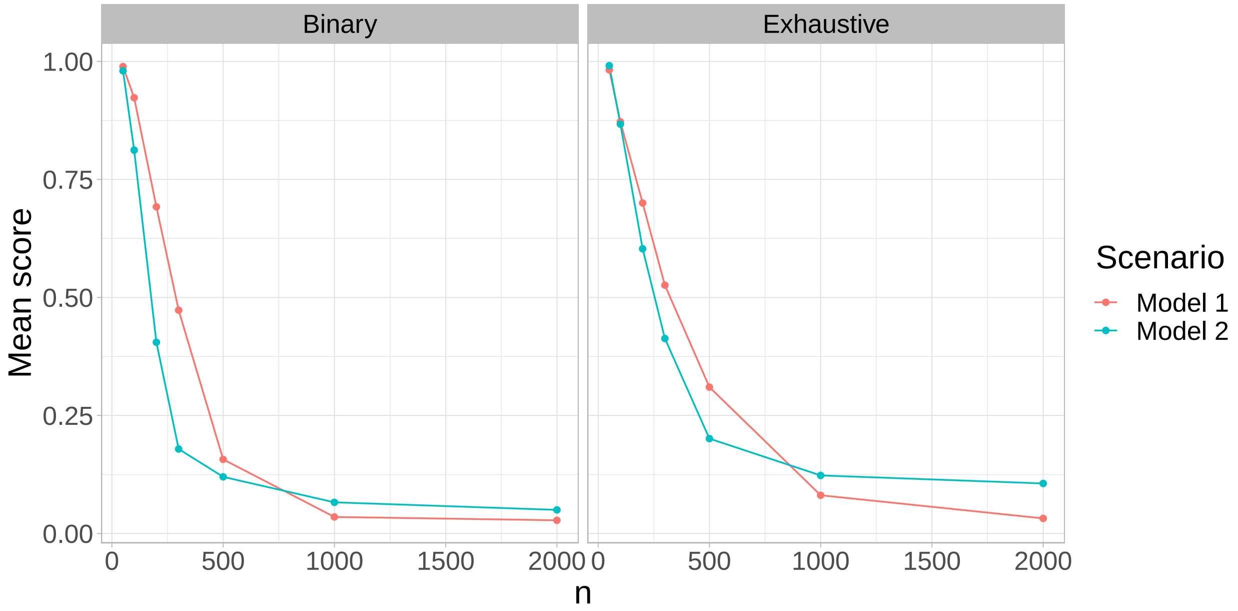


Figure 4: Performance of binary and exhaustive algorithms for Models 1 and 2.

Figure 4 shows the comparison between binary and exhaustive algorithms for these two models and Figure 5 presents the error rate of each algorithm for the models considered. Again, these rates are computed as proportion of replications in which the estimated set differs from the true set of independence. Mean time spent in seconds for each algorithm is shown in Figure 6. As in the Gaussian independent scenarios, both algorithms reach the true set of independence as n increases in the two models.

5 Independent blocks in the São Francisco River

Rivers are constantly moving and there are many factors, both natural and human-induced that cause streams to change, for instance, runoff from rainfall and snow melt, ground-water discharge from aquifers, river-flow regulation for hydropower and navigation, surface-water withdrawals and transbasin diversions, irrigation, among others.

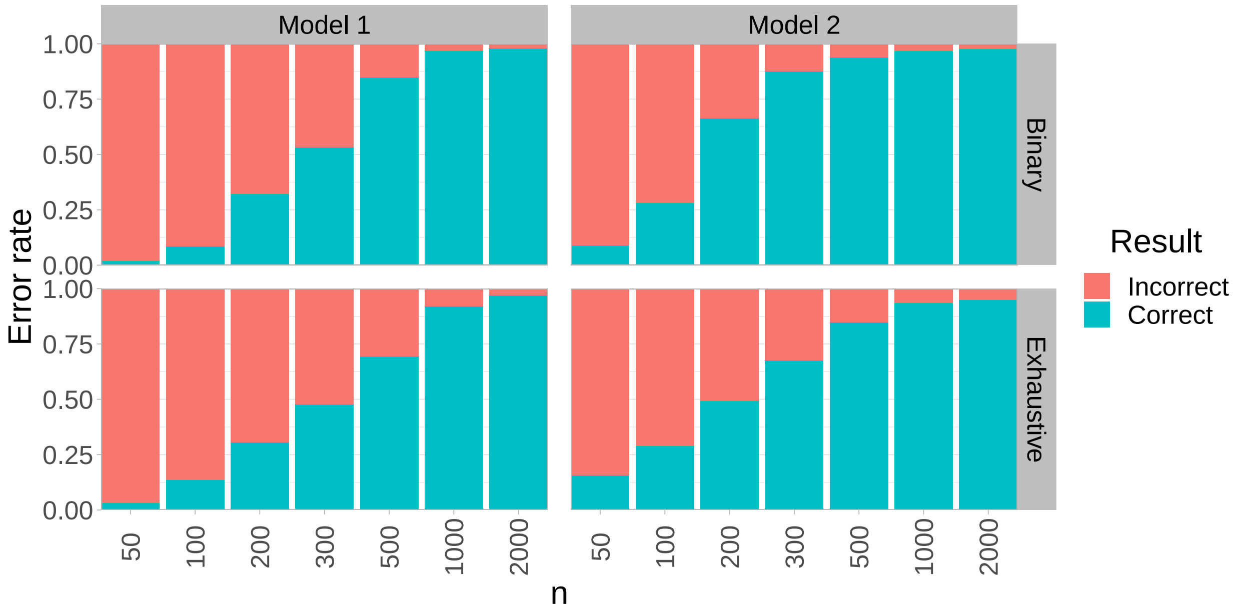


Figure 5: Error rate for binary and exhaustive algorithms in Models 1 and 2.

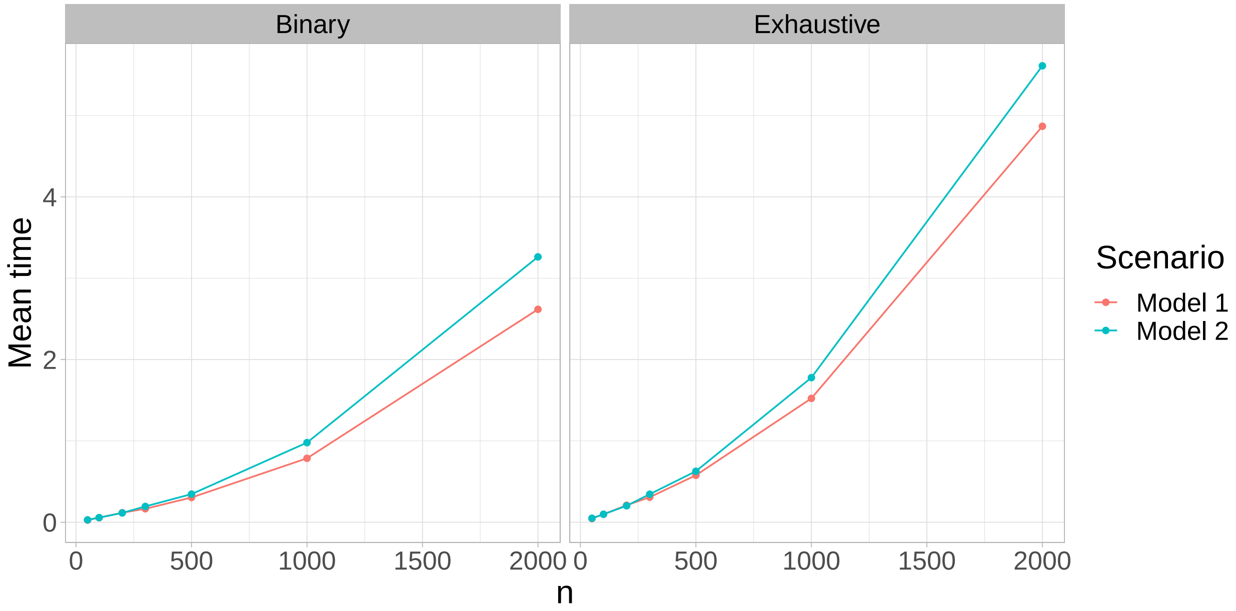
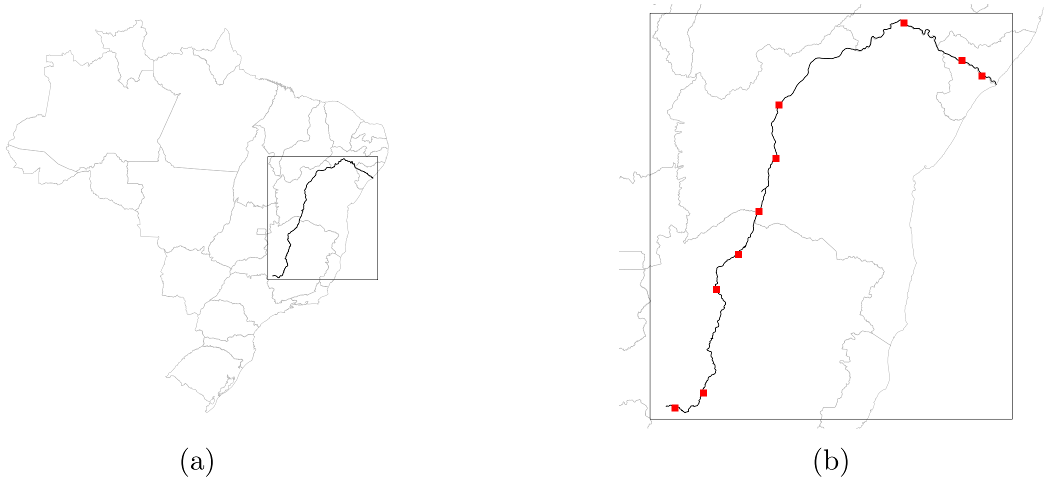


Figure 6: Mean processing time (in seconds) for Models 1 and 2 using binary and exhaustive algorithms.

In this section we study the volumetric discharge in the São Francisco River in Brazil by taking into account measurements at $d = 10$ gauges located along the course of the river (they are numbered according to the order in which they appear on the river). The São Francisco River is the longest river that runs entirely in the Brazilian territory, with a length of 2,914 kilometers. Its headwaters originates in the Canastra mountain range, in central part of Brazil, and runs north towards the northeast Brazilian region. Figure 7a shows the



(a)

(b)

Figure 7: (a) Geographic border of Brazil and its states limits. The rectangle highlights the area where the São Francisco River is located; (b) A zoom of the boxed area in (a), containing the São Francisco River. Red points represent the ten stream flow gauges considered in our analysis, numbered in increasing order from bottom to top.

river's course within Brazil and Figure 7b points out where the gauges are located. We retain our attention solely to the monthly mean of stream flows at 10 stations located along the course of São Francisco River registered between January 1977 and January 2016. This data form the $\mathbf{X}^{(i)}$ vector described above. Therefore our aim is to determine the set of independence among the stream gauges.

The course of the river can be divided into four sections: the high part (where stations 1 and 2 are located), from its source to Pirapora city; the upper middle part (stations 3, 4, 5, 6, and 7), from Pirapora to Sobradinho dam, the navigable part; the lower middle part from Sobradinho dam to Itaparica dam (station 8); and the low part, from Itaparica dam to the river mouth (stations 9 and 10). The flow of the river at different points can also be affected by the period of the year. The wet season, which holds nearly 60% of the yearly precipitation, begins in November and goes until January, while the dry season is from June to August.

We consider $n = 358$ observations consisting of monthly averages of the registered data, in m^3/s. Both the exact and the binary splitting algorithms with $\lambda_n = n^{0.25}$ estimated the same set of independence $\widehat{U}_n = \widehat{U}_n^{\text{bin}} = \{7\}$. It is important to note that this finding can be explained by the fact that between stations 7 and 8 is located the Sobradinho hydroelectric dam, the biggest along the course of the São Francisco River. Figure 8 shows boxplots of the stream measurements at the considered gauges and the point of independence given by our approach. We observe that at point 7 there is a qualitative change of regime in the boxplots, and this can be due to the effect of the hydroelectric in the flow of the river, showing that the independence obtained by the algorithm can be in some sense expected at this point.

One characteristic of this dataset is that it is not stationary by nature, that means in our context that data on each month can have a different distribution. But even in this case, the method can still be effective to detect the common points of independence, that is, the points of independence shared by all the distributions. To investigate more about this

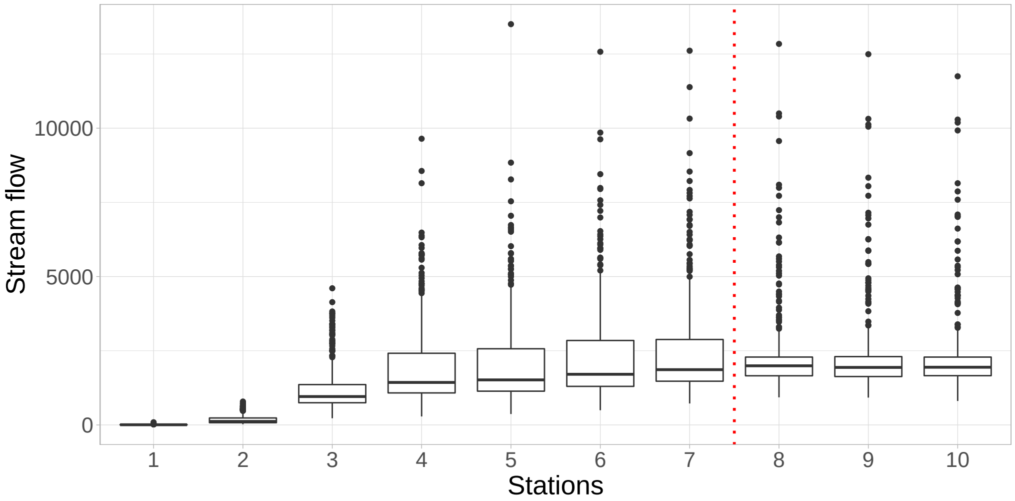


Figure 8: Stream flows measured at the ten stations in the São Francisco River. The red dotted line represents the point at which both the exact and the binary algorithms estimated a point of independence for the random vector.

issue, we applied both algorithms to the subsets of the data corresponding to each month, using the same tuning parameters. As expected, both algorithms estimated more points of independence, but in general the point 7 was detected in the majority of the months. Due to space limitations, the results for the different months are compiled in the supplementary material to this article.

6 Discussion

In this paper we introduced a model selection approach to detect independent blocks in multivariate time series. The method is based on a penalized criterion and on a general estimator of the cumulative distribution function. We proved the convergence to the true set of points of independence, in a iid scenario and in a dependent mixing setting for discrete processes. We also introduced a more efficient binary splitting algorithm to be used when the computation of the exact estimator is computationally time demanding. We proved that the approximation given by this algorithm also converges to the true set of points of independence. These results could be extended to other scenarios, as for example the case of dependent gaussian processes or more general continuous processes. In these cases, the penalization factor λ_n should be chosen depending on the rate of convergence of the selected estimator for the distribution function F.

From the simulations we concluded that both estimators have a very good performance, even for relatively small sample size, and the performance is better when higher is the correlation between the dependent variables. It is worth noticing that the simulations were implemented with a fixed value for the penalty λ_n and it remains as an open problem how to select the tuning parameter λ_n in an efficient way. In the supplementary material we included a simulation study considering different values for the penalizing constant and we can see that, as expected, the exact algorithm seems to outperform the binary search algorithm on a larger set of penalizing constants.

In this work we focused on the identification of a block structure, but we think our method is sufficiently general to be possibly adapted to other interesting structures of a random vector, as for example the identification of the interaction graph in a graphical model or Markov random field.

Acknowledgements

This work was produced as part of the activities of the Research, Innovation and Dissemination Center for Neuromathematics, grant FAPESP 2013/07699-0, FAPESP's project "Model selection in high dimensions: theoretical properties and applications", grant 2019/17734-3, São Paulo Research Foundation, Brazil; Grant 20020170100330BA from Universidad de Buenos Aires and PICT-201-0377 / PICT-2018-02842 from ANPYCT, Argentina. M.T.F. Severino is supported by a CNPq's scholarship and F. Leonardi was partially supported by a CNPq's research fellowship (309964/2016-4).

7 Proofs of theoretical results

In the sequel, we use $\widehat{F}_n$ in lieu of $\widehat{F}_{\mathbf{X}^n}$, $\widehat{F}_{n,u:v}$ for its $u:v$ marginal distribution and $\widehat{F}_{n,U}$ for its U-product.

Proof of Theorem 1: We show that, eventually almost surely as $n \to \infty$,

$$\mathrm{PL}(U^*, \mathbf{X}^n) < \mathrm{PL}(U, \mathbf{X}^n) \ , \quad \text{for all } U \subseteq \{1, \dots, d-1\} \ , \tag{12}$$

which means that $\widehat{U}_n = U^*$. In order to do so, note that

$$|\ell(U, \widehat{F}_{\mathbf{X}^n}) - \ell(U, F)| \leq \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_{n,U}(\mathbf{x}) - F_U(\mathbf{x})| + \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})|.$$

On the other hand, $\sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_{n,u:v}(\mathbf{x}) - F_{u:v}(\mathbf{x})| \leq \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})|$. Thus, under condition (7), eventually almost surely as $n \to \infty$,

$$|\ell(U, \widehat{F}_{\mathbf{X}^n}) - \ell(U, F)| \leq |U| a_n + a_n \leq d\, a_n. \tag{13}$$

Now, to prove (12), first consider U which is not contained in U^*: $U \not\subset U^*$. In such a case, $\ell(U, F) > \alpha > 0$. Therefore, using (13), we get that

$$\begin{array}{rcl} \mathrm{PL}(U, \mathbf{X}^n) - \mathrm{PL}(U^*, \mathbf{X}^n) & = & \ell(U, \widehat{F}_{\mathbf{X}^n}) - \ell(U^*, \widehat{F}_{\mathbf{X}^n}) + \lambda_n \left\{ (|U| + 1)^{-1} - (|U^*| + 1)^{-1} \right\} \\ & \geq & -d\, a_n + \alpha - d\, a_n + \lambda_n \left\{ (|U| + 1)^{-1} - (|U^*| + 1)^{-1} \right\} . \end{array}$$

Since $\alpha > 0$ and both λ_n and a_n converge to zero, eventually almost surely as $n \to \infty$, we get that

$$\mathrm{PL}(U, \mathbf{X}^n) - \mathrm{PL}(U^*, \mathbf{X}^n) > 0 \ . \tag{14}$$

If $U^* = \emptyset$, no other case should be considered. If not, take U such that U is strictly contained in U^*, that is $U \subset U^*$. Therefore, $\ell(U, F) = \ell(U^*, F) = 0$ and thus,

$$\begin{aligned} \mathrm{PL}(U, \mathbf{X}^n) - \mathrm{PL}(U^*, \mathbf{X}^n) \qquad &= \ell(U, \widehat{F}_{\mathbf{X}^n}) - \ell(U^*, \widehat{F}_{\mathbf{X}^n}) + \lambda_n \left\{(|U|+1)^{-1} - (|U^*|+1)^{-1}\right\} \\ &\geq -2d\, a_n + \lambda_n \left\{(|U|+1)^{-1} - (|U^*|+1)^{-1}\right\}. \end{aligned}$$

Finally, since $U \subset U^*$, we have that $|U| + 1 \leq |U^*|$ and thus

$$\frac{1}{|U|+1} - \frac{1}{|U^*|+1} \;\geq\; \frac{1}{d^*(d^*+1)} > \frac{1}{d(d+1)}.$$

Since $d\, a_n/\lambda_n \to 0$, we conclude that, for n large enough, $\frac{1}{d(d+1)} > \frac{2a_n}{\lambda_n}$, which implies that

$$\mathrm{PL}(U, \mathbf{X}^n) - \mathrm{PL}(U^*, \mathbf{X}^n) > 0\ ,$$

eventually almost surely as $n \to \infty$. □

Proof of Corollary 3: Adler and Brown (1986) proved that

$$P(\sqrt{n} \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})| > \lambda) \leq C\lambda^{2(d-1)} e^{-2\lambda^2},$$

for all $\lambda > 0$ and n large enough ($n > n_\lambda$). Therefore, if $0 < \delta < 1/2$, we have that

$$P(\sqrt{n} \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})| > n^\delta) \leq Ce^{-2\{n^{2\delta} - \delta(d-1)\ln n\}} = C_n^{-2\gamma_n},$$

where $\gamma_n = \frac{n^{2\delta}}{\ln n} - \delta(d-1)$. Since $\gamma_n \to \infty$, there exists n_1 such that for $n > n_1$, $\gamma_n > 1$ and thus

$$P(\sqrt{n} \sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})| > n^\delta) < Cn^{-2},$$

which shows that $\sum_{n=1}^\infty P(\sup_{\mathbf{x} \in \mathbb{R}^d} |\widehat{F}_n(\mathbf{x}) - F(\mathbf{x})| > n^{\delta - 1/2}) < \infty$. This guarantees that condition (7) is satisfied with $a_n = n^{\delta - \frac{1}{2}}$, for any $\delta < 1/2$. Finally, given $\xi \in (0, 1/2)$ and choosing $\delta < 1/2 - \xi$, we conclude that a_n and $\lambda_n = cn^{-\xi}$ fulfills the conditions of Theorem 1. □

Proof of Corollary 4: According to Csiszár (2002, Theorem 1), there exists a constant C (depending on the size of the alphabet A^d) such that eventually almost surely as $n \to \infty$

$$|\widehat{F}_{\mathbf{X}^n}(\mathbf{x}) - F(\mathbf{x})| \;\leq\; \sqrt{\frac{C \log^2 n}{np(\mathbf{x})}}\,,$$

for all $\mathbf{x}$ with $p(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}) \geq C \log^2 n/n$. Define

$$p_{\min} = \inf\{p(\mathbf{x})\colon \mathbf{x} \in A^d \text{ and } p(\mathbf{x}) > 0\}\,.$$

Then for all $\mathbf{x} \in A^d$ we have

$$|\widehat{F}_{\mathbf{X}^n}(\mathbf{x}) - F(\mathbf{x})| \;\leq\; \sqrt{\frac{C \log^2 n}{np_{\min}}}.$$

If we take $\lambda_n = cn^{-\xi}$, with $\xi \in (0, 1/2)$ we have that $\lambda_n \to 0$ and $a_n/\lambda_n \to 0$ as $n \to \infty$ and we are in the hypothesis of Theorem 1. Then, $\widehat{U}_n = U^*$ eventually almost surely when $n \to \infty$. □

Proof of Proposition 5: First consider the case $U^*(F) = \emptyset$. Then by (3) we have that $\ell(U, F) > \alpha$ for all $U \neq \emptyset$, in particular for all U with a single point. By the same arguments used in the proof of Theorem 1 that lead to (14) we obtain that $h(1\!:\!d, \mathbf{X}^n) = d$ eventually almost surely as $n \to \infty$ and therefore $\widehat{U}_n^{\text{bin}} = U^*(F)$. Now suppose there is at least one point of independence in $U^*(F)$. Consider the candidate sets U having a single point, that is $U = \{u\}$. By (3) we have that $\ell(\{u\}, F) = 0$ for all $u \in U^*(F)$ while $\ell(\{v\}, F) > \alpha$ for all $v \notin U^*(F)$. One more time, by the same arguments used in the proof of Theorem 1 we have that eventually almost surely $h(1\!:\!d, \mathbf{X}^n) = u$ for some $u \in U^*(F)$. Now the criterion is repeated in the sub-intervals $1 : u$ and $(u + 1) : d$. Note that if $u \in U^*(F)$ we have that $U^*(F) \cap (1\!:\!u) = U^*(F_{1:u})$, that is, the points of independence of the marginal $F_{1:u}$ are exactly the points in the intersection $U^*(F) \cap (1\!:\!u)$ (and the same is true for the complement $(u+1)\!:\!d$). Then the criterion can be consistently iterated on both sub-intervals $1\!:\!u$ and $(u+1)\!:\!d$, with F replaced by the marginals $F_{1:u}$ and $F_{(u+1):d}$, respectively. If the set $U^*(F)$ is finite, by this iterative procedure $\widehat{U}_n^{\text{bin}}$ will converge almost surely to $U^*(F)$. □

8 Data Availability Statement

In this work we consider information provided by the Brazilian National Water Agency (Agência Nacional de Águas, in Portuguese), see Sistema Nacional de Informações sobre Recursos Hídricos (2019). Its database is publicly available and stores daily meteorological and hydrological measurements, such as river levels, flows, rainfall, climatology, water quality, and sediment. The data analysed in Section 5, corresponding to monthly mean of stream flows at 10 stations located along the course of São Francisco River, are available as supplementary material to this article and can also be found on `https://www.ime.usp.br/~gpeca/data/sao_francisco_river.csv`.

References

Adams, T. M., Nobel, A. B., et al. (2010). Uniform convergence of Vapnik–Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367.

Adler, R. J. and Brown, L. D. (1986). Tail behaviour for suprema of empirical processes. *The Annals of Probability*, pages 1–30.

Castro, B. M., Lemes, R. B., Cesar, J., Hünemeier, T., and Leonardi, F. (2018). A model selection approach for multiple sequence segmentation and dimensionality reduction. *Journal of Multivariate Analysis*, 167:319–330.

Csiszár, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48(6):1616–1628. Special issue on Shannon theory: perspective, trends, and applications.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11(1):102–113.

Dugué, D. (1975). Sur des tests d'indépendance "indépendants de la loi". *C. R. Acad. Sci. Paris Sér. A-B*, 281(24):Aii, A1103–A1104.

Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369.

Gretton, A. and Gyorfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(Apr):1391–1423.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Shen, C., Priebe, C. E., and Vogelstein, J. T. (2019). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*, pages 1–22.

Sistema Nacional de Informações sobre Recursos Hídricos (2019). Portal HidroWeb `http://www.snirh.gov.br/hidroweb/`. Accessed: 2019-09-03.

Székely, G. and Rizzo, M. (2009). Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.