Seleção de modelos para processos estocásticos

Florencia Leonardi

Instituto de Matemática e Estatística
Universidade de São Paulo

Processo estocástico

Família de variáveis aleatórias X_t indexadas num conjunto T; i.e. $\{X_t\colon\ t \in T\}$, assumindo valores em algum espaço S.

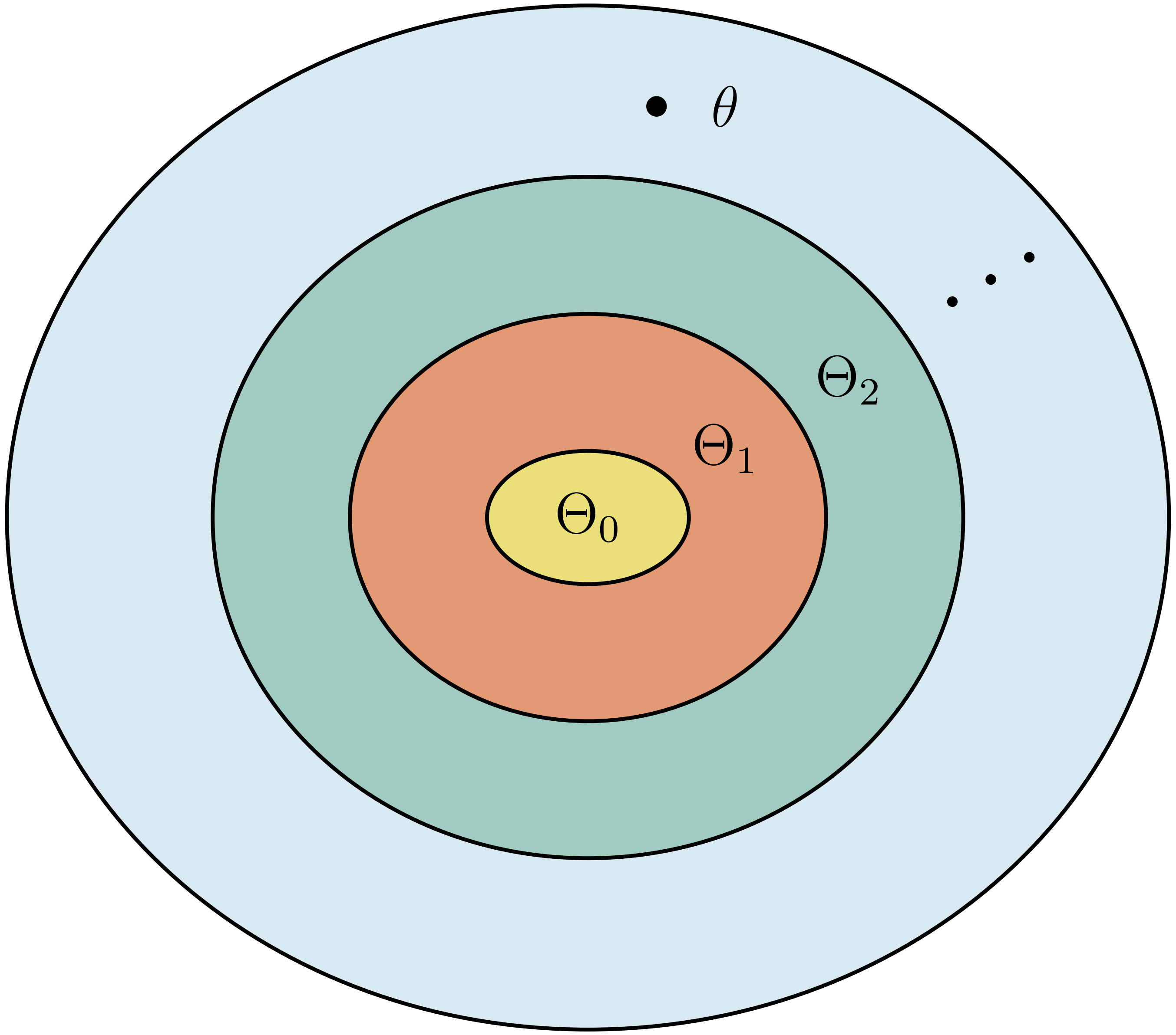
Usualmente T denota "tempo", $T = \mathbb{N}$ (tempo discreto) o $T = \mathbb{R}$ (tempo contínuo), mas pode representar outros índices (por ex. espaço $\mathbb{Z}^2$ ou $\mathbb{R}^2$)

Processo estocástico

A descrição do modelo muitas vezes pode ser feita por um conjunto de parâmetros $\theta \in \Theta_k$, com dimensão variável.

Exemplo: cadeias de Markov sobre $\{0,1\}$ com memória $k = 0,1,2,...$

Θ_k: classe das cadeias de Markov de memória k



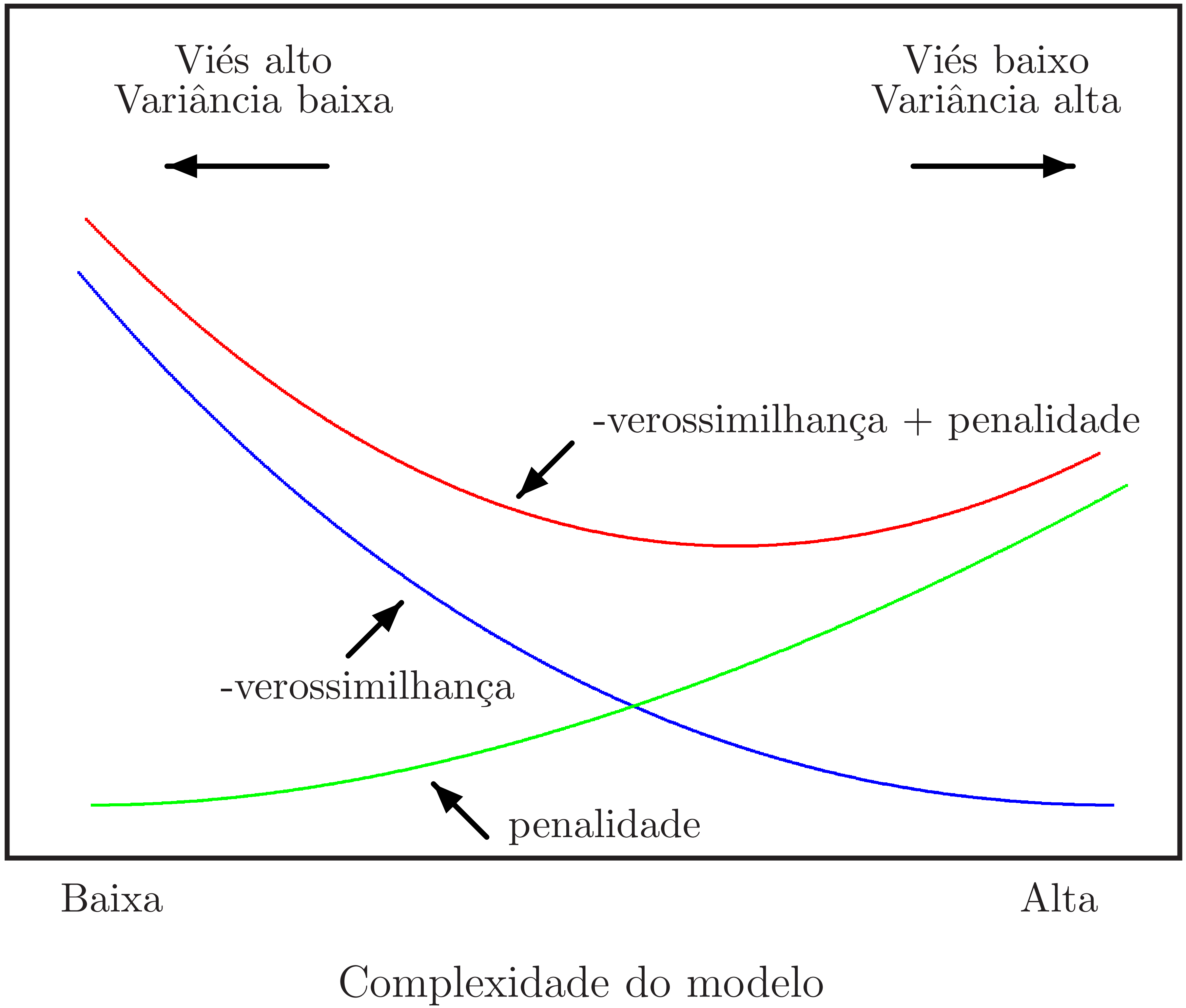
Seleção de modelos

Para cada processo, há uma dimensão mínima" (básica). Essa dimensão é a "ordem" do processo

Se observamos uma amostra de um processo de dimensão "mínima" (ordem) k, como podemos identificar esta dimensão?

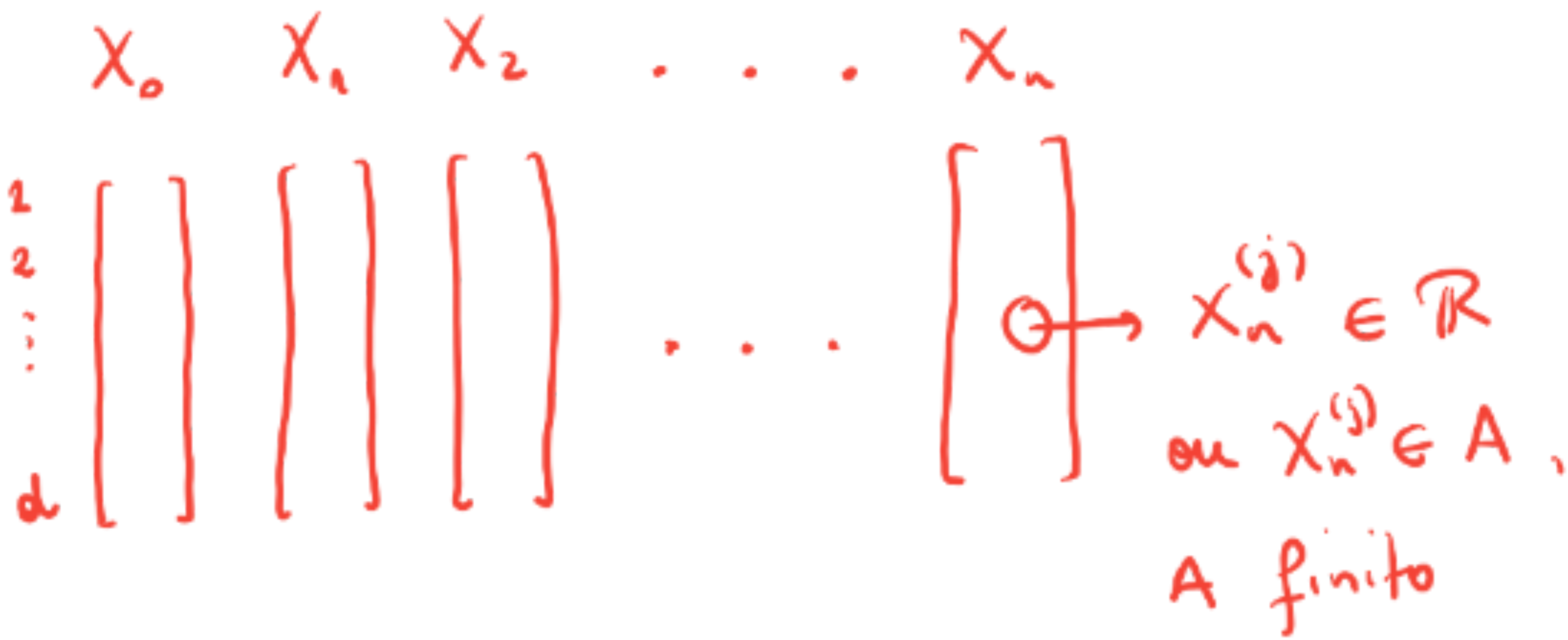
Em geral, dado um k fixo, sabemos estimar os parâmetros $\hat{\theta} \in \Theta_k$ (ex: máxima verossimilhança)

Estimadores regularizados



Exemplo 1

Série temporal $X_t \in \mathbb{R}^d$ com $t \in \mathbb{N}$.



Independent block identification in multivariate time series

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

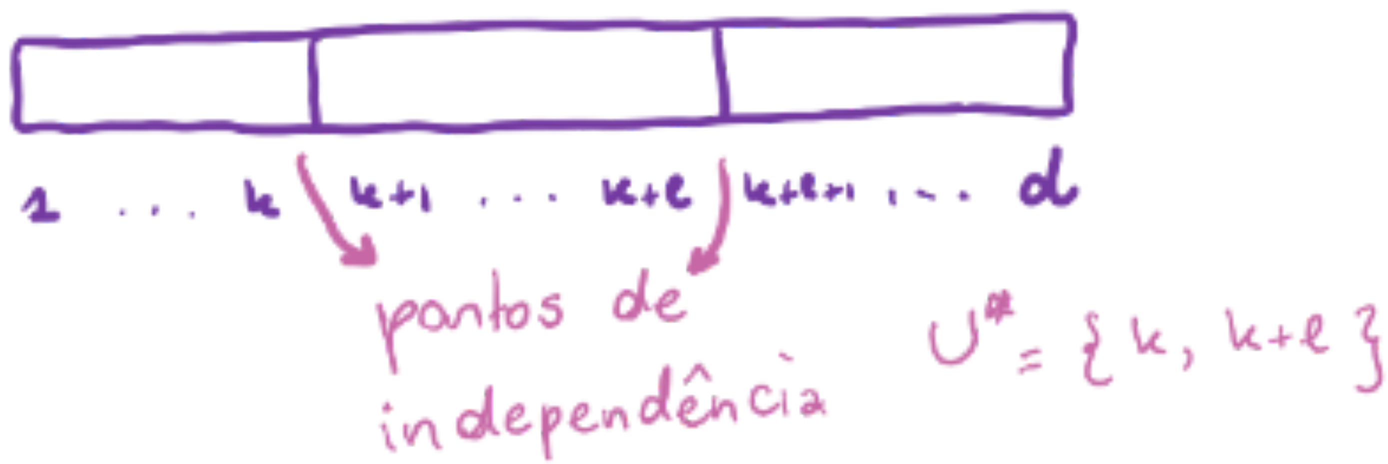
June 9, 2020

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 stationary 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.

Aceito em *Journal of Time Series Analysis, https://doi.org/10.1111/jtsa.12553*

Estrutura - Blocos independentes



Função de distribuição:

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

Quanto maior o número de pontos de independência, menor o sobreajuste do modelo

Estrutura - Blocos independentes

Função de distribuição:

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

Função de “custo":

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

Propriedade fundamental:

$$\ell(U, F) = 0 \quad \text{if } U \subseteq U^*(F), \text{ while} \quad \ell(U, F) > \alpha \quad \text{if } U \nsubseteq U^*(F).$$

Como estimar U^*?

Função de distribuição empírica:

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

Função de "custo" penalizada:

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

Estimador:

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

Consistência

Theorem 1 *Assume that*

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

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

Caso i.i.d:

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* $n \to \infty$.

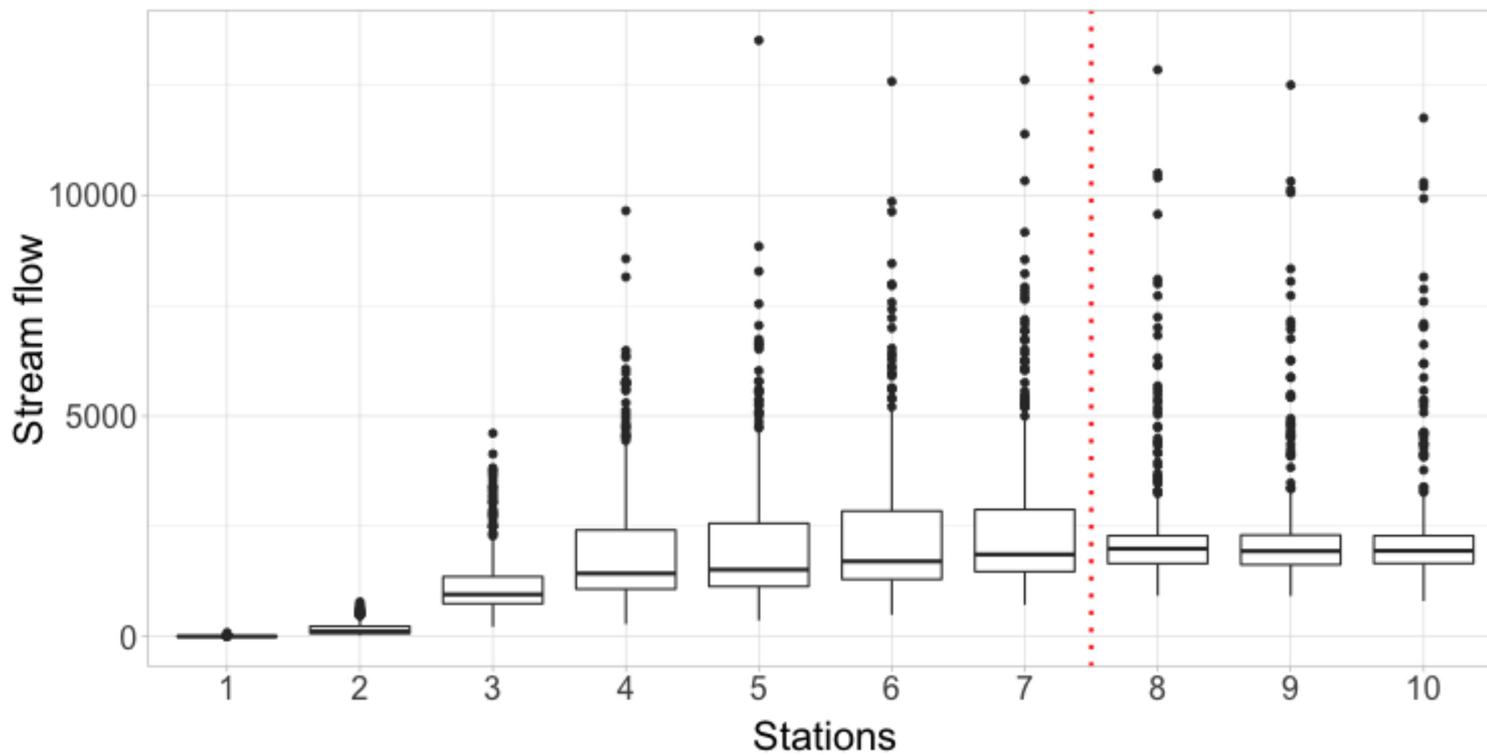
Propriedade de mistura para processos discretos:

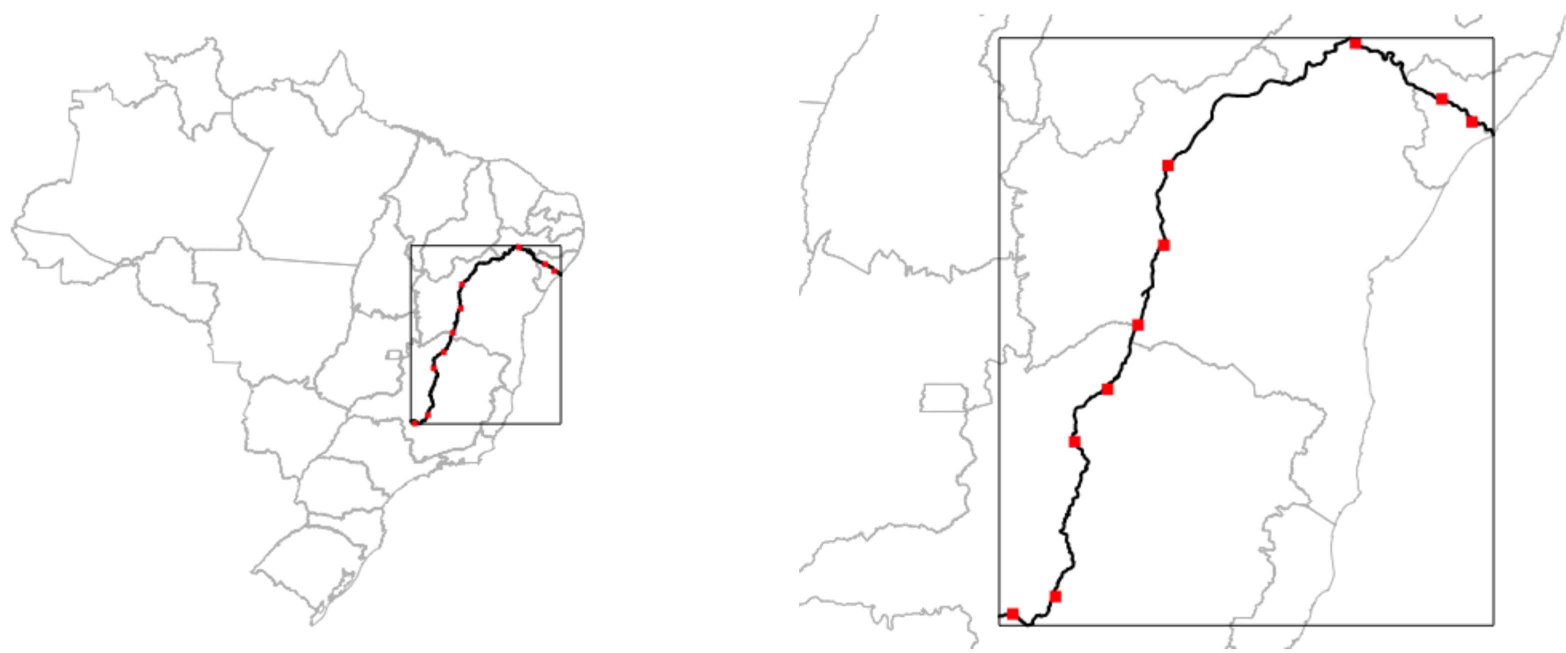
$\{\mathbf{X}^{(i)} : i \geq 1\}$ satisfies a mixing condition with rate $\{\psi(\ell)\} \downarrow 0$ as $\ell \to \infty$ if for each k, m and each $\mathbf{x}_1^k \in A^k$, $\mathbf{x}_1^m \in A^m$ with $\mathbb{P}(\mathbf{X}^{(1:m)} = \mathbf{x}_1^m) > 0$ we have

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

Caso discreto, “mixing":

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* $n \to \infty$.



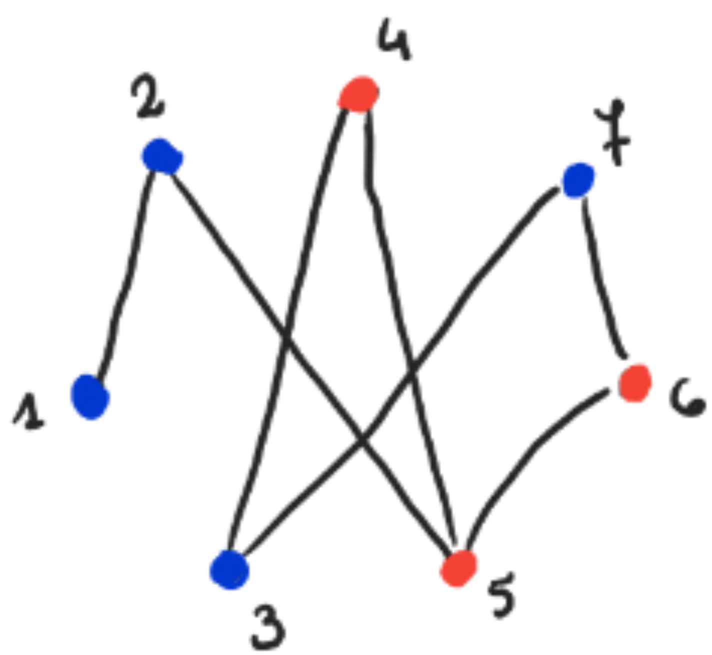


(a) (b)

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 circles represent the ten stream flow gauges considered in our analysis, numbered in increasing order from bottom to top.

Exemplo 2

Rede aleatória $A_{n\times n}$ com $n \in \mathbb{N}$



$$A_{2\times 2} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$A_{3\times 3} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Estimation of the Number of Communities in the Stochastic Block Model

Andressa Cerqueira and Florencia Leonardi

***Abstract*—In this paper we introduce an estimator for the number of communities in the Stochastic Block Model (SBM), based on the maximization of a penalized version of the so-called Krichevsky-Trofimov mixture distribution. We prove its eventual almost sure convergence to the underlying number of communities, without assuming a known upper bound on that quantity. Our results apply to both the dense and the sparse regimes. To our knowledge this is the first consistency result for the estimation of the number of communities in the SBM in the unbounded case, that is when the number of communities is allowed to grow with the same size.**

***Index Terms*—Model selection, SBM, Krichevsky-Trofimov distribution, Minimum Description Length, Bayesian Information Criterion**

Aceito para publicação em *IEEE Transactions on Information Theory*

Estrutura - Número de comunidades

Espaço de parâmetros do SBM

$$\Theta^k = \Big\{(\pi, P)\colon \pi \in (0,1]^k, \sum_{a=1}^{k} \pi_a = 1, P \in [0,1]^{k\times k},$$
$$P \text{ is symmetric}\Big\}.$$

Distribuição *a priori*

$$\nu_k(\pi, P) = \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k} \prod_{1\le a\le k} \pi_a^{-\frac{1}{2}}$$
$$\times \prod_{1\le a\le b\le k} \frac{1}{\Gamma\left(\frac{1}{2}\right)^2} P_{a,b}^{-\frac{1}{2}} (1-P_{a,b})^{-\frac{1}{2}}$$

Estrutura - Número de comunidades

Distribuição de Krichevsky-Trofimov

$$\begin{aligned}\mathbf{KT}_k(\mathbf{a}_{n\times n}) &= \mathbb{E}_{\nu_k}[\,\mathbb{P}_{\pi,P}(\mathbf{a}_{n\times n})\,] \\ &= \int_{\Theta^k} \mathbb{P}_{\pi,P}(\mathbf{a}_{n\times n})\nu_k(\pi,P)d\pi dP\,,\end{aligned}$$

Estimador

$$\hat{k}_{\mathrm{KT}}(\mathbf{a}_{n\times n}) = \underset{1\le k\le n}{\arg\max}\{\,\log \mathbf{KT}_k(\mathbf{a}_{n\times n}) - \mathrm{pen}(k,n)\,\}\,,$$

com a penalidade

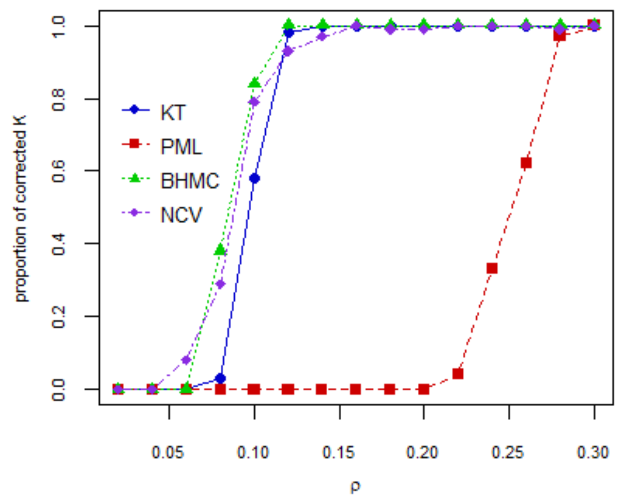
$$\mathrm{pen}(k,n) = \Big[\tfrac{k(k-1)(2k-1)}{12} + \tfrac{k(k-1)}{2} + \tfrac{(1+\epsilon)(k-1)}{2}\Big]\log n$$

Consistência

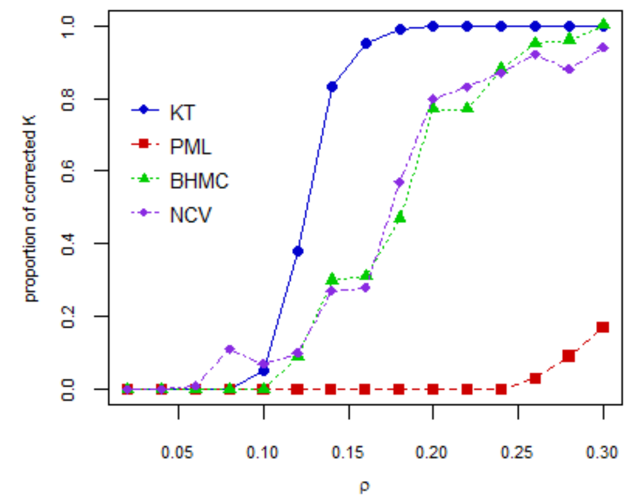
Theorem 2. *Suppose the SBM has order* k_0 *with parameters* (π^0, P^0)*, and suppose pen*(k,n) *is given by* (8)*. Then we have that*

$$\hat{k}_{KT}(\mathbf{a}_{n\times n}) = k_0$$

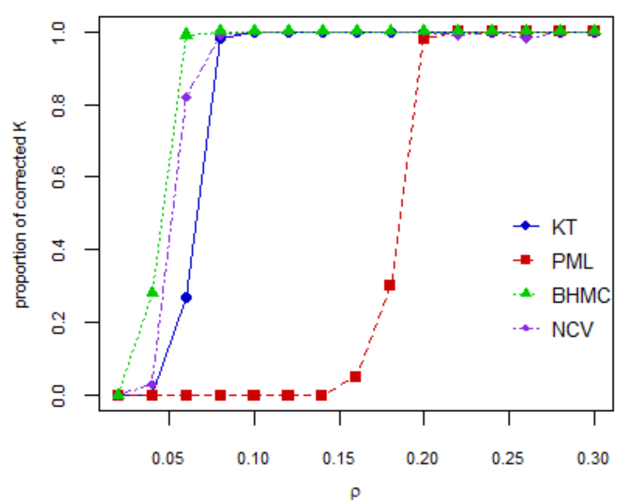
eventually almost surely as $n \to \infty$.



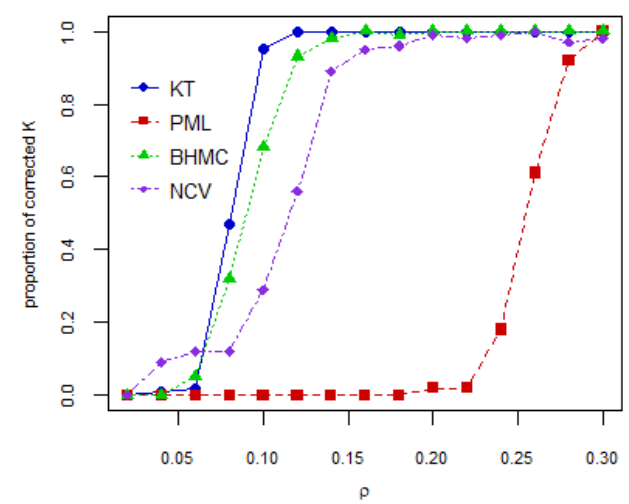
(a) $n = 300$ and $\pi = (1/3, 1/3, 1/3)$



(b) $n = 300$ and $\pi = (0.2, 0.5, 0.3)$



(c) $n = 500$ and $\pi = (1/3, 1/3, 1/3)$



(d) $n = 500$ and $\pi = (0.2, 0.5, 0.3)$

Fig. 1: Proportion of correct estimates for k_0 using the methods: Krichevsky-Trofimov (KT), Beth-Hessian matrix with moment correction (BHMC), network cross-validation (NCV) and penalized maximum likelihood (PML). We consider the model with $k_0 = 3$, $P_0 = \rho S_0$, where S_0 has diagonal entries equal to 2 and off-diagonal entries equal to 1. The tuning parameter in PML was chosen as $\lambda = 0.1$ and in KT as $\epsilon = 1$.

Perguntas em aberto (trabalhos futuros)

- Modelos com “alta dimensão” $(d, k \to \infty)$

- Desigualdades de grandes desvios (velocidade de convergência de estimadores)

- Outras estruturas de dependência (modelos gráficos)

- Desigualdades de tipo “oráculo” (quando não assumimos um modelo “correto”)

Obrigada !!