KYBERNETIKA — VOLUME 55 (2019), NUMBER 5, PAGES 782–801

GRAPHICAL MODEL SELECTION FOR A PARTICULAR CLASS OF CONTINUOUS-TIME PROCESSES

Mattia Zorzi

Graphical models provide an undirected graph representation of relations between the components of a random vector. In the Gaussian case such an undirected graph is used to describe conditional independence relations among such components. In this paper, we consider a continuous-time Gaussian model which is accessible to observations only at time T. We introduce the concept of infinitesimal conditional independence for such a model. Then, we address the corresponding graphical model selection problem, i. e. the problem to estimate the graphical model from data. Finally, simulation studies are proposed to test the effectiveness of the graphical model selection procedure.

Keywords: sparse inverse covariance selection, regularization, graphical models, entropy, optimization

Classification: 93B30, 65K10

1. INTRODUCTION

In many fields, e. g. biology, medicine, and econometrics, there are applications in which there is a large number of variables and it is crucial to understand the interactions among them. Such interactions can be described by directed and undirected graphs, [23, 24, 26]. In the present paper, we consider the undirected case. Consider a Gaussian random vector $x \in \mathbb{R}^n$ with zero mean and covariance matrix Σ. Let x_i denote the ith component of x. We can attach to it an undirected graph $\mathcal{G}$ with n nodes, one for each component x_i, $i = 1 \dots n$, and there is an edge connecting nodes i and j if and only if the components x_i and x_j are *conditionally dependent* given the other components of x. Let $(A)_{ij}$ denote the entry in position (i, j) of a matrix A. The conditional independence property can be characterized in terms of the covariance matrix: x_i and x_j are conditionally independent with respect to the other components if and only if

$$(\Sigma^{-1})_{ij} = 0. \tag{1}$$

This characterization allows to include conditional independence relations while estimating Σ. Dempster in [14] proposed the following *covariance selection* problem:

$$\begin{aligned} \Sigma^\circ = & \operatorname*{argmax}_{\Sigma \succ 0} \log\det(\Sigma) \\ & \text{subject to } (\Sigma)_{ij} = (\hat{\Sigma})_{ij} \; \forall\, (i,j) \in \Omega \end{aligned} \tag{2}$$

DOI: 10.14736/kyb-2019-5-0782

where $\Sigma \succ 0$ means that Σ is required to be positive definite. Here $\hat{\Sigma}$ is the sample covariance matrix, $\Omega \subseteq \{(i,j) \text{ s.t. } i,j = 1 \dots n\}$, such that $(i,i) \in \Omega$ for any $i = 1 \dots n$, and the objective function is the differential entropy of x, [12]. Dempster has proven that the solution Σ° of (2) is such that the support of $(\Sigma^\circ)^{-1}$ does coincide with Ω. Accordingly, such a covariance selection problem provides a covariance matrix corresponding to a graphical model with topology defined by Ω. Furthermore, such a paradigm represents a maximum entropy problem and thus it maximizes the information that can be encoded in the random vector x. Covariance selection problems received a great deal of attention, [8, 9, 10, 15], as well as their dynamic generalizations which are limited to discrete-time models [1, 2, 3, 6, 7, 16, 27, 31, 33, 36, 37]. Interestingly, the optimal solution of (2) is such that $\Sigma^\circ = S^{-1}$ where S is given by solving the dual problem:

$$\begin{aligned} &\min_{S \succ 0} -\log\det(S) + \operatorname{tr}(S\hat{\Sigma}) \\ &\text{subject to } (S)_{ij} = 0 \; \forall\, (i,j) \notin \Omega \end{aligned} \tag{3}$$

that is Ω represents the set of conditionally dependent pairs of components for the optimal solution.

In practice the topology of the graph, i. e. Ω, is not known and needs to be estimated from data. So, heuristic methods for topology selection have been introduced. The latter consider a relaxed version of (3) which is referred to as *topology selection* problem:

$$\min_{S \succ 0} -\log\det(S) + \operatorname{tr}(S\hat{\Sigma}) + \gamma h(S) \tag{4}$$

where the constraint with Ω is replaced by a ℓ_1-type penalty term $h(S)$ which favors a sparse solution S, [13, 17, 20, 25]. Looking at the support of the optimal solution for S, we obtain an estimate of Ω. Finally, $\gamma > 0$ is the regularization parameter whose optimal value is given by using a model selection criterium with complexity terms, e. g. AIC, BIC, [28, 37].

In biology and medicine the underlying variables are modeled by differential equations. One could estimate the interactions among those variables by using discrete-time models, provided that the sampling frequency (hereafter called empirical frequency) is high enough. On the other hand, in such applications the time series are sampled considerably slower than the empirical frequency. This leads to an unsuitable model parametrization and thus unsatisfactory estimates of the topology. It is then necessary to develop continuous-time estimation paradigms which exploit a suitable model parametrization for estimating such graphs. Within this framework it is worth mentioning the estimation procedures proposed in [30, 22] for continuous-time directed graphs.

In the present paper we consider a continuous-time graphical model selection problem. First, we introduce the definition of infinitesimal conditional independence for the corresponding continuous-time process, see Section 2. Then, we derive the corresponding covariance selection and topology selection problems, see Section 3. Moreover, we show how these problems are connected with the usual covariance selection problem (2) introduced above, see Section 4. Finally, we test the proposed graphical model selection procedure with some Monte Carlo studies showing that the notion of infinitesimal conditional independence leads to the correct parametrization in the continuous-time case, see Section 5.

Further notation. Let I denote the identity matrix. The vector space of symmetric matrices of dimension n is denoted by $\mathcal{Q}_n$. The matrix $\mathrm{diag}(d_1 \dots d_n)$ denotes a diagonal matrix whose elements in the main diagonal are $d_1 \dots d_n$. Let $Q = UDU^T$ be the eigenvalue decomposition of a positive definite matrix Q, i. e. $D = \mathrm{diag}(d_1 \dots d_n) \succ 0$ and U orthogonal; then Q^c, with $c \in \mathbb{R}$, and $\log(Q)$ are defined as $Q^c = U\mathrm{diag}(d_1^c \dots d_n^c)U^T$ and $\log(Q) = U\mathrm{diag}(\log d_1 \dots \log d_n)U^T$, respectively. The matrix exponential of $S \in \mathcal{Q}_n$ is denoted as e^Q. Given a matrix $S \in \mathcal{Q}_n$, $\mathrm{diag}(S)$ denotes the diagonal matrix whose main diagonal coincides with one of S. Given a matrix $S \in \mathcal{Q}_n$, $\mathcal{I}_S := \{(i,j) \text{ with } i,j = 1 \dots n \text{ s.t. } i \neq j,\ (S)_{ij} = 0\}$. $\mathcal{I}_S^c$ denotes the complement of set $\mathcal{I}_S$. Let $\mathrm{card}(\mathcal{I}_S)$ denote the cardinality of $\mathcal{I}_S$. The notation $x \sim (\mu, \Sigma)$ means that x is a Gaussian random vector with mean vector μ and covariance matrix Σ. The symbol $\mathbb{E}[\cdot]$ denotes the expectation operator.

2. INFINITESIMAL CONDITIONAL INDEPENDENCE

Consider the continuous-time zero mean stochastic process defined in the interval $[0, T]$

$$\begin{aligned} \dot{x}(t) &= -\frac{1}{2} S x(t), \quad t \in [0,T] \\ x(0) &\sim \mathcal{N}(0, I) \end{aligned} \tag{5}$$

where $S \in \mathcal{Q}_n$. Note that, at the final time T we have $x(T) = e^{-\frac{1}{2}ST}x(0)$. Thus, $x(T)$ is a Gaussian random vector with covariance matrix

$$\Sigma := \mathbb{E}[x(T)x(T)^\top] = e^{-ST}. \tag{6}$$

In what follows, we shall show that the support of matrix S reflects the presence of infinitesimal conditional dependence relations among the components of $x(t)$. In order to introduce formally the definition of infinitesimal conditional independence, we consider the sampled zero mean process $x^k(t)$, $t \in [0, T)$, such that $x^k(t) = x_d^k(l)$ for any $lk^{-1}T \leq t < (l+1)k^{-1}T$, $l \in \{0, 1 \dots k-1\}$, $T > 0$, $k \in \mathbb{N}$. $x_d^k(l)$ is the discrete-time stochastic process

$$\begin{aligned} x_d^k(l+1) &= (I + k^{-1}TS)^{-\frac{1}{2}} x_d^k(l), \quad l \in \{0, 1 \dots k-1\} \\ x_d^k(0) &\sim \mathcal{N}(0, I) \end{aligned} \tag{7}$$

where $I + k^{-1}TS \succ 0$. Note that,

$$\mathbb{E}[x^k(T/k)x^k(T/k)^\top] = (I + k^{-1}TS)^{-1}.$$

Let $\hat{x}_{i,j}^k(T/k) = \mathbb{E}[x_i^k(T/k) | x_l^k(T/k),\ l \neq i,j]$ be the conditional expectation of $x_i^k(T/k)$ given $x_l^k(T/k)$ with $l \neq i,j$ and $i \neq j$. The residual error is defined as $\varepsilon_{i,j}^k(T/k) = x_i^k(T/k) - \hat{x}_{i,j}^k(T/k)$. Note that, $\varepsilon_{i,j}^k(T/k)$ is a Gaussian random vector with zero mean. It is well known that $x_i^k(T/k)$ and $x_j^k(T/k)$ are conditionally independent given $x_l^k(T/k)$, with $l \neq i,j$ and $i \neq j$, if and only if $\varepsilon_{i,j}^k(T/k)$ and $\varepsilon_{j,i}^k(T/k)$ are independent, i. e. $\mathbb{E}[\varepsilon_{i,j}^k(T/k)(\varepsilon_{j,i}^k(T/k))^\top] = 0$.

Proposition 2.1. The following relation holds:

$$\mathbb{E}[\varepsilon^k_{i,j}(T/k)(\varepsilon^k_{j,i}(T/k))^\top] = \frac{-k^{-1}T(S)_{ij}}{(1+k^{-1}T(S)_{ii})(1+k^{-1}T(S)_{jj}) - (k^{-1}T(S)_{ij})^2}. \tag{8}$$

P r o o f. For a given (i,j), with $i \neq j$, let P be a permutation matrix such that

$$\tilde{x}^k(t) = Px^k(t) = [\,(y^k(t))^\top\,(s^k(t))^\top\,]^\top$$

where $y^k(t) = [\,x^k_i(t)\,x^k_j(t)\,]^\top$ and $s^k(t)$ is formed by the remaining components ordered by their indices. Then, the covariance matrix of $\tilde{x}^k(T/k)$ is $\Gamma_k = P(I+k^{-1}TS)^{-1}P^T = (I+k^{-1}TPSP^\top)^{-1}$ and we consider the following partition according to y^k and s^k:

$$\Gamma_k = (I+k^{-1}TPSP^\top)^{-1} = \begin{bmatrix} \Gamma_{k;yy} & \Gamma_{k;ys} \\ \Gamma_{k;sy} & \Gamma_{k;ss} \end{bmatrix}.$$

Consider the residual error vector

$$\begin{aligned} \varepsilon^k_y(T/k) &= y^k(T/k) - \mathbb{E}[y^k(T/k)\,|\,s^k(T/k)] \\ &= \begin{bmatrix} \varepsilon^k_{i,j}(T/k) \\ \varepsilon^k_{j,i}(T/k) \end{bmatrix} \end{aligned}$$

which is Gaussian and with zero mean. Then,

$$\begin{aligned} \mathbb{E}[\varepsilon^k_y(T/k)(\varepsilon^k_y(T/k))^\top] &= \Gamma_{k;yy} - \Gamma_{k;ys}\Gamma^{-1}_{k;ss}\Gamma_{k;sy} \\ &= \begin{bmatrix} 1+k^{-1}T(S)_{ii} & k^{-1}T(S)_{ij} \\ k^{-1}T(S)_{ij} & 1+k^{-1}T(S)_{jj} \end{bmatrix}^{-1} \\ &= \frac{1}{(1+k^{-1}T(S)_{ii})(1+k^{-1}T(S)_{jj}) - (k^{-1}T(S)_{ij})^2} \times \\ &\qquad \times \begin{bmatrix} 1+k^{-1}T(S)_{jj} & -k^{-1}T(S)_{ij} \\ -k^{-1}T(S)_{ij} & 1+k^{-1}T(S)_{ii} \end{bmatrix}. \end{aligned} \tag{9}$$

In position (1,2) of the equality above we have (8). □

In view of (8), the conditional independence property between $x^k_i(T/k)$ and $x^k_j(T/k)$ is given by condition $(S)_{ij} = 0$. More precisely, $\mathcal{I}_S$ describes the conditional independent pairs of variables in x^k at time T/k assuming that $x^k(0)$ has independent components. It is worth noting that when $k = 1$, in model (7) there is no dynamics and $\mathcal{I}_S$ describes the conditional independence pairs of variables in $x^1(T)$. Accordingly, model (7) with $k = 1$ coincides with the graphical model in [17], i. e. the one described in the Introduction. Note that, at the final time T we have $x^k(T) = (I+k^{-1}TS)^{-\frac{k}{2}}x^k(0)$. Thus, $x^k(T)$ is a zero mean Gaussian random vector with covariance matrix

$$\Sigma_k := (I+k^{-1}TS)^{-k}. \tag{10}$$

Next we consider the case when k approaches infinity. From (7), we have

$$\frac{x^k(t+T/k) - x^k(t)}{T/k} = \frac{(I+k^{-1}TS)^{-\frac{1}{2}} - I}{T/k}x^k(t). \tag{11}$$

Proposition 2.2. We have:

$$\lim_{k\to\infty} \frac{(I + k^{-1}TS)^{-\frac{1}{2}} - I}{T/k} = -\frac{1}{2}S.$$

Proof. Let $S = UDU^\top$ be the eigenvalue decomposition of S, that is $UU^\top = I$ and $D = \mathrm{diag}(d_1 \dots d_n)$. Then,

$$\lim_{k\to\infty} \frac{(I + k^{-1}TS)^{-\frac{1}{2}} - I}{T/k} = U\mathrm{diag}(f_{1,k} \dots f_{n,k})U^\top \tag{12}$$

where $f_{j,k} = ((1 + k^{-1}Td_j)^{-\frac{1}{2}} - 1)/(T/k)$. It is not difficult to see that $f_{j,k} \to -d_j/2$ as $k \to \infty$. Accordingly, the right hand side of (12) tends to $-\frac{1}{2}U\mathrm{diag}(d_1 \dots d_n)U^\top$ which concludes the proof. □

In view of Proposition 2.2, taking the limit of (11) as $k \to \infty$ we obtain

$$\dot{x}(t) = -\frac{1}{2}Sx(t)$$

where $x(t) := \lim_{k\to\infty} x^k(t)$ for any $t \in [0, T]$. Taking the limit of (8) as $k \to \infty$, the covariance of $\varepsilon^k_{i,j}(T/k)$ and $\varepsilon^k_{j,i}(T/k)$ tends to zero, that is the components of $x(0^+)$ are independent. On the other hand, from (8) we have

$$\lim_{k\to\infty} \frac{k\mathbb{E}[\varepsilon^k_{i,j}(T/k)(\varepsilon^k_{j,i}(T/k))^\top]}{T} = (S)_{ij} \tag{13}$$

where the term on the left hand side can be understood as the infinitesimal covariance between $\hat{\varepsilon}_{i,j}(0^+)$ and $\hat{\varepsilon}_{j,i}(0^+)$ where $\hat{\varepsilon}_{i,j}(t) = x_i(t) - \mathbb{E}[x_i(t)|x_l(t),\, l \neq i, j]$. More precisely, $\mathcal{I}_S$ describes the infinitesimal conditional independent pairs of x at time 0^+ assuming that $x(0)$ has independent components.

Definition 2.3. Consider the model (5). We say that x_i and x_j, with $i \neq j$, are infinitesimally conditionally independent given x_l with $l \neq i, j$, if and only if

$$\lim_{k\to\infty} \frac{k\mathbb{E}[\varepsilon^k_{i,j}(T/k)(\varepsilon^k_{j,i}(T/k))^\top]}{T} = 0.$$

As a consequence, x_i and x_j are infinitesimally conditionally independent given x_l with $l \neq i, j$ if and only if $(i, j) \in \mathcal{I}_S$. It is worth noting that matrix S does not describe a Bayesian network because it does not represent a directed acyclic graph. This means that, if we consider a directed graph for S then we are not able to understand "who causes who", making the interpretation of the directed graph ambiguous.

Definition 2.4. The graphical model for (5) corresponds to an undirected graph $\mathcal{G}$ with n nodes, one for each component of $x(t)$, and there is an edge connecting nodes i and j if and only if the components x_i and x_j are infinitesimally conditionally dependent given the other components of x.

3. GRAPHICAL MODEL SELECTION

Consider the model in (5), and assume we have the sample covariance $\hat{\Sigma} \succ 0$ of $x(T)$. Such a sample covariance matrix is obtained by considering N independent realizations of x and measuring $x(t)$ only at time T. Moreover, to ease the exposition we assume that $T = 1$. The aim of this section is to show how to select the graphical model of (5) from the sample covariance of x at the final time T. In doing that, we introduce the covariance and topology selection problems, and finally we outline the procedure to select the graphical model. It is worth noting that we could consider the suggested problem as simply a reparametrization of the graphical lasso problem in (4), i. e. find a matrix S sparse such that $\hat{\Sigma} \approx e^{-ST}$. However, the latter does not correspond to a covariance selection problem.

3.1. Covariance selection problem

Recall that the covariance matrix of $x(T)$ with $T = 1$ in (5) is $\Sigma = e^{-S}$ with S symmetric. In what follows, we shall show that the solution of the following covariance selection problem

$$\begin{aligned} \Sigma^{\circ} = & \underset{\Sigma \succeq 0}{\operatorname{argmax}} \, \mathbb{H}(\Sigma) \\ & \text{subject to } (\Sigma)_{ij} = (\hat{\Sigma})_{ij} \; \forall \, (i,j) \in \Omega \end{aligned} \tag{14}$$

is such that $\Sigma^{\circ} = e^{-S}$ with S symmetric and $\mathcal{I}_S^c = \Omega$ and thus in accordance with model (5). Here, $\mathbb{H}(\Sigma) = -\operatorname{tr}(\Sigma \log(\Sigma))$ is the *von Neumann* entropy [12] and $\Omega \subseteq \{(i,j) \text{ s.t. } i,j = 1 \dots n\}$ is such that $(i,i) \in \Omega$ for any $i = 1 \dots n$. It is worth noting that, in contrast with the usual entropy in (2), $\mathbb{H}$ is well defined also in the case that Σ is positive semidefinite and singular.

We analyze the solution of (14) by using the duality theory. We define the operator $P_{\Omega} : \mathcal{Q}_n \to \mathcal{Q}_n$ as follows

$$(P_{\Omega}(\Sigma))_{ij} = \begin{cases} (\Sigma)_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Then, we can rewrite the constraint in (14) as $P_{\Omega}(\hat{\Sigma} - \Sigma) = 0$. The Lagrangian function is

$$L(\Sigma, Q) = -\operatorname{tr}(\Sigma \log(\Sigma)) + \operatorname{tr}(Q P_{\Omega}(\hat{\Sigma} - \Sigma)) \tag{16}$$

where $Q \in \mathcal{Q}_n$ is the Lagrange multiplier. Moreover,

$$L(\Sigma, Q) = -\operatorname{tr}(\Sigma \log(\Sigma)) + \operatorname{tr}(P_{\Omega}(Q)(\hat{\Sigma} - \Sigma)) \tag{17}$$

where we exploited the fact that P_{Ω} is a self-adjoint operator. The first and the second variation (i. e. Gateaux derivative) of $L(\Sigma, Q)$ with respect to Σ in direction $\delta\Sigma \in \mathcal{Q}_n$

are, respectively,

$$\begin{aligned}\delta L(\Sigma, Q; \delta\Sigma) &= \operatorname{tr}((-I - \log(\Sigma) - P_\Omega(Q))\delta\Sigma)\\ \delta^2 L(\Sigma, Q; \delta\Sigma) &= -\operatorname{tr}\left(\int_0^\infty (\Sigma + tI)^{-1}\delta\Sigma(\Sigma + tI)^{-1}\mathrm{d}t\delta\Sigma\right) \qquad (18)\\ &= -\operatorname{tr}\int_0^\infty (\Sigma + tI)^{-\frac{1}{2}}\delta\Sigma(\Sigma + tI)^{-1}\delta\Sigma(\Sigma + tI)^{-\frac{1}{2}}\mathrm{d}t. \qquad (19)\end{aligned}$$

Hence, $L(\cdot, Q)$ is strictly concave because $\delta^2 L(\Sigma, Q; \delta\Sigma) < 0$ for any $\delta\Sigma \neq 0$. Indeed, the integrand of the second variation is a nonnull positive semi-definite matrix. Accordingly, its maximum point is given by setting equal to zero its first variation for any $\delta\Sigma \in \mathcal{Q}_n$:

$$\delta L(\Sigma, Q; \delta\Sigma) = \operatorname{tr}((-I - \log(\Sigma) - P_\Omega(Q))\delta\Sigma) = 0$$

which implies that $-I - \log(\Sigma) - P_\Omega(Q) = 0$, accordingly the optimal form for Σ is

$$\Sigma^\circ = e^{-I - P_\Omega(Q)}. \qquad (20)$$

In view of (6), we define

$$S := I + P_\Omega(Q), \qquad (21)$$

moreover we have that $(S)_{ij} = 0$ for any $(i, j) \notin \Omega$.

The dual function is

$$\begin{aligned}J(S) &= L(e^{-S}, Q)\\ &= -\operatorname{tr}(e^{-S}\log(e^{-S})) + \operatorname{tr}(S(\hat{\Sigma} - e^{-S})) - \operatorname{tr}(\hat{\Sigma} - e^{-S})\\ &= \operatorname{tr}(Se^{-S}) + \operatorname{tr}(S\hat{\Sigma}) - \operatorname{tr}(Se^{-S}) - \operatorname{tr}(\hat{\Sigma}) + \operatorname{tr}(e^{-S})\\ &= \operatorname{tr}(e^{-S}) + \operatorname{tr}(S\hat{\Sigma}) - \operatorname{tr}(\hat{\Sigma}).\end{aligned}$$

Thus the dual problem is

$$\begin{aligned}&\min_{S \in \mathcal{Q}_n} \operatorname{tr}(e^{-S}) + \operatorname{tr}(S\hat{\Sigma}) - \operatorname{tr}(\hat{\Sigma})\\ &\text{subject to } (S)_{ij} = 0 \;\forall\, (i, j) \notin \Omega. \qquad (22)\end{aligned}$$

Theorem 3.1. Problem (22) admits a solution S° which is also unique. Thus, the optimal solution to problem (14) is $\Sigma^\circ = e^{-S^\circ}$.

P r o o f. The first and the second variation of $J(S)$ in direction $\delta S \in \mathcal{Q}_n$ are, respectively,

$$\begin{aligned}\delta J(S; \delta S) &= \operatorname{tr}((-e^{-S} + \hat{\Sigma})\delta S)\\ \delta^2 J(S; \delta S) &= \operatorname{tr}\int_0^1 e^{-(1-\tau)S}\delta S e^{-\tau S}\delta S\mathrm{d}\tau\\ &= \operatorname{tr}\int_0^1 e^{-\frac{1}{2}(1-\tau)S}\delta S e^{-\tau S}\delta S e^{-\frac{1}{2}(1-\tau)S}\mathrm{d}\tau.\end{aligned}$$

Since the integrand of $\delta^2 J(S;\delta S)$ is positive semi-definite and different from zero for $\delta S \neq 0$, we have that $\delta^2 J(S;\delta S) > 0$ for $\delta S \neq 0$. Accordingly, $J(S)$ is strictly convex. Thus, if the solution to the dual problem exists, then it is unique.

We proceed to prove the existence. We have to minimize J over the set $\mathcal{C} := \{\, S \in \mathcal{Q}_n \text{ s.t. } (S)_{ij} = 0 \,\forall\, (i,j) \notin \Omega\}$ which is unbounded. We show that the search of the minimum can be restricted over a closed and bounded set. To do that, we take a sequence $S_k \in \mathcal{C}$ with $k \in \mathbb{N}$ and such that $\|S_k\| \to \infty$ as $k \to \infty$. Since S_k is symmetric this implies that the matrix has some eigenvalues which diverge as $k \to \infty$. We have two possible cases. *First case.* S_k has some eigenvalues tending to infinity and the others are bounded. Since $\hat{\Sigma} \succ 0$, there exists $\mu > 0$ such that $\hat{\Sigma} \succeq \mu I$. We have that $\mathrm{tr}(S_k\hat{\Sigma}) \to \infty$, while the other terms are bounded. Thus, $J(S_k) \to \infty$. *Second case.* S_k has some eigenvalues tending to minus infinity and the other eigenvalues are bounded or tend to infninty. Clearly $\mathrm{tr}(e^{-S_k}) \to \infty$ and $\mathrm{tr}(S_k\hat{\Sigma})$ can tend to a finite value or to $\pm\infty$. Since the former dominates the latter, $J(S_k) \to \infty$. Therefore, we can restrict the search of the minimum over a closed and bounded set. Since J is a continuous function, by the Weierstrass' Theorem we conclude that the dual problem admits solution. □

Remark 3.2. Let

$$\mathcal{P}_\Omega := \{\, e^{-S} \text{ s.t. } S \in \mathcal{Q}_n,\, (S)_{ij} = 0 \,\forall\, (i,j) \notin \Omega \,\}. \tag{23}$$

We can rewrite (22) in terms of $\Sigma \in \mathcal{P}_\Omega$ as follows:

$$\Sigma^\circ = \underset{\substack{\Sigma \in \mathcal{P}_\Omega \\ \text{s.t. } \Sigma \succ 0}}{\mathrm{argmin}}\ \mathrm{tr}(\Sigma) - \mathrm{tr}(\hat{\Sigma}\log\Sigma) \tag{24}$$

where we have exploited the relations $\Sigma = e^{-S}$ and $S = -\log\Sigma$. In the objective function of (24), we can add the terms $\mathrm{tr}(\hat{\Sigma}\log\hat{\Sigma}) - \mathrm{tr}(\hat{\Sigma})$ not depending on Σ:

$$\begin{aligned}\Sigma^\circ &= \underset{\substack{\Sigma \in \mathcal{P}_\Omega \\ \text{s.t. } \Sigma \succ 0}}{\mathrm{argmin}}\ \mathrm{tr}(\Sigma) - \mathrm{tr}(\hat{\Sigma}\log\Sigma) + \mathrm{tr}(\hat{\Sigma}\log\hat{\Sigma}) - \mathrm{tr}(\hat{\Sigma}) \\ &= \underset{\substack{\Sigma \in \mathcal{P}_\Omega \\ \text{s.t. } \Sigma \succ 0}}{\mathrm{argmin}}\ \mathcal{D}(\hat{\Sigma}\|\Sigma)\end{aligned} \tag{25}$$

where $\mathcal{D}(\hat{\Sigma}\|\Sigma)$ is the Umegaki-von Neumann's relative entropy [33]:

$$D(\hat{\Sigma}\|\Sigma) = \mathrm{tr}[\hat{\Sigma}(\log\hat{\Sigma} - \log\Sigma) - \hat{\Sigma} + \Sigma]. \tag{26}$$

Therefore, Σ° is the closest covariance matrix to $\hat{\Sigma}$, in the Umegaki-von Neumann metric, which belongs to $\mathcal{P}_\Omega$.

Corollary 3.3. The optimal solution of problem (14) has the structure $\Sigma^\circ = e^{-S^\circ}$ where $(S^\circ)_{ij} = 0$ for any $(i,j) \notin \Omega$ that is Σ° represents the covariance matrix at the final time for the model (5) wherein the infinitesimal conditional dependence pairs are specified by Ω.

It is worth noting that Corollary 3.3 is equivalent to the result by Dempster for the case $k = 1$, [14].

3.2. Topology selection problem

In practice in the problem (22) Ω is not known and we have to estimate it from the sample covariance of $x(T)$. Therefore, we consider the following regularized version of problem (22) inducing sparsity on S:

$$\min_{S\in\mathcal{Q}_n} \operatorname{tr}(e^{-S}) + \operatorname{tr}(S\hat{\Sigma}) + \gamma h(S) \tag{27}$$

where $\gamma > 0$ is the regularization parameter and h is the ℓ_1-type penalty function, [29], defined as

$$h(S) = \sum_{i\neq j} |(S)_{ij}|. \tag{28}$$

Note that, h does not penalize the entries in the main diagonal of S indeed we already know that Ω contains the pairs (i, i), with $i = 1 \dots n$.

To show that problem (27) does admit solution we exploit the duality theory. First, we rewrite it by introducing a new variable $Y \in \mathcal{Q}_n$:

$$\begin{aligned} &\min_{S,Y\in\mathcal{Q}_n} \operatorname{tr}(e^{-S}) + \operatorname{tr}(S\hat{\Sigma}) + \gamma h(Y) \\ &\text{subject to } S = Y. \end{aligned} \tag{29}$$

The Lagrangian is

$$L(S, Y, Z) = \operatorname{tr}(e^{-S}) + \operatorname{tr}(S\hat{\Sigma}) + \gamma h(Y) + \operatorname{tr}(Z(S - Y)) \tag{30}$$

where $Z \in \mathcal{Q}_n$ is the Lagrange multiplier. The term in L depending on Y is $\gamma h(S) - \operatorname{tr}(ZY)$ which is bounded from below if and only if

$$\operatorname{diag}(Z) = 0, \;\; |(Z)_{ij}| \leq \gamma \;\; i \neq j \tag{31}$$

and it that case the corresponding term takes value equal to zero. Thus, we have

$$\inf_Y L(S, Y, Z) = \begin{cases} \operatorname{tr}(e^{-S}) + \operatorname{tr}(S(\hat{\Sigma} + Z)) & \text{if (31) holds} \\ -\infty & \text{othervise.} \end{cases} \tag{32}$$

The remaining term is strictly convex in S, to see that it is sufficient to apply the same reasoning exploited in the proof of Theorem 3.1. Thus, the point of minimum for $L(\cdot, Y, Z)$ is given by setting equal to zero the first variation for any $\delta S \in \mathcal{Q}_n$:

$$\delta L(S, Y, Z; \delta S) = \operatorname{tr}((-e^{-S} + \hat{\Sigma} + Z)\delta S) = 0 \tag{33}$$

which implies that $S = -\log(\hat{\Sigma} + Z)$ under the assumption that

$$\hat{\Sigma} + Z \succ 0. \tag{34}$$

Thus, the dual function, under conditions (31) and (34), is

$$\begin{aligned} J(Z) &= L(-\log(\hat{\Sigma} + Z), Y, Z) \\ &= \operatorname{tr}(\hat{\Sigma} + Z - (\hat{\Sigma} + Z)\log(\hat{\Sigma} + Z)). \end{aligned}$$

Therefore, the dual problem is

$$\max_{Z\in\mathcal{Q}_n} \operatorname{tr}(\hat{\Sigma}+Z-(\hat{\Sigma}+Z)\log(\hat{\Sigma}+Z)))$$
$$\text{subject to } \operatorname{diag}(Z)=0,\ \ |(Z)_{ij}|\le\gamma\ \ i\neq j,\ \ \hat{\Sigma}+Z\succ 0. \tag{35}$$

Theorem 3.4. Problem (35), and thus also problem (27), admits a unique solution.

P r o o f . The first and the second variation of J are, respectively:

$$\begin{aligned}\delta J(Z;\delta Z) &= -\operatorname{tr}(\log(\hat{\Sigma}+Z)\delta Z)\\ \delta^2 J(Z;\delta Z) &= -\operatorname{tr}\left(\int_0^\infty (\hat{\Sigma}+Z+tI)^{-1}\delta Z(\hat{\Sigma}+Z+tI)^{-1}\delta Z\mathrm{d}t\right)\\ &= -\operatorname{tr}\left(\int_0^\infty (\hat{\Sigma}+Z+tI)^{-\frac{1}{2}}\delta Z(\hat{\Sigma}+Z+tI)^{-1}\delta Z(\hat{\Sigma}+Z+tI)^{-\frac{1}{2}}\mathrm{d}t\right)\end{aligned}$$

where $\hat{\Sigma}+Z+tI\succ 0$. Then, $\delta^2 J(Z;\delta Z)<0$ for any $\delta Z\neq 0$ because the integrand is nonnull and negative semi-definite for any $\delta Z\neq 0$. Thus, the $J(Z)$ is strictly concave, accordingly the solution, if it exists, is unique.

We proceed to prove the existence. First, we observe that the objective function is maximized over the set $\mathcal{C}:=\{Z\in\mathcal{Q}_n \text{ s.t. } \operatorname{diag}(Z)=0,\ \ |(Z)_{ij}|\le\gamma\ \ i\neq j,\ \ \hat{\Sigma}+Z\succ 0\}$ which is open and bounded. We show that we can restrict the search over a closed and bounded set. Then, the existence of the solution follows by applying the Weierstrass' Theorem, because $J(Z)$ is a continuos function over $\mathcal{C}$. Let $\partial\mathcal{C}$ denote the boundary not contained in $\mathcal{C}$. The latter contains matrices Z such that condition (31) holds and $\hat{\Sigma}+Z\succeq 0$ is singular. It is not difficult to see that $J(Z)=\sum_{i=1}^n d_i-d_i\log d_i$ where d_i denotes the ith eigenvalue of $\hat{\Sigma}+Z$. Since $\lim_{d\to 0^+} d-d\log d=0$, we conclude that $\lim_{Z\to\partial\mathcal{C}} J(Z)$ takes finite values. On the other hand, we know that the gradient of $J(Z)$ is $\nabla J(Z)=-\log(\hat{\Sigma}+Z)$ and its eigenvalues are $-\log d_i$. Since $\lim_{d\to 0^+}-\log d=\infty$, this means that there exists at least one direction in $\nabla J(Z)$ for which the corresponding eigenvalue tends to infinity as $Z\to\partial\mathcal{C}$. Accordingly, the maximum cannot be in $\partial\mathcal{C}$, and thus we can restrict the search of the maximum over a closed and bounded set. □

We exploited problem (35) to show the existence of a unique solution of problem (27). A remarkable difference between (27) and (35) is that in the former the objective function is not differentiable while in the latter it is. Accordingly, the importance of problem (35) is also that it allows to compute the optimal solution of (27) by applying gradient descent methods, [5].

3.3. Graphical model selection procedure

We propose the procedure displayed in Algorithm 1 to select the graphical model given the sample covariance matrix $\hat{\Sigma}$ of $x(T)$. It is worth noting that also problem (35) provides an estimate of Σ. However, the latter is biased by the penalty term h. For this reason, we refine it by solving problem (14) as suggested in [29]. In ℓ_{AIC}, the first two terms form, up to a constant factor, the negative Whittle log-likelihood, while the last

Algorithm 1 Graphical model selection procedure

1: Select a regularization path (i. e. an ordered sequence of values for the regularization parameter) γ_m, with $m = 1 \dots M$, such that $\gamma_{m+1} > \gamma_m > 0$.
2: For each γ_m solve problem (35); Let S_m be the optimal solution, then set $\Omega_m = \mathcal{I}_{S_m}^c$.
3: Solve problem (14) with Ω_m; let Σ_m° be the optimal solution.
4: Among all the models $\mathcal{M}_m = (\Omega_m, \Sigma_m^\circ)$ choose the one that minimizes the AIC score function:

$$\ell_{AIC}(\mathcal{M}_m) = N \log\det(\Sigma_m^\circ) + N \operatorname{tr}(\hat{\Sigma}(\Sigma_m^\circ))^{-1} + 2\operatorname{card}(\Omega_m) \tag{36}$$

where N is the number of independent realizations of $x(T)$ used to compute $\hat{\Sigma}$.

one is a complexity term which favors sparse solutions. Another possible choice is to consider the one given by the BIC criterium

$$\ell_{BIC}(\mathcal{M}_m) = N \log\det(\Sigma_m^\circ) + N \operatorname{tr}(\hat{\Sigma}(\Sigma_m^\circ))^{-1} + \operatorname{card}(\Omega_m) \log N. \tag{37}$$

3.4. The case with deterministic input

Consider the continuous-time stochastic process defined in the interval $[0, T]$

$$\begin{aligned} \dot{x}(t) &= -\frac{1}{2} S x(t) + B u(t), \quad t \in [0, T] \\ x(0) &\sim \mathcal{N}(0, I) \end{aligned} \tag{38}$$

where $S \in \mathcal{Q}_n$ and $B \in \mathbb{R}^{n \times p}$. The stochastic process (38) could find application in modeling noise-free brain networks with external stimuli [19]: $x(t)$ represents the magnitude of neurophysiological activity of the brain regions at time t; $-\frac{1}{2}S$ is the symmetric weighted adjacency matrix whose elements indicate the connections among brain regions; B identifies the control points in the brain; $u(t)$ represents the external stimuli. In such an application data are the so called "functional magnetic resonance imaging" (fMRI) data. The latter are characterized by a high sampling time, e. g. $T = 2$ sec, [21]. In other words, we can observe $x(T)$ only at time T.

We assume that both $B \in \mathbb{R}^{n \times p}$ and $u(t)$, with $t \in [0, T]$, are known. We assume to collect N independent realizations of x and measuring $x(t)$ only at time T. We want to estimate the matrix S.

It is worth noting that $x(T)$ is a Gaussian random vector and such that

$$x(T) = e^{-\frac{1}{2}ST} x(0) + \int_0^T e^{-\frac{1}{2}S(T-t)} B u(t) \mathrm{d}t. \tag{39}$$

Let

$$\mu_S := \int_0^T e^{-\frac{1}{2}S(T-t)} B u(t) \mathrm{d}t. \tag{40}$$

Then,

$$\bar{x}(T) := x(T) - \mu_S = e^{-\frac{1}{2}ST} x(0) \tag{41}$$

and

$$\mathbb{E}[\bar{x}(T)\bar{x}(T)^\top] = e^{-ST}. \tag{42}$$

Accordingly, we can estimate S by using an iterative procedure: given S, we compute μ_S; we compute the sample covariance matrix $\hat{\Sigma}$ of $\bar{x}(T)$; we apply Algorithm 1 to update S; we repeat the procedure until S does not change. The complete procedure is displayed in Algorithm 2.

Algorithm 2 Graphical model selection procedure with deterministic input

1: Select a regularization path γ_m, with $m = 1 \dots M$, such that $\gamma_{m+1} > \gamma_m > 0$.
2: **for** each γ_m **do**
3: **Initialize:** $S_{m,0} = -I$.
4: $l = 0$
5: **repeat**
6: Compute $\mu_{S_{m,l}} = \int_0^T e^{-\frac{1}{2}S_{m,l}(T-t)} Bu(t)\mathrm{d}t$.
7: Detrend the process $\bar{x}(T) = x(T) - \mu_{S_{m,l}}$ and compute the sample covariance matrix of $\bar{x}(T)$, say $\hat{\Sigma}_{m,l}$.
8: Solve Problem (35) with γ_m and the sample covariance matrix $\hat{\Sigma}_{m,l}$. Let $\Omega_{m,l}$ denote the topology of the optimal solution.
9: Solve Problem (14) with $\Omega_{m,l}$ and $\hat{\Sigma}_{m,l}$. Let $S_{m,l+1}$ be the corresponding dual solution.
10: $l = l + 1$
11: **until** $\|S_{m,l} - S_{m,l-1}\| \leq \varepsilon$ with ε fixed small constant.
12: Set $\Omega_m = \Omega_{m,l}$ and $\Sigma_m^\circ = e^{-S_{m,l}T}$.
13: **end for**
14: Among all the models $\mathcal{M}_m = (\Omega_m, \Sigma_m^\circ)$ choose the one that minimizes the AIC or BIC score function defined in (36) and (37), respectively.

4. CONNECTION WITH THE CLASSICAL PROBLEM

In Section 2 we showed that model (7) connects the usual graphical model without dynamics, $k = 1$, with the continuous-time model, $k = \infty$. Next we show that it is possible to set up a graphical model selection problem also for (7), that is there exists a connection between the usual graphical model selection problem and the one of Section 3. We consider the following family of entropies indexed by $k > 1$

$$\mathbb{H}_k(\Sigma) = \frac{k}{k-1}\operatorname{tr}(\Sigma^{\frac{k-1}{k}}) - k\operatorname{tr}(\Sigma) - n\frac{k}{k-1}. \tag{43}$$

Proposition 4.1. $\mathbb{H}_k(\Sigma) \leq 0$ and equality holds if and only if $\Sigma = I$. Moreover, $\mathbb{H}_k$ can be extended by continuity as follows:

$$\lim_{k\to 1} \mathbb{H}_k(\Sigma) = \log\det(\Sigma) - \operatorname{tr}(\Sigma) + n$$
$$\lim_{k\to \infty} \mathbb{H}_k(\Sigma) = -\operatorname{tr}(\Sigma\log(\Sigma)) + \operatorname{tr}(\Sigma) - n.$$

P r o o f . The first point follows from the fact that $\mathbb{H}_k(\Sigma) = -\mathcal{D}(\Sigma\|I)$, where $\mathcal{D}$ denotes the beta-divergence, [34, 35], with parameter $\beta = k^{-1}+1$. The two limits can be derived by using the following result, [32, Proposition 3.1]: Given $X \succ 0$, we have

$$\lim_{c\to 0} c^{-1}(X^c - I) = \log(X). \tag{44}$$

□

It is worth noting that $\mathbb{H}_\infty(\Sigma)$ is different from the von Neumann entropy $\mathbb{H}(\Sigma)$ because of the term $\mathrm{tr}(\Sigma) - n$. On the other hand, in problem (14) the elements in the main diagonal of Σ are fixed, accordingly the term $\mathrm{tr}(\Sigma) - n$ plays no role in the optimization problem. We conclude that $\mathbb{H}_\infty$ and $\mathbb{H}$ are equivalent for our purposes. The same observation applies to $\mathbb{H}_1(\Sigma)$ in problem (2).

Let $\hat{\Sigma} \succ 0$ be the sample covariance matrix of $x^k(T)$ defined in (10) where we have dropped the subscript k in order to ease the notation. Then the corresponding covariance selection problem is

$$\begin{aligned} \Sigma^\circ = &\operatorname*{argmax}_{\Sigma\succeq 0} \mathbb{H}_k(\Sigma) \\ &\text{subject to } (\Sigma)_{ij} = (\hat{\Sigma})_{ij} \ \forall\, (i,j) \in \Omega. \end{aligned} \tag{45}$$

It is not difficult to see that the dual problem of (45) is

$$\begin{aligned} &\min_{S\in\mathcal{Q}_n} \frac{k}{k-1}\,\mathrm{tr}((I+k^{-1}S)^{1-k}) + k\,\mathrm{tr}((I+k^{-1}S)\hat{\Sigma}) \\ &\text{subject to } I + k^{-1}S \succ 0. \end{aligned} \tag{46}$$

Theorem 4.2. Problem (46) and problem (45) admit a unique solution. In particular, $\Sigma^\circ = (I + k^{-1}S)^{-k}$ where S° is such that $(S^\circ)_{ij} = 0$ for any $(i,j) \notin \Omega$.

P r o o f . The proof follows the same lines of the one of Theorem 3.1. □

Problem (45) selects a covariance matrix Σ° wherein the conditionally dependent pairs of variables in x^k at time T/k, assuming that $x^k(0)$ has independent components, are defined by Ω.

To estimate Ω from $\hat{\Sigma}$, we consider the following regularized version of problem (46):

$$\begin{aligned} &\min_{S\in\mathcal{Q}_n} \frac{k}{k-1}\,\mathrm{tr}((I+k^{-1}S)^{1-k}) + k\,\mathrm{tr}((I+k^{-1}S)\hat{\Sigma}) + \gamma h(S) \\ &\text{subject to } I + k^{-1}S \succ 0 \end{aligned} \tag{47}$$

where $h(S)$ has been defined in (28).

Theorem 4.3. Problem (47) admits a unique solution.

P r o o f . The proof is similar to the one of Theorem 3.4. □

Finally, the graphical model selection procedure is similar to the one in Section 3.3. It is worth noting these discrete-time versions can be useful in the case that the sampling time, i. e. k/T, of the underlying model is not known. In this case k is not known. Let Σ_k° denote the optimal solution given by the graphical model selection procedure using k. We solve the latter for different values of k, then the optimal k is the one minimizing ℓ_{AIC} or ℓ_{BIC}.

5. NUMERICAL EXPERIMENTS

The aim of this section is to show that if the underlying model is of type (5) then the correct parametrization of the model is the one induced by the notion of infinitesimal conditional independence. Accordingly, such a parametrization leads to an estimation algorithm outperforming the classical lasso approach in (4).

5.1. Case without input

In each study we consider 200 models of type (5) with $n = 10$ and for each them:

- S is randomly chosen in such a way that the percentage of nonnull entries is 20%; the value of each nonnull entry has been drawn from a Gaussian random variable with zero mean and variance 0.2; recall that $\mathcal{I}_S^c$ denotes the support of S;
- $N = 500$ independent realizations $\mathrm{x}(T,1)\dots\mathrm{x}(T,N)$ realizations of $x(T)$ are generated; from the latter we compute the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{x}(T,k)\mathrm{x}(T,k)^T;$$

- we apply the procedure of Section 3.3 denoted by $\mathbf{P}_\infty$. The regularization path is constituted by $M = 20$ points in such a way that γ_1 gives a non-sparse solution while γ_{20} gives a very sparse solution; finally we compute the quantity

$$e = \frac{n(n-1) - \mathrm{card}(\mathcal{I}_S^c \cap \hat{\Omega})}{n(n-1)}$$

  which represents the relative error in reconstructing the support of S;
- we apply the procedure of Section 4, denoted by $\mathbf{P}_k$, with $k \in \{1,2,3,5\}$ and we compute the previous relative error.

The optimization problems have been solved by using the CVX package [18]; however, the implementation of such algorithms could be improved by using the ideas in [11]. We have considered three Monte Carlo studies with different final times: $T = 1$, $T = 2$ and $T = 3$. In Figure 1 the boxplots of the error of the estimate $\hat{\Omega}$ using ℓ_{AIC} in the three cases are depicted. It is evident the superiority of $\mathbf{P}_\infty$ in respect to the others in all the three cases. This numerical evidence can be explained as follows. Consider the general

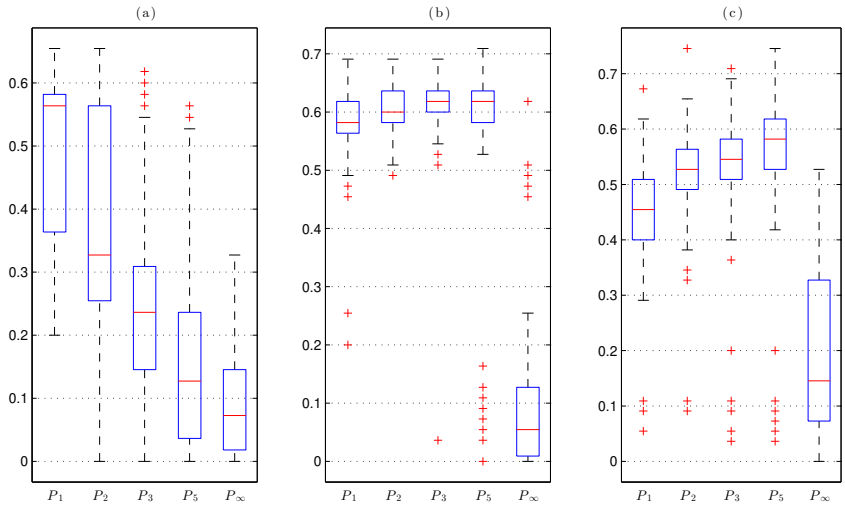

Fig. 1. Boxplots of the relative error in reconstructing the support of S using $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$, $\mathbf{P}_5$ and $\mathbf{P}_\infty$ with ℓ_{AIC} and $T = 1$ (a), $T = 2$ (b), $T = 3$ (c).

case, i. e. when T is not necessarily equal to one. These system identification procedures hinge on the regularized problem

$$\begin{cases} \min\limits_{I+TS\succ 0} -\log\det(I+TS) + T\,\mathrm{tr}(S\hat{\Sigma}) + \gamma Th(S), & k=1 \\ \min\limits_{I+k^{-1}TS\succ 0} \frac{k}{k-1}\,\mathrm{tr}((I+k^{-1}TS)^{1-k}) + k\,\mathrm{tr}((I+k^{-1}TS)\hat{\Sigma}) + \gamma Th(S), & 1<k<\infty \\ \min\limits_{S} \mathrm{tr}(e^{-TS}) + T\,\mathrm{tr}(S\hat{\Sigma}) + \gamma Th(S), & k=\infty \end{cases} \tag{48}$$

and the corresponding regularized covariance matrix has the structure

$$\Sigma_k^R = \begin{cases} (I+k^{-1}TS)^{-k}, & 1\le k<\infty \\ e^{-TS}, & k=\infty. \end{cases} \tag{49}$$

where we made explicit the dependence upon k. These problems have an objective function which is composed by a fit term (e. g. for $k=\infty$ it is $\mathrm{tr}(e^{-TS})+T\,\mathrm{tr}(S\hat{\Sigma})$) and a complexity term (i. e. $\gamma Th(S)$). Data have been generated by the model in (5), i. e. $\hat{\Sigma}\approx e^{-TS}$ where S is the weighted adjacency matrix of the true model and it is sparse. For $k=1$ the fit term would select a matrix $\hat{S}_{fit}$ such that $\hat{\Sigma}\approx(I+T\hat{S}_{fit})^{-1}$. Such a $\hat{S}_{fit}$ is not close to be sparse in general. Thus, the corresponding complexity term $\gamma Th(S_{fit})$ takes a large value in general. Although the solution $\hat{S}$ minimizing (48) with

$k = 1$ finds a compromise between the fit and the complexity term, it may happen that: i) $(I + T\hat{S})^{-1}$ is not so close to $\hat{\Sigma}$; ii) $\hat{S}$ is sparse, but its sparsity pattern is different than S. On the contrary, the solution with $k = \infty$ finds a good compromise with the two terms in the objective function because it considers the correct model parametrization (i. e. in terms of infinitesimal conditional dependence relations). A similar argument holds for the case $1 < k < \infty$. In Figure 2

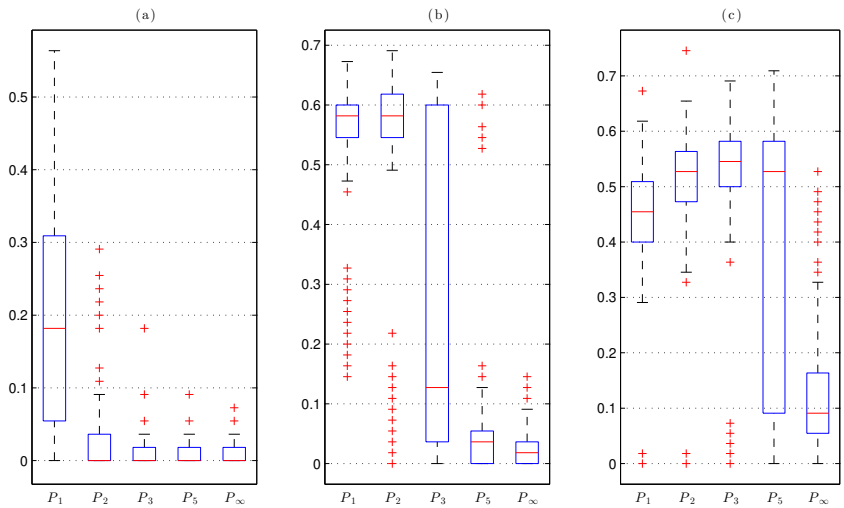

Fig. 2. Boxplots of the relative error in reconstructing the support of S using $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$, $\mathbf{P}_5$ and $\mathbf{P}_\infty$ with ℓ_{BIC} and $T = 1$ (a), $T = 2$ (b), $T = 3$ (c).

the boxplots of the error of the estimate $\hat{\Omega}$ using ℓ_{BIC} in the three cases are depicted. The performance of all the estimators is better than the one using AIC. In the case $T = 1$ $\mathbf{P}_3$, $\mathbf{P}_5$ and $\mathbf{P}_\infty$ performs in a similar way, in particular they outperforms $\mathbf{P}_1$ $\mathbf{P}_2$. In the case $T = 2$ $\mathbf{P}_3$ starts to perform poorly. In the case $T = 3$ also $\mathbf{P}_5$ starts to perform poorly. We conclude that the superiority of $\mathbf{P}_\infty$ is more evident in the case that the final time is sufficiently large. In other words, the value of k for which it is possible to see the performance improvement is proportional to T. This fact can be explained as follows. Notice that:

$$(\Sigma_k^R)^{-1} = \begin{cases} (I + k^{-1}TS)^k, & 1 \leq k < \infty \\ e^{TS}, & k = \infty. \end{cases} \tag{50}$$

It is not difficult to see that:

$$(\Sigma_k^R)^{-1} = \begin{cases} \sum_{l=0}^{k} \begin{pmatrix} k \\ l \end{pmatrix} \frac{1}{k^l} T^l S^l, & 1 \leq k < \infty \\ \sum_{l=0}^{\infty} \frac{1}{l!} T^l S^l, & k = \infty \end{cases} \tag{51}$$

where

$$\begin{pmatrix} k \\ l \end{pmatrix} = \frac{k!}{l!(k-l)!} \tag{52}$$

is the binomial coefficient. In particular, notice that

$$\lim_{k\to\infty} \begin{pmatrix} k \\ l \end{pmatrix} \frac{1}{k^l} = \frac{1}{l!} \lim_{k\to\infty} \frac{k(k-1)\dots(k-l+1)}{k^l} = \frac{1}{l!}. \tag{53}$$

In view of (51) and (53), $(\Sigma_k^R)^{-1}$ represents an approximation of the kth order expansion of $(\Sigma_\infty^R)^{-1}$. Accordingly, the smaller T is, the more $(\Sigma_k^R)^{-1}$ will be similar to the one with $k = \infty$.

5.2. Case with input

We consider a Monte Carlo study composed by 100 models of type (38) with $n = 10$, $T = 3$ and for each them:

- S is generated as before; B is equal to the identity matrix and $u(t)$ is extracted from an i.i.d. Gaussian process with zero mean and covariance matrix equal to the identity;
- $N = 500$ independent realizations $\mathrm{x}(T,1)\dots\mathrm{x}(T,N)$ realizations of $x(T)$ are generated;
- we apply the procedure of Section 3.4 denoted by $\mathbf{P}_\infty$ with $\varepsilon = 10^{-3}$. The regularization path is designed as before; then, we compute the relative error e in reconstructing the support of S;

The right panel of Figure 3 shows the boxplot of the error of the estimate $\hat{\Omega}$ using ℓ_{BIC}. The performance is comparable with the one without input (see the left panel of Figure 3). Thus, the proposed procedure is effective also in this scenario. In particular, the exit condition in the iterative part (Step 11) of Algorithm 2 has been always satisfied.

6. CONCLUSIONS

In this paper, the graphical model selection problem for a continuous-time process has been addressed: we have introduced the concept of infinitesimal conditional independence, the covariance selection problem, the topology selection problem as well as the graphical model selection problem. Moreover, we have shown the connection between the classic graphical model selection problem and the one for the continuous-time case. Finally, we have tested the effectiveness of the proposed procedure through some examples.

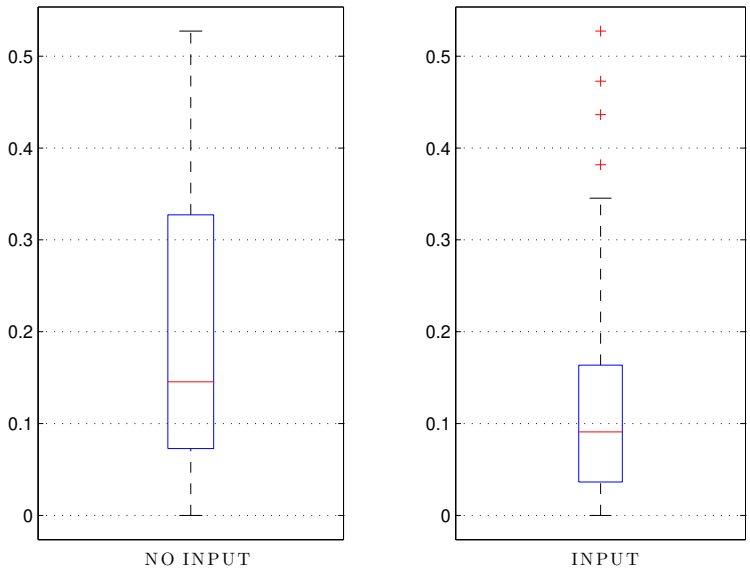

Fig. 3. Boxplots of the relative error in reconstructing the support of S in the case with (right panel) and without (left panel) deterministic input using $\mathbf{P}_\infty$ with ℓ_{BIC} and $T = 3$. Notice that the left panel corresponds to $\mathbf{P}_\infty$ in Figure 2(c).

ACKNOWLEDGEMENTS

Part of this work has been supported by the University of Padova through the Project BIRD162411/16 "Statistical learning methods for estimating the effective connectivity of human brain networks" and by the MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

(Received March 8, 2019)

REFERENCES

[1] D. Alpago, M. Zorzi, and A. Ferrante: Identification of sparse reciprocal graphical models. IEEE Control Systems Lett. *2* (2018), 4, 659–664. DOI:10.1109/lcsys.2018.2845943

[2] E. Avventi, A. Lindquist, and B. Wahlberg: ARMA identification of graphical models. IEEE Trans. Automat. Control *58* (2013), 1167–1178. DOI:10.1109/tac.2012.2231551

[3] G. Baggio: Further results on the convergence of the Pavon-Ferrante algorithm for spectral estimation. IEEE Trans. Automat- Control *63* (2018), 10, 3510–3515. DOI:10.1109/tac.2018.2794407

[4] O. Banerjee, L. El Ghaoui, and A. d'Aspremont: Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Machine Learning Res. *9* (2008), 485–516.

[5] S. Boyd and L. Vandenberghe: Convex Optimization. Cambridge Univ. Press, Cambridge 2004. DOI:10.1017/cbo9780511804441

[6] C. Byrnes, S. Gusev, and A. Lindquist: A convex optimization approach to the rational covariance extension problem. SIAM J. Optim. *37* (1998), 211–229. DOI:10.1137/s0363012997321553

[7] C. I. Byrnes, T. T. Georgiou, and A. Lindquist: A new approach to spectral estimation: A tunable high-resolution spectral estimator. IEEE Trans. Signal Process. *48* (2000), 3189–3205. DOI:10.1109/78.875475

[8] E. Candes and Y. Plan: Matrix completion with noise. Proc. IEEE *98* (2010), 925–936. DOI:10.1109/jproc.2009.2035722

[9] E. Candes and B. Recht: Exact matrix completion via convex optimization. Comm. ACM *55* (2012), 111–119. DOI:10.1145/2184319.2184343

[10] V. Chandrasekaran, P. Parrilo, and A. Willsky: Latent variable graphical model selection via convex optimization. Ann. Statist. *40* (2010), 1935–2013. DOI:10.1214/12-aos1020

[11] V. Chandrasekaran and P. Shah: Relative entropy optimization and its applications. Math. Program. *161* (2017), (1–2), 1–32. DOI:10.1007/s10107-016-0998-2

[12] T. Cover and J. Thomas: Information Theory. Wiley, New York 1991. DOI:10.1002/0471200611

[13] A. d'Aspremont, O. Banerjee, and L. El Ghaoui: First-order methods for sparse covariance selection. SIAM J. Matrix Analysis Appl. *30* (2008), 56–66. DOI:10.1137/060670985

[14] A. Dempster: Covariance selection. Biometrics *28* (1972), 157–175. DOI:10.2307/2528966

[15] A. Ferrante and M. Pavon: Matrix completion à la Dempster by the principle of parsimony. IEEE Trans. Inform. Theory *57* (2011), 3925–3931. DOI:10.1109/tit.2011.2143970

[16] A. Ferrante, M. Pavon, and F. Ramponi: Hellinger versus Kullback-Leibler multivariable spectrum approximation. IEEE Trans. Autom. Control *53* (2008), 954–967. DOI:10.1109/tac.2008.920238

[17] J. Friedman, T. Hastie, and R. Tibshirani: Sparse inverse covariance estimation with the graphical lasso. Biostatistics *9* (2008), 432–441. DOI:10.1093/biostatistics/kxm045

[18] M. Grant and S. Boyd: CVX: Matlab software for disciplined convex programming, version 2.1. 2014.

[19] S. Gu, R. Betzel, M. Mattar, M. Cieslak, P. Delio, S. Grafton, F. Pasqualetti, and D. Bassett: Optimal trajectories of brain state transitions. NeuroImage *148* (2017), 305–317. DOI:10.1016/j.neuroimage.2017.01.003

[20] J. Huang, N. Liu, M. Pourahmadi, and L. Liu: Covariance matrix selection and estimation via penalised normal likelihood. Biometrika *93* (2006), 85–98. DOI:10.1093/biomet/93.1.85

[21] N. Huotari, L. Raitamaa, H. Helakari, J. Kananen, V. Raatikainen, A. Rasila, T. Tuovinen, J. Kantola, V. Borchardt, V. Kiviniemi, and V. Korhonen: Sampling rate effects on resting state fMRI metrics. Frontiers Neurosci. *13* (2019), 279. DOI:10.3389/fnins.2019.00279

[22] A. Jalali and S. Sanghavi: Learning the dependence graph of time series with latent factors. In: International Conference on Machine Learning Edinburgh 2012.

[23] D. Koller and N. Friedman: Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

[24] S. Lauritzen: Graphical Models. Oxford University Press, Oxford 1996.

[25] N. Meinshausen and P. Bühlmann: High-dimensional graphs and variable selection with the lasso. Annals Statist. *34* (2006), 1436–1462. DOI:10.1214/009053606000000281

[26] J. Pearl: Graphical models for probabilistic and causal reasoning. In: Quantified representation of uncertainty and imprecision, Springer 1998, pp. 367–389. DOI:10.1007/978-94-017-1735-9_12

[27] A. Ringh, J. Karlsson, and A. Lindquist: Multidimensional rational covariance extension with approximate covariance matching. SIAM J. Control Optim. *56* (2018), 2, 913–944. DOI:10.1137/17m1127922

[28] J. Songsiri, J. Dahl, and L. Vandenberghe: Graphical models of autoregressive processes. In: Convex Optimization in Signal Processing and Communications (D. Palomar and Y. Eldar, eds.), Cambridge Univ. Press, Cambridge 2010, pp. 1–29.

[29] J. Songsiri and L. Vandenberghe: Topology selection in graphical models of autoregressive processes. J. Machine Learning Res. *11* (2010), 2671–2705.

[30] Z. Yue, J. Thunberg, L. Ljung, and J. Gonçalves: Identification of sparse continuous-time linear systems with low sampling rate: Exploring matrix logarithms. arXiv preprint arXiv:1605.08590, 2016.

[31] B. Zhu and G. Baggio: On the existence of a solution to a spectral estimation problem a la Byrnes-Georgiou-Lindquist. IEEE Trans. Automat. Control *64* (2019), 2, 820–825. DOI:10.1109/tac.2018.2836984

[32] M. Zorzi: A new family of high-resolution multivariate spectral estimators. IEEE Trans. Automat. Control *59* (2014), 892–904. DOI:10.1109/tac.2013.2293218

[33] M. Zorzi: Rational approximations of spectral densities based on the Alpha divergence. Math. Control Signals Systems *26* (2014), 259–278. DOI:10.1007/s00498-013-0118-2

[34] M. Zorzi: An interpretation of the dual problem of the THREE-like approaches. Automatica *62* (2015), 87–92. DOI:10.1016/j.automatica.2015.09.023

[35] M. Zorzi: Multivariate Spectral Estimation based on the concept of Optimal Prediction. IEEE Trans. Automat. Control *60* (2015), 1647–1652. DOI:10.1109/tac.2014.2359713

[36] M. Zorzi: Empirical Bayesian learning in AR graphical models. Automatica *109* (2019), 108516. DOI:10.1016/j.automatica.2019.108516

[37] M. Zorzi and R. Sepulchre: AR identification of latent-variable graphical models. IEEE Trans. Automat. Control *61* (2016), 2327–2340. DOI:10.1109/tac.2015.2491678

Mattia Zorzi, Dipartimento di Ingegneria dell'Informazione, Università degli studi di Padova, via Gradenigo 6/B, 35131 Padova. Italy.

e-mail: zorzimat@dei.unipd.it