Journal of Machine Learning Research (2017) Submitted 03/16; Revised 11/16; Published ??/??

Reconstructing Undirected Graphs from Eigenspaces

Yohann De Castro YOHANN.DECASTRO@MATH.U-PSUD.FR
Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, F-91405 Orsay, France

Thibault Espinasse ESPINASSE@MATH.UNIV-LYON1.FR
Institut Camille Jordan (CNRS UMR 5208), Université Claude Bernard Lyon 1, F-69622 Villeurbanne, France

Paul Rochet PAUL.ROCHET@UNIV-NANTES.FR
Laboratoire de Mathématiques Jean Leray (CNRS UMR 6629), Université de Nantes, F-44322 Nantes, France

Editor: Amir Globerson

Abstract

We aim at recovering the weighted adjacency matrix W of an undirected graph from a perturbed version of its eigenspaces. This situation arises for instance when working with stationary signals on graphs or Markov chains observed at random times. Our approach relies on minimizing a cost function based on the Frobenius norm of the commutator $\mathsf{AB}-\mathsf{BA}$ between symmetric matrices A and B. We describe a particular framework in which we have access to an estimation of the eigenspaces and provide support selection procedures from theoretical and practical points of view. In the Erdős-Rényi model on N vertices with no self-loops, we show that identifiability (*i.e.*, the ability to reconstruct W from the knowledge of its eigenspaces) follows a sharp phase transition on the expected number of edges with threshold function $N \log N/2$. Simulated and real life numerical experiments assert our methodology.

Keywords: Support recovery; Identifiability; Stationary signal processing; Graphs; Backward selection algorithm

1. Presentation

Networks have become a natural and popular way to model interactions in applications such as information technology (Rossi and Latouche, 2013), social life (Jiang et al., 2013; Matias et al., 2015), genetics (Giraud et al., 2012) or ecology (Thomas et al., 2015; Miele and Matias, 2017). In this paper, we investigate the reconstruction of an undirected weighted graph of size N from incomplete information on its set of edges (for instance, one knows that the target graph has no self-loops) and an estimation of the eigenspaces of its adjacency matrix W. This situation depicts any model where one knows in advance a linear operator K that commutes with W.

For instance, several authors (Espinasse et al., 2014; Girault, 2015; Perraudin and Vandergheynst, 2016; Marques et al., 2016) have introduced a definition of stationarity for signal processing on graphs. In the Gaussian framework, they have shown that this definition implies that the covariance operator K is jointly diagonalizable with the Laplacian (Perraudin and Vandergheynst, 2016) or some weighted symmetric adjacency matrix W supported on the graph (Espinasse et al., 2014; Marques et al., 2016).

Another framework adapted to our methodology concerns time-varying Markov processes, which are used to model numerous phenomena such as chemical reactions (Anderson and Kurtz, 2011) or waiting lines in queuing theory (Gaver Jr, 1959), see also Pittenger (1982); MacRae (1977); Barsotti et al. (2014). In some cases, one may observe at random times a Markov chain with transition matrix P. The transition matrix Q of the resulting Markov chain can be shown

to be a function of P. Thus, the transitions on the original process can be recovered from an estimation of Q given that P and Q commute. Several models are presented in Section 3 while the general model is given in Section 2.1.

Section 2.2 is concerned with identifiability issues, *i.e.*, the capacity to solve such problems. We exhibit sufficient and necessary conditions on the ability to reconstruct an undirected graph with no self-loops from the knowledge of the eigenspaces of W. These conditions allow us to derive a sharp phase transition on identifiability in the Erdős-Rényi model.

In Section 4.1, we introduce and theoretically assert new estimation schemes based on the Frobenius norm of the commutator $\mathsf{AB} - \mathsf{BA}$ between symmetric matrices A and B. More precisely, we assume that we have access to an estimation $\widehat{\mathsf{K}}$ of K and we consider the empirical contrast given by the commutator, namely $\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|$, where $\|\cdot\|$ denotes the Frobenius norm. Using backward-type procedures based on this empirical contrast, we build in Section 4 an estimator of the graph structure, *i.e.*, its set of edges $S^\star$ referred to as the support. Numerical experiments on simulated data (Section 5) and actual data (Section 6) assess the performances of our new estimation method. A discussion and related questions are presented in Section 7.

Related topics encompass spectral, least-squares and moment methods for graph reconstruction (Verzelen et al., 2015; Guédon and Vershynin, 2015; Klopp et al., 2017; Bubeck et al., 2016), Graphical Models (Verzelen, 2008; Giraud et al., 2012), or Vectorial AutoRegressive process (Hyvärinen et al., 2010). In the specific cases of Ornstein-Uhlenbeck processes and non-linear diffusions, the interesting papers Bento et al. (2010) and Bento and Ibrahimi (2014) tackle a related problem which is to estimate W along a trajectory, see Section 3.6 for further details. Note that the framework of the present paper addresses processes observed at random times—with possibly unknown distribution—which are not covered by Bento et al. (2010) and Bento and Ibrahimi (2014).

2. Model and Identifiability

2.1 The Model

Consider a symmetric matrix $\mathsf{W} \in \mathbb{R}^{N\times N}$, viewed as the weighted adjacency matrix of an undirected graph with N vertices. We investigate the eigenspaces of W in a situation where we have no direct information on the spectrum of the graph. Depicting this situation, we assume that the information on the target W stems from an unknown transformation $\mathsf{K} = f(\mathsf{W}) \in \mathbb{R}^{N\times N}$ or, in more realistic scenarios, from a perturbed version $\widehat{\mathsf{K}}$ of K. Precisely, let $f : x \mapsto \sum_{n=0}^{\infty} a_n x^n$ be an injective function, analytical on the spectrum of W, the matrix K is given by

$$\mathsf{K} = f(\mathsf{W}) := \sum_{n=0}^{\infty} a_n \mathsf{W}^n. \tag{1}$$

In this setting, the transformation $\mathsf{K} = f(\mathsf{W})$ preserves the eigenspaces. In particular, W and K commute, *i.e.*, $\mathsf{WK} = \mathsf{KW}$, since they share the same eigenspaces.

Our goal is to reconstruct W from a perturbed observation of its eigenspaces, provided by an estimator $\widehat{\mathsf{K}}$ of K. The key point is then to use extra information given by the location of some zero entries of W. Hence, we assume that one knows in advance a set $F \subset [1, N]^2$ of "forbidden" entries such that

$$\forall (i, j) \in F, \quad \mathsf{W}_{ij} = 0 \tag{$\mathbf{H}_\mathrm{F}$}$$

Equivalently, the set F is disjoint from the set of edges of the target graph. Throughout this paper, a special interest is given to the case $F = F_{\mathrm{diag}} := \{(i, i) \ : \ 1 \leq i \leq N\}$ conveying that there are no self-loops in W.

2.2 Identifiability

For $S \subseteq [1, N]^2$, denote by $\mathcal{E}(S)$ the set of symmetric matrices A whose support is included in S, which we write $\mathrm{Supp}(\mathsf{A}) \subseteq S$. Given the set F of forbidden entries defined via ($\mathbf{H}_{\mathrm{F}}$), the matrix of interest W is sought in the set $\mathcal{E}(\overline{F})$ with $\overline{F}$ the complement of F. In some cases, typically for F sufficiently large, most matrices $\mathsf{W} \in \mathcal{E}(\overline{F})$ are uniquely determined by their eigenspaces. For those $\mathsf{W} \in \mathcal{E}(\overline{F})$, there is no matrix $\mathsf{A} \in \mathcal{E}(\overline{F})$ non collinear with W that commutes with W. This property is encapsulated by the notion of *F-identifiability* as follows.

Definition 1 (F-identifiability) *We say that a symmetric matrix W is F-identifiable if, and only if, the only solutions A with $\mathrm{Supp}(\mathsf{A}) \subseteq \overline{F}$ to $\mathsf{AW} = \mathsf{WA}$ are of the form $\mathsf{A} = t\mathsf{W}$ for some $t \in \mathbb{R}$. Equivalently,*

$$\left\{\mathsf{A} \in \mathbb{R}^{N\times N} \ : \ \mathsf{A} = \mathsf{A}^\top,\ \mathsf{AW} = \mathsf{WA} \text{ and } \mathrm{Supp}(\mathsf{A}) \subseteq \overline{F}\right\} = \left\{t\mathsf{W} \ : \ t \in \mathbb{R}\right\} \tag{2}$$

A matrix W is F-identifiable if the set of symmetric matrices with the same eigenvectors as W and whose support is included in $\overline{F}$ is the line spanned by W.

Remark 2 *The dimension of the commutant, defined by*

$$\mathrm{Com}(\mathsf{W}) := \left\{\mathsf{A} \in \mathbb{R}^{N\times N} \ : \ \mathsf{A} = \mathsf{A}^\top,\ \mathsf{AW} = \mathsf{WA}\right\},$$

is entirely determined by the multiplicity of the eigenvalues of W. Indeed, letting $\lambda_1, \dots, \lambda_s$ denote the different eigenvalues of W and $\ell_1, \dots, \ell_s$ their multiplicities, one can show that

$$N \leq \dim\left(\mathrm{Com}(\mathsf{W})\right) = \sum_{j=1}^{s} \frac{\ell_j(\ell_j+1)}{2} \leq \frac{N(N+1)}{2}\,.$$

Now, the F-identifiability of W can be stated equivalently as $\dim\left(\mathrm{Com}(\mathsf{W}) \cap \mathcal{E}(\overline{F})\right) = 1$, observing that the left hand side of (2) is exactly $\mathrm{Com}(\mathsf{W}) \cap \mathcal{E}(\overline{F})$. Using a simple inclusion/exclusion formula, one can check that the condition

$$|F| \geq \dim\left(\mathrm{Com}(\mathsf{W})\right) - 1$$

is necessary for the F-identifiability, where $|F|$ denotes the cardinality of F. In particular, a matrix W with repeated eigenvalues requires a large set F of forbidden entries to be F-identifiable.

Proposition 3 (Lemma 2.1 in Barsotti et al. (2014)) *Let $S \subseteq \overline{F}$, the set of F-identifiable matrices in $\mathcal{E}(S)$ is either empty or a dense open subset of $\mathcal{E}(S)$.*

This proposition conveys that the F-identifiability of a matrix W is essentially a condition on its support S. The proof uses the fact that non F-identifiable matrices in $\mathcal{E}(S)$ can be expressed as the zeros of a particular analytic function, we refer to Barsotti et al. (2014) for further details. By abuse of notation, we say that a support $S \subseteq \overline{F}$ is *F-identifiable* if almost every matrix in $\mathcal{E}(S)$ is F-identifiable.

Characterizing the F-identifiability appears to be a challenging issue since it can be viewed as understanding the eigen-structure of graphs through their common support. The special case of the diagonal F_{diag} as the set of forbidden entries turns out to be particularly interesting. Indeed, the F_{diag}-identifiability, or diagonal identifiability, can be reasonably assumed in many practical situations since it entails that W lives on a simple graph, with no self-loops. In Theorem 16 (see Appendix A.1), we introduce necessary and sufficient conditions on the target support $\mathrm{Supp}(\mathsf{W})$ for diagonal identifiability. Defining the *kite* graph ∇_N of size $N \geq 3$ as the graph (V, E) with vertices $V = [1, N]$ and edges $E = \{(k, k+1),\ 1 \leq k \leq N-1\} \cup \{(N-2, N)\}$ (see Figure 1), one simple sufficient condition on diagonal identifiability reads as follows—a proof in given in Section A.2.

Proposition 4 *If the graph $G = ([1, N], S)$ contains the kite graph ∇_N as a subgraph, then S is diagonally identifiable.*

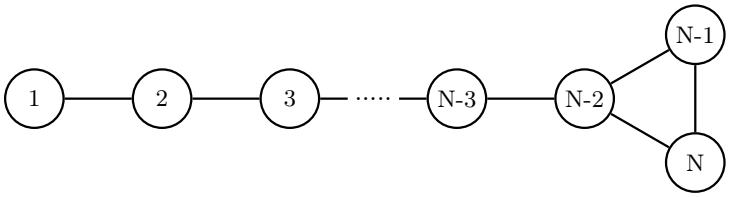

Figure 1: The kite graph ∇_N on N vertices.

Denote $G(N, p)$ the Erdős-Rényi model on graphs of size N where the edges are drawn independently with respect to the Bernoulli law of parameter p. Using Theorem 16, one can prove that $\log N/N$ is a sharp threshold for diagonal identifiability in the Erdős-Rényi model (see Section A.4). This can be stated as follows.

Theorem 5 *Diagonal identifiability in the Erdős-Rényi model occurs with a sharp phase transition with threshold function $\log N/N$: for any $\varepsilon > 0$, it holds*

- *If $p_N \geq (1+\varepsilon)\log N/N$ and $G_N \sim G(N, p_N)$ then the probability that $\mathrm{Supp}(G_N)$ is diagonally identifiable tends to 1 as N goes to infinity.*
- *If $p_N \leq (1-\varepsilon)\log N/N$ and $G_N \sim G(N, p_N)$ then the probability that $\mathrm{Supp}(G_N)$ is diagonally identifiable tends to 0 as N goes to infinity.*

In practice, one may expect that any target graph of size N with no self-loops and degree bounded from below by $\log N$ is diagonally identifiable. In this case, it might be recovered from its eigenspaces. Conversely, small degree graphs (*i.e.*, graphs with some vertices of degree much smaller than $\log N$) may not be identifiable. In this case, there is no hope to reconstruct it from its eigenspaces since there exists another small degree undirected weighted graph with the same eigenspaces.

3. Some Concrete Models

3.1 Markov chains

We begin with an example treated in the companion papers Barsotti et al. (2014, 2016). Consider a Markov chain $(X_n)_{n\in\mathbb{N}}$ with finite state space $[1, N]$ and transition matrix $\mathsf{P} \in \mathbb{R}^{N\times N}$. Let $(T_k)_{k\geq 1}$ be a sequence of random times such that the time gaps $\tau_k := T_{k+1} - T_k$ are i.i.d random variables independent of $(X_n)_{n\in\mathbb{N}}$. One can show that the sequence $Y_k = X_{T_k}$ is also a Markov chain with transition matrix $\mathsf{Q} = \mathbb{E}[\mathsf{P}^{\tau_1}] =: f(\mathsf{P})$ where f is the generating function of τ_k. Indeed, this follows from noticing that

$$\begin{aligned}
\mathbb{P}[Y_{k+1} = j | Y_k = i] &= \mathbb{P}[X_{T_{k+1}} = j | X_{T_k} = i] \\
&= \sum_{t\geq 0} \mathbb{P}[X_{T_k+t} = j, \tau_k = t | X_{T_k} = i] \\
&= \sum_{t\geq 0} \mathbb{P}[X_t = j | X_0 = i]\mathbb{P}[\tau_k = t] \\
&= \sum_{t\geq 0} \mathbb{P}[\tau_1 = t](\mathsf{P}^t)_{ij}.
\end{aligned}$$

Under regularity conditions, $\mathsf{Q} = f(\mathsf{P})$ can be estimated from the Y_k's and one may recover P from Q without any information on the distribution of the time gaps τ_k.

3.2 Vectorial AutoRegressive process

Consider a stationary Vectorial AutoRegressive process of order one $(X_n)_{n\in\mathbb{Z}}$ verifying

$$X_{n+1} = \mathsf{W}X_n + \varepsilon_n\,,$$

with ε_i i.i.d. centered random variables. Define as above $Y_k = X_{T_k}$ where again T_k are random times such that the time gaps $\tau_k = T_{k+1} - T_k$ are i.i.d. with generating function f and independent of $(X_n)_{n\in\mathbb{Z}}$. Then, it holds

$$\mathbb{E}[Y_{k+1}|Y_k] = \mathbb{E}[\mathbb{E}[Y_{k+1}|Y_k,\tau_k]|Y_k] = \sum_{j=0}^{\infty} \mathsf{W}^j Y_k \mathbb{P}(\tau_k = j) = f(\mathsf{W})Y_k,$$

which allows us to estimate $\mathsf{K} = f(\mathsf{W})$ and ultimately recover W.

3.3 Ornstein-Uhlenbeck process

The same property holds for the continuous time version of this process, namely a vectorial Ornstein-Uhlenbeck process observed at random times verifying

$$\mathrm{d}X_t = \mathsf{W}X_t\mathrm{d}t + \mathrm{d}B_t.$$

In this case, one can check that the random process $Y_k := X_{T_k}$ where the T_k's are random times with i.i.d. gaps $\tau_k = T_{k+1} - T_k$ satisfies

$$\mathbb{E}[Y_{k+1}|Y_k] = f(\mathsf{W})Y_k\,,$$

for f the Laplace transform of τ_1, that is, $f(\mathsf{W}) = \mathbb{E}[\exp(-\tau_1\mathsf{W})]$. Indeed, note that

$$\forall t,u\in\mathbb{R},\quad \mathbb{E}[X_{t+u}|X_u] = \exp(-t\mathsf{W})X_u$$

so that

$$\mathbb{E}[Y_{k+1}|Y_k] = \mathbb{E}[\mathbb{E}[X_{T_k+\tau_k}|X_{T_k},\tau_k]|X_{T_k}] = \mathbb{E}[\exp(-\tau_k\mathsf{W})X_{T_k}|X_{T_k}] = \mathbb{E}[\exp(-\tau_k\mathsf{W})]Y_k\,,$$

by independence of τ_k and Y_{k-1}.

3.4 Seasonal VAR structure

Consider a seasonal VAR structure: let T be a positive integer, $(u_k)_{k\in\mathbb{Z}}, (v_k)_{k\in\mathbb{Z}}$ periodic sequences of period T and

$$\forall k\in\mathbb{Z}\,,\quad Y_{k+1} = u_kY_k + v_k\mathsf{W}Y_k + \varepsilon_k\,,$$

where ε_k are independent and centered random variables. We may observe the model only at time gap intervals T with some error, *i.e.*, $X_t = Y_{tT+k_0} + \eta_t$ with η_t centered and independent random variables. This falls into the general frame

$$\mathbb{E}[X_t|X_{t-1}] = f(\mathsf{W})X_{t-1}\quad\text{where}\quad f(x) := \prod_{k=1}^{T}(u_k - v_kx)\,.$$

In this case, $\mathsf{K} = f(\mathsf{W})$ can be estimated from the observations.

3.5 Gaussian Graphical models

Our model is related to Gaussian Graphical models for which an overview can be found in the thesis Verzelen (2008). The reader may also consult the pioneering paper Friedman et al. (2008). One may consider the target W as the precision matrix, *i.e.*, the inverse of the covariance matrix K, having some non zero entries described by a graph of dependencies. Using $f(x) = x^{-1}$, this falls into our setting, trying to recover the "dependency" graph given by the precision matrix W from the estimation of the covariance matrix K. Of course, in this case, it is better to use the knowledge of f, which certainly improves estimation. Nevertheless, our procedure allows us to estimate the function f and heuristically validate the hypothesis $f(x) = x^{-1}$.

3.6 Spatial AutoRegressive Gaussian fields

Note that Gaussian AutoRegressive processes on $\mathbb{Z}$ verify that the precision operator may be written as a polynomial of the adjacency operator of $\mathbb{Z}$. One natural way to extend this property (see for instance Espinasse et al. (2014)) is to define centered Gaussian AutoRegressive fields on a graph through the same relation between the covariance operator K and the adjacency operator W (or the discrete Laplacian, depending on the framework) : $\mathsf{K}^{-1} = P(\mathsf{W})$, with P a polynomial of degree p. In this framework, Graphical models methods will infer the graph of path of length p, whereas our method aims at recovering W. Note that this framework extends to ARMA spatial fields where K writes as a rational fraction of W, and the property of commutativity between W and K still holds.

In the previous cases, we assumed that we can not estimate directly W. For spatio-temporal processes, this means that we do not have access to a full trajectory. It may be the case when the sample is drawn at random times, or when we sample with respect to the stationary measure of the process—for instance when observation times are a lot larger than the typical evolution time's scale of the process. If the whole trajectory is available, it is better to use this extra information, see for instance Bento et al. (2010) for the Ornstein-Uhlenbeck case and Bento and Ibrahimi (2014) for the non-linear diffusion case.

4. Estimating the Support

4.1 Empirical Contrast: the Commutator

The methodology presented in the paper relies on the fact that the target matrix W commutes with the matrix K, in view of $\mathsf{K} := f(\mathsf{W})$, as defined in Eq. (1). Because W is symmetric, it has real eigenvalues $\lambda_1, ..., \lambda_N$ (here listed with repetitions, if any) and is diagonalizable in an orthogonal basis. That is, letting Λ denote the diagonal matrix with diagonal entries λ_k, there exists an orthogonal matrix U such that $\mathsf{W} = \mathsf{U}\Lambda\mathsf{U}^\top$. With this notation, one verifies easily from Eq. (1) that $\mathsf{K} = \mathsf{U}\mathsf{D}\mathsf{U}^\top$, where $\mathsf{D} := f(\Lambda)$ is the diagonal matrix with diagonal entries $f(\lambda_i), i = 1, \dots, N$. Since f is assumed one-to-one on the spectrum of W, the matrices W and K share the same eigenspaces associated to λ_k and $f(\lambda_k)$ respectively :

$$\{v \in \mathbb{R}^N : \mathsf{W}v = \lambda_k v\} = \{v \in \mathbb{R}^N : \mathsf{K}v = f(\lambda_k)v\},$$

for all λ_k. Moreover, the dimension of each eigenspace is equal to the multiplicity of the corresponding eigenvalue λ_k in the spectrum of W. Additionally, when F-identifiability holds, the only solutions A with $\mathrm{Supp}(\mathsf{A}) \subseteq \overline{F}$ to $\mathsf{A}\mathsf{K} = \mathsf{K}\mathsf{A}$ are of the form $\mathsf{A} = t\mathsf{W}$ for some $t \in \mathbb{R}$.

Remark 6 (Matrix perturbation theory) *In practice, we do not observe* K *but an estimation* $\widehat{\mathsf{K}}$ *(which we assume symmetric) with perturbed eigen-decomposition* $\widehat{\mathsf{K}} = \widehat{\mathsf{U}}\widehat{\mathsf{D}}\widehat{\mathsf{U}}^\top$. *Nevertheless, by continuity of the eigen-decomposition, we know that* $\widehat{\mathsf{K}}$ *being close to* K *implies*

simultaneously the proximity of their eigenvalues and eigenspaces (see for instance Mirsky's inequality (Stewart and Sun, 1990, Corollary 4.12) and Wedin's $\sin(\theta)$ *theorem (Stewart and Sun, 1990, P. 260) for the details). Thus, if we consider* A *such that* $\mathsf{A}\widehat{\mathsf{K}} = \widehat{\mathsf{K}}\mathsf{A}$*, then* A *has the same eigenspaces as* $\widehat{\mathsf{K}}$ *which in turns, are expected to be close to the eigenspaces of* W*.*

Given an estimator $\widehat{\mathsf{K}}$ of K, remark that W verifies

$$\frac{\|\mathsf{W}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{W}\|}{\|\mathsf{W}\|} = \frac{\|\mathsf{W}(\widehat{\mathsf{K}} - \mathsf{K}) - (\widehat{\mathsf{K}} - \mathsf{K})\mathsf{W}\|}{\|\mathsf{W}\|} \leq 2\|\widehat{\mathsf{K}} - \mathsf{K}\|\,. \tag{3}$$

Hence, in view of (3) and the discussion above, we aim to estimate W by minimizing the following cost function

$$\mathsf{A} \mapsto \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|}{\|\mathsf{A}\|}, \quad \mathsf{A} \in \mathcal{E}(\overline{F}) \setminus \{0\}.$$

This empirical criterion was first used in Barsotti et al. (2014), in a Markov Chain context, to reflect that W is expected to nearly commute with $\widehat{\mathsf{K}}$.

4.2 The ℓ_0-approach

Given an estimator $\widehat{\mathsf{K}}$ of K build from a sample of size n and a set of forbidden entries F reflecting ($\mathbf{H}_{\mathrm{F}}$), we construct an estimator $\widehat{S}$ of the target support $S^\star := \mathrm{Supp}(\mathsf{W})$ as a minimizer of the criterion Q given by

$$\forall S \subseteq \overline{F}, \quad Q(S) := \min_{\mathsf{A} \in \mathcal{E}(S)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|}{\|\mathsf{A}\|} + \lambda_n |S|,$$

for some tuning parameter $\lambda_n > 0$ and defining the minimum of an empty set as ∞. Recall that $\mathcal{E}(S)$ is the set of symmetric matrices A such that $\mathrm{Supp}(\mathsf{A}) \subseteq S$. Our estimator of the true support $S^\star$ is defined as

$$\widehat{S} \in \arg\min_{S \subseteq \overline{F}} Q(S)$$

Furthermore, we assume that the estimator $\widehat{\mathsf{K}}$ converges toward K in probability with

$$\forall t > 0\,, \quad \mathbb{P}\big\{\|\widehat{\mathsf{K}} - \mathsf{K}\| \geq t\big\} \leq R_n(t), \tag{$\mathbf{H}_2$}$$

where $R_n, n \in \mathbb{N}$ is a sequence of non-increasing functions that converge pointwise toward 0 as n goes to ∞.

Theorem 7 *Assume that* ($\mathbf{H}_2$) *and* ($\mathbf{H}_{\mathrm{F}}$) *hold. If* W *is* F*-identifiable, then*

$$\mathbb{P}\big\{\widehat{S} \neq S^\star\big\} \leq R_n\Big(\frac{c_0(S^\star) - \lambda_n |S^\star|}{4}\Big) + R_n\Big(\frac{\lambda_n}{2}\Big)\,,$$

where

$$c_0(S^\star) := \min_{\substack{S \neq S^\star \\ |S| \leq |S^\star|}} \min_{\mathsf{A} \in \mathcal{E}(S)} \frac{\|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\|}{\|\mathsf{A}\|} > 0\,. \tag{4}$$

A proof of Theorem 7 is given in Section B.1.

Corollary 8 *Under the assumptions of Theorem 7, if*

$$\lambda_n \to 0 \quad \text{and} \quad \sum_{n \in \mathbb{N}} R_n\Big(\frac{\lambda_n}{2}\Big) < +\infty\,,$$

then $\widehat{S} \to S^\star$ *almost surely.*

Note that, based on the upper bound in Theorem 7, a good scaling may be $\lambda_n^\star = \frac{c_0(S^\star)}{|S^\star|+4}$ which interestingly does not depend on n. This leads to the upper bound

$$\mathbb{P}\{\widehat{S} \neq S^\star\} \leq 2R_n\Big(\frac{c_0(S^\star)}{2|S^\star|+8}\Big) \xrightarrow{n\to\infty} 0$$

which is optimal up to a constant less than 2. This oracle choice $\lambda_n^\star$ is of course irrelevant in practice since both $c_0(S^\star)$ and $|S^\star|$ are unknown. Alternatively, we may choose a sequence λ_n decreasing slowly to 0 to ensure both conditions of Corollary 8.

4.3 Edge significance based on the commutator criterion

The exponential complexity of the ℓ_0-approach making it generally infeasible in practice, a backward methodology provides a computationally feasible alternative to the support reconstruction problem. Starting from the maximal acceptable support $\overline{F}$, the idea of the backward procedure is to remove the least significant entries one at a time and stop when every entry is significant. Using the corresponding small case letter to denote the vectorization of a matrix, *e.g.*, $a = \mathrm{vec}(\mathsf{A}) = (\mathsf{A}_{11}, ..., \mathsf{A}_{N1}, ..., \mathsf{A}_{1N}, ..., \mathsf{A}_{NN})^\top$, significancy can be leveraged using the Frobenius norm of the commutator operator $a \mapsto \Delta(\mathsf{K})a = \mathrm{vec}(\mathsf{KA} - \mathsf{AK})$, where

$$\Delta(\mathsf{K}) = \mathrm{I} \otimes \mathsf{K} - \mathsf{K} \otimes \mathrm{I} \quad \in \mathbb{R}^{N^2 \times N^2}$$

and $\otimes$ denotes the Kronecker product. Indeed, searching for the target W in the commutant of K reduces to searching for $w = \mathrm{vec}(\mathsf{W})$ in $\ker(\Delta(\mathsf{K}))$, the kernel of $\Delta(\mathsf{K})$. Because the Frobenius norm coincides with the Euclidean norm of the vectorization, the functions $\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|^2$ and $a \mapsto \|\Delta(\widehat{\mathsf{K}})a\|^2$ can be used indistinctly as cost functions.

Assumptions

Assume the three following hypotheses ($\mathbf{H}_\Sigma$), ($\mathbf{H}_1$) and ($\mathbf{H}_{\mathrm{Id}}$).

∘ Deriving the asymptotic law of least-squares estimators, we may assume that the estimate $\widehat{\mathsf{K}}$ is such that

$$\sqrt{n}(\widehat{k} - k) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma), \tag{$\mathbf{H}_\Sigma$}$$

where Σ is a $N^2 \times N^2$ covariance matrix (either known or that can be estimated). For instance, one can think of $\widehat{\mathsf{K}}$ as the empirical covariance when observing a sample of vectors of covariance K. This condition is verified for instance in the framework considered in Barsotti et al. (2014, 2016). Note that asymptotic normality is a standard ground base investigating any least-squares procedure.

∘ In order to exclude the trivial solution $a = 0$, the target W is assumed normalized

$$\mathbf{1}^\top w = 1, \tag{$\mathbf{H}_1$}$$

where $\mathbf{1}$ has all its entries equal to one. Because the available information on W is of spectral nature and as such, is scale-invariant, a normalization of some kind is crucial for the reconstruction. Here, the condition $\mathbf{1}^\top w = 1$ achieves two goals: preventing the null matrix form being a solution and making the problem identifiable.

Remark 9 *The main drawback of this normalization concerns the situation where the entries of* W *sum up to zero, in which case the normalization is impossible. If the context suggests that the solution may be such that* $\mathbf{1}^\top w = 0$*, a different affine normalization* $\mathbf{v}^\top w = 1$ *(with any fixed vector* $\mathbf{v}$*) must be used, without major changes in the methodology. In practice, one may consider the vector* $\mathbf{v}$ *at random (for instance with isotropic law), so that* ($\mathbf{H}_1$) *is almost surely fulfilled for any fixed target* w*. Finally, observe that if* W *has non-negative entries, then the normalization* ($\mathbf{H}_1$) *is always feasible.*

◦ For S a support included in $\overline{F}$, we aim at a solution in the affine space

$$\mathcal{A}_S := \{a = \mathrm{vec}(\mathsf{A}) : \mathrm{Supp}(\mathsf{A}) \subseteq S,\ \mathsf{A} = \mathsf{A}^\top\ ,\ \mathbf{1}^\top a = 1\}.$$

with linear difference space given by

$$\mathcal{L}_S := \{a = \mathrm{vec}(\mathsf{A}) : \mathrm{Supp}(\mathsf{A}) \subseteq S,\ \mathsf{A} = \mathsf{A}^\top\ ,\ \mathbf{1}^\top a = 0\}\,.$$

By abuse of notation, $\mathcal{A}_S$ may refer both to the space of matrices or their vectorizations. To find the target support $S^\star$, one must exploit the fact that the vector w lies in the intersection of $\ker(\Delta(\mathsf{K}))$ and $\mathcal{A}_{\overline{F}}$. Actually, w can be recovered if the intersection is reduced to the singleton $\{w\}$. In this case, the matrix W and its support $S^\star$ are F-identifiable. Hence, we assume that

$$\ker(\Delta(\mathsf{K})) \cap \mathcal{L}_{\overline{F}} = \{0\}, \tag{$\mathbf{H}_{\mathrm{Id}}$}$$

which is implied by F-identifiability, see Definition 1.

Asymptotic normality and significance test

The framework under consideration can be viewed as a heteroscedastic linear regression model with noisy design for which $w = \mathrm{vec}(\mathsf{W})$ is the parameter of interest. Indeed, consider for each support $S \subseteq \overline{F}$ a full-ranked matrix $\Phi_S \in \mathbb{R}^{N^2 \times \dim(\mathcal{A}_S)}$ whose column vectors form a basis of $\mathcal{L}_S$. Assuming that W is F-identifiable and taking $S \subseteq \overline{F}$, the operator $\Delta(\mathsf{K})\Phi_S$ is one-to-one. In this case, evaluating the commutator $a \mapsto \Delta(\mathsf{K})a$ over $\mathcal{A}_S$ reduces to considering the map

$$b \mapsto \Delta(\mathsf{K})(a_0 - \Phi_S b)\,, \quad b \in \mathbb{R}^{\dim(\mathcal{A}_S)},$$

with a_0 chosen arbitrarily in $\mathcal{A}_S$. When replacing the unknown $\Delta(\mathsf{K})$ with its estimate $\Delta(\widehat{\mathsf{K}})$, the minimization of the criterion $a \mapsto \|\Delta(\widehat{\mathsf{K}})a\|^2$ over $\mathcal{A}_S$ can be written similarly as a linear regression framework where the parameter of interest is estimated by

$$\widehat{\beta}_S \in \arg\min_{b \in \mathbb{R}^{\dim(\mathcal{A}_S)}} \|\Delta(\widehat{\mathsf{K}})(a_0 - \Phi_S b)\|^2. \tag{5}$$

We recognize a linear model with response $y = \Delta(\widehat{\mathsf{K}})a_0$ and noisy design matrix $X = \Delta(\widehat{\mathsf{K}})\Phi_S$. In this setting, remark that $w = a_0 - \Phi_S\beta$ with β the unique solution to $\Delta(\mathsf{K})(a_0 - \Phi_S\beta) = 0$. Denoting by $\mathsf{M}^\dagger$ the pseudo-inverse of a matrix M, we deduce the following result.

Theorem 10 *If $S^\star \subseteq S$, the estimator $\widehat{\beta}_S$ is asymptotically Gaussian with*

$$\sqrt{n}(\widehat{\beta}_S - \beta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Omega_S),$$

where $\Omega_S := \big(\Phi_S^\top \Delta(\mathsf{K})\big)^\dagger \Delta(\mathsf{W})\Sigma\Delta(\mathsf{W})\big(\Delta(\mathsf{K})\Phi_S\big)^\dagger$.

We then have

$$\widehat{w}_S = \mathrm{vec}(\widehat{\mathsf{W}}_S) = \arg\min_{a \in \mathcal{A}_S} \|\Delta(\widehat{\mathsf{K}})a\|^2 = a_0 - \Phi_S\widehat{\beta}_S\,. \tag{6}$$

The asymptotic distribution of $\widehat{w}_S$ follows directly from Theorem 10,

$$\sqrt{n}(\widehat{w}_s - w) \xrightarrow[n\to\infty]{d} \mathcal{N}\big(0, \Phi_S\Omega_S\Phi_S^\top\big). \tag{7}$$

The limit covariance matrix is unknown, but plugging the estimates $\widehat{\mathsf{W}}_S$, $\widehat{\mathsf{K}}$ and $\widehat{\Sigma}$ yields an estimator $\Phi_S\widehat{\Omega}_S\Phi_S^\top$, which is consistent under the F-identifiability assumption. In particular,

the diagonal entry of $\Phi_S \widehat{\Omega}_S \Phi_S^\top$ associated to the (i,j)-entry of W, which we denote $\widehat{\sigma}^2_{S,ij}$, provides a consistent estimator for the asymptotic variance of $\widehat{\mathsf{W}}_{S,ij}$. As a result, the statistic

$$\tau_{ij}(S) := \sqrt{n}\frac{\widehat{\mathsf{W}}_{S,ij}}{\widehat{\sigma}_{S,ij}} \tag{8}$$

can be used to measure the relative significance of the estimated entry $\widehat{\mathsf{W}}_{S,ij}$. The backward support selection procedure is then implemented by the recursive algorithm as follows.

Algorithm 1: Backward algorithm for support selection

Data: A set of forbidden entries F, a matrix $\widehat{\mathsf{K}}$.
Result: A sequence of estimators $\widehat{\mathsf{W}}_{S_1}, \widehat{\mathsf{W}}_{S_2}, ...$ with nested supports $S_1 \supset S_2 \supset ....$

1: Start with the maximal acceptable support $S_1 = \overline{F}$,
2: At each step k, compute the statistics $\tau_{ij}(S_k)$ for all $(i,j) \in S_k$,
3: Remove the least significant edge (i,j) which minimizes $|\tau_{ij}(S_k)|$ for $(i,j) \in S_k$, and set $S_{k+1} = S_k \setminus \{(i,j),(j,i)\}$,
4: Stop when all edges have been removed.

The backward algorithm produces a sequence of nested supports that one can choose to stop once all the edges are judged significant, that is, when all the statistics $\tau_{ij}(S_k)$, $(i,j) \in S_k$ exceed in absolute value some fixed threshold τ_0. Owing to the asymptotic normality of $\widehat{\mathsf{W}}_{S,ij}$ shown in Eq. (7), the $(1-\frac{\alpha}{2})$-quantile of the standard Gaussian distribution would appear as a reasonable choice for the threshold τ_0, as it boils down to performing an asymptotic significance test of level α. However, due to the slow convergence to the limit distribution and the tendency to overestimate the variance for small sample sizes (see Figure 2), a threshold based on the Gaussian quantile inevitably leads to an overly large estimated support. Nevertheless, we show that an adaptive calibration of the threshold can be achieved by considering the overall behavior of the commutator $\Delta(\widehat{\mathsf{K}})\widehat{w}_{S_m}$ computed over the nested sequence of active supports.

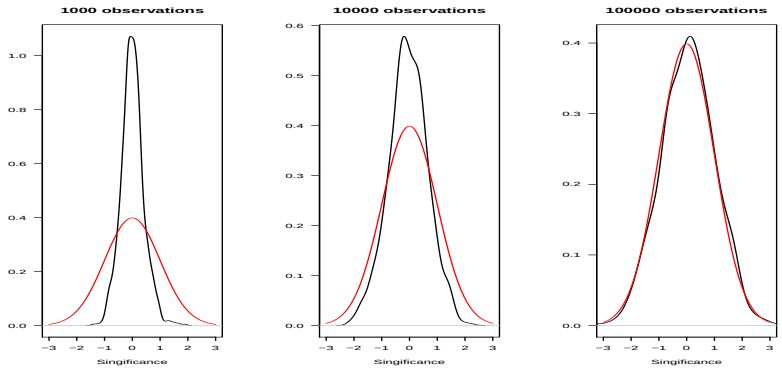

Figure 2: Estimated density of the statistic $\tau_{ij}(S)$ for an edge $(i,j) \in S \setminus S^\star$ compared to its theoretical Gaussian limit distribution, for samples of size $n = 1000$ (left), $n = 10000$ (center) and $n = 100000$ (right).

Calibration of the threshold by cross-validation

By removing the least significant edge at each step, the backward algorithm generates a sequence of nested active supports $S_1 \supset \cdots \supset S_\ell$, that we refer to as a "trajectory". Along this trajectory, we compute the empirical contrast defined by

$$\forall S \subseteq \overline{F}, \quad S \mapsto \mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}}) := \frac{\|\widehat{\mathsf{W}}_S\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\widehat{\mathsf{W}}_S\|}{\|\widehat{\mathsf{W}}_S\|}\,. \tag{9}$$

Note that computing this criterion boils down to compute $\widehat{\mathsf{W}}_S$ which is a simple projection onto $\mathcal{A}_S$ as shown in (6).

When the true support $S^\star$ lies in the trajectory, one expects to observe a "gap" in the sequence $j \mapsto \mathrm{Crit}(\widehat{\mathsf{W}}_{S_j}, \widehat{\mathsf{K}})$ when S_j goes from $S^\star$ to a smaller support. Indeed:

- For $S^\star \subseteq S$, the target W is consistently estimated by $\widehat{\mathsf{W}}_S$ so that $\mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}})$ tends to zero at rate $\sqrt{n}$,
- For $S \subsetneq S^\star$, the lower bound $\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\| \geq \|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\| - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|\|\mathsf{A}\|$ yields

$$\mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}}) = \frac{\|\widehat{\mathsf{W}}_S\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\widehat{\mathsf{W}}_S\|}{\|\widehat{\mathsf{W}}_S\|} \geq c(S) - 2\|\widehat{\mathsf{K}} - \mathsf{K}\| \tag{10}$$

  with $c(S) := \min_{A \in \mathcal{A}_S} \|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\|/\|\mathsf{A}\|$ a positive constant. In particular, one has

$$\min_{S \subsetneq S^\star} c(S) \geq \min_{\substack{S \neq S^\star \\ |S| \leq |S^\star|}} c(S) = c_0(S^\star) > 0$$

  where $c_0(S^\star)$ is defined in (4).

In some way, $c_0(S^\star)$ measures the amplitude of the signal: one expects to be able to recover the target W when the estimation error $\|\widehat{\mathsf{K}} - \mathsf{K}\|$ reaches at least the same order as $c_0(S^\star)$. The true support $S^\star$ then corresponds to a transitional gap in the contrast curve that can be captured by a suitably chosen threshold $t > 0$. Since $\widehat{\mathsf{K}}$ converges toward K in probability, any threshold $0 < t < c_0(S^\star)$ will work with probability one asymptotically.

Remark 11 *The condition that $S^\star$ lies in the trajectory of nested supports is crucial to detect the commutation gap, although seldom verified in practice due to the tremendous amount of testable supports. This issue is specifically targeted by the bagging version of the backward algorithm discussed in Section 4.4.*

An obstacle to the detection of the commutation gap is the increasing behavior of the commutator over the nested trajectory $S_1 \supset \cdots \supset S_\ell$. This phenomenon, indirectly caused by the dependence between the trajectory and $\widehat{\mathsf{K}}$, can be annihilated when considering the empirical contrast over a trajectory built from a training sample. In fact, the monotonicity can even be "reversed" before reaching the true support if the $\widehat{\mathsf{W}}_{S_j}$ are estimated independently from $\widehat{\mathsf{K}}$. This can be explained as follows. Consider the ideal scenario where estimators $\widetilde{\mathsf{W}}_{S_1}, ..., \widetilde{\mathsf{W}}_{S_\ell}$ are built from the backward algorithm using an estimator $\widetilde{\mathsf{K}}$ independent from $\widehat{\mathsf{K}}$. We assume moreover that the true support $S^\star$ lies in the trajectory $S_1 \supset ... \supset S_\ell$. The trick is to write

$$\Delta(\widehat{\mathsf{K}})\widetilde{w}_{s_j} = \Delta(\widetilde{\mathsf{K}})w + \Delta(\mathsf{K})\widetilde{w}_{S_j} + \Delta(\widehat{\mathsf{K}} - \mathsf{K})(\widetilde{w}_{S_j} - w),$$

and to analyze the three terms separately:

- The term $\Delta(\widehat{\mathsf{K}})w$ has no influence as it is common to all supports in the trajectory.

- The term $\Delta(\mathsf{K})\widetilde{w}_{S_j}$ approaches zero as $\widetilde{w}_{S_j}$ gets closer to w. Heuristically, the variance of $\widetilde{w}_{S_j}$, and incidentally that of $\Delta(\mathsf{K})\widetilde{w}_{S_j}$, is larger for over-fitting supports $S \supseteq S^\star$. This results in the sequence $j \mapsto \Delta(\mathsf{K})\widetilde{w}_{S_j}$ being stochastically decreasing as S_j approaches $S^\star$ from above. On the other hand, the bias of order $O(1)$ is expected to dominate once the trajectory passes through the true value $S^\star$, making the remaining of the sequence $\Delta(\mathsf{K})\widetilde{w}_{S_j}$ increase stochastically.

- The term $\Delta(\widehat{\mathsf{K}} - \mathsf{K})(\widetilde{w}_{S_j} - w)$ is negligible for $S \supseteq S^\star$, as both $\widehat{\mathsf{K}} - \mathsf{K}$ and $\widetilde{w}_{S_j} - w$ tend to zero independently. We emphasize that this argument no longer holds without the independence of $\widetilde{w}_{S_j}$ and $\widehat{\mathsf{K}}$. This is precisely why we use a training sample.

Thus, the sequence $j \mapsto \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_j}, \widehat{\mathsf{K}}) = \|\Delta(\widehat{\mathsf{K}})\widetilde{w}_{S_j}\|/\|\widetilde{w}_{S_j}\|$ is expected to achieve its minimum for the best estimator $\widetilde{w}_{S_j}$ in the trajectory, that is for $S_j = S^\star$. Furthermore, beyond the true support (for small active supports), $\widetilde{w}_{S_j}$ is not a consistent estimator of w so that the criterion no longer approaches zero, resulting in the so-called commutation gap.

The "reversed" monotonicity provides an easy way to calibrate the threshold in the backward algorithm. Indeed, since $S_j \mapsto \Delta(\widehat{\mathsf{K}})\widetilde{w}_{S_j}$ is expected to decrease when approaching the true support (coming from larger active supports along a trajectory), the estimated support can be heuristically chosen as the last time the criterion is below an adaptive threshold, see Figure 3. In particular, $\mathrm{Crit}(\widetilde{\mathsf{W}}_{S_1}, \widehat{\mathsf{K}})$ can be used as an adaptive threshold for the backward algorithm when the estimator $\widehat{\mathsf{K}}$ and the trajectory $S_1 \supset \cdots \supset S_\ell$ are obtained from independent samples.

Of course, to afford splitting the sample to build the $\widetilde{\mathsf{W}}_{S_j}$ independent from $\widehat{\mathsf{K}}$ may be unrealistic. Nevertheless, the numerical study suggests that the independence is well mimicked when $\widehat{\mathsf{K}}$ is built from the whole dataset but the backward algorithm sequence $\widetilde{\mathsf{W}}_{S_1}, ..., \widetilde{\mathsf{W}}_{S_\ell}$ is obtained from a learning sub-sample, as illustrated in Figure 3. Empirically, the optimal size of training samples could be calibrated in function of the number of observations using the robustness of the outputs of the algorithm. In this paper, we always draw training samples by taking each observation with probability 1/2, with no consideration regarding the size of the whole sample.

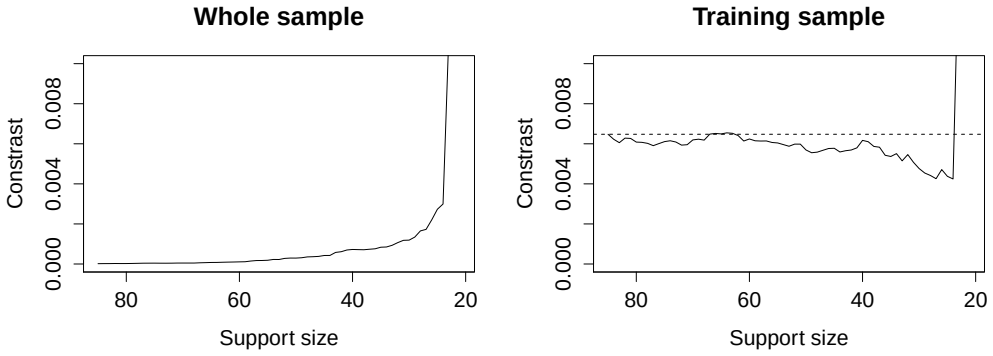

Figure 3: The contrast sequence $j \mapsto \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_j}, \widehat{\mathsf{K}})$ computed in the example of Section 5.2. The nested support sequence and estimators $\widetilde{\mathsf{W}}_{S_j}$ are obtained from the backward algorithm implemented on the whole sample (left) and on a training sample of half size (right). In both cases, $\widehat{\mathsf{K}}$ is constructed from the whole sample. Using a training sample manages to reverse the monotonicity in the first part of the sequence, thus making the commutation gap easier to locate. The initial value of the sequence $t = \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_1}, \widehat{\mathsf{K}})$ then provides a tractable adaptive choice for the threshold.

4.4 Improving the backward algorithm by bagging

The main weakness of the backward procedure remains that it requires the true support $S^\star$ to lie in the trajectory $S_1 \supset \cdots \supset S_\ell$ obtained from removing the least significant edge one at a time. In practice, this condition is rarely verified, especially with small datasets. A way to solve this issue is to replicate the backward algorithm over a collection of random sub-samples, a process commonly known to as Bootstrap Aggregating, or bagging. The description of this algorithm is given in Algorithm 2.

Algorithm 2: Bagging backward algorithm

Data: A set of forbidden entries F, a sample X.
Result: A collection of estimated supports $\widehat{S}_m, m = 1, ..., M$.

1: Build M bootstrapped samples without replacement.
2: For each sub-sample $m = 1, ..., M$, build an estimator $\widetilde{\mathsf{K}}_m$ of K.
3: For all m, run Algorithm 1 without stopping condition and return M trajectories $S_{1m} \supset \cdots \supset S_{\ell m}$ and the corresponding estimators $\widetilde{\mathsf{W}}_{S_{km}}$.
4: Evaluate the empirical contrast $\mathrm{Crit}(\widetilde{\mathsf{W}}_{S_{km}}, \widehat{\mathsf{K}})$ over each trajectory with the estimator $\widehat{\mathsf{K}}$ calculated from the whole sample.
5: For each trajectory, return the estimated support $\widehat{S}_m := S_{\hat{k}_m m}$ as the last support whose contrast lies below the initial value:

$$\hat{k}_m := \max\left\{k = 1, ..., \ell : \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_{km}}, \widehat{\mathsf{K}}) \leq \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_{1m}}, \widehat{\mathsf{K}})\right\}.$$

The bagging algorithm produces a collection of estimated supports in a way to make the final decision more robust. At this point, several solutions are possible: select the most represented support among the $\widehat{S}_m$'s, keep the edges present in the most supports etc... A preliminary detection of the outliers among the $\widehat{S}_m$'s, e.g. by removing beforehand the supports $\widehat{S}_m$'s that are either too big or too small, might also considerably improve the method, as we illustrate on actual examples in Section 5.

5. Numerical study

5.1 Toy example

In the previous section, we have introduced different algorithms. To emphasize the motivation of the bagging algorithm, we consider a simple example, and implement the different algorithms for support recovery. To check the performances of the ℓ_0 procedure, we need to consider a graph with a small number of vertices (since the ℓ_0 complexity grows with $2^{N(N-1)/2}$ where N denotes the number of vertices). Here, we consider the graph G_1 represented in Figure 4, the kite graph on 5 vertices.

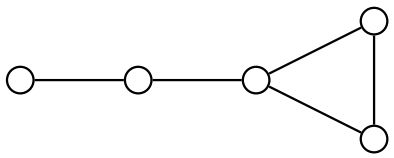

Figure 4: The kite graph $G_1 = \nabla_5$.

We choose W as the (normalized) adjacency matrix of G_1 then draw a sample of size $n = 500$ of centered Gaussian vectors $X_1, \cdots X_n$ of $\mathbb{R}^5$ with covariance matrix $\mathsf{K} = \exp(\mathsf{W})$. We assume

known that G_1 contains no self-loop so that we take $F = F_{diag}$ as the set of forbidden values. In this simple example, the constant $c_0(S^\star)$ (see Eq. (4)) can be calculated explicitly, yielding $c_0(S^\star) \approx 0.12$. In comparison, for $n = 500$, $\mathbb{E}\|\widehat{\mathsf{K}} - \mathsf{K}\|$ is evaluated to approximately 0.27 by Monte-Carlo. We expect to be able to recover the true support when the noise level drops below the signal amplitude. Based on the bound of Eq. (10), this occurs as soon as $\|\widehat{\mathsf{K}} - \mathsf{K}\| \leq c_0(S^\star)/2$. However, because this bound is not sharp, a lesser level of precision is required in practice.

We compare the following algorithms:

1. *Contrast penalized ℓ_0 minimization with optimal penalization constant.* We compute

$$\widehat{S} = \arg\min_{S \subseteq \overline{F}_{diag}} \left\{ \min_{\mathsf{A} \in \mathcal{E}(S)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|}{\|\mathsf{A}\|} + \lambda|S| \right\}.$$

   The constant λ is chosen as the best possible value, minimizing the oracle error $\delta(\widehat{S})$ measured by the symmetric difference between $\widehat{S}$ and $S^\star$, namely

$$\delta(\widehat{S}) := |\widehat{S} \cup S^\star \setminus \widehat{S} \cap S^\star| \,.$$

   Note that the calibration parameter λ is chosen optimally. Hence, the numerical performances of the method can be expected to be overestimated compared to a fully data-driven procedure.

2. *Thresholding contrast minimization with optimal threshold.* The target matrix W is estimated over the maximal acceptable support $\overline{F}_{diag}$. We then compute

$$\widehat{S} = \{(i,j) : |\widehat{\mathsf{W}}_{ij}| > t\},$$

   where the threshold t is chosen so as to minimize the oracle error $\delta(\widehat{S})$. Here again, the performances are expected to be overestimated compared to a data-driven threshold.

3. *Backward algorithm.* We generate a training sample by taking each observation with probability $1/2$ independently, from the whole sample. The estimator of K in this subsample is denoted $\widetilde{\mathsf{K}}$. We implement Algorithm 1 on $\widetilde{\mathsf{K}}$, yielding a trajectory $S_1 \supset ... \supset S_\ell$ of nested supports whose sizes vary from $|S_1| = 20$ (the full off-diagonal support) to $|S_\ell| = 12$ (the minimal size required for diagonal identifiability), along with the associated estimators $\widetilde{\mathsf{W}}_{S_k}, k = 1, ..., \ell$. Remark that the supports are symmetric, hence two entries are removed at each step so that $\ell = 5$ in this case. We then compute the threshold $t = \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_1}, \widehat{\mathsf{K}})$ corresponding to the initial value of the contrast. The estimated support $\widehat{S}$ is defined as the smallest support S in the trajectory such that $\mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}}) \leq t$. Here, the choice of t is adaptive in a fully data driven manner.

4. *Bagging backward algorithm.* The previous algorithm is implemented over $M = 100$ training samples drawn keeping observations with probability $1/2$. For each $m = 1, ..., M$, we retain

   - the threshold $t_m = \mathrm{Crit}(\widehat{\mathsf{W}}_{S_{1m}}, \widehat{\mathsf{K}})$ corresponding to the initial value of the contrast,
   - the estimated support, that is, the smallest support $\widehat{S}_m$ in the trajectory such that $\mathrm{Crit}(\widehat{\mathsf{W}}_{\widehat{S}_m}, \widehat{\mathsf{K}}) \leq t_m$.

   The estimated support $\widehat{S}$ is obtained as follows. Only a proportion q of the training samples m with a small initial contrast t_m are kept, they are expected to provide better results—in the whole paper, we chose $q = 2/\sqrt{M}$ empirically. Then, the smallest support among the remaining candidates is retained, breaking ties arbitrarily.

Remark 12 *Using a ℓ_1-penalized contrast*

$$\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|^2 + \lambda\|\mathsf{A}\|_1$$

(*subject to a normalizing condition so as to rule out the trivial solution $\mathsf{A} = 0$*) *tends to overestimate the support. In fact, any conservative choice of λ will lead to false positives in the estimated support (typically, a full support matrix may commute with $\widehat{\mathsf{K}}$ while still having a small ℓ_1 norm). Hence, when aiming for support recovery, the typical solution is to vanish the small entries of the minimizer, making it no more efficient than the thresholded ℓ_2 procedure considered in Algorithm 2. For this reason, the numerical performances of the Lasso procedure are not included in the study.*

The next table compares the performances of the four algorithms. We calculated the Monte-Carlo estimated mean error $\mathbb{E}(\delta(\widehat{S}))$ and probability of exact recovery $\mathbb{P}\{\widehat{S} = S^\star\}$ for 1000 repetitions of the experiment. The average computational time (obtained with the function `timer` of Scilab) on a processor Intel Xeon @2.6 GHz are shown, using the oracle values of λ and t for the first two algorithms (the calibration of these parameters is thus not accounted for in the computation time).

Algorithm	ℓ_0	ℓ_2-thresholding	Backward	Bagging Backward
Mean Error	0.45	0.37	1.95	0.68
Exact recoverery	68%	75%	23%	61%
CPU time (s)	0.32	0.002	0.009	0.59

In this example, the first two algorithms are the more accurate. The percentage of successful recoveries for the bagging backward algorithm is nonetheless competitive given that the first two procedures have been calibrated optimally for each experiment, which would be highly infeasible in practice. Finally, we observe that although it is much more expensive computationally, the bagging version of the backward algorithm yields an undeniable improvement.

Upper bounds for the time and space complexity of the algorithms are given in the next table. The time complexity is calculated as the number of different supports S considered to lead to the solution in function of the size N of the graph and the number M of training samples. The spatial complexity measures the memory size needed to compute the solution. In this setting, it is the main limitation for applying the procedures to large graphs. The N^4 comes from the computation of $\Delta(\mathsf{K}) = \mathsf{K} \otimes \mathrm{I} - \mathrm{I} \otimes \mathsf{K}$ in the solver. Admittedly, the complexity could be improved by using sparse matrix encoding although this was not implemented.

Algorithm	ℓ_0	ℓ_2-thresholding	Backward	Bagging Backward
Space Complexity	$O(N^4)$	$O(N^4)$	$O(N^4)$	$O(N^4)$
Time Complexity	$O(2^{N(N-1)/2})$	$O(1)$	$O(N^2)$	$O(N^2.M)$

On the current version, the bagging backward algorithm contains scalability issues for big graphs due to its space complexity. Leads to reduce the spatial complexity include using sparse matrix encoding or the use of cheap approximations of the criterion. These shall be investigated in future works.

5.2 A diagonally identifiable matrix

The advantages of the bagging backward algorithm are highlighted for larger graphs. In the next example, we consider the graph G_2 on $N = 15$ vertices represented in Figure 5. The experimental conditions are similar to that of the previous example, a sample of size $n = 10000$ is drawn from

a centered Gaussian vector of variance $\mathsf{K} = \exp(\mathsf{W})$ where W is the normalized adjacency matrix of G_2, with normalizing constant chosen such that $\mathbf{1}^\top w = 1$. The implementation of the different algorithms follow the description of the previous example.

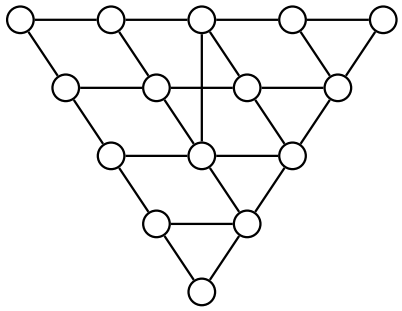

Figure 5: The graph G_2 is diagonally identifiable.

In this case, the number of possible supports is too large for the ℓ_0 method to be implementable while the accuracy of the thresholded ℓ_2 drops considerably compared to smaller cases. We summarize the results in the following table.

Algorithm	ℓ_2-thresholding	Backward	Bagging Backward
Mean Error	10	25	1
Exact recoverery	22%	26%	69%
CPU time (s)	0.04	2.5	256

A drawback of the bagging backward algorithm is the larger computational time: it takes around 4 minutes in average to estimate the support. Being essentially $M = 100$ repetitions of the backward algorithm, the numerical complexity of the bagging version is roughly M times that of the simple backward algorithm, although the improvement is, here again, clear.

To illustrate the influence of the unknown function f, we consider $f : t \mapsto (1-t)^{-2}$ and reproduce the numerical study for $\mathsf{K} = f(\mathsf{W})$. The results for various sample sizes are gathered in the next table, for $M = 100$ bagging runs.

n	10000	5000	2000	1000
Exact recovery	97%	87%	83%	13%
Mean error	0.05	0.33	0.9	8.5

The probability of recovering the true support appears to be greater than in the previous example (97% against 69% previously for $n = 10000$). This sheds lights on another important factor in the efficiency of the methods which is the separability of the spectrum of K. Indeed, in this framework, the information needed to recover W lies in its eigenspaces, which are estimated via $\widehat{\mathsf{K}}$. The accuracy of these estimates depends on the distance between the different eigenvalues (see e.g. Corollary 4.12 in Stewart and Sun (1990) and Wedin's $\sin(\theta)$ theorem in Stewart and Sun (1990)). Thus, for the spectrum $\lambda_1, ..., \lambda_N$ of W, the ability to recover W from $\mathsf{K} = f(\mathsf{W})$ essentially relies on how far the $f(\lambda_i)$'s are from each other. For the sake of comparison, the spectrum of W which lies in the interval $[-0.5, 0.5]$ is more "spread" by the function $t \mapsto (1-t)^{-2}$ than by the exponential, as we can see in Figure 6.

Remark 13 *We also implemented the procedure in a random setting where* W *is drawn from an Erdös-Rényi graph with binomial entries. The conclusions obtained in this case are similar to those already discussed and shall not be presented to avoid redundancy.*

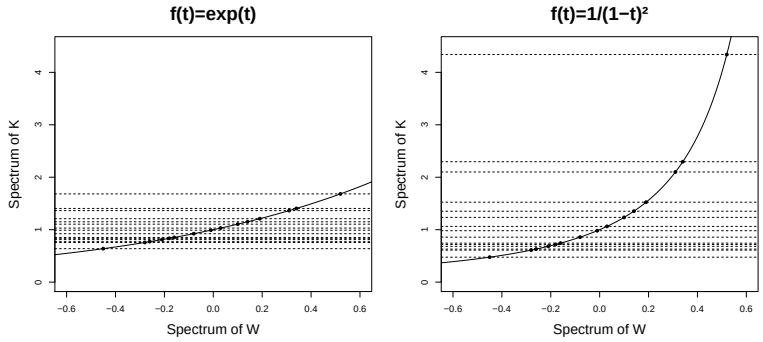

Figure 6: Separability of the spectrum of $\mathsf{K} = f(\mathsf{W})$ for $f : t \mapsto \exp(t)$ (left) and $f : t \mapsto (1-t)^{-2}$ (right). The eigenvalues of K are more separated in the second case, making it easier to approximate its eigenspaces from the estimator $\widehat{\mathsf{K}}$.

6. Real life application

We now implement the bagging backward algorithm on real life data provided by Météorage and Météo France. The data contain the daily number of lightnings during a 3 year period in 16 regions of France localized on a 4×4 grid. We expect to recover the spatial structure of the graph from the dependence of the lightning occurrences between the regions.

The data are refined as follows. We first eliminate the days without any lighting, leaving 950 vectors X_i of length 16 that contain the number of impacts at day i in each of the 16 regions. This numbers are highly non Gaussian, contain many zeros, and show a clear south-east/north-west tendency, with much more lightning in the south east. As a pre-processing, we apply the transformation $x \mapsto \log(1+x)$ to the data and subtract the spatial tendency estimated by linear regression. The process is then normalized in such manner that the conditional variance at each vertex conditionally to all the others is 1. This way, the precision matrix K^{-1}, with K the covariance matrix, has diagonal 1.

We model the resulting process as a spatial AutoRegressive process of order $p > 1$ on a 4×4 grid, as described in Section 3.6. In this setting, K^{-1} writes as a degree p polynomial of a matrix W supported on the 4×4 grid. Using only the information that W has zero diagonal, we aim to recover this dependency from the empirical estimator $\widehat{\mathsf{K}}$ of the covariance, using that K and W commute. Remark that, because the target graph is bipartite, the model is not diagonally identifiable—for instance W^3 also has zero diagonal. However, we still manage to recover the support from the fact that W is the *sparsest* matrix that commute with K. We run our algorithm 100 times and we show in Figure 7 the most frequent edges appearing in the output graph.

To the best of our knowledge, this exact framework has not been tackled in the literature. For this reason, it is difficult to compare the performances of our algorithm to other existing methods. For instance, while being a reference in Graphical Models, the package `GGMselect` (see Giraud et al. (2012)) fails to uncover the dependence structure in this case, as we see in Figure 8. The reason is simple: the package `GGMselect` aims to estimate the precision matrix, which is a polynomial of W. Thus, while GGM inference may be more stable, faster ($0.3s$ for `GGMselect` and $400s$ for our algorithm) and easily interpretable, it is only adapted to recover the dependence structure if $p = 1$, in which case K^{-1} is an affine function of W.

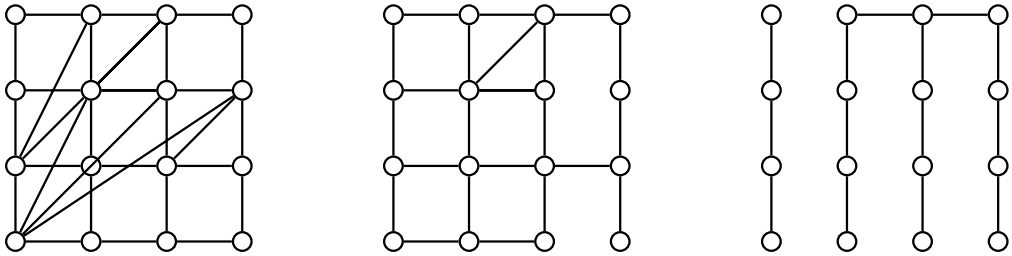

Figure 7: Edges that appear 30%, 50%, and 70% of the time when running the bagging backward algorithm 100 times. The spatial dependency becomes apparent around the 50% mark.

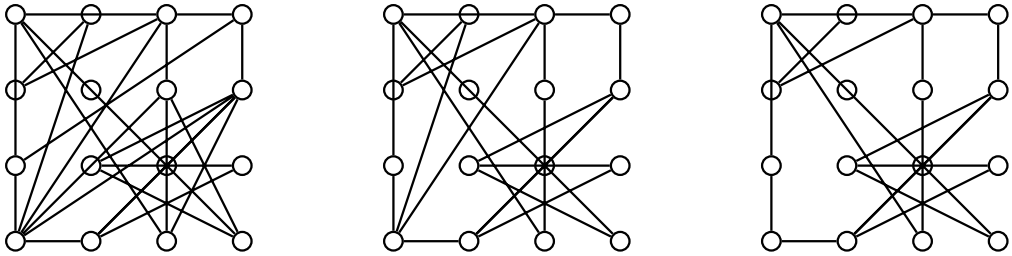

Figure 8: Edges that appear 30%, 50%, and 70% of the time using the `GGMselect` package with family `CO1` and maximal degree `dmax` $= 5$.

7. Discussion

In this paper, we develop a new method to recover hidden graphical structures in different models. We consider a general framework in which we have access to an approximation of the eigen-structure of an unknown graph, via an operator that commutes with its weighted adjacency matrix W. We are able to recover the support of W from an estimate of $\mathsf{K} = f(\mathsf{W})$ with f an unknown function, under the sole assumption that the location of some zeros of W are known. We tackle two situations where this condition may arise: Markov processes observed at random times and stationary signals on graphs. We focus on the particular case of a weighted adjacency matrix W with zero diagonal, indicating that the underlying graph has no self-loop.

A main limitation of our method lies in the large amount of data necessary to recover the true support with high probability. Arguably, this limitation is intrinsic to our model and is a matter of comparison between the estimation error $\|\widehat{\mathsf{K}} - \mathsf{K}\|$ and the signal strength, measured by the constant $c_0(S^\star)$ in (4) and (10). Even under the assumption that the estimation error has order $1/\sqrt{n}$, the difficulties stem from the constant $c_0(S^\star)$ being extremely small in some cases.

For practical issues, there remain three main challenges to be addressed. The first one concerns the symmetry of W, that once relaxed, would offer a wider range of applications. The second concerns the generalization and applications to identifiability conditions other than W having zero diagonal. Finally, our algorithm is greedy when the size of the graph increases. For large graphs, it remains to find a way to compute more efficiently the criterion and significance of the variables.

Acknowledgement

We would like to warmly thank Météo France and Météorage for providing us the data used in Section 6. We also thank Dieter Mitsche for fruitful discussions, as well as the Universidad de la Habana (Cuba) and the Centro de Modelamiento Matematico (Chile) for their hospitality.

Appendix A. Asserting the Diagonal Identifiability

A.1 Necessary and sufficient conditions

In this section, we focus on the F-identifiability in the special case where the set of forbidden entries is the diagonal $F_{\text{diag}} := \{(i,i) : i \in [1,N]\}$. Recall that a support S is F_{diag}-identifiable, or simply diagonally identifiable (DI), if for almost every matrix $\mathsf{A} \in \mathcal{E}(S)$,

$$\mathsf{BA} = \mathsf{AB}\ ,\ \operatorname{diag}(\mathsf{B}) = 0\ ,\ \mathsf{B} = \mathsf{B}^\top \implies \mathsf{B} = \lambda \mathsf{A}$$

for some $\lambda \in \mathbb{R}$. In other words, a support S is diagonally identifiable if almost every symmetric matrix A with support in S is uniquely determined, up to scaling, by its eigenspaces among symmetric matrices with zero diagonal. In this section, we provide both sufficient and necessary conditions on a support S to ensure the F_{diag}-identifiability. For this, we consider a simple undirected graph $G_S = ([1,N], S)$ on N vertices with edge set S.

Definition 14 (Induced subgraph) *For $V \subseteq [1,N]$, the induced subgraph $G_S(V) = (V, S(V))$ is the graph on V with edge set $S(V) = S \cap V^2$.*

Proposition 15 *For all support $S \subseteq [1,N]^2$, the set of invertible matrices in $\mathcal{E}(S)$ is either empty or a dense open subset of $\mathcal{E}(S)$.*

The proof is straightforward when writing the determinant of $\mathsf{A} \in \mathcal{E}(S)$ as a polynomial in its entries. Observe that by this property, finding one invertible matrix A in $\mathcal{E}(S)$ guarantees that almost every matrix in $\mathcal{E}(S)$ is invertible. In this case, we say that the graph G_S is invertible. Similarly, we say that G_S is diagonally identifiable if S is diagonally identifiable.

Theorem 16 (Conditions for F_{diag}-identifiability) *Let $S \subseteq \overline{F}_{\text{diag}}$ and $G_S = ([1,N], S)$.*

1. ***Necessary condition:*** *If S is diagonally identifiable then there exists a sequence of subsets $V_3, ..., V_{N-1} \subset [1,N]$ such that $|V_k| = k$ and $G_S(V_k)$ is invertible for all $k = 3, ..., N-1$.*
2. ***Sufficient condition:*** *If there exists a nested sequence $V_3 \subset ... \subset V_{N-1} \subset [1,N]$ with $|V_k| = k$ such that $G_S(V_k)$ is invertible for all $k = 3, ..., N-1$, then S is diagonally identifiable.*

The gap between the sufficient and necessary conditions lies in the fact that the sequence $V_3, ..., V_{N-1}$ need to be nested for the sufficient condition.

Proof We proceed by contradiction. For the necessary condition, let $k \geq 3$ be such that $G_S(V_k)$ is not invertible, for all $V_k \subset [1,N]$ of size k. For $\mathsf{A} \in \mathcal{E}(S)$, denote by $\psi_0(\mathsf{A}), \psi_1(\mathsf{A}), \ldots, \psi_N(\mathsf{A})$ the coefficients of the characteristic polynomial

$$\det(z\,\mathrm{I} - \mathsf{A}) = \sum_{j=0}^{N} \psi_j(\mathsf{A})\ z^j, \quad z \in \mathbb{R}.$$

Consider the matrix $M_k(\mathsf{A}) := \sum_{j=0}^{k} \psi_j(\mathsf{A})\ \mathsf{A}^j$. By Eq. (14) in Espinasse and Rochet (2016), we see that the (i,i)-entry of $M_k(\mathsf{A})$ equals the sum of all minors of size k that do not contain the

vertex i. Since for all subset V_k of size k, $G_S(V_k)$ is not invertible, this implies that $M_k(\mathsf{A})$ has zero diagonal. On the other hand, the non-zero entries of $M_k(\mathsf{A})$ are degree k polynomials in the variables $\mathsf{A}_{ij}, (i,j) \in \mathrm{Supp}(\mathsf{A})$. Therefore, the equality $M_k(\mathsf{A}) = \lambda \mathsf{A}$ for some $\lambda \in \mathbb{R}$ occurs for at most a countable number of $\mathsf{A} \in \mathcal{E}(S)$. Since $M_k(\mathsf{A})$ commutes with A, we deduce that S is not diagonally identifiable.

For the sufficient condition, we will need the following lemma.

Lemma 17 *If there exists a subset $V' \subset [1, N]$ of size $N-1$ such that $G_S(V')$ is both DI and invertible, then G_S is DI.*

Proof We may assume that $V' = [1, N-1]$ without loss of generality. Let M' denote a symmetric $(N-1)\times(N-1)$ matrix indexed on V' that is both invertible and diagonally identifiable, *i.e.*, for all non-zero matrix $\mathsf{A}' \neq \lambda \mathsf{M}'$,

$$\mathsf{M}'\mathsf{A}' = \mathsf{A}'\mathsf{M}' \implies \mathrm{diag}(\mathsf{A}') \neq 0.$$

To prove that G_S is DI, it suffices to find a symmetric matrix M with support S that is diagonally identifiable. Consider M defined by

$$\mathsf{M} = \begin{bmatrix} \mathsf{M}' & 0 \\ 0 & 0 \end{bmatrix}.$$

Let A be a matrix with zero diagonal that commutes with M and write

$$\mathsf{A} = \begin{bmatrix} \mathsf{A}' & a \\ a^\top & 0 \end{bmatrix}$$

for some $a \in \mathbb{R}^{N-1}$, with $\mathrm{diag}(\mathsf{A}') = 0$. The condition $\mathsf{MA} = \mathsf{AM}$ can be stated equivalently as

$$\begin{cases} \mathsf{M}'\mathsf{A}' = \mathsf{A}'\mathsf{M}' \\ \mathsf{M}'a = 0 \end{cases}$$

Since M' is invertible by assumption, $a = 0$ and the only matrix A with zero diagonal that commutes with M is the null matrix. Thus, M is diagonally identifiable. ■

We now go back to prove the sufficient condition in Theorem 16. Assume that G_S is not diagonally identifiable, then by Lemma 17, neither is $G_S(V_{N-1})$. By iterating the argument, we conclude that $G_S(V_3)$ is not diagonally identifiable. However, the only invertible graph on three vertices is the triangle graph, which is diagonally identifiable, leading to a contradiction. ■

Remark 18 *The proof of Theorem 16 combines the results of Lemma 2.1 in Barsotti et al. (2014) and Eq. (14) in Espinasse and Rochet (2016). The first one is of topological flavor proving that the set of identifiable matrices is either dense or empty in the set of matrices with prescribed support. The second ingredient is Eq. (14) in Espinasse and Rochet (2016) which contains a key combinatorial computation on the adjugate matrix of weighted graphs. Admittedly, its purpose was to provide the first step to prove the necessary condition for identifiability in the present article (precisely, that $M_k(\mathsf{A})$ has zero diagonal). The proof of the sufficient condition does not, however, involve this result.*

A.2 Proof of Proposition 4

From Claim (*ii*) in Theorem 16 and considering the nested sequence $V_{N-1} \supset ... \supset V_3$ obtained by removing the last vertex on the tail of the kite at each step, we deduce a simple and tractable sufficient condition for a graph G_S to be diagonally identifiable, namely that G_S contains the kite graph as a vertex covering (possibly not induced) subgraph.

A.3 Existence of kites

The condition on containing the kite graph ∇_N as a subgraph is mild in the sense that it is satisfied in the dense regime $\log n/n$ by random graphs, as depicted in the following proposition.

Proposition 19 *The existence of kite graphs in the Erdős-Rényi model occurs as follows. For any $\omega(N) \to \infty$ and for $G_N \sim G(N, p_N)$, if $p_N \geq (1/N)(\log N + \log\log N + \omega(N))$ then $\mathbb{P}\{G_N$ has a kite of length $N\}$ tends to 1 as N goes to infinity.*

The proof makes use of the existence of a hamiltonian cycle which is a standard result in Random Graph Theory, see Corollary 8.12 in Bollobás (1998) for instance. This results shows that in the regime $(\log N + \log\log N)/N$ an Erdős-Rényi graph is diagonally identifiable.

Proof We now present the proof of this fact. Let $\omega(n) \to \infty$ and set

$$p_1 := (1/n)(\log n + \log\log n + \omega(n)/2),$$
$$p_2 := \omega(n)/(2n)\,.$$

Let $G^{(1)}$ and $G^{(2)}$ be two independent Erdős-Rényi graphs such that

$$G_n^{(1)} \sim G(n, p_1) \quad \perp\!\!\!\perp \quad G_n^{(2)} \sim G(n, p_2)\,.$$

As shown in Corollary 8.12 in Bollobás (1998) for instance, $\mathbb{P}\{G_n^{(1)}$ is hamiltonian$\}$ tends to 1 as n goes to infinity. Given a hamiltonian cycle C_n of length n in $G^{(1)}$ one can construct a kite of length n using edges of $G^{(2)}$ to connect a pair of vertices at distance 2 on the cycle C_n. Invoke the independence of $G^{(1)}$ and $G^{(2)}$ to get that this latter probability is

$$\mathbb{P}\{\{k, k+2\} \text{ is an edge of } G^{(2)} \text{ for some } k\} = \mathbb{P}\{B(n, p_2) > 0\}\,,$$

where $B(n, p_2)$ denotes the binomial law. Using Poisson approximation one gets that this probability tends to 1 as n goes to infinity. We deduce that the probability that the graph $G = G_n^{(1)} + G_n^{(2)}$ has at least a kite tends to 1. Observe that G is an Erdős-Rényi graph of size n and parameter $p = p_1 + p_2 - p_1 p_2 \leq p_n$ which concludes the proof. ■

A.4 Proof of Theorem 5

Combining Proposition 19 and Theorem 16, we deduce the first point. From the necessary condition in Theorem 16, we see that it is sufficient to find two isolated vertices to prove non-identifiability. Indeed, in this case, the kernel of the adjacency matrix has co-dimension at least 2 showing that all sub-graphs of size $N-1$ are not invertible. Furthermore, one knows (see Theorem 3.1 in Bollobás (1998) for instance) that the event "there is at least two isolated points" has sharp threshold function $\log n/n$. This proves the second point.

Appendix B. Support reconstruction

B.1 Proof of Theorem 7

Define $\mathcal{S}_1 := \{S \in \mathcal{S} : |S| \leq |S^\star|, S \neq S^\star\}$ and $\mathcal{S}_2 := \{S \in \mathcal{S} : |S| > |S^\star|\}$, clearly it holds $\mathcal{S} = \{S^\star\} \cup \mathcal{S}_1 \cup \mathcal{S}_2$. We want to control the terms $\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\}$ and $\mathbb{P}\{\widehat{S} \in \mathcal{S}_2\}$ separately and conclude in view of

$$\mathbb{P}\{\widehat{S} \neq S^\star\} = \mathbb{P}\{\widehat{S} \in \mathcal{S}_1\} + \mathbb{P}\{\widehat{S} \in \mathcal{S}_2\}\,.$$

Since the Frobenius norm is sub-multiplicative, it holds, for all $\mathsf{A} \in \mathcal{E}(\overline{F})$,

$$\|\mathsf{A}(\widehat{\mathsf{K}} - \mathsf{K}) - (\widehat{\mathsf{K}} - \mathsf{K})\mathsf{A}\| \leq \|\mathsf{A}(\widehat{\mathsf{K}} - \mathsf{K})\|_2 + \|(\widehat{\mathsf{K}} - \mathsf{K})\mathsf{A}\| \leq 2\|\mathsf{A}\|\|\widehat{\mathsf{K}} - \mathsf{K}\|\,.$$

Thus, the quantity $\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|$ for $\mathsf{A} \in \mathcal{E}(\overline{F})$ can be bounded from below and above by

$$\|\mathsf{AK} - \mathsf{KA}\| - 2\|\mathsf{A}\|\|\widehat{\mathsf{K}} - \mathsf{K}\| \leq \|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\| \leq \|\mathsf{AK} - \mathsf{KA}\| + 2\|\mathsf{A}\|\|\widehat{\mathsf{K}} - \mathsf{K}\|. \tag{11}$$

To bound the term $\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\}$, we use (11) to remark that for all $S \in \mathcal{S}_1$,

$$Q(S) = \min_{\mathsf{A} \in \mathcal{E}(S)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|}{\|\mathsf{A}\|} + \lambda_n|S| \geq \min_{\mathsf{A} \in \mathcal{E}(S)\setminus\{0\}} \frac{\|\mathsf{AK} - \mathsf{KA}\|}{\|\mathsf{A}\|} - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|\,.$$

It follows

$$\min_{S \in \mathcal{S}_1} Q(S) \geq \min_{S \in \mathcal{S}_1} \min_{\mathsf{A} \in \mathcal{E}(S)\setminus\{0\}} \frac{\|\mathsf{AK} - \mathsf{KA}\|}{\|\mathsf{A}\|} - 2\|\widehat{\mathsf{K}} - \mathsf{K}\| = c_0(S^\star) - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|. \tag{12}$$

The constant $c_0(S^\star)$ is positive by F-identifiability of W. Moreover, observe that

$$Q(S^\star) = \min_{\mathsf{A} \in \mathcal{E}(S^\star)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|}{\|\mathsf{A}\|} + \lambda_n|S^\star| \leq \frac{\|\mathsf{W}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{W}\|}{\|\mathsf{W}\|} + \lambda_n|S^\star| \leq 2\|\widehat{\mathsf{K}} - \mathsf{K}\| + \lambda_n|S^\star|, \tag{13}$$

where we used both Eq. (11) and the fact that $\mathsf{WK} - \mathsf{KW} = 0$. Combining (12) and (13), we get

$$\mathbb{P}\{\widehat{S} \in \mathcal{S}_1\} \leq \mathbb{P}\Big\{\min_{S \in \mathcal{S}_1} Q(S) \leq Q(S^\star)\Big\} \leq \mathbb{P}\Big\{\|\widehat{\mathsf{K}} - \mathsf{K}\| \geq \frac{c_0(S^\star) - \lambda_n|S^\star|}{4}\Big\}\,.$$

To control the term $\mathbb{P}(\widehat{S} \in \mathcal{S}_2)$, we use that $\min_{S \in \mathcal{S}_2} Q(S) \geq \lambda_n \min_{S \in \mathcal{S}_2} |S| \geq \lambda_n(|S^\star| + 1)$. By Eq. (13), it follows

$$\begin{aligned}
\mathbb{P}\big\{\widehat{S} \in \mathcal{S}_2\big\} &\leq \mathbb{P}\Big\{\min_{S \in \mathcal{S}_2} Q(S) \leq Q(S^\star)\Big\} \\
&\leq \mathbb{P}\Big\{\lambda_n(|S^\star| + 1) \leq 2\|\widehat{\mathsf{K}} - \mathsf{K}\| + \lambda_n|S^\star|\Big\} \\
&= \mathbb{P}\Big\{\|\widehat{\mathsf{K}} - \mathsf{K}\| \geq \frac{\lambda_n}{2}\Big\}\,.
\end{aligned}$$

The proof of Theorem 7 follows directly by ($\mathbf{H}_2$). The corollary is a direct consequence using Borel-Cantelli's Lemma.

B.2 Proof of Theorem 10

Since $\Delta(\mathsf{K})\Phi_S$ is of full rank, the value $\widehat{\beta}_S = \big(\Delta(\widehat{\mathsf{K}})\Phi_S\big)^\dagger \Delta(\widehat{\mathsf{K}})a_0$ is the unique solution to Eq. (5) with probability tending to one asymptotically. Since the value of $\widehat{\beta}_S$ does not depend on $a_0 \in \mathcal{A}_S$, one can take $a_0 = w$ in view of $S^\star \subseteq S$. We obtain

$$\widehat{\beta}_S = \big(\Delta(\widehat{\mathsf{K}})\Phi_S\big)^\dagger \Delta(\widehat{\mathsf{K}})w = -\big(\Delta(\widehat{\mathsf{K}})\Phi_S\big)^\dagger \Delta(\mathsf{W})\widehat{k}.$$

The result follows from Slutsky's lemma, using that $(\Delta(\widehat{\mathsf{K}})\Phi_S)^\dagger$ converges in probability towards $(\Delta(K)\Phi_S)^\dagger$ and

$$\sqrt{n}\left(\Delta(\mathsf{W})\widehat{k} - \Delta(\mathsf{W})k\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\big(0, \Delta(\mathsf{W})\Sigma\Delta(\mathsf{W})^\top\big).$$

References

David F Anderson and Thomas G Kurtz. Continuous time markov chain models for chemical reaction networks. In *Design and Analysis of Biomolecular Circuits*, pages 3–42. Springer, 2011.

Flavia Barsotti, Yohann De Castro, Thibault Espinasse, and Paul Rochet. Estimating the transition matrix of a Markov chain observed at random times. *Statistics & Probability Letters*, 94:98–105, 2014.

Flavia Barsotti, Anne Philippe, and Paul Rochet. Hypothesis testing for markovian models with random time observations. *Journal of Statistical Planning and Inference*, 173:87–98, 2016.

José Bento and Morteza Ibrahimi. Support recovery for the drift coefficient of high-dimensional diffusions. *IEEE Transactions on Information Theory*, 60(7):4026–4049, 2014.

José Bento, Morteza Ibrahimi, and Andrea Montanari. Learning networks of stochastic differential equations. In *Advances in Neural Information Processing Systems*, pages 172–180, 2010.

Béla Bollobás. *Random graphs*. Springer, 1998.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 2016.

Thibault Espinasse and Paul Rochet. Relations between connected and self-avoiding hikes in labelled complete digraphs. *Graphs and Combinatorics*, to appear, 2016.

Thibault Espinasse, Fabrice Gamboa, and Jean-Michel Loubes. Parametric estimation for gaussian fields indexed by graphs. *Probability Theory and Related Fields*, 159(1-2):117–155, 2014.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Donald P Gaver Jr. Imbedded markov chain analysis of a waiting-line process in continuous time. *The Annals of Mathematical Statistics*, pages 698–720, 1959.

Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

Benjamin Girault. Stationary graph signals using an isometric graph translation. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1516–1520. IEEE, 2015.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *The Journal of Machine Learning Research*, 11:1709–1731, 2010.

Jing Jiang, Christo Wilson, Xiao Wang, Wenpeng Sha, Peng Huang, Yafei Dai, and Ben Y Zhao. Understanding latent interactions in online social networks. *ACM Transactions on the Web (TWEB)*, 7(4):18, 2013.

Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Elizabeth Chase MacRae. Estimation of time-varying Markov processes with aggregate data. *Econometrica*, 45(1):183–198, 1977. ISSN 0012-9682.

Antonio G Marques, Santiago Segarra, Geert Leus, and Alejandro Ribeiro. Stationary graph processes and spectral estimation. *arXiv preprint arXiv:1603.04667*, 2016.

Catherine Matias, Tabea Rebafka, and Fanny Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *arXiv preprint arXiv:1512.07075*, 2015.

Vincent Miele and Catherine Matias. Revealing the hidden structure of dynamic ecological networks. *arXiv preprint arXiv:1701.01355*, 2017.

Nathanaël Perraudin and Pierre Vandergheynst. Stationary signal processing on graphs. *arXiv preprint arXiv:1601.02522*, 2016.

Arthur O. Pittenger. Time changes of Markov chains. *Stochastic Process. Appl.*, 13(2):189–199, 1982. ISSN 0304-4149. doi: 10.1016/0304-4149(82)90034-5. URL http://dx.doi.org/10.1016/0304-4149(82)90034-5.

Fabrice Rossi and Pierre Latouche. Activity date estimation in timestamped interaction networks. In *21-th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 113–118, Apr. 2013.

Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press Boston, 1990.

Mathieu Thomas, Nicolas Verzelen, Pierre Barbillon, Oliver T Coomes, Sophie Caillon, Doyle McKey, Marianne Elias, Eric Garine, Christine Raimond, Edmond Dounias, et al. A network-based method to detect patterns of local crop biodiversity: Validation at the species and infra-species levels. *Advances in Ecological Research*, 53:259–320, 2015.

Nicolas Verzelen. *Gaussian graphical models and Model selection*. PhD thesis, Université Paris Sud-Paris XI, 2008.

Nicolas Verzelen, Ery Arias-Castro, et al. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.