Reconstructing undirected graphs from eigenspaces

Yohann De Castro* Thibault Espinasse† Paul Rochet‡

February 29, 2016

Abstract

In this paper, we aim at recovering an undirected weighted graph of N vertices from the knowledge of a perturbed version of the eigenspaces of its adjacency matrix W. Our approach is based on minimizing a cost function given by the Frobenius norm of the commutator $\mathsf{AB}-\mathsf{BA}$ between symmetric matrices A and B. In the Erdős-Rényi model 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 $N \log N/2$. When an estimation of the eigenspaces is at hand, we provide backward-type support selection procedures from theoretical and practical point of views.

1 Presentation

We consider a set of problems where one aims at recovering an undirected weighted graph 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. Several models (including Markov chain, stationary Vectorial AutoRegressive process, vectorial Ornstein-Uhlenbeck process for instance) 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 problem. 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.

Then, 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, see Section 4.1. Using backward-type procedures, Section 4 derives estimators of the graph structure (i.e. its support) from a perturbed observation of its eigenspaces. A numerical approach developed in Section 5 assesses the performances of our new estimation method. Discussion and related questions are presented in Section 7.

*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France
†Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France
‡Laboratoire de Mathématiques Jean Leray, Université de Nantes, 44322 Nantes, France

2 Model and identifiability

2.1 The model

Consider a symmetric matrix $\mathsf{W} \in \mathbb{R}^{N\times N}$ with some zero entries, where nonzero entries describe the intensity of a link of any form of local interaction. One may understand W as the adjacency matrix of an undirected weighted graph with N vertices. We focus on the eigenspaces of W examining models where we have no information on the spectrum of the graph. Depicting this situation, we assume that we consider a matrix $\mathsf{K} \in \mathbb{R}^{N\times N}$ such that $\mathsf{KW} = \mathsf{WK}$ or, in more realistic scenarios, we may observe a perturbed version $\widehat{\mathsf{K}}$ of K. In particular, we may assume that there exists an unknown injective function f on the real line such that $\mathsf{K} =: f(\mathsf{W})$ where $f(\mathsf{W})$ denotes a symmetric matrix with the same eigenspaces as W applying f to each of the underlying eigenvalues.

Our goal is to uncover W and f from the knowledge of an estimator $\widehat{\mathsf{K}}$ of K, namely reconstruct W from a perturbed observation of its eigenspaces. 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 to the set of edges of the target graph. Throughout this paper, a special case of interest is given by $F = F_{\mathrm{diag}} := \{(i,i),\ 1 \le i \le N\}$ meaning that there are no self-loops in W.

2.2 Identifiability under ($\mathbf{H}_{\mathrm{F}}$)

For $S \subseteq [1,N]^2$, denote by $\mathscr{E}(S)$ the set of symmetric matrices A whose support is included in S, namely $\mathrm{Supp}(\mathsf{A}) \subseteq S$. Given the set F of forbidden entries defined via ($\mathbf{H}_{\mathrm{F}}$), the matrix of interest W is sought in the set $\mathscr{E}(\overline{F})$ where $\overline{F}$ denotes the complement of F. In some cases, most matrices $\mathsf{W} \in \mathscr{E}(\overline{F})$ are uniquely determined by their eigenspaces. More precisely, for each of those $\mathsf{W} \in \mathscr{E}(\overline{F})$, there is no matrix $\mathsf{A} \in \mathscr{E}(\overline{F})$ different from 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 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 $\mathsf{A} = \lambda \mathsf{W}$ for some $\lambda \in \mathbb{R}$.*

Interestingly, we have the following proposition.

Proposition 1 (Lemma 2.1 in [BDCER14]) *Let $S \subseteq \overline{F}$, the set of F-identifiable matrices in $\mathscr{E}(S)$ is either empty or a dense open subset of $\mathscr{E}(S)$.*

The proof uses the fact that non F-identifiable matrices in $\mathscr{E}(S)$ can be expressed as the zeroes of a particular analytic function, we refer to [BDCER14] for further details. This proposition shows that the F-identifiability of a matrix W is essentially a condition on its support S. By abuse of notation, we say that a support $S \subseteq \overline{F}$ is *F-identifiable* if almost every matrix in $\mathscr{E}(S)$ are F-identifiable.

Characterizing the F-identifiability appears to be a challenging issue since it can be viewed as understanding the eigenstructure of a graph through its support. The particular case of the diagonal F_{diag} as the set of forbidden entries is given a particular attention in this paper. 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 10 (see Appendix A), we introduce necessary and sufficient conditions on the target support Supp(W) for diagonal identifiability. Defining the *kite* graph ∇_N of size N 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 for instance), one simple sufficient condition on diagonal identifiability reads as follows.

Proposition 2 *If the kite graph ∇_N of size N is a subgraph of the graph of size N and edges S then S is diagonally identifiable.*

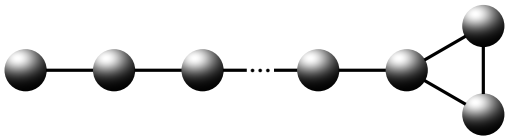

Figure 1: The kite graph ∇_N on a set of vertices of size N.

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.

Proposition 3 *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.*

A proof of this statement can be found in Section A.2. In particular, 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 [Bol98] for instance. This theorem shows that in the regime $(\log N + \log\log N)/N$ an Erdős-Rényi graph is diagonally identifiable. Looking at a pair of isolated points and using Theorem 10, one can prove that $\log N/N$ is a sharp threshold for diagonal identifiability in the Erdős-Rényi model (see Section A.3), it can be stated as follows.

Theorem 4 *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* 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* 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 exist another small degree undirected weighted graph with the same eigenspaces.

3 Some concrete models

Markov chains

We begin with an example treated in the companion papers [BDCER14, BPR16]. Consider a Markov chain $(X_n)_{n\in\mathbb{N}}$ with finite state space $[1,N]$ and transition matrix $\mathsf{Q} \in \mathbb{R}^{N\times N}$. Let $(T_k)_{k\geq 1}$ be a sequence of random times such that $T_{k+1} - T_k$ are i.i.d random variables independent of $(X_n)_{n\in\mathbb{N}}$. Denoting $Y_k = X_{T_k}$, remark that Y is also a Markov chain with transition matrix $f(\mathsf{Q})$ where f is the generating function of $T_k - T_{k-1}$. Therefore $\mathsf{K} = f(\mathsf{Q})$ may be estimated and one may recover Q from K. That was one purpose of the paper .

Vectorial AutoRegressive process

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

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

with ε_i i.i.d. random variables. Define as above $Y_k = X_{T_k}$ where T_k are random times such that the time gaps $T_k - T_{k-1}$ are i.i.d. with generating function f. Then, it holds

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

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

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, if $Y_k = X_{T_k}$ where T_k are random times such that the time gaps $T_k - T_{k-1}$ are i.i.d., then

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

where f is the generating function of the the time gaps $T_k - T_{k-1}$.

Graphical models

Our model may be related to Graphical models. Indeed, one may consider W as the precision matrix, which is the inverse of the covariance matrix, having some non zero entries described by a graph of dependancies. Using $f(x) = x^{-1}$, this falls into our setting, trying to recover W from the estimation of the covariance matrix. Of course, in this case, it is better to use the knowledge of f, which certainly improves estimation. However, our procedure allows us to estimate the function f and heuristically validate the hypothesis $f(x) = x^{-1}$, see Section 5.

Seasonal VAR **structure**

We can also consider a toy example looking at a seasonal VAR structure without any randomness on the times of observations. Let T be a positive integer, and $(u_k)_{k\in\mathbb{Z}}, (v_k)_{k\in\mathbb{Z}}$ be some periodic sequences of period T. Consider the following model

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

We may observe the model only at time gap intervals T with some error

$$X_t = Y_{tT+k_0} + \eta_t .$$

This falls into the general frame

$$X_t = f(\mathsf{W}) X_{t-1} + \mu_t ,$$

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

4 Estimating the support

4.1 The commutator

The methodology presented in the paper relies on the fact that the matrix of interest W commutes with K. However, to be able to recover W from the estimator $\widehat{\mathsf{K}}$, one needs additional information which comes from the identifiability condition ($\mathbf{H}_{\mathrm{F}}$). We use the following minimization criterion

$$\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|_2 , \quad \mathsf{A} \in \mathscr{E}(\overline{F}),$$

where $\|.\|_2$ denotes the Frobenius norm. This criterion was first used in [BDCER14] in a similar context to reflect that W is expected to nearly commute with $\widehat{\mathsf{K}}$, provided that $\widehat{\mathsf{K}}$ is sufficiently close to its true value K.

4.2 The ℓ_0-approach

Given an estimator $\widehat{\mathsf{K}} = \widehat{\mathsf{K}}_n$ 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 $\widehat{S} = \widehat{S}_n$ of the target support $S^\star := \mathrm{Supp}(\mathsf{W})$ as a minimizer of the criterion Q given by

$$\forall S \subseteq \overline{F}, \quad Q(S) := \min_{\mathsf{A} \in \mathscr{E}(S)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|_2}{\|\mathsf{A}\|_2} + \lambda_n |S|,$$

for some tuning parameter $\lambda_n > 0$ and defining the minimum of an empty set as ∞. Recall that $\mathscr{E}(S)$ is the set of matrices A such that $\mathrm{Supp}(\mathsf{A}) \subseteq S$. Furthermore, we assume that the estimator $\widehat{\mathsf{K}}_n$ converges toward K in probability R_n, namely

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

where $t \mapsto R_n(t)$ is non-increasing and such that, for all $t > 0$, $R_n(t) \to 0$ as n goes to ∞.

Theorem 5 *Assume that* $(\mathbf{H}_2)$ *and* $(\mathbf{H}_{\mathrm{F}})$ *hold. If* W *is* F*-identifiable, then*

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

where

$$c_0 := \min_{\substack{S \neq S^\star \\ |S| \leq |S^\star|}} \min_{\mathsf{A} \in \mathscr{E}(S)} \frac{\|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\|_2}{\|\mathsf{A}\|_2} > 0.$$

Corollary 6 *Under the assumptions of Theorem 5, if it holds*

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

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

Note that, based on the upper bound in Theorem 5, the optimal scaling should be

$$\lambda_n^\star = \frac{4c_0}{|S^\star| + 2},$$

which interestingly does not depend on n. However, this calibration is not relevant since both c_0 and $|S^\star|$ are unknown. Nevertheless, we may choose a sequence λ_n decreasing slowly to 0 to ensure both conditions of Corollary 6.

4.3 Backward support selection based on commutator criterion

The ℓ_0-approach meets with the curse of dimensionality, especially since the size of the support increases quadratically with the dimension. 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 $\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|_2^2$ and $a \mapsto \|\Delta(\widehat{\mathsf{K}})a\|_2^2$ can be used indistinctly as cost functions. Minimizing this criterion over model spaces of decreasing size, we may consider sequences of least-squares estimates in the sequel.

Assumptions

In this section, we may assume the three following hypothesis $(\mathbf{H}_\Sigma)$, $(\mathbf{H}_1)$ and $(\mathbf{H}_{\mathrm{Id}})$.

∘ Deriving the asymptotic law of least-squares estimators, we may assume that the estimate $\widehat{\mathsf{K}}$ is built from a sample of size n growing to infinity and asymptotically Gaussian with asymptotic covariance matrix either known or that can be estimated. One can write

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

where Σ is a $N^2 \times N^2$ covariance matrix. This condition is verified for instance in the framework considered in [BDCER14, BPR16]. Note that asymptotic normality is a standard ground base investigating any least-square procedure.

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

$$\mathbf{1}^\top w = 1, \qquad (\mathbf{H}_1)$$

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 identifiability. Here, the condition $\mathbf{1}^\top w = 1$ achieves two goals: preventing the null matrix form being a solution and making the problem identifiable. 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 $\mathbf{v}$ as random (for instance with isotropic law), so that $(\mathbf{H}_1)$ is almost surely fulfilled for any fixed target w.

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

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

with linear difference space given by

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

By abuse of notation, $\mathscr{A}_S$ 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 $\mathscr{A}_{\overline{F}}$. Actually, w can then 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 may assume that

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

which is the definition of F-identifiability, see Definition 1.

Asymptotic normality and a 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 $\Phi_S \in \mathbb{R}^{N^2 \times \dim(\mathscr{A}_S)}$ whose column vectors form a basis of $\mathscr{L}_S$. Since W is F-identifiable and $S \subseteq \overline{F}$, the operator $\Delta(\mathsf{K})\Phi_S$ is one-to-one. In this case, evaluating the commutator $a \mapsto \Delta(\mathsf{K})a$ over $\mathscr{A}_S$ reduces to considering the map

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

with a_0 chosen arbitrarily in $\mathscr{A}_S$. When replacing the unknown $\Delta(\mathsf{K})$ with its estimate $\Delta(\widehat{\mathsf{K}})$, the minimization of the criterion $a \mapsto \|\Delta(\widehat{\mathsf{K}})a\|_2^2$ over $\mathscr{A}_S$ can be written similarly as a linear regression framework where the parameter of interest is estimated by

$$\widehat{\beta}_S \in \arg\min_{b \in \mathbb{R}^{\dim(\mathscr{A}_S)}} \|\Delta(\widehat{\mathsf{K}})(a_0 - \Phi_S b)\|_2^2. \tag{1}$$

We recognize a linear model with response variable $y = \Delta(\widehat{\mathsf{K}})a_0$ 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 $\Delta(\mathsf{K})(a_0 - \Phi_S\beta) = 0$. Denote by $\mathsf{M}^\dagger$ the pseudo-inverse of a matrix M, we deduce the following result.

Theorem 7 *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 $\Omega_S = (\Phi_S^\top \Delta(\mathsf{K}))^\dagger \Delta(\mathsf{W})\Sigma\Delta(\mathsf{W})(\Delta(\mathsf{K})\Phi_S)^\dagger$.

We then have

$$\widehat{w}_S = \mathrm{vec}(\widehat{\mathsf{W}}_S) = \arg\min_{a \in \mathscr{A}_S} \|\Delta(\widehat{\mathsf{K}})a\|_2^2 = a_0 - \Phi_S\widehat{\beta}_S.$$

The asymptotic distribution of $\widehat{w}_S$ follows directly from Theorem 7,

$$\sqrt{n}(\widehat{w}_s - w) \xrightarrow[n\to\infty]{d} \mathcal{N}\big(0, \Phi_S\Omega_S\Phi_S^\top\big). \tag{2}$$

The limit covariance matrix is unknown, but plugging the estimates $\widehat{\mathsf{W}}_S$, $\widehat{\mathsf{K}}$ and $\widehat{\Sigma}$ yields a consistent estimator $\Phi_S\widehat{\Omega}_S\Phi_S^\top$. So, letting $\widehat{\sigma}^2_{S,ij}$ denote the diagonal entry of $\Phi_S\widehat{\Omega}_S\Phi_S^\top$ associated to $\widehat{\mathsf{W}}_{S,ij}$, the (i,j) entry is judged significant if the statistic

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

exceeds in absolute value some quantile of the standard Gaussian distribution. The backward support selection procedure is then implemented by the following recursive algorithm.

1. Start with the maximal acceptable support $S_1 = \overline{F}$
2. At each step m, compute the statistics $\tau_{ij}(S_m)$ for all $(i,j) \in S_m$
3. If the minimal value $|\tau_{ij}(S_m)|$ is smaller than some threshold $t > 0$, set $S_{m+1} = S_m \setminus \{(i,j),(j,i)\}$
4. Stop when all entries are judged significant, i.e. when all $|\tau_{ij}(S_m)|$ are greater than t.

Based on Eq. (2), the quantile $q_{1-\frac{\alpha}{2}}$ of the standard Gaussian distribution appears as a reasonable choice for the threshold, as it boils down to performing an asymptotic significativity test of level α. However, due to the slow convergence to the limit distribution and the tendency to underestimate the variance for small sample sizes (see Figure 3), the numerical study shows that a better choice for the threshold depends on the overall behavior of the commutator $\Delta(\widehat{\mathsf{K}})\widehat{w}_{S_m}$ computed over the nested sequence of supports. The details are discussed in Section 5 where the backward procedure is computed on numerical examples, along with a more robust bootstrap-assisted version.

Similarly as the Fisher test in the linear regression model, the previous procedure can be extended to test the significance of a whole set of entries at once. For Γ a symmetric positive semi-definite matrix, we note $\chi^2(\Gamma)$ the distribution of the generalized chi-square random variable $\varepsilon^\top\Gamma\varepsilon$ where ε is a standard Gaussian vector. Moreover, for any pair S', S of nested supports with $S' \subseteq S \subseteq \overline{F}$, we define the linear space

$$E_{S',S} := \mathrm{Im}\big(\Delta(\mathsf{K})\Phi_S\big) \cap \ker\big(\Phi_{S'}^\top\Delta(\mathsf{K})\big),$$

and denote by $\Pi_{E_{S',S}}$ the orthogonal projector onto $E_{S',S}$.

Proposition 8 *Let S', S be nested supports with $S^\star \subseteq S' \subseteq S$,*

$$T(S',S) := n\bigg(\min_{a\in\mathscr{A}_{S'}} \|\Delta(\widehat{\mathsf{K}})a\|^2 - \min_{a\in\mathscr{A}_S} \|\Delta(\widehat{\mathsf{K}})a\|^2\bigg) \xrightarrow[n\to\infty]{d} \chi^2(\Gamma_{S',S})$$

where $\Gamma_{S',S} = \Pi_{E_{S',S}}\Delta(\mathsf{W})^\top\Sigma\Delta(\mathsf{W})\Pi_{E_{S',S}}$.

The proof, given in Appendix B.3, follows the same guidelines as Theorem 3.1 in [BPR16].

5 Simulations

In this section, we provide some simulations to show the performances of the stepwise backward algorithm.

5.0.1 Studied cases

First, we explain two different frameworks we will work into. In Subsection 5.0.3, we will consider the graph G_1 showed by Figure 2. We chose W fixed as the normalized adjacency matrix of G_1. For the function f, we took arbitrarily $f(t) = \frac{1+2t^2}{(1-t)^3}e^{-t}$, which happens to be positive on the spectrum of W.

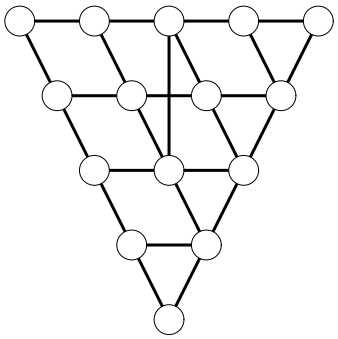

Figure 2: Graph G_1

From $K = f(W)$, we draw n i.i.d. realizations of a Gaussian centered vector X with covariance matrix K. We use empirical estimates $\hat{K}_n$ for K and $\hat{\Sigma}_n$ for Σ, the covariance matrix of $\hat{K}$.

Then, we will take a look at the performances of the algorithm in a second framework where we add randomness to the graph, the entries of W, and the function f. More precisely,

- The graph G is taken randomly as a symmetric Erdös-Rényi graph (N, p_N)
- The matrix W is chosen randomly, drawing sign with probability $\frac{1}{2}$ and entries as binomial variables increased by a constant (in order to prevent them from beeing closed to zero), and then normalized such that $\|W\|_2 = 1$, where $\|W\|_2$ denotes the Frobenius norm of W (this normalization is here to ensure the positivity of the function on the spectrum of W).
- The function f is chosen randomly uniformly among the 3 following functions, chosen

arbitrarily as

$$f_1(t) = \frac{1}{(1-t)^4} e^{-t}$$
$$f_2(t) = \frac{1+2t^2}{(1-t)^3} e^{t}$$
$$f_3(t) = (1-t)^3 (1+t^2) e^{-t},$$

Then, as previously, a Gaussian centered sample X is drawn from the covariance matrix $K = f(W)$, and we use X to recover the graph G.

5.0.2 The backward algorithm

The backward algorithm explained in Section 4.3 provides from X a sequence of nested supports. Recall its main steps :

Algorithm 1

Backward algorithm

1: *Begin with the full support* $S = F^C$
2: *Compute* $\hat{W}_S := \arg\min_{w \in \mathscr{A}^S} \|\Delta(\hat{K}_n) w\|^2$.
3: *Compute the significancy* τ_{ij}, *for all* $i, j \in [1, N]$.
4: *Remove from* S *the less significant entry* (i, j).
5: *Back to step* 2.

end

Figure 3 shows the speed of convergence of $\hat{W}_{12}$ to the normal distribution while increasing the number of observations. Here, we make no constraints on the support. That may be an explanation of the fact that the distribution seems tighter when n decreases.

In practice, the stopping condition given in Section 4.3 happens to overestimate the support. Indeed, it does not take into account the fact that the same sample is used to remove variables, and to select the best support among the sequence of candidates given by the backward procedure. Furthermore, the non Gaussianity of the τ_{ij} for small samples may be a problem. We could perform another threshold based on multiple tests, but thresholding directly the commutator works even better.

Observing that the backward algorithm builds a sequence of nested support $S_1 \supset \cdots \supset S_l, 0 \leq l \leq \frac{N(N-2)}{2}$, we can choose the best support by taking a look at the behavior of the minimized commutator $\|\Delta(\hat{K}_n)\hat{W}_{S_i}\|^2$ in function of i, when the true support $S^\star$ lies in the chosen sequence. While the transition seems obvious for large samples, it remains quite noticeable for smaller samples, if we look at it at the good scale. Figure 4 summarizes this fact (we choose realizations for which the true support $S^\star$ lies in the sequence selected by the backward algorithm, which is not always the case.).

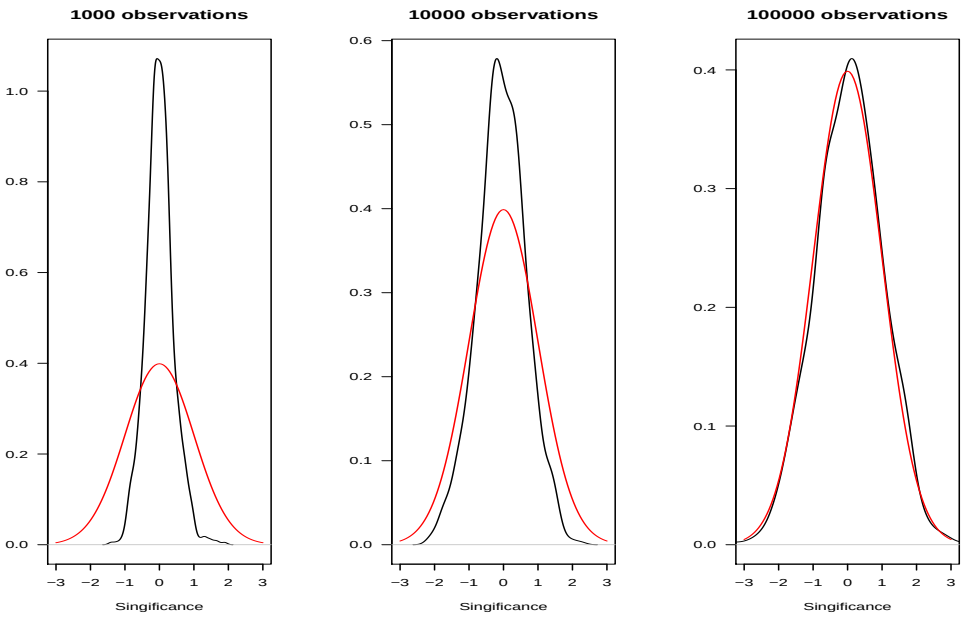

Figure 3: Convergence of the significance to the normal distribution

Therefore, the very simple idea to apply a threshold to the commutator is very efficient from a practical point of view. To understand the level of thresholding that should be applied, notice that $\|\Delta(\hat{K}_n)\hat{W}_{S_i}\|^2$ is of order maximal $\frac{N^2}{n}$ whenever $S_i \supset S^\star$ and that $\|\Delta(\hat{K}_n)\hat{W}_{S_i}\|^2 > c_0$, where the unknown constant $c_0 > 0$ is given in Section 4.2. This constant is unknown but can be upper bound by the value $c_1 = \|\Delta(\hat{K}_n)\hat{W}_{S_i}\|^2$, for a known support S_l that does not contain $S^\star$, and has less edges. From this simple observation, a lot of ideas for the threshold level should work. Here comes some useful observations:

- For large samples, speed as $\frac{1}{\sqrt{n}}$ works perfectly, as any speed that goes to zero, but slower than $\frac{1}{n}$.
- For smaller samples, notice that $\mathrm{Tr}(\Sigma)$ bounds the variance of $\hat{K} - K$ and can be estimated.
- Furthermore, any additional information about the number of edges of $S^\star$ can be efficiently used in the choice of the threshold, since it reduce the value of the estimated constant c_1.

In the following, we took as a threshold $\frac{\hat{c}_1 \log(n)}{n}$, where $\hat{c_1}$ has been estimated just by taking the minimized commutator over the last model of the sequence with 14 edges (and the assumption of identifiability ensures that S^* contains at least 15 edges).

The following table summarizes the first results, varying n from 100000 to 100. Remark that the graph G_1 has $N = 15$ vertices, and 25 edges among 105 possible. The table shows the proportion of successful runs (the support has been perfectly recovered), the mean of the number of false edges (either wrongly removed or wrongly kept). It gives also the number

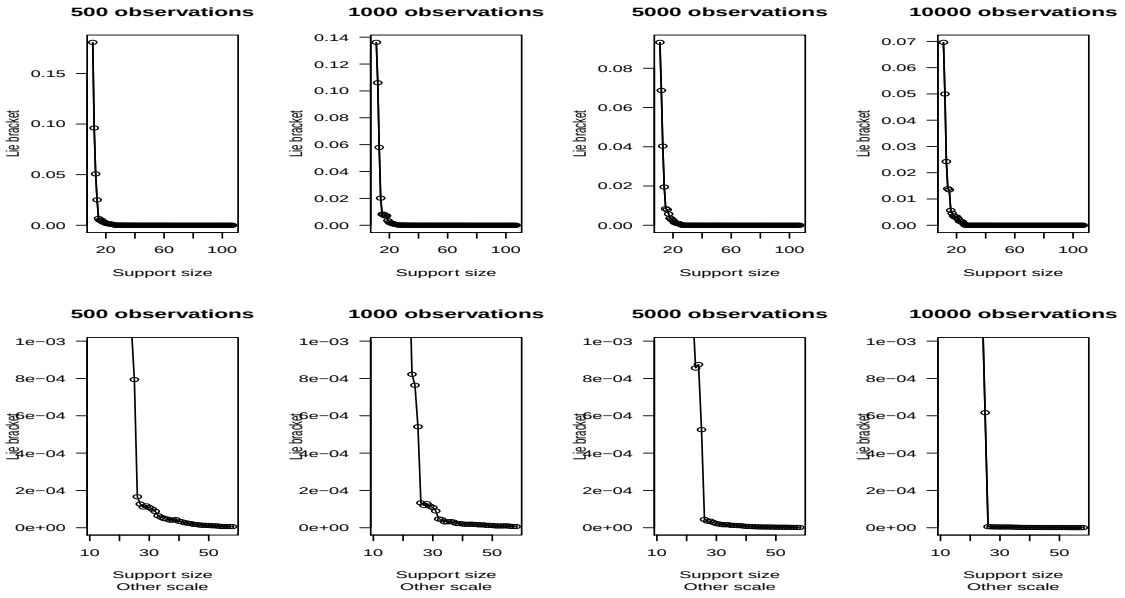

Figure 4: Observed commutator values

of trajectories (i.e. sequences of support) that contains $S^\star$. Theses numbers are computed over 1000 runs.

n	50000	20000	10000	5000	2000	1000	500	200	100
Succesfully recovered	$> 99\%$	97%	92%	87%	78%	31%	14%	4%	$< 1\%$
Mean error	0.15	1.3	3.2	4.6	6.4	9.7	12.3	20	25
Good trajectories	$> 99\%$	97%	92%	87%	80%	69%	58%	28%	5%

Even if it was not the purpose of this paper, Figure 5 shows some estimations of f, on the spectral domain of W, computed with 1000 values, on the first simulated case in which the support has been recovered.

Yet, for smaller samples, the good threshold is very dependent into the size of the true unknown support. Even if the transition appear to be obvious in Figure 4, it is mainly because we choose the good scale to look at it. Actually, with less than 2000 observations, the optimal window for the thresholding is not very large, and we should calibrate it adaptively. This is the purpose of the next section.

5.0.3 How to perform a kind of cross validation

The main problem for small samples remains robustness of the algorithm. Actually, we synthesized all available informations in $\hat{K}_n$, so we throw away a lot of useful information. When we have the sample X at hand, we could run many times the previous algorithm on learning samples to improve robustness. The best practical choice seems to do it without

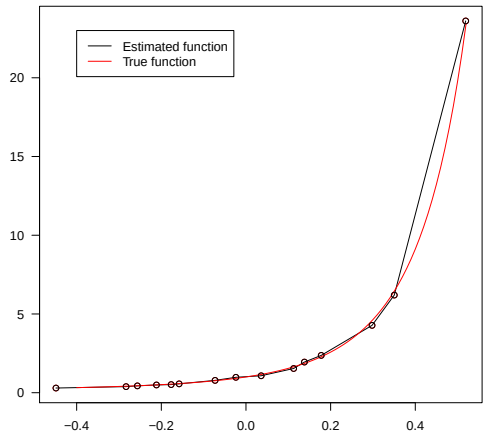

Figure 5: Estimation of the unknown function f

replacement. Doing that, it appears that if we compute at each step i the value r_i of the commutator between $\hat{W}_{S_i}$ and $\hat{K}_n$ (where the estimator $\hat{W}_{S_i}$ of W is obtained with the learning sample, and constraints on support S_i, and the value of $\hat{K}_n$ is estimated with the whole sample), then this sequence is quite constant as soon as S_i contains $S^\star$, and even decreases close to $S^\star$, as shown if Figure 6.

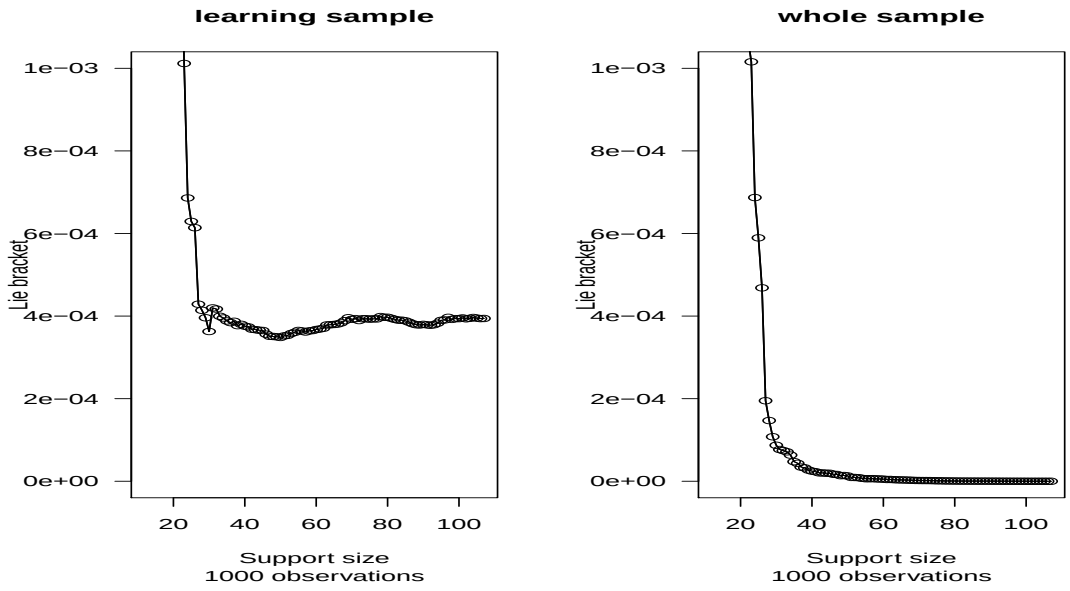

Figure 6: Comparison between errors for a learning sample and the whole sample

It allows us to efficiently calibrate the threshold only using the value of the commutator on the first model (for large sample, it can be multiplied by a constant between 1 and 2 to even improve the results). To add robustness to this method, and to avoid to select a model that did wrong at the begging of the algorithm, we may remove a proportion q of the trajectories with the largest value r_1.

We end with many estimators $\hat{S}_j$, each one build with a different learning sample. But since we can be as much conservative as we want choosing q close to 1, we can now take the smallest support selected without risk of choosing a too small support.

Let us summarize the algorithm now :

Algorithm 2

CV algorithm

1: *Compute $\hat{K}_n$ on the whole sample*
2: *Build m learning samples without replacement.*
3: *For each learning sample, run the backward algorithm without stopping condition, return m sequences (of length $\frac{N(N-2)}{2}$) of nested supports, and the values $(r_{ij})_{i=1:\frac{N(N-2)}{2},j=1:m}$ of the commutator between $\hat{W}_{ij}$ and $\hat{K}_n$, where $\hat{W}_{ij}$ is computed with support S_{ij} and the learning sample j and $\hat{K}_n$ is computed with the whole sample.*
4: *Remove a proportion q of sample with largest r_{1j}.*
5: *Compute the estimated support $\hat{S}_j = S_{i^\star j}$ for all learning sample j by choosing $i^\star$ as the last index i such that $r_{ij} > C r_{1j}$.*
6: *Return $\hat{S}$ as the smallest of the $\hat{S}_j$ in terms of number of edges*

end

Here we made the conservative choice $C = 1$, but $C > 1$ also works fine. Actually, this constant does not need to be calibrated, but it defines the ratio "signal over noise" needed to be able to recover the graph.

The following table summarizes the improvements. We chose arbitrarily 10 learning samples drawn from X with a probability decreasing with n, to avoid over-learning (and maximal of $\frac{3}{4}$ for very small samples). We removed 60% of the trajectories, with the largest r_{1j}. These numbers are computed over only 1000 runs.

n	50000	20000	10000	5000	2000	1000	500	200	100
Succesfully recovered	100%	99.6%	99.4%	94.6%	83.4%	75.4%	40.2%	8.8%	$< 1\%$
Mean error	0	0.21	0.3	0.88	1.85	2.7	5.9	19.6	38.9

For computational time issues, we took only 10 learning samples in the previous table, but the following table shows the improvement of the results, for $n = 500, 1000$, only increasing the number of trajectories, and the proportion q of removed trajectories. (This time, the frequencies are estimated with only 200 runs).

Number of trajectories	10	20	50	100
q	0.6	0.7	0.8	0.85
Succesfully recovered, $N = 1000$	75.4%	85%	91%	93%
Mean error, $N = 1000$	2.7	0.4	0.2	0.1
Succesfully recovered, $N = 500$	40%	47%	60%	61%
Mean error, $N = 500$	5.9	3.2	1.3	1.1

Now, we can run the algorithm in the second framework, where the graph, the entries of W, and the function f are all random. We choose $N = 12, p_N = 0.4$, and removed 60% of the trajectories. The following table summarizes the results :

n	50000	20000	10000	5000	2000	1000	500
Succesfully recovered	%	%	81%	77%	58%	38%	%
Mean Error	%	%	3.3	4.2	8.2	14.8	%

6 Real life application

In this section, we will apply our algorithm on real life datas for which the graph is known. Indeed, we consider spatial data on a 4×4 grid over France. At each of the 16 vertices, we observed the daily number of lightning during 3 years in the corresponding region. From theses observations, we try to recover an underlying graph that should make the grid appear.

To this aim, we first eliminate day without any lighting all over France and we obtain some observations Y_i, $i = 1 \cdots, 950$, where Y_i is a vector of length 16 giving the number of impacts day i in each of the 16 regions. This numbers are highly non Gaussian, contains a lot of zeros, and show a clear south-east/north-west tendency (with much more lightning in the south east). Therefore, we look at the numbers at the log scale and we substracted a tendency.

Now, it remains a strong inhomogeneity, that should violate the assumption that the underlying graph has no self-loop (i.e. the diagonal of W is zero). To overcome this problem, we normalize the process in such manner that the conditional variance at each vertex conditionally to all the other is 1.

Finally, considering that the covariance matrix of the process commute with an underlying graph, we applied our algorithm with default parameters : we draw 50 learning samples keeping all observations with probability $\frac{3}{4}$, and removed 60% of the 50 trajectories. We do not obtain exactly the same graph at each run, which means that the results are not perfect. However, the graphs obtained are most of the time very satisfying.

To show the results, we ran 100 times the algorithm, and memorized, for every edge, the proportion of the time that this edge appears. This is summarized if Figure 7.

To compare our method with other, we only used the package *GGM* used to infer Graphical Models. This package is very efficient, and powerful even for samples with more vertices than observations. It is not designed exactly for our case, so we do not pretend that our

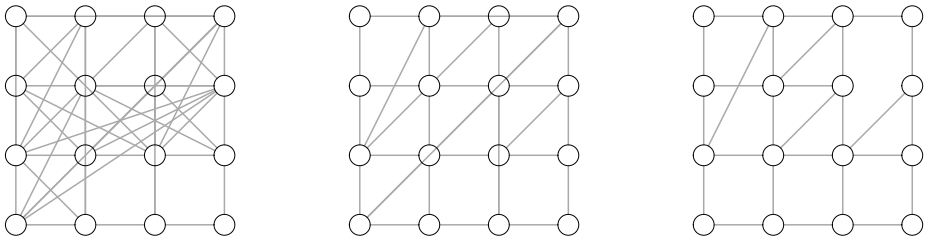

Figure 7: Results with our methods : Edges that appears 30%, 50%, and 70% of the time.

method makes better than this algorithm. Furthermore, we did not tune the parameters, and used rather the default parameters, only specifying the maximal degree of each vertex as $dmax = 4$. The results are given in Figure 8.

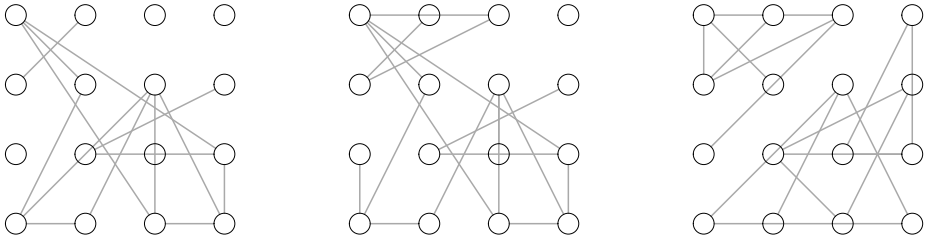

Figure 8: Results with the GGM package, with families “LA”, “C01” and “QE”

The results show that, in this case, our method seems to work at least as fine as an inference of a graphical model, modelling the datas as a Gaussian Markov Field.

We insist that we do not claim this fact to be general. In particular, we need much more observations than the methods developed in this package. But we pretend that, in different contexts, and with enough observations, we can be as good as other methods. Indeed, our method yet present one advantage : the process does not need to be Markov, and for instance, we could infer spatial autoregressive process of any order (whereas graphical model inference can only recover underlying graphs for AR_1 spatial processes, which are Markov). But this advantage turns into a problem when the process is truly Markov, because we do not use the knowledge of the function f, which can be taken as $\frac{1}{x}$ in the Markov case.

7 Discussion

In this paper, we devellop a new method to recover hidden graphical structures, in different models that shares the fact that, one way or another, we access to an approximation of

the eigenstructure of the graph, through an estimation of an operator that commutes with a weighted adjacency matrix of this unkown graph. This is noticeable that we do not need any sparsity assumption to make the method work, and even with the large number of unknown parameters ($K = f(W)$, with the support, the function f, and the non-null entries of W all unknown), we can perfectly recover the support when enough observations are available.

We only assume that we know the location of some zeros. The most interesting case is when the known zeros are localized onto the diagonal, because it only means that the process is well normalized, in a sense, because all self-loop have same weights.

We can also explain why the ℓ^1 penalty has not been tackled in this paper. First note that in our framework, the design itself is noised, and that any ℓ^1 penalty without further constraints leads to the null matrix. Therefore, we need to add a condition to avoid 0, for instance that $\|W\|_2 = 1$. We aim at recovering the exact support when the number of observations is large. But using the ℓ^1 penalty tends to overestimate a lot the support. It can be understood by seeing that a full matrix may commute with $\hat{K}$, and in the same time have a smaller ℓ^1 norm.

Furthermore, we can note that there is a number of observations behind which the algorithm always provides a wrong support. Furthermore, this fact can be data-observed, because almost all learning samples will lead to different supports. This limit is intrinsic to our model and is only the observation of a balance between the noise and the signal. The noise is only the estimation error of K, and has order $\frac{1}{\sqrt{n}}$, whereas the signal is of order c_0, unknown. Yet, being under this limit can be data validated only verifying that the error behaves as in the left picture of Figure 4.

Finally, it is really surprising that this model is almost surely identifiable, as soon as the graph is not too sparse. The limit $(1+\epsilon)\frac{\log(N)}{N}$ for p_N in Erdös-Rényi model $G(N, p_N)$ is as low as we could expect it to be.

For practical issues, it remains 3 challenges that have to be bypassed. The first one concern the assumption about the symmetry of W, that should be released for real practical interest. The second concerns the assumption that W has a null diagonal. It remains to find an effective way to normalize the process when this assumption does not hold (the normalization used in Section 6 assume an autoregressive structure). Finally, our algorithm is greedy when the size of the graphs increases, and for large graphs, it would be really interesting to find a way to minimize the commutator, and to compute the significance of the variable without multiplying $N^2 \times N^2$ matrices.

Acknowledgement

The authors would like to thank the Universidad de la Habana (Cuba) and the Centro de Modelamiento Matematico (Chile) for their hospitality. We would like to thank Dieter Mitsche for fruitful discussions.

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 $\mathsf{A} \in \mathscr{E}(S)$,

$$\mathsf{BA} = \mathsf{AB} \implies \operatorname{diag}(\mathsf{B}) \neq 0.$$

In other words, a support S is diagonally identifiable if almost every symmetric matrix A with support in S is uniquely determined 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 2 (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 9 *For all support $S \subseteq [1,N]^2$, the set of invertible matrices in $\mathscr{E}(S)$ is either empty or a dense open subset of $\mathscr{E}(S)$.*

The proof is straightforward when writing the determinant of $\mathsf{A} \in \mathscr{E}(S)$ as a polynomial in its entries. Observe that by this property, finding one invertible matrix A in $\mathscr{E}(S)$ guarantees that almost every matrix in $\mathscr{E}(S)$ 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 10 (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 $V_3, \dots, V_{N-1} \subset [1,N]$ such that $|V_k| = k$ and $G_S(V_k)$ is invertible for all $k = 3, \dots, N-1$.*
2. ***Sufficient condition:*** *If there exists a nested sequence $V_3 \subset \dots \subset V_{N-1} \subset [1,N]$ with $|V_k| = k$ such that $G_S(V_k)$ is invertible for all $k = 3, \dots, N-1$, then S is diagonally identifiable.*

The gap between the sufficient and necessary conditions lies essentially in the fact that the sequence $V_3, \dots, V_{N-1}$ 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 subset $V_k \subset [1,N]$ of size k. For $\mathsf{A} \in \mathscr{E}(S)$, denote by $\psi_0(\mathsf{A}), \psi_1(\mathsf{A}), \dots, \psi_N(\mathsf{A})$ 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 $M_k(\mathsf{A}) := \sum_{j=0}^{k} \psi_j(\mathsf{A})\, A^j$. By Eq. (14) in [ER15], 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. Thus, the condition that $G_S(V_k)$ is not invertible for all subset V_k of size k 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 $\mathsf{A} \in \mathscr{E}(S)$. 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 11 *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 A',

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

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

$$\mathsf{A} = \begin{bmatrix} \mathsf{A}' & a \\ a^T & 0 \end{bmatrix}$$

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

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

Since M' is invertible by assumption, $\alpha = 0$ 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 10. Assume that G_S is not diagonally identifiable, then by Lemma 11, 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. □

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 (possible not induced) subgraph. This property is given in Proposition 2. The proof is a direct consequence of Theorem 10, considering the nested sequence $V_{N-1} \supset \ldots \supset V_3$ obtained by removing the end vertex of the kite at each step.

A.2 Proof of Proposition 3

The condition of 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 Proposition 3. 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 [Bol98] for instance, one knows that $\mathbb{P}\{G_n^{(1)} \text{ is hamiltonian}\}$ 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.3 Proof of Theorem 4

Combining Proposition 3 and Theorem 10, we deduce the first point. In view of the first point of Theorem 10, 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 [Bol98] for instance) that the event "there is at least two isolated points" has sharp threshold function $\log n/n$. It proves the second point.

B Support reconstruction

B.1 Proof of Theorem 5

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

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

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

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

Thus, the quantity $\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|_2$ for $\mathsf{A} \in \mathscr{E}(\overline{F})$ can be bounded from below and above by

$$\|\mathsf{AK} - \mathsf{KA}\|_2 - 2\|\mathsf{A}\|_2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2 \leq \|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|_2 \leq \|\mathsf{AK} - \mathsf{KA}\|_2 + 2\|\mathsf{A}\|_2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2. \tag{3}$$

To bound the term $\mathbb{P}\{\widehat{S} \in \mathscr{S}_1\}$, we use (3) to remark that for all $S \in \mathscr{S}_1$,

$$Q(S) = \min_{\mathsf{A} \in \mathscr{E}(S)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|_2}{\|\mathsf{A}\|_2} + \lambda_n|S| \geq \min_{\mathsf{A} \in \mathscr{E}(S)\setminus\{0\}} \frac{\|\mathsf{AK} - \mathsf{KA}\|_2}{\|\mathsf{A}\|_2} - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2 .$$

It follows

$$\min_{S \in \mathscr{S}_1} Q(S) \geq \min_{S \in \mathscr{S}_1} \min_{\mathsf{A} \in \mathscr{E}(S)\setminus\{0\}} \frac{\|\mathsf{AK} - \mathsf{KA}\|_2}{\|\mathsf{A}\|_2} - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2 = c_0 - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2. \tag{4}$$

The constant c_0 is positive by F-identifiability of W. Moreover, observe that

$$Q(S^\star) = \min_{\mathsf{A} \in \mathscr{E}(S^\star)\setminus\{0\}} \frac{\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\|_2}{\|\mathsf{A}\|_2} + \lambda_n|S^\star| \leq \frac{\|\mathsf{W}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{W}\|_2}{\|\mathsf{W}\|_2} + \lambda_n|S^\star| \leq 2\|\widehat{\mathsf{K}} - \mathsf{K}\|_2 + \lambda_n|S^\star|, \tag{5}$$

where we used both Eq. (3) and the fact that $\mathsf{WK} - \mathsf{KW} = 0$. Combining (4) and (5), we get

$$\mathbb{P}\{\widehat{S} \in \mathscr{S}_1\} \leq \mathbb{P}\Big\{\min_{S \in \mathscr{S}_1} Q(S) \leq Q(S^\star)\Big\} \leq \mathbb{P}\Big\{\|\widehat{\mathsf{K}} - \mathsf{K}\|_2 \geq \frac{c_0 - \lambda_n|S^\star|}{4}\Big\}.$$

To control the term $\mathbb{P}(\widehat{S} \in \mathscr{S}_2)$, we use that $\min_{S \in \mathscr{S}_2} Q(S) \geq \lambda_n \min_{S \in \mathscr{S}_2} |S| \geq \lambda_n(|S^\star| + 1)$. By Eq. (5), it follows

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

The proof of Theorem 5 follows directly by ($\mathbf{H}_2$). The corollary is a direct consequence using Borel-Cantelli's Lemma.

B.2 Proof of Theorem 7

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. (1) with probability tending to one asymptotically. Since the value of $\widehat{\beta}_S$ does not depend on $a_0 \in \mathscr{A}_S$, 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

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

B.3 Proof of Proposition 8

For a support S such that $S^* \subseteq S$, remark that $\mathscr{A}_S = \{w - \Phi_S b : b \in \mathbb{R}^{\dim(\mathscr{A}_S)}\}$. Thus, by least square minimization, we get

$$\min_{a \in \mathscr{A}_S} \|\Delta(\widehat{\mathsf{K}})a\|^2 = \min_{b \in \mathbb{R}^{\dim(\mathscr{A}_S)}} \|\Delta(\widehat{\mathsf{K}})(w - \Phi_S b)\|^2 = \|(\mathrm{I} - \Pi_{\mathrm{Im}(\Delta(\widehat{\mathsf{K}})\Phi_S)})\Delta(\widehat{\mathsf{K}})w\|^2.$$

Let $\widehat{E}_{S',S} = \mathrm{Im}(\Delta(\widehat{\mathsf{K}})\Phi_S) \cap \mathrm{Im}(\Delta(\widehat{\mathsf{K}})\Phi_{S'})^\perp = \mathrm{Im}(\Delta(\widehat{\mathsf{K}})\Phi_S) \cap \ker(\Phi_{S'}^\top \Delta(\widehat{\mathsf{K}}))$, we obtain

$$\min_{a \in \mathscr{A}_{S'}} \|\Delta(\widehat{\mathsf{K}})a\|^2 - \min_{a \in \mathscr{A}_S} \|\Delta(\widehat{\mathsf{K}})a\|^2 = \|\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2.$$

Write $\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w = \Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w + (\Pi_{\widehat{E}_{S',S}} - \Pi_{E_{S',S}})\Delta(\widehat{\mathsf{K}})w$. We have by assumption

$$\sqrt{n}\, \Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w = -\sqrt{n}\, \Pi_{E_{S',S}} \Delta(\mathsf{W})(\widehat{k} - k) \xrightarrow{d} \mathscr{N}\big(0, \Gamma_{S',S}\big).$$

It follows that $n\|\Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 \xrightarrow{d} \chi^2(\Gamma_{S',S})$. So, to prove the result, it remains to show that the approximation of $\Pi_{\widehat{E}_{S',S}}$ instead of $\Pi_{E_{S',S}}$ is negligible with

$$n\big(\|\Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 - \|\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2\big) = o_P(1).$$

Write

$$\begin{aligned}\|\Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 - \|\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 &\leq \|(\Pi_{E_{S',S}} - \Pi_{\widehat{E}_{S',S}})\Delta(\widehat{\mathsf{K}})w\| \left(\|\Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w\| + \|\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w\|\right) \\ &\leq |||\Pi_{E_{S',S}} - \Pi_{\widehat{E}_{S',S}}||| \times 2\|\Delta(\widehat{\mathsf{K}})w\|^2,\end{aligned}$$

where $|||.|||$ denotes the operator norm. Because S and S' both contain the true model $S^\star$, $\Delta(K)\Phi_{S'}$ and $\Delta(K)\Phi_S$ are of full rank and $\Pi_{\widehat{E}_{S',S}}$ converges in probability to $\Pi_{E_{S',S}}$. Since $\sqrt{n}\|\Delta(\widehat{\mathsf{K}})w\| = O_P(1)$, it follows that $\|\Pi_{E_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 - \|\Pi_{\widehat{E}_{S',S}} \Delta(\widehat{\mathsf{K}})w\|^2 = o_P(n^{-1})$, which ends the proof.

References

[BDCER14] 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.

[Bol98] Béla Bollobás. *Random graphs*. Springer, 1998.

[BPR16] Flavia Barsotti, Anne Philippe, and Paul Rochet. Hypothesis testing for markovian models with random time observations. *Journal of Statistical Planning and Inference*, page to appear, 2016.

[ER15] Thibault Espinasse and Paul Rochet. On relations between connected and self-avoiding walks on a graph. *arXiv preprint arXiv:1505.05725v2*, 2015.