Gaussian Graphical Model exploration and selection in high dimension low sample size setting

Thomas Lartigue, Simona Bottani, Stephanie Baron, Olivier Colliot, Stanley Durrleman, Stéphanie Allassonnière

Gaussian Graphical Model exploration and selection in high dimension low sample size setting

Thomas Lartigue, Simona Bottani, Stéphanie Baron, Olivier Colliot, Stanley Durrleman, Stéphanie Allassonnière for the Alzheimer's Disease Neuroimaging Initiative

Abstract—Gaussian Graphical Models (GGM) are often used to describe the conditional correlations between the components of a random vector. In this article, we compare two families of GGM inference methods: the nodewise approach of [1] and [2] and the penalised likelihood maximisation of [3] and [4]. We demonstrate on synthetic data that, when the sample size is small, the two methods produce graphs with either too few or too many edges when compared to the real one. As a result, we propose a composite procedure that explores a family of graphs with a nodewise numerical scheme and selects a candidate among them with an overall likelihood criterion. We demonstrate that, when the number of observations is small, this selection method yields graphs closer to the truth and corresponding to distributions with better KL divergence with regards to the real distribution than the other two. Finally, we show the interest of our algorithm on two concrete cases: first on brain imaging data, then on biological nephrology data. In both cases our results are more in line with current knowledge in each field.

1 Introduction

Dependency networks are a prominent tool for the representation and interpretation of many data types as, for example, gene co-expression [2], interactions between different regions of the cortex [5] or population dynamics. In those examples, the number of observations n is often small when compared to the number of vertices p in the network.

Conditional correlation networks are graphs where there exists an edge between two vertices if and only if the random variables on these nodes are correlated conditionally to all others. This structure can be more interesting than a regular correlation graph. Indeed, in real life, two phenomena, like the atrophy in two separate areas of the brain or two locations of bird migration, are very likely to be correlated. There almost always exists a "chain" of correlated events that "link", ever so slightly, any two occurrences. As a result, regular correlation networks tend to be fully connected and mostly uninformative. On the other hand, when intermediary variables explain the totality of the co-variations of two vertices, then these two are conditionally uncorrelated, removing their edge from the conditional correlation graph. The conditional correlation structure captures only the direct, explicit interactions between vertices. In our analyses, these interactions are the ones of most interest.

A Gaussian Graphical Model (GGM) is a network whose values on the p vertices follow a Centred Multivariate Normal distribution in $\mathbb{R}^p$: $X \sim \mathcal{N}(0_p, \Sigma)$. This assumption is almost systematic when studying conditional correlation networks for three main reasons. First, it ensures that each conditional correlation $corr(X_i, X_j|(X_k)_{k\neq i,j})$ is a constant and not a function of the $p-2$ dimensional variable $(X_k)_{k\neq i,j}$; a crucial property allowing us to talk about a single graph and not a function graph. Second, it equates the notions of independence and un-correlation, in particular: $corr(X_i, X_j|(X_k)_{k\neq i,j}) = 0 \iff X_i \perp\!\!\!\perp X_j|(X_k)_{k\neq i,j}$. This makes interpretation much clearer. Finally, under the GGM assumption, we have the explicit formula: $corr(X_i, X_j|(X_k)_{k\neq i,j}) = -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}$, where $K := \Sigma^{-1}$ is the inverse of the unknown covariance matrix. This means that the conditional correlations graph between the components of X is entirely described by a single matrix parameter, K. Moreover the graph and K have the exact same sparsity structure. With this property in mind, the author of [6] introduced the idea of Covariance Selection which consists of inferring - under a Gaussian assumption - a sparse estimation $\widehat{K}$ of K and interpreting its sparsity structure as a conditional dependency network.

Subsequently, many authors have proposed their own

- *T. Lartigue is with the CMAP, CNRS, École polytechnique and Aramis project-team, Inria. E-mail: thomas.lartigue@inria.fr*
- *S. Bottani is with the Aramis project-team, Inria, Institut du Cerveau et de la Moelle épinière, Sorbonne University, Inserm U1127, CNRS UMR 7225. E-mail: simona.bottani@icm-institute.org*
- *S. Baron is with Hôpital Européen Georges-Pompidou AP-HP. E-mail: stephanie.baron@aphp.fr*
- *O. Colliot and S. Durrleman are with the Aramis project-team, Inria, Institut du Cerveau et de la Moelle épinière, Sorbonne University, Inserm U1127, CNRS UMR 7225.*
  E-mail: olivier.colliot@upmc.fr & stanley.durrleman@inria.fr
- *S. Allassonnière is with the Centre de Recherche des Cordeliers, Université de Paris, INSERM, Sorbonne Université. E-mail: stephanie.allassonniere@parisdescartes.fr*

- *Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*

estimators $\widehat{K}$. In [1], a local edge selection approach that solves a LASSO problem on each node is introduced. It was noticeably followed by [2], [7], who developed the GGMselect algorithm, a practical implementation of this approach coupled with a model selection procedure. We call these methods "local", since they focus on solving problems independently at each node, and evaluating performances with an aggregation of nodewise metrics. Other works within the local paradigm have proposed Dantzing selectors [8], constrained l_1 minimisation [9], scaled LASSO [10], or merging all linear regression into a single problem [11]. On a different note, the authors of [3] and [4] considered a more global paradigm where the estimator is solution of a single l_1-penalised log-likelihood optimisation problem, that has the form of Eq. (1).

$$\widehat{K} := \underset{\widetilde{K} \succ 0}{\operatorname{argmax}} \mathcal{L}\left(\widetilde{K}\right) - \rho \sum_{i<j} \left|\widetilde{K}_{ij}\right| . \tag{1}$$

We call this point of view "global" since the likelihood estimates at once the goodness of fit of the whole proposed matrix. The introduction of problem (1) generated tremendous interest in the GGM community, and in its wake, many authors developed their own numerical methods to compute its solution efficiently. A few notable examples are block coordinate descent for the Graphical Lasso algorithm (GLASSO) of [12], Nesterov's Smooth gradient methods [13], Interior Point Methods (IPM) [14], Alternating Direction Methods of Multipliers (ADMM) [15], [16], Newton-CG primal proximal point [17], Newton's method with sparse approximation [18], Projected Subgradient Methods (PSM) [19], and multiple QP problems for the DP-GLASSO algorithm of [20]. The theoretical properties of the solutions to Eq. (1) are studied in [21], [22] and in [23]. Other methods within the global paradigm include [24], with penalties other than l_1 in (1), and [25], with a RKHS estimator.
More recent works have proposed more involved estimators, defined as modifications of already existing solutions and possessing improved statistical properties, such as asymptotic normality or better element-wise convergence. The authors of [26] and [27] adapted solutions of local regression problems including [1], whereas [28] modified the solutions of (1). In [29], the two approaches are unified with a de-biasing method applied to both local and global estimators.

In our applications - where the number of observations n is a fixed small number, usually smaller than the number of vertices p - we did not find satisfaction with the state of the art methods from either the local or the global approach. On one hand, GGMselect yields surprisingly too sparse graph, missing many of the important already known edges. On the other hand, the only solutions from the penalised likelihood problem (1) that are a decent fit for real distribution have so many edges that the information is hidden. To interpret a graph, one would prefer an intermediary number of edges. Additionally, the low sample size setting requires a method with non-asymptotic theoretical properties.

In this paper, we design a composite method, combining the respective strengths of the local and global approaches, with the aim of recovering graphs with a more reasonable amount of edges, that also achieves a better quantitative fit with the data. We also prove non-asymptotic oracle bounds in expectation and probability on the solution.

To measure the goodness of fit, many applications are interested in recovering the true graph structure and focus on the "sparsistency". In our case, the presence or absence of an edge is not sufficient information. The correlation amplitude is of equal interest. Additionally, we need the resulting structure to make sense as a whole, that is to say: describe a co-variation dynamic as close as possible to the real one despite being a sparse approximation. This means that edgewise coefficient recovery - as assessed by the l_2 error $\left\|K - \widehat{K}\right\|_F^2 = \sum_{i,j}(K_{i,j} - \widehat{K}_{i,j})^2$ for instance - which does not take into account the geometric structure of the graph as a whole is not satisfactory either. We want the distribution function described by the proposed matrix to be similar to the original distribution. The natural metric to describe proximity between distribution functions is Cross Entropy (CE) or, equivalently, the Kullback-Leibler divergence (KL). In the end, the CE between the original distribution and the proposed one - $\mathcal{N}\left(0, \widehat{K}^{-1}\right)$ - is our metric of choice. Other works, such as [30] and [31], have focused on the KL in the context of GGM as well.

In the following, we quantify the shortcomings of the literature's local and global methods when the data is not abundant. The GGMselect graphs are very sparse, but consistently and substantially outperform the solutions of Eq. (1) in terms of KL, regardless of the penalisation intensity ρ. In the KL/sparsity space, the solutions of GGMselect occupy a spot of high performing, very sparse solutions that the problem (1) simply does not reach. Additionally, the better performing solutions of (1) are so dense that they are excessively difficult to read. Subsequently, we demonstrate that despite its apparent success, the GGMselect algorithm is held back by its model selection criterion which is far too conservative and interrupts the graph exploration process too early. This results in graphs that are not only difficult to interpret but also perform sub-optimally in terms of KL.
With those observations in mind, we design a simple nodewise exploration numerical scheme which, when initialised at the GGMselect solution, is able to extract a family of larger, better performing graphs. We couple this exploration process with a KL-based model selection criterion to identify the best candidates among this family. This algorithm is composite insofar as it combines a careful local graph construction process with a perceptive global evaluation of the encountered graphs.
We prove non-asymptotic guarantees on the solution of the model selection procedure. We demonstrate with experiments on synthetic data that this selection procedure satisfies our stated goals. Indeed, the selected graphs are both substantially better in terms of distribution reconstruction (KL divergence), and much closer to the original graph than any other we obtain with the state of the art methods. Then, we put our method to the test with two

experiments on real medical data. First on a neurological dataset with multiple modalities of brain imaging data, where $n < p$. Then on biological measures taken from healthy nephrology test subjects, with $p < n$. In both cases, the results of our method correspond more to the common understanding of the phenomena in their respective fields.

2 Covariance Selection within GGM

2.1 Introduction to Gaussian Graphical Models

Let S_p^+ and S_p^{++} be respectively the spaces of positive semi-definite and positive definite matrices in $\mathbb{R}^{p\times p}$. We model a phenomenon as a centred multivariate normal distribution in $\mathbb{R}^p$: $X \sim \mathcal{N}(0_p, \Sigma)$. To estimate the unknown covariance matrix $\Sigma \in S_p^{++}$, we have at our disposal an iid sample $\left(X^{(1)}, ..., X^{(n)}\right)$ assumed to be drawn from this distribution. We want our estimation to bring interpretation on the conditional correlations network between the components of X. No real network is truly sparse, yet it is natural to propose a sparse approximation. Indeed, this means recovering in priority the strongest direct connections and privileging a simpler explanation of the phenomenon, one we can hope to infer even with a small amount of data. Sparsity in the conditional correlations structure is equivalent to sparsity in the inverse covariance matrix $K := \Sigma^{-1}$. Namely $K_{ij} = 0 \iff \text{Corr}\left(X_i, X_j | (X_k)_{k\neq i,j}\right) = 0$. As a consequence, our goal is to estimate from the dataset a covariance matrix $\widehat{\Sigma} \in S_p^{++}$ with both a good fit and a sparse inverse $\widehat{K}$. We say that $\widehat{\Sigma} := \widehat{K}^{-1}$ is "inverse-sparse".

In the following, we use the Cross Entropy to quantify the performances of a proposed matrix $\widehat{K}$. The CE, $H(p,q) = -\mathbb{E}_p\left[\log q(X)\right] = \int_x -p(x) ln(q(x)) \mu(\mathrm{d}x)$, is an asymmetric measure of the deviation of distribution q with regards to distribution p. The CE differs from the KL-divergence only by the term $H(p,p)$, which is constant when the reference distribution p is fixed. In GGM, the score $H(f_\Sigma, f_{\widehat{\Sigma}})$ represents how well the normal distribution with our proposed covariance $\widehat{\Sigma}$ is able to reproduce the true distribution $\mathcal{N}(0,\Sigma)$. We call this score the True CE of $\widehat{\Sigma}$. This metric represents a global paradigm where we explicitly care about the behaviour of the matrix as a whole. This is in contrast to a coefficient-wise recovery, for instance, which is a summation of local, nodewise, metrics. After removal of the additive constants, we get the simple formula (2) for the CE between two centred multivariate normal distributions $\mathcal{N}(0,\Sigma_1)$ and $\mathcal{N}(0,\Sigma_2)$.

$$H(\Sigma_1, \Sigma_2) := H(f_{\Sigma_1}, f_{\Sigma_2}) \equiv \frac{1}{2}\left(tr(\Sigma_1 K_2) - ln(|K_2|)\right). \tag{2}$$

In the general case, the CE between a proposed distribution f_θ and an empirical distribution $\hat{f}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{x=X^{(i)}}$ defined from data is the opposite of the log-likelihood: $H(\hat{f}_n, f_\theta) = -\frac{1}{n}\log p_\theta(X^{(1)}, ..., X^{(n)})$. In the GGM case, we denote the observed data $\underline{X} := \left(X^{(1)}, ..., X^{(n)}\right)^T \in \mathbb{R}^{n\times p}$, and set $S := \frac{1}{n}\underline{X}^T\underline{X} \in S_p^+$, the empirical covariance matrix. The opposite log-likelihood of any centred Gaussian $\mathcal{N}(0,\Sigma_2)$ satisfies:

$$H(S, \Sigma_2) := H\left(\hat{f}_n, f_{\Sigma_2}\right) \equiv \frac{1}{2}\left(tr(SK_2) - ln(|K_2|)\right), \tag{3}$$

similar to Eq. (2). As a result, we adopt an unified notation. Details on calculations to obtain these formulas can be found in Section 7.1.

We use the following notations for matrix algebra, let A be a square real matrix, then: $|A|$ denotes the determinant, $\|A\|_* := tr\left(\left(A^TA\right)^{\frac{1}{2}}\right)$ the nuclear norm, $\|A\|_F := tr\left(\left(A^TA\right)\right)^{\frac{1}{2}} = \left(\sum_{i,j} A_{ij}^2\right)^{\frac{1}{2}}$ the Frobenius norm and $\|A\|_2 := \sup_x \frac{\|Ax\|_2}{\|x\|_2} = \lambda_{max}(A)$ the spectral norm (operator norm 2) which is also the highest eigenvalue. We recall that when A is symmetrical positive, then $\|A\|_* = tr(A)$ and $\|A\|_F = tr(A^2)^{\frac{1}{2}}$. We also consider the scalar product $\langle A, B\rangle := tr\left(B^TA\right)$ on $\mathbb{R}^{p\times p}$.

2.2 Description of the state of the art

After its introduction, problem (1) became the most popular method to infer graphs from data with a GGM assumption. Reducing the whole inference process to a single loss optimisation is convenient. What is more, the optimised loss is a penalised version of the likelihood - which is an estimator of the True CE - hence the method explicitly takes into account the global performances of the solution. However, even though the l_1 penalty mechanically induces sparsity in the solution, it does not necessarily recover the edges that best reproduce the original distribution, especially when the data is limited. Indeed, the known "sparsitency" dynamics of the solutions of (1), see [22], always involve a large number of observations tending towards infinity. We demonstrate in this paper that, when the sample size is small, other methods recover consequently more efficient sparse structures, inaccessible to the l_1 penalised problem (1).
On the other hand, the local approach of [1] carefully assesses each new edge, focusing on making the most efficient choice at each step. We confirm that the latter approach yields better performance by comparing the solutions of problem (1) and GGMselect [2] on both synthetic and real data (Sections 4 and 5). However, the loss optimised in GGMselect, $Crit(\mathcal{G})$, see (4), is an amalgam of local nodewise regression score, with no explicit regard for the overall behaviour of the matrix:

$$Crit(\mathcal{G}) := \sum_{a=1}^p \left[\left\|X_a - \underline{X}\left[\widehat{\theta}_{\mathcal{G}}\right]_a\right\|_2^2 \left(1 + \frac{pen(d_a(\mathcal{G}))}{n - d_a(\mathcal{G})}\right)\right], \tag{4}$$

where pen is a specific penalty function, $d_a(\mathcal{G})$ is the degree of the node a in the graph $\mathcal{G}$, X_a are all the observed values at node a, such that $\underline{X} = (X_1, ..., X_p) \in \mathbb{R}^{n\times p}$ is the full

data, and:

$$\begin{aligned}\hat{\theta}_{\mathcal{G}} &:= \underset{\theta\in\Lambda_{\mathcal{G}}}{\operatorname{argmin}} \|\underline{X}(I_p-\theta)\|_F^2 \\ &= \underset{\theta\in\Lambda_{\mathcal{G}}}{\operatorname{argmin}} \sum_{a=1}^{p} \|X_a - \underline{X}\,[\theta]_a\|_2^2 \\ &= \left\{ \underset{\theta_a\in\Lambda_{\mathcal{G}}^a}{\operatorname{argmin}} \|X_a - \underline{X}\theta_a\|_2^2 \right\}_{a=1}^{p},\end{aligned} \tag{5}$$

where $\Lambda_{\mathcal{G}}$ is the set of $p\times p$ matrices θ such that $\theta_{i,j}$ is non zero if and only if the edge (i,j) is in $\mathcal{G}$, and $\Lambda_{\mathcal{G}}^a$ is the set of vectors $\theta_a \in \mathbb{R}^p$ such that $(\theta_a)_i$ is non zero if and only if the edge (i,a) is in $\mathcal{G}$. Note that by convention, auto-edges (i,i) are never in the graph $\mathcal{G}$, and, in our work, $\mathcal{G}$ is always undirected. The full expression of pen can be found in Eq. 3 of [2]. It depends on a dimensionless hyper-parameter called K which the authors recommend to set equal to 2.5. We first tried other values without observing significant change, and decided to use the recommended value in every later experiment.

The expression (5) illustrates that each nodewise coefficients $\left[\hat{\theta}_{\mathcal{G}}\right]_a$ in the GGMselect loss are obtained from independent optimisation problems which each involve only the local sparsity of the graph in the vicinity of the node a, as seen in the definition of $\Lambda_{\mathcal{G}}^a$. In each parallel optimisation problem $\underset{\theta_a\in\Lambda_{\mathcal{G}}^a}{\operatorname{argmin}} \|X_a - \underline{X}\theta_a\|_2^2$, the rest of the graph is not constrained, hence is implicitly fully connected. In particular, the solutions of such problems involve an estimation of the covariance matrix between the rest of the vertices that is not inverse-sparse. This can bias the procedure towards the sparser graphs since it actually implicitly measures the performances of more connected graphs. Finally, the GGMselect model selection criterion (GGMSC) explicitly penalises the degree of each node in the graph making it so that string-like structures are preferred over hubs. Empirically, we observe that with low amounts of data, graphs with hubs are consistently dismissed by the GGMSC. Overall, we expect the selected solutions to be excessively sparse, which experiments on both synthetic and real data in Sections 4 and 5 confirm.

2.3 Graph constrained MLE

Even though a covariance matrix Σ uniquely defines a graph with its inverse K, the reciprocal is not true. To a given graph $\mathcal{G} := (V,E)$, with vertex set V and edge set E, corresponds a whole subset $\Theta_{\mathcal{G}}$ of S_p^{++}:

$$\Theta_{\mathcal{G}} := \left\{ \widetilde{\Sigma}\in S_p^{++} \middle| \forall i\neq j,\ (i,j)\notin E \Rightarrow \left(\widetilde{\Sigma}^{-1}\right)_{ij} = 0 \right\}.$$

When data is available, the natural matrix representing $\mathcal{G}$ is the constrained MLE:

$$\widehat{\Sigma}_{\mathcal{G}} := \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmax}}\, p_{\widetilde{\Sigma}}(X^{(1)},...,X^{(n)}) = \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S,\widetilde{\Sigma}\right). \tag{6}$$

The existence of the MLE is not always guaranteed (see [6], [32]). When $n<p$, no MLE exists for the more connected graphs. However, in this paper, we design a procedure that can propose a MLE for any n and any graph without computation errors. To tackle the issue of existence, we add a very small regularisation term to the empirical covariance matrix S. This leads to solving:

$$\widehat{\Sigma}_{\mathcal{G},\lambda} := \underset{\widetilde{\Sigma}\in\Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S+\lambda I_p,\widetilde{\Sigma}\right). \tag{7}$$

λ is not a true hyper parameter of the model. Its value is set once and for all, and as small as possible as long as the machine still recognises $S+\lambda I_p$ as invertible. Typical values range between 10^{-7} and 10^{-4}. This trick changes little for the already existing solutions. Indeed, if $\widehat{\Sigma}_{\mathcal{G}}$ solution of Eq. (6) exists, we observe empirically that for small values of λ: $\widehat{\Sigma}_{\mathcal{G}} \simeq \widehat{\Sigma}_{\mathcal{G},\lambda}$. On the other hand, if no solution $\widehat{\Sigma}_{\mathcal{G}}$ to Eq. (6) exists, then we now are able to propose a penalised MLE $\widehat{\Sigma}_{\mathcal{G},\lambda}$, thus avoiding degenerated computations. From now on, the MLE we use are always solutions of (7). We will omit the index λ and keep the notation $\widehat{\Sigma}_{\mathcal{G}}$ for the sake of simplicity.

2.4 Our composite algorithm

The exploration steps of our method are a variation of the local paradigm of [1]. First, we use the GGMselect solution as initialisation. Then we add edges one by one: at each step, for each vertex independently, we run a sparse linear regression using as predictors the vertices that are not among its neighbours yet, and as target the residual of the linear regression between the value on the vertex and its neighbours. With these regressions, each vertex proposes to add to the current graph an edge between them and their new best predictor. Here however, we deviate from the local paradigm by using a global criterion - the out of sample likelihood of the whole resulting new matrix - to evaluate each proposition and select one edge among these candidates. We end this exploration procedure after a fixed number of steps, the result is a family of gradually more connected graphs. The final selection step is done with a global metric: we pick, among the so constructed family, the graph minimising the Cross Validated (with fresh data) Cross Entropy. See Fig. 1 for the details.

In the spirit of [26], [27], [28], [29], this method is designed to complete an already existing efficient, but sparse, solution. As a result, it is sensitive to the initial graph.

3 Oracle bounds on the model selection procedure

In this Section, we give non-asymptotic guarantees on the model selection step of our algorithm. We prove these results in Section 7. Using the statistical properties of our model selection criterion, in particular the absence of bias and convergence towards the oracle criterion, we describe the difference between the performance of the selected model and the oracle best performance ("regret"). This regret is dependent on the convergence of a Wishart random variable towards its expectation. As a result, we are able to prove non-asymptotic upper bounds in expectation and probability for the regret.

Inputs: The *train* set are all the observations available for graph inference, Nb of steps T fixed in advance.
Start:
• Run GGMselect on the *train* set to get the initial graph $\mathcal{G}_0 = (V, E_0)$;
Partition the *train* set into a *validation* set and *exploration* set;
for $t = 1, .., T$ **do**
Partition randomly the *exploration* set into a *learning* set and an *evaluation* set;
Compute the empirical covariance S^t_{eval} from the *evaluation* set;
 # We then "ask" each node for its desired next neighbour:
for $a \in V$ vertex of $\mathcal{G}_{t-1}$ **do**
• Let $N_{t-1}(a)$ be the set of neighbours of a in $\mathcal{G}_{t-1}$ and $F_{t-1}(a) := V \setminus \{N_{t-1}(a) \cup \{a\}\}$ the remaining vertices;
• Run on the *learning* set the linear regression with the vector X_a of the values on a as the target, and the vectors $\{X_s | s \in N_{t-1}(a)\}$ on the neighbour nodes as predictors. Let $\tilde{X}_a$ be the residual of this regression;
• Run on the *learning* set one step of the LARS algorithm of [33], with $\tilde{X}_a$ as the target, and the remaining $\{X_s | s \in F_{t-1}(a)\}$ as predictors. Call $c_t(a) \in F_{t-1}(a)$ the index of the feature chosen by LARS;
end for
 # We now have p potential new edges $\{(a, c_t(a))\}_{a \in V}$ *some of which can be identical*
 # We give priority to mutual selections: when $c_t(c_t(a)) = a$
if $\{(a, c_t(a))\}_{c_t(c_t(a))=a} \neq \emptyset$ **then**
Let $\mathcal{C} = \{(a, c_t(a))\}_{c_t(c_t(a))=a}$ be our set of candidate edges;
 # We keep only the mutual selections
else
Let $\mathcal{C} = \{(a, c_t(a))\}_{a \in V}$;
 # No mutual selection ⇒ keep the whole set
end if
for $c \in \mathcal{C}$ **do**
Compute, with the *learning* set, the MLE $\widehat{\Sigma}^c_t$ from each new potential graph $\mathcal{G}^c_t := \mathcal{G}_{t-1} \cup c$;
end for
• $c^* := \underset{c \in \mathcal{C}}{\operatorname{argmin}} H\left(S^t_{eval}, \widehat{\Sigma}^c_t\right)$;
• $\mathcal{G}_t := \mathcal{G}^{c^*}_t$;
Compute, with the *exploration* set, the MLE $\widehat{\Sigma}_t$ from $\mathcal{G}_t$;
end for
Compute, with the *exploration* set, the MLE $\widehat{\Sigma}_0$ from $\mathcal{G}_0$;
Compute the empirical covariance S_{val} from the *validation* set;
• $t^* := \underset{t=0,\dots,T}{\operatorname{argmin}} H\left(S_{val}, \widehat{\Sigma}_t\right)$;
• $\widehat{\mathcal{G}} := \mathcal{G}_{t^*}$;
Return: Inferred graph $\widehat{\mathcal{G}}$.

Fig. 1. Composite GGM estimation. We respectively identify with green • or orange • bullets the steps adhering to a local or global paradigm. Comments are in *italics*.

3.1 Framework

In this Section we define or recall the relevant concepts and notations. We recall and rephrase the definition, given in Eq. (7), of the constrained Maximum Likelihood Estimator we build from a given graph $\mathcal{G}$:

$$\begin{aligned}\widehat{\Sigma}_{\mathcal{G}}(S) &= \underset{\widetilde{\Sigma} \in \Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S + \lambda I_p, \widetilde{\Sigma}\right) \\ &= \underset{\widetilde{\Sigma} \in \Theta_{\mathcal{G}}}{\operatorname{argmin}} H\left(S, \widetilde{\Sigma}\right) + \frac{\lambda}{2}\left\|\widetilde{K}\right\|_* .\end{aligned}$$

We use the Cross Validated Cross Entropy (CVCE) $H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ as a criterion to pick a graph $\widehat{\mathcal{G}}_{CV}$ among the ones encountered. This Cross Validated criterion uses the partition of the *training* set into a *validation* set - used to build the estimation S_{val} of the true matrix Σ - and an *exploration* set - used for the graph exploration process and to build the constrained MLE $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$ for each encountered graph $\mathcal{G}$. We compare the graph $\widehat{\mathcal{G}}_{CV}$ selected with CVCE with $\widehat{\mathcal{G}}^*$ selected with the True Cross Entropy $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ of the matrix $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$. We define formally those graphs: in Eq. (8) and (9):

$$\widehat{\mathcal{G}}^* \in \underset{\mathcal{G} \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)\right], \tag{8}$$

$$\widehat{\mathcal{G}}_{CV} \in \underset{\mathcal{G} \in \mathcal{M}}{\operatorname{argmin}} \left[H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)\right], \tag{9}$$

where we call $\mathcal{M}$ the family of graphs uncovered by the Composite algorithm.

Remark With the data available, the ideal model selection would be made with True Cross Entropy $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right)$ of the matrix $\widehat{\Sigma}_{\mathcal{G}}(S_{train})$ built from the whole *train* set. Comparing ourselves to this criterion would allow to quantify the importance of having a balanced split between *validation* and *exploration* set. This is outside the scope of this Section. We just compare our $H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$ to $H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{expl})\right)$. In this case, the convergence of S_{val} towards Σ is the only dynamic that matters.

3.2 Basic control

In this Section, we show a general upper bound on the regret, using only the properties of the model selection criterion, and not yet the properties of the estimators. From this point on, we generally do not highlight the dependency of $\widehat{\Sigma}_{\mathcal{G}}$ in S_{expl} to simplify notation. First of all, note that by definition we always have the lower bound on the difference of CE:

$$0 \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) .$$

The rest of the guarantees focus on the upper bounds for this difference.
From the observation that $H\left(\Sigma, \widehat{\Sigma}\right) = H\left(S, \widehat{\Sigma}\right) + \frac{1}{2}\left\langle \Sigma - S, \widehat{K}\right\rangle$, we get the control (10) on the regret $H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)$:

$$H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) \leq \frac{1}{2}\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}} - \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle, \tag{10}$$

where all the MLE $\widehat{\Sigma}_{\mathcal{G}}$ depend only on $\mathcal{G}$ and S_{expl}. The random variable $\widehat{\mathcal{G}}^*$ is a function of S_{expl} only, whereas $\widehat{\mathcal{G}}_{CV}$ depends on both S_{val} and S_{expl}. Since S_{val} and S_{expl} are independent, then:

$$\mathbb{E}\left[\left\langle S_{val}, \widehat{K}_{\widehat{\mathcal{G}}^*}(S_{expl})\right\rangle \middle| S_{expl}\right] = \left\langle \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}(S_{expl})\right\rangle .$$

In the end, with $e := \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right]$ the expected regret, we have:

$$0 \leq e \leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right] . \tag{11}$$

3.3 Control in expectation

In this Section, we use the sparsity properties of the estimator $\widehat{K}_{\widehat{\mathcal{G}}_{CV}}$ as well as the statistical properties of $\Sigma - S_{val}$ to obtain a more explicit control on the expected regret. In addition, we use a known concentration result to obtain an alternative control in expectation. The result (11) is completely agnostic of the way the matrices $\widehat{K}_{\mathcal{G}} \in S_p^{++}$ are defined as long as they depend on S_{expl} only. To get an order of this control, however, we use the assumption that $\widehat{\Sigma}_{\mathcal{G}}$ is the graph constrained MLE defined in (7). Let us first notice that we can ensure $\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}$ thanks to our penalised definition of (7). Let $\Sigma_\infty := \max_{i,j} |\Sigma_{ij}|$. We call $E_{\max}$ the union of the maximal edge sets in $\mathcal{M}$, and $d_{\max} = |E_{max}| \leq \frac{p(p-1)}{2}$ its cardinal. We underline here that, by convention, conditional correlation graphs do not contain self loops, hence the edge sets E never include any of the pairs $\{(i,i)\}_{i=1,\dots,p}$. We then get the control (12) by using Cauchy-Schwartz's inequality in (11).

***Proposition*:** *With the previously introduced notations, if the set E_{max} is independent of the exploration empirical matrix S_{expl}, we have:*

$$0 \leq e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{(p + 2d_{max})^{\frac{1}{2}}\, p}{\sqrt{n_{val}}} . \tag{12}$$

In the case of our Composite procedure, by construction E_{max} is a random variable depending on the exploration set. However (12) still holds by replacing d_{max} with $\mathbb{E}[d_{max}]$:

$$0 \leq e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{(p + 2\mathbb{E}[d_{max}])^{\frac{1}{2}}\, p}{\sqrt{n_{val}}} . \tag{13}$$

We can get an alternative order of the control by using known concentrations inequalities.

***Proposition*:** *By using the Theorem 4 of [34], we get:*

$$0 \leq e \leq c\frac{\lambda_{max}(\Sigma)}{\lambda} p \left(\sqrt{\frac{p}{n_{val}}} \vee \frac{p}{n_{val}}\right) . \tag{14}$$

Where c is a constant independent of the problem.

In the end, with (13) and (14), we have two different upper bounds on e and can use the minimum one depending on the situation.

3.4 Control in probability

In this Section, we use the sparsity properties of the estimator $\widehat{K}_{\widehat{\mathcal{G}}_{CV}}$ as well as the concentration properties of $\Sigma - S_{val}$ around 0 to obtain a control in probability (concentration inequality) on the regret. In addition to the controls in expectation we got in (11) and (12), there is in the CVCE a concentration dynamic based on the convergence rate of a Wishart random matrix towards its average. We call $\Pi_{\max}$ the orthogonal projection on the set of edges $E_{\max} \cup \{(i,i)\}_{i=1}^p$. That is to say, for any matrix $M \in \mathbb{R}^{p\times p}$, $\Pi_{\max}(M)_{i,j} = M_{i,j}\mathbb{1}_{(i,j)\in E_{\max}\cup\{(i,i)\}_{i=1}^p}$. Let $W := K^{\frac{1}{2}} S_{val} K^{\frac{1}{2}}$. Then $n_{val}\, W \sim \mathcal{W}_p(n_{val}, I_p)$ is a standard Wishart random variable depending only on the *validation* data, hence independent of every matrix $\widehat{K}_{\mathcal{G}}$. Let $P := \mathbb{P}\left(\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\right)$ be the probability that the regret is small. We get two different lower bounds (15) and (16) on P.

***Proposition*:** *With the previously introduced notations, the two following inequalities hold:*

$$P \geq \mathbb{P}\left(\|W - I_p\|_F \leq \frac{\delta}{\max_{\mathcal{G}} \left\|\Sigma^{\frac{1}{2}}\widehat{K}_{\mathcal{G}}\Sigma^{\frac{1}{2}}\right\|_F}\right) , \tag{15}$$

$$P \geq \mathbb{P}\left(\|\Pi_{\max}(S_{val} - \Sigma)\|_F \leq \frac{\delta}{\max_{\mathcal{G}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F}\right) . \tag{16}$$

Moreover, the results (15) *and* (16) *hold when every probability is taken conditionally to the exploration data or, equivalently here, conditionally to S_{expl}.*

If we work conditionally to the *exploration* data, then $\max_{\mathcal{G}} \left\|\Sigma^{\frac{1}{2}}\widehat{K}_{\mathcal{G}}\Sigma^{\frac{1}{2}}\right\|_F$, $\max_{\mathcal{G}} \left\|\widehat{K}_{\mathcal{G}}\right\|_F$ and E_{max} are constants of the problem. In that case, the lower bound in (15) only depends on the dynamic of a standard Wishart $\mathcal{W}_p(n_{val}, I_p)$. Similarly, the lower bound in (16) only depends on the convergence dynamic of some coefficients of S_{val} towards the corresponding ones in Σ.
The bound in (16) has a less general formulation than (15), since the $S_{val} \mapsto \Sigma$ is a more specific dynamic than $W \mapsto I_p$. On the other hand, only the diagonal coefficients and those in $E_{\max}$ need to be close, which can make a huge difference if p is very large and $\mathcal{M}$ contains only sparse graphs and make the bound (16) tighter.

4 Experiments on synthetic data

We show in this Section the shortcomings of the global problem (1) of [3] and [4] and of the local approach of [1] and [2] on synthetic data. We demonstrate that - when the data is not abundant - the solutions of GGMselect consistently reproduce the true distribution much better than any solution of the global problem (1). In addition to being outperformed in KL divergence, the best solutions of (1) are also very connected, consequently more than the real graph. However, we also illustrate that the solutions of GGMselect are always very sparse, regardless of the real graph. In the end, we demonstrate that our selection criterion improves both the distribution reproduction and the graph recovery of the previous two methods.

4.1 The solutions missed by the global paradigm: a comparison of GLASSO and GGMselect

We start by comparing the two state of the art global and local paradigms, and show that the global paradigm misses crucial solutions when the number of observations is small. We use the scikit learn, see [35], implementation of the GLASSO of [12] to solve problem (1) for any penalisation level ρ and the R implementation of GGMselect, see [2], to represent the [1] approach.

We use an inverse-sparse covariance matrix Σ fixed once and for all to generate a matrix of observations $\underline{X}$. The same observations are provided to the two methods. On Fig. 2, we compare the True CE $H(\Sigma, \widehat{\Sigma})$ of each estimated matrix as a function of the number of non-zero, off diagonal coefficients in their inverse $\widehat{K}$ (complexity of the model). The green dot is the MLE - computed as in (7) - under the constraints of the GGMselect graph. In the case of GLASSO, different solutions are obtained by changing the level of penalisation ρ in Eq. (1). We call those solutions $\widehat{\Sigma}_{\rho}$, indexed by their penalisation intensity ρ. They are represented by the blue curve on Fig. 2. All of them are inverse-sparse and define a graph we call $\mathcal{G}(\rho)$. The orange curve is the path of the MLEs $\widehat{\Sigma}_{\mathcal{G}(\rho)}$ - computed as in (7) - refitted from those same graphs without the l_1 penalty of problem (1). They have the same inverse-sparsity as their raw solution counterparts, but do not have the extra-penalisation on the non-zero coefficients that every LASSO solution bears.

The three columns correspond to graphs with different connectivity - illustrated by a random example on top of each column - and the two rows have different graph sizes, $p = 30$ and $p = 50$ respectively. For each simulation, the two methods were given the same $n = 30$ observations to work with, and each figure represents the average and standard deviation of 100 simulations.

We notice that the GGMselect solution is always very sparse. When the true graph is sparse, GGMselect outperforms the penalised likelihood problem (1) regardless of the penalty intensity. For large connected graphs, the most connected solutions of (1) can perform better than the GGMselect solution. However GGMselect is consistently better than the equally sparse problem (1) solution. The failure of GLASSO to reach the spot of GGMselect in the performances/complexity with any penalisation intensity - even when the MLE is refitted from the GLASSO graph without penalty - indicates that when n is small, the l_1 penalised likelihood problem (1) has difficulties selecting the most efficient edges. Additionally, the better performing solutions of GLASSO have many edges - usually much more than the real graph - which draws the focus away from the relevant ones and makes it difficult to get a qualitative reading of the graph.

When the number of observations is small, it seems that GGMselect's numerical scheme allows it to find high performing sparse graphs that problem (1) never can. This is the type of solution we want, and the main reason why we choose to initialise our composite method from this point.

4.2 Conservativeness of the GGMselect criterion: an example with a hub

We identified that GGMselect produced high quality, very sparse solutions. We argue here that they might be too sparse for their own good.

As discussed in Section 2.2, the numerical scheme of the GGMselect algorithm is based on a nodewise approach, and so is its model selection criterion. It penalises independently the degree of every node in the proposed graph. This makes it very unlikely to select graphs with a hub, i.e. a central node connected to many others. However recovering hubs is very important in conditional correlation networks. Genetic regulation networks for instance often feature hubs. With synthetic data, $n = 30, p = 30$, we encounter a "soft cap" effect, where it becomes very hard for GGMselect to propose a graph including a node of degree higher than 3. The penalty for such a node being too large to be compensated by the improved goodness of fit. On the other hand, we see on Fig. 3 that the Cross Validated Cross Entropy selects a graph which features the entire hub, and is in addition closer to the real graph regarding the remaining edges. Indeed, in the example of Fig. 3, other edges than the ones forming the hub are also ignored by GGMselect. With such a behaviour of the model selection criterion when the number of observations n is small, the GGMselect graphs are hard to interpret, with many key connections potentially missing.

Such observations motivated us to replace the GGMselect criterion with the Cross Validated Cross Entropy for graph selection. The next subsection proposes a quantitative comparison of the graphs selected by these two metrics.

4.3 The short-sightedness of the local model selection: a comparison of the GGMselect criterion and the CVCE

In this Section, we compare solely the model section metrics - and not the graph exploration schemes - on a fixed, shared, family of graphs. We demonstrate that our global approach to model selection yields graphs much closer to the original one and that reproduces the true distribution much better than the GGMselect criterion, which rejects the better, more connected graphs.

We compare the graphs selected by our Cross Validated CE (CVCE) and the GGMSC when shown the same family of candidate graphs. We consider a given true graph ($p = 30$). We compute once and for all one GGMselect solution with $n = 30$ observations drawn from this graph. With these key graphs in hand, we build manually (without the exploration scheme of Fig. 1) a deterministic sequence of graphs. Starting from the Fully Sparse with no edges, we add one by one, and in an arbitrary order, the edges needed to reach the GGMselect graph. From there, in the same manner, we add the missing edges and remove the excess edges to reach the true graph. Finally, we add - still one by one, still in an arbitrary order - the remaining edges until the Fully Connected graph, with all possible edges. All the encountered graphs in this sequence constitute the fixed family of candidates to be assessed by the model selection criteria. For each simulation, we generate n observations and use them to compute the GGMSC and CVCE along the path. We make 1000 of those simulations. The GGMSC uses the full data freely, while the CVCE must split the n points

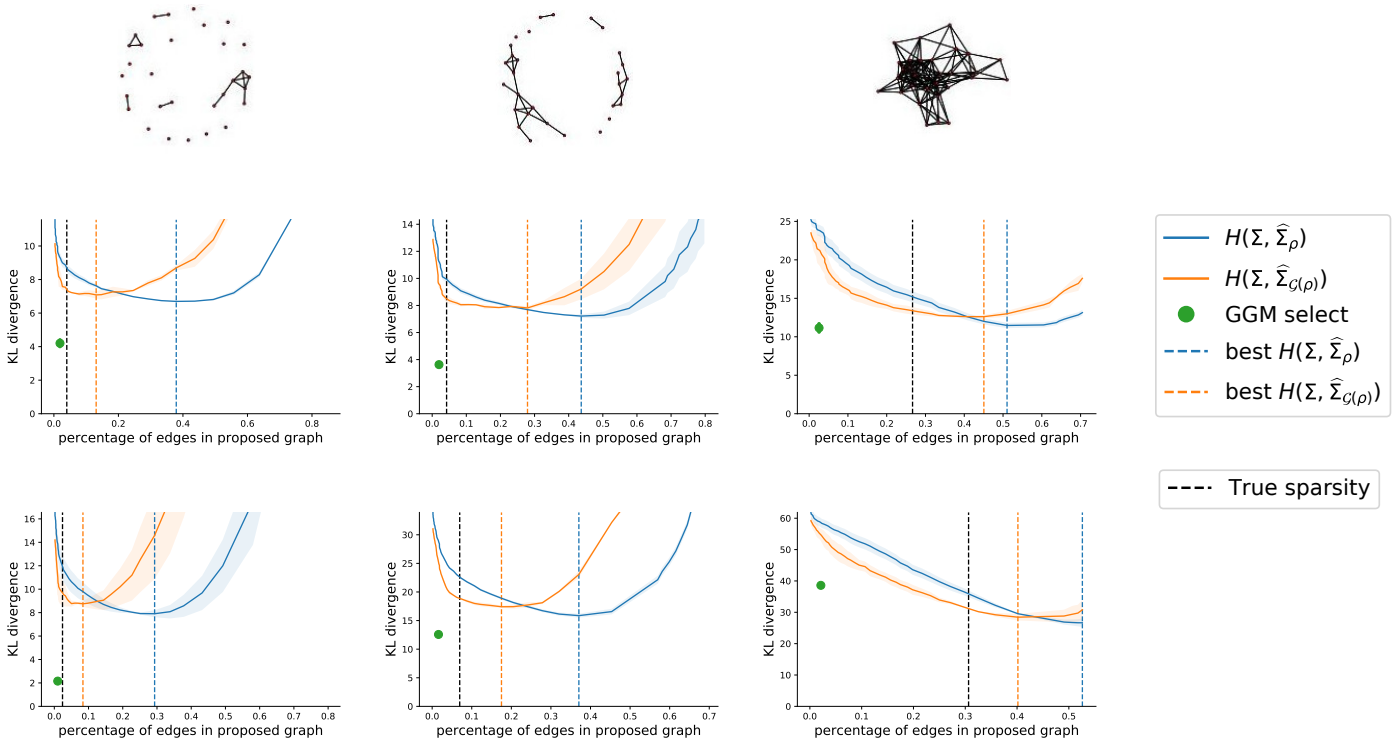


Fig. 2. Average performances as a function of the complexity for: the MLE from the GGMselect graph (green), the GLASSO solutions (blue) and the MLEs from the GLASSO graphs (orange). The average is taken over 100 simulations. In each simulation, $n = 30$ data points are simulated from a given true graph, different for each subfigure. The two rows of subfigures correspond to two different graph sizes, $p = 30$ and $p = 50$ vertices respectively. The three columns correspond to true graphs with different connectivity. At the top of each column, a graph illustrates the typical connectivity of the true graphs in said column.

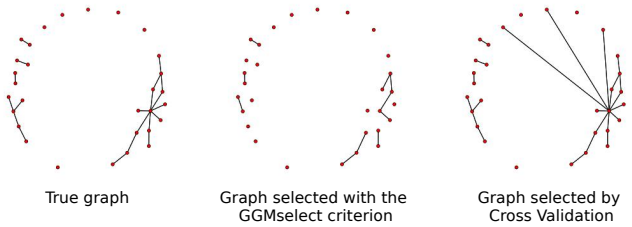


Fig. 3. Graph selection in the presence of a hub. The first figure is the true graph. The second and third are the graphs respectively selected by the GGMSC and CVCE on the same fixed graph path going from the fully sparse to the fully connected, via the GGMselect graph and the true graph

into the *exploration* covariance S_{expl}, to compute the graph constrained MLE $\widehat{\Sigma}_{\mathcal{G}}(S_{expl})$, and a *validation* covariance S_{val} to evaluate them. This leads to different results depending on the split size. Let S_{train} be the empirical covariance matrix built with the full data. We assess the performances of each graph $\mathcal{G}$ with the True CE (TCE) of the MLE built from S_{train} under the constraints of $\mathcal{G}$: $H(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. Since there is a known true Σ we actually compute the True KL $KL(\Sigma, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. This metric differs from the TCE only by a constant, hence is equivalent when ranking methods, but offers a sense of scale since the proximity to 0 in KL is meaningful. Fig. 4 illustrates the behaviour on one simulation. The most noticeable trend is that the GGMSC (in green) advocates a much earlier stop than the CVCE (in red), which stops almost on the same graph as the TCE (in blue). Additionally, on that run, the graph selected by the CVCE is actually the true graph (in grey). Fig. 5 represents the results over all simulations. We compare the average and standard deviation of the performances (true KL, on the y axis) and complexity (number of edges, x axis) of the models selected by the CVCE with different *exploration/validation* splits (in shades of red), GGMSC (in green) and with the TCE (in blue). The three columns represent different number of available observations ($n = 25, 40, 100$) and the second row is a zoomed in view of the first. This quantitative analysis confirms that the GGMSC selects graphs that are way too sparse even when shown more complex graphs with better performances. With the performances measured in KL, relative improvement is meaningful, and we see the CVCE improving the GGMSC choice by a factor from 2 to 5, and being much closer to the oracle solution in terms of KL. Additionally, the graphs selected by CVCE are also much closer to the original one. This is especially true when a large fraction of the data (35% or 40% of the *training* data) is kept in the *validation* set. The same results are observed with two other oracle metrics: the l_2 recovery of the True Σ, $\left\|\Sigma - \widehat{\Sigma}_{\mathcal{G}}(S_{train})\right\|_F$, and the oracle nodewise regression l_2 recovery $\left\|\Sigma^{\frac{1}{2}}(I_p - \Theta_{\mathcal{G}}(\underline{X}_{train}))\right\|_F$ (the oracle metric of the GGMselect authors [2]). Those metrics also reveal that when the *validation* set is small (20%), the variance of the performances of CVCE increases and it can become less

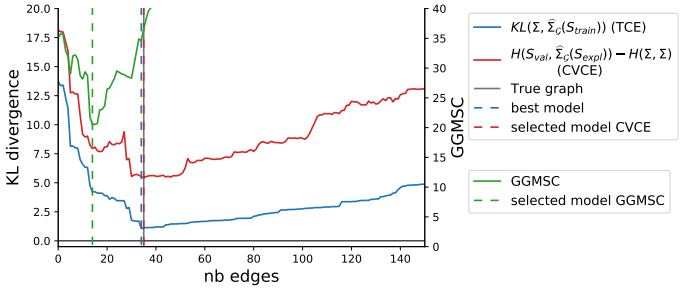


Fig. 4. On a single simulation: evolution of and model selected by GGMSC (green), CVCE (red) and TCE (blue) along the fixed deterministic path. The true graph's position on that path is represented by a vertical grey line. GGMSC stops early whereas CVCE selects the true graph (the vertical grey line and the dashed red one are the same). Moreover, the CVCE graph is very close to the best graph in terms of True Cross Entropy.

reliable depending on the metric. The Figures and details on these two metrics can be found in supplementary materials. This experiment illustrated how the model selection criterion of GGMselect can actually be very conservative, and even though the numerical scheme of the method explores interesting graph families, the model selection criterion might dismiss the more complex, better performing ones on them. This leads us to believe we can make substantial improvements by using the CVCE on a path built using the GGMselect solution as initialisation.

4.4 Execution time comparison

In this Section, we compare the runtimes of GLASSO, GGMselect and the Composite method for several values of p. For each p, 20 simulation are made, with $n = p/2$ observations each. This number of observations is an arbitrary heuristic to have both $n < p$ and n increasing with p. TABLE 1 synthesises the results. The runtime and complexity of the Composite method depend linearly on the number of steps chosen by the user. As seen in Fig. 1, this number of steps is the number of graphs that are constructed and evaluated. Ideally, this sequence of graphs should be just long enough to see the Oracle (or Out of Sample) performance improve as much as they can, and stop when they start deteriorating, when the point of overfitting is reached. In this experiment, the number of steps is chosen according to an heuristic depending on the number of edges in the initialisation graph with regards to p. The average number of steps over the simulations is also recorded in TABLE 1.
The Composite method and GGMselect both include a model selection step, however GLASSO just returns one solution of Eq. (1) for one given value of the penalty parameter ρ. As a result, all three methods are not strictly comparable. This was corrected in this experiment: for every simulation, the GLASSO is run on a grid of ρ with as many values as the number of estimated graphs by the Composite method. We call this the "grid GLASSO".

TABLE 1 shows that GGMselect is faster than the other two methods by 1 and 2 orders of magnitude in average. The Composite method is faster than the grid of GLASSOs when the dimension is small, but suffers when the dimension goes above $p = 100$. The Composite algorithm has indeed a high complexity in p, it runs $p \times n_{steps}$ ordinary linear regression with $p - 2$ features and computes then evaluates $(p + 1) \times n_{steps}$ graph constrained MLE of size $p \times p$ each.

TABLE 1
Average and (standard deviation) of the execution times of different GGM methods. The grid GLASSO compute solutions for as many values of the penalty parameter ρ as there are estimated graphs (steps) in the Composite method. The last column presents the average of this number of steps/number of estimated graphs. The number of observations is $n = p/2$.

p	GGMsel (fast)	grid GLASSO	Composite	nb steps
30	0.19 (0.07)	14.9 (8.60)	3.09 (1.80)	8.4
50	0.39 (0.03)	62.1 (32.9)	16.6 (8.20)	14.9
100	1.66 (0.66)	247 (135)	226 (138)	26.3
300	25.8 (1.04)	1470 (775)	6847 (1453)	40

The algorithmic of GGMselect and GLASSO were very well optimised by their respective authors. This shows in the very fast GGMselect computations, making it a very efficient initialisation for our Composite method. However, the implementation of the Composite, see Fig. 1, is naive and sequential. By running the linear regressions and LARS in parallel, and not re-calculating the MLE for the same graph several times, the performance would be greatly improved and closer to GLASSO.

5 Experiments on real data with the Composite GGM estimation algorithm

In this Section, we present two experiments with our composite method on real data. First, we demonstrate on brain imaging data from a cohort of Alzheimer's Disease patients that it recovers the known structures better than the classical local and global methods, while also having a better Out of Sample goodness of fit with the data. Then, we showcase how it is able to describe known dynamics between factors involved in Adrenal steroid synthesis on a database of Nephrology test subjects.

5.1 Experiment on Alzheimer's Disease patients

We first confirm our previous observations and demonstrate the performances of the complete numerical scheme of our composite procedure on real medical data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We have $p = 343$ features, $n = 92$ different patients. The first 240 features are measures of atrophy (MRI) and glucose consumption (PET) in the 120 areas of the cortex defined by the AAL2 map. The next 98 are two descriptors of the diffusion, fractional anisotropy and mean diffusivity, followed in the 49 regions of the JHU ICBM-DTI-81 white matter atlas. The rest of the features are basic descriptions of the patient.

5.1.1 *Experiment*

First we need a new evaluation metric. Indeed, with real data, we do not know the real covariance matrix. So we cannot anymore compute the True Cross Entropy to evaluate

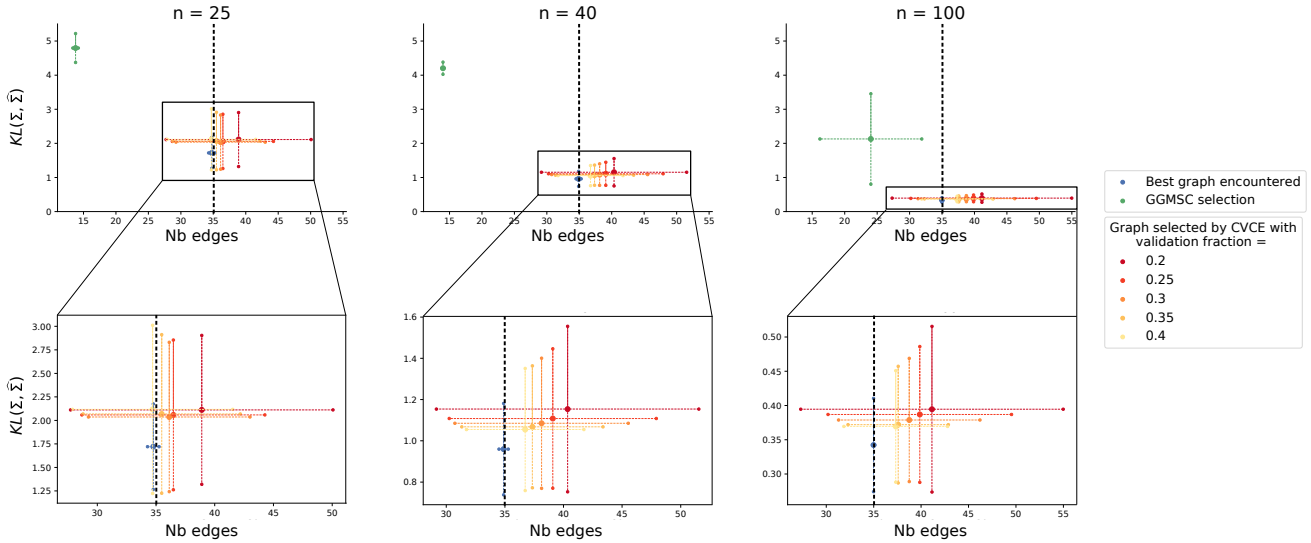


Fig. 5. Average KL divergence (y axis) and complexity (x axis) of the models selected with GGMSC (green), CVCE (shades of red) and TCE (blue) on synthetic data. The sparsity level of the true graph is represented by a black dashed vertical line. The second row offers a zoomed in view of the boxed areas to focus on the CVCE and TCE models. The graphs selected by the CVCE are much closer to the best in True Cross Entropy in terms of performance and edge structure than the GGMSC one. Moreover, they are also very close to the true graph used in the simulation, even when the sample size is small.

the inferred matrices. To replace the TCE, we keep $n = 18$ patients aside as a *test* set to define a test empirical covariance matrix S_{test}, whereas the $n = 74$ patients left constitute the *train* set, used to define S_{train}. We evaluate an inverse-sparse covariance matrix built from S_{train} with the negative Out of Sample Likelihood (OSL): $H(S_{test}, \widehat{\Sigma}_{\mathcal{G}}(S_{train}))$. The OSL is less absolute than the True CE, but still quantifies with no bias the goodness of fit for real data. Additionally, we cannot use a KL divergence for scale reference anymore, see Section 7.1 for more details.

The experiment run on the ADNI database is very simple: we compute the GGMselect solution and build our Composite GGM estimation procedure from it. To be fair, we also evaluate every graph our procedure encounters with the GGMSC, giving GGMselect a chance to change its mind if one of the new graphs were to fit its criterion better. In addition, we used the GLASSO algorithm of [12] to get the solutions of (1) for different penalty intensity.

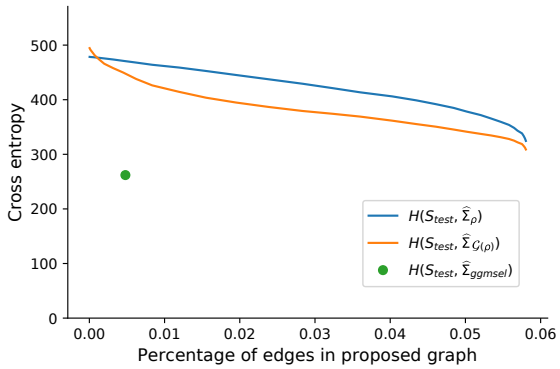


Fig. 6. Out of sample performances as a function of the complexity for: the MLE from the GGMselect graph (green), the GLASSO solutions (blue) and the MLEs refitted from the GLASSO graphs (orange).

5.1.2 Comparison of GLASSO and GGMselect

We confirm the observations and conclusions of Section 4.1. Fig. 6 shows that, even with varying penalty intensity, GLASSO does not encounter any solution with an OSL as good as GGMselect. This indicates that the optimisation problem (1) cannot find high-performing sparse graphs in this concrete setting either. The path of GLASSO is interrupted before its completion as we have computational error with the scikit learn package at low penalty levels. We encounter such errors eventually no matter how we regularise and precondition the empirical covariance S. This means we do not get to see the more connected solutions of the GLASSO. This is not a problem since we already go far enough in the GLASSO path to reach unacceptably complex graphs: 6% of the $\sim$ 59000 possible edges, i.e. 3500 edges for a graph with 343 nodes. By stopping early, we only consider the reasonable solutions of the GLASSO. In that case, GGMselect has a clear advantage, proposing a solution with a better Out of Sample fit with the data and only 281 edges.

5.1.3 Comparison of GGMselect and the Composite GGM estimation algorithm

We represent the selected graphs on left panel of Fig. 7, with the same conventions as Fig. 5. Once again the GGMSC (green) selects a sparse model, with 281 edges over the $\sim$ 60k possible. All the reasonable *validation* fractions (from 10% to 30%) of the CVCE (shades of red) select one out of two graphs, with both better OSL than the GGMSC one and closer to the OSL-optimum on the path (blue). Those two

graphs have 589 or 813 edges respectively. This indicates that many conditional correlations were potentially missed by GGMSC, and that the CVCE graphs may propose a more complete interpretation.

For a full comparison of the thee methods, the right panel of Fig. 7 is a zoomed out view that also includes the best model obtainable with problem (1) in terms of OSL (purple point). As we have seen, it is a very complex model with many edges. We visualise the successive improvements in Out of Sample Likelihood made first by GGMselect, with a sparser solution, then with our Composite GGM estimation procedure, with a more complete model. This experiment demonstrates the quantitative benefits of running the Composite algorithm in a High Dimension Low Sample Size setting.

In addition to those *quantitative* improvements, our method allows for a better *qualitative* interpretation of the disease. Fig. 8 represents, using the Colin 27 brain image of [36] and the MRView software of [37], the graphs selected by GGMSC and CVCE (589 edges version), as well as the best GLASSO graph in OSL ($\sim$ 3500 edges). We recall that each of the methods estimates a large graph with $p = 343$ vertices, a mix of different modalities measured in different areas of the cortex. The full graph cannot be displayed on an image of the cortex. For the sake of clarity, we only represent sub-parts of this one graph. On Fig. 8, only edges in-between the 120 MRI measures are represented. Additional views of the cortex can be found in supplementary materials. The GGMselect network is mostly composed of inter-hemispheric connections between symmetrical areas (hidden by the perspective in Fig. 8, see the supplementary materials for different views). These mainly reflect the symmetry of the atrophy pattern and are less informative for understanding disease process. The intra-hemispheric connections have a better interpretation potential to explain the pathology. Our algorithm reveals many more of these correlations - for instance in parietal areas, which are thought to be key hubs in the disease process - promising a more interesting description of the pathology. The GLASSO solution on the other hand, proposes many edges, making even this simple sub-graph unreadable. Similar observations can be made for connections in-between PET measures (see supplementary materials).
Additionally, Fig. 9 shows that the GGMselect graph features absolutely no edge between MRI and PET measures, effectively proposing a model in which there is no correlation whatsoever between anatomical and functional variables, a very unlikely and unsatisfactory description. Our method on the contrary recovers a reasonable amount of edges between those two modalities. GLASSO recovers a similar number of edges in this sub-part of the graph. However, Fig. 8 shows that it does so while having an extremely large number of edges in other regions of the graphs. Sparser GLASSO solution on the other hand, behave similarly to GGMselect and recover no edge linking MRI and PET measures, see supplementary materials. Of all these solutions, the Composite method proposes the most balanced.
These results suggest that our approach could be an interesting tool to study inter-regional and inter-modality dependencies in Alzheimer's Disease. This would need to be confirmed with larger populations of patients and more extensive experiments, which is out of the scope of the present paper and is left for future work.

5.2 Experiments on neprhology patients

In this Section, we compare qualitatively the methods in an environment with $p < n$. Although the Composite procedure was developed specifically for the case $n < p$, we demonstrate here that it still holds up to the state of the art outside of its intended application framework. We use a dataset of variables relevant to the adrenal steroidogenesis on a cohort of healthy test subjects.
Adrenal steroid synthesis in childhood is a complex process involving an enzymatic cascade that transforms cholesterol into mineralocorticoids, glucocorticoids or androgens, depending on the enzymatic equipment of each zona of the adrenal gland. Even though most important ways of adrenal steroidogenesis are known, we now assess new related metabolite that may ask new questions regarding adrenal steroidogenesis. Thus, we analysed a pediatric cohort of $n = 172$ healthy volunteers aged from 3 months to 16 years old with blood count and LC-MS/MS adrenal steroid profile analysis ($p = 35$).

Fig. 10 represents the matrices of pairwise conditional correlations corresponding to the GGMselect solution (left), the Composite solution (middle) and a sparse GLASSO solution (right). The rest of the path of GLASSO solution can be found in the supplementary materials. The other solutions contain many more edges than any of the three matrices here.
The models proposed by the three matrices have been compared to literature data for hematological parameters and steroidogenesis analysis. Regarding hematological analysis, both the Composite and GGMselect models confirm well known relations such as strong direct positive links between hemoglobin concentration (Hb) and red cells count (RBC); between hemoglobin concentration and mean corpuscular volume (WCV); between white cells (WBC) and platelet counts (PC); and a strong negative link between red cells count and mean corpuscular volume; between white cells count and age. The GLASSO solution did not show any of them.
Regarding steroid metabolism, 11-β1 hydroxylase (11 Ohase B1) and 21 hydroxylase (21 Ohase) activities, the Composite method and GGMselect reach the same conclusion: there is a strong positive direct link between enzymatic activities and the concentration of their corresponding alternate product. This is in accordance with common description of adrenal steroidogenesis process: decreased activity leads to an accumulation product of the alternative pathway. The GLASSO solution failed to show these relations. In the same way, GGMselect and the Composite method exhibit a negative link between the lack of 11-β HSD type 2 (11b HSD2) activity (that catabolizes cortisol into cortisone) and the concentration of its product, cortisone (e). The sparse GLASSO fails to underline this link. All

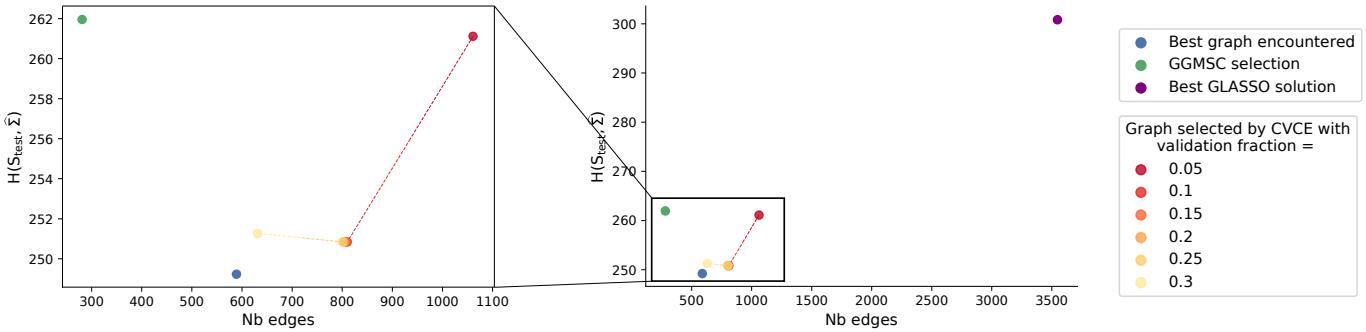


Fig. 7. Out of Sample Likelihood (y axis) and complexity (x axis) of models selected by GGMSC (green), CVCE (shades of red) and OSL (blue) on real data. The right picture offers a zoomed out view to include the model selected by OSL on the GLASSO path (purple). The left figure corresponds to the boxed area of the right figure.

these data tend to show a better interpretation of steroids profile with the GGMselect and Composite solutions. Interestingly, these models also underline a new link: a strong positive link between 18-hydroxycorticosterone (18ohb) and 18-hydroxycortisol (18ohf) concentrations, two steroids that are supposed to be independently produced in two different zonas of the adrenal gland. This result could imply an alternative pathway in adrenal steroidogenesis that needs to be explored.
The GGMselect and Composite graphs are mostly identical, although some of the conditional correlations are weaker in the Composite matrix. Among the subtle differences, two edges that are coherent with the state of the art, and are present in the GGMselect graph, were alleviated in the Composite matrix (resulting in invisible connections in Fig. 10): the link between the 18-oxocortisol (18oxof) and cortisol (f) concentrations, and the very strong negative link between the ratio cortisol/18-oxocortisol (F/18oxof) and 18-oxocortisol. The other very few additions and removals in the Composite model are hard to validate or disprove with the current state of the art.
From a medical analysis point of view, all these results are preliminary and will have to be confirmed by more in depth studies. From a purely machine learning point of view, this example illustrates that the Composite method behaves appropriately when $p < n$. In this example, the GGMselect solution seems already acceptable, and the Composite procedure does not deviate too much from it.

To summarise these experimental studies, Section 5.1 showed the quantitative and qualitative improvements made by the Composite method on real data, in the High Dimension Low Sample size setting ($n < p$) the method was designed for. In this Section, with enough data available ($p < n$), hence outside the intended area of application, the qualitative analysis suggests that, running the Composite procedure does not provide additional benefits, but does not cause any loss either.

6 Conclusion

When it came to inferring conditional covariance graphs from a small number of observations, we were dissatisfied with the state of the art GGM methods. In this paper, we quantified the shortcomings in terms of goodness of fit, distribution reconstruction and interpretability of the local approach of [1] and the global optimisation problem of [3], [4]. *We proposed a method composed of a structure learning algorithm coupled with model selection criterion*. In the latter, the structure learning steps are a variation of the parallel nodewise linear regressions of [1] and the model selection steps guided by out of sample versions of the likelihood optimised in [3] and [4]. The validity of our method was demonstrated on synthetic and real data when $n < p$. Quantitatively, it consistently reached consequently lower KL divergences and better sparsistency than the aforementioned state of the art paradigms. A qualitative analysis on a neurological data set of real data, revealed that it better recovered the known dynamics of the field. An additional real data experiment, with $p < n$, suggested that the method did not cause any loss when used outside the intended scope of application. In the future, optimising the numerical scheme will allow us to make further quantitative improvements. Such as lower execution times and better performances with less reliance on the initialisation.

7 Proofs of the main results

7.1 Basic Cross Entropy calculus for Gaussian vectors

In this Section, we offer details and commentary on the Cross Entropy manipulation with normal distributions and prove (2) and (3).
The formula of the Cross Entropy $H(p,q)$ is given by:

$$H(p,q) := -\mathbb{E}_p\left[\log q(X)\right] = \int_x -p(x)ln(q(x))\mu(\mathrm{d}x)\,.$$

The likelihood p_θ of a parametric distribution f_θ with iid observations $(X^{(1)},...,X^{(n)})$ is given by:

$$p_\theta(X^{(1)},...,X^{(n)}) = \prod_{i=1}^{n} f_\theta(X^{(i)})\,.$$

Let $\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x=X^{(i)}}$ be the empirical distribution of the sample $(X^{(1)},...,X^{(n)})$, we see the connection between CE and likelihood:

$$H\left(\hat{f}_n, f_\theta\right) = -\frac{1}{n}\sum_{i=1}^{n}\log(f_\theta(X_i)) = -\frac{1}{n}\log p_\theta(x_1,...,x_n)\,.$$

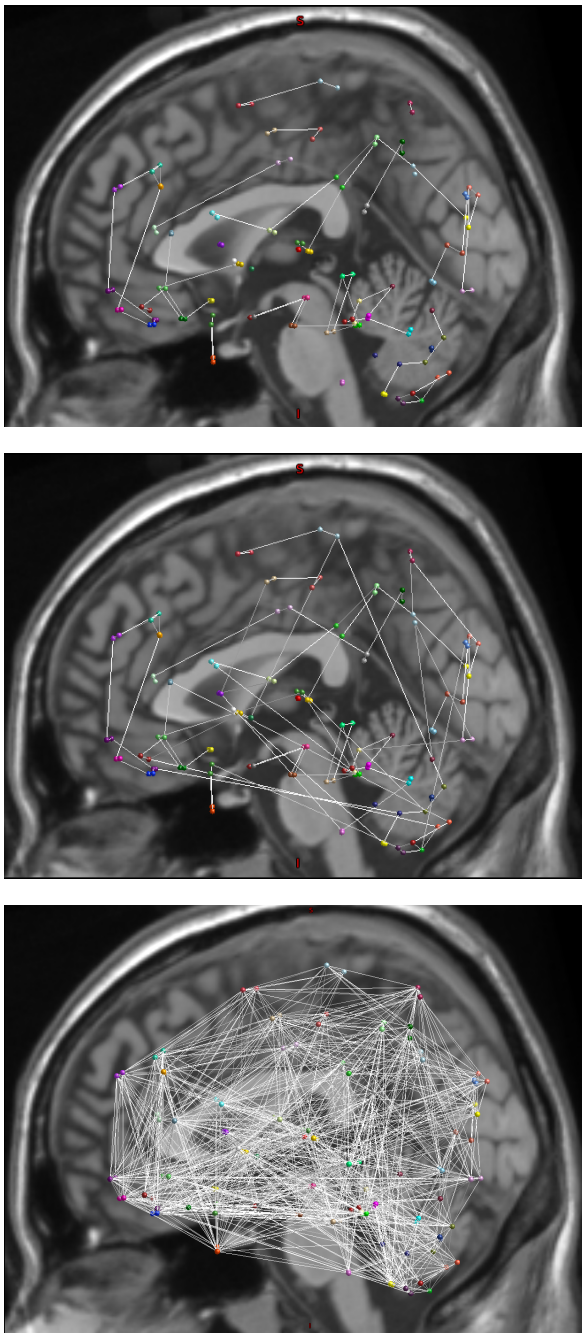


Fig. 8. Selected edges by GGMselect (up), our Composite method (mid) and the best Out of Sample GLASSO (down) in-between MRI measures. The perspective of the sagittal view hides the many edges between symmetrical regions. GLASSO proposes too many to allow for interpretation.

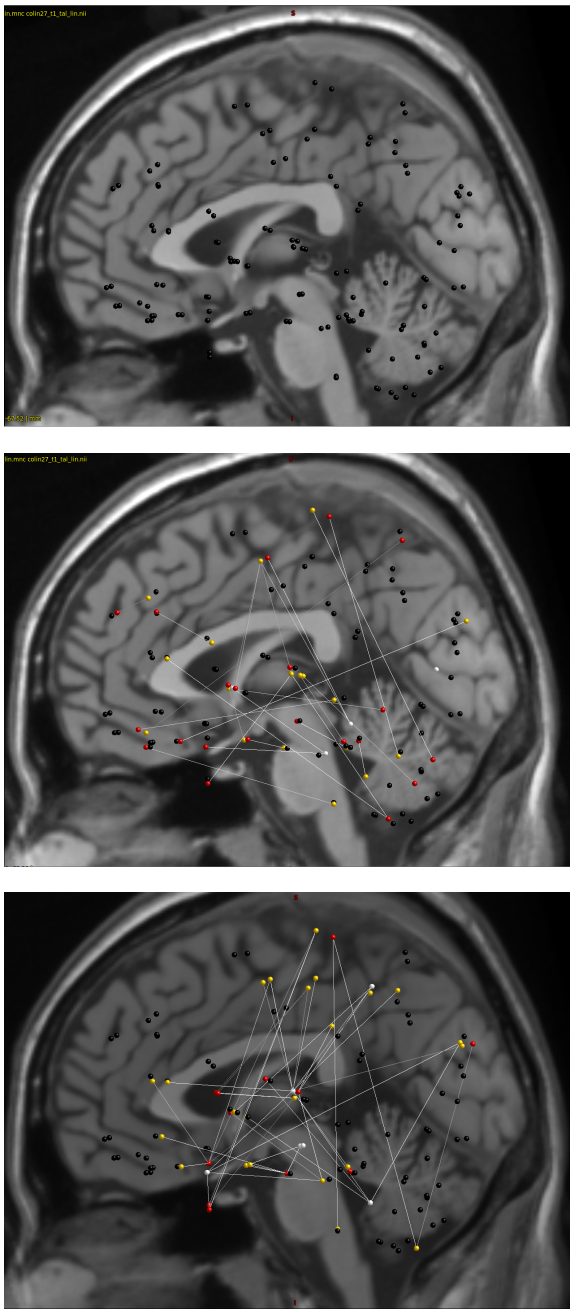


Fig. 9. Selected edges by GGMselect (up), our Composite method (mid) and the best Out of Sample GLASSO (down) between PET (yellow) and MRI (red) measures. GGMselect finds no connection in this sub-part of the graph, although one may expect some.

Proof of (2) *and* (3) *:* In the case of Centered Multivariate Gaussians, let $H\left(\Sigma_1, \Sigma_2\right) := H\left(f_{\Sigma_1}, f_{\Sigma_2}\right)$ and let us omit the constant $\frac{p}{2} ln(2\pi)$ from the calculations:

$$\begin{aligned} H\left(\Sigma_1, \Sigma_2\right) &\equiv \int_X f_{\Sigma_1}(x) \left(-\frac{1}{2} ln(|K_2|) + \frac{1}{2} X^T K_2 X\right) \mathrm{d}X \\ &= -\frac{1}{2} ln(|K_2|) + \frac{1}{2} \int_X f_{\Sigma_1}(x) \left\langle XX^T, K_2 \right\rangle \mathrm{d}X \\ &= -\frac{1}{2} ln(|K_2|) + \frac{1}{2} \left\langle \int_X f_{\Sigma_1}(x) XX^T \mathrm{d}X, K_2 \right\rangle \\ &= -\frac{1}{2} ln(|K_2|) + \frac{1}{2} \left\langle \Sigma_1, K_2 \right\rangle \ . \end{aligned}$$

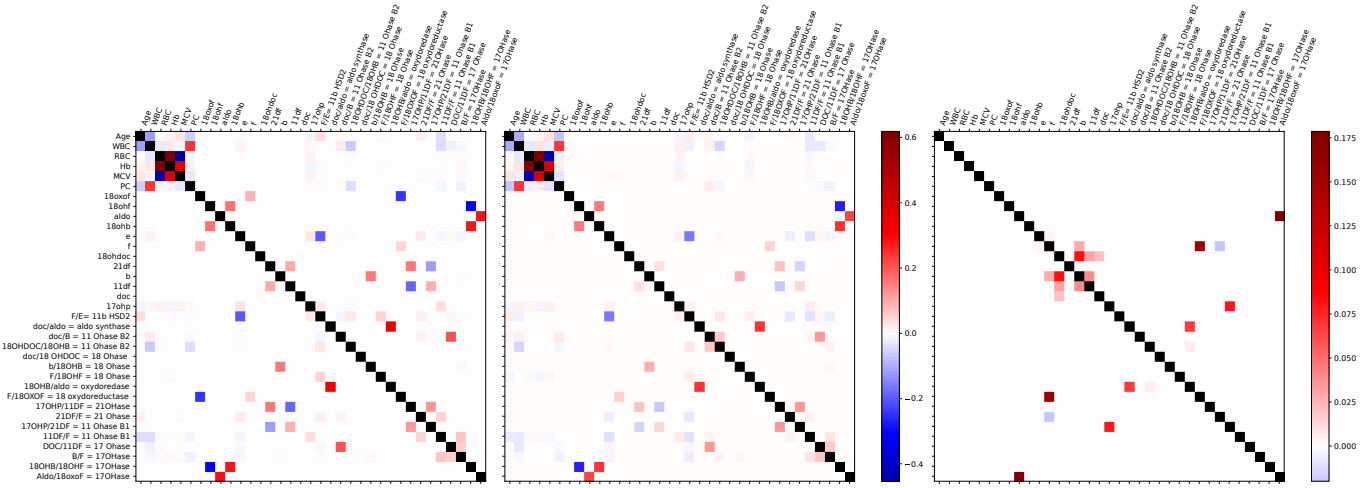


Fig. 10. Conditional covariance matrix between the 35 variables measured on the cohort. The positive correlations are in red and the negative in blue. The diagonal coefficients are ignored in this study. GGMselect (left) and Composite (middle) share the same colour scale. The rightmost figure corresponds to one of the sparsest GLASSO solution.

In the end, we get (2):

$$\boxed{H\left(\Sigma_1, \Sigma_2\right) \equiv \frac{1}{2}\left(\langle \Sigma_1, K_2 \rangle - ln(|K_2|)\right) .}$$

With the observed data $\underline{X} := (X_1, ..., X_n)^T \in \mathbb{R}^{n \times p}$, let $S := \frac{1}{n}\underline{X}\,\underline{X}^T \in S_p^+$, the empirical covariance matrix. The log likelihood of any centred Gaussian distribution f_{Σ_2} is given by:

$$\begin{aligned} H\left(\hat{f}_n, f_{\Sigma_2}\right) &\equiv \frac{1}{2n}\sum_{i=1}^{n}\left(-ln(|K_2|) + X_i^T K_2 X_i\right) \\ &= -\frac{1}{2}ln(|K_2|) + \left\langle \sum_{i=1}^{n}\frac{X_i X_i^T}{2n}, K_2 \right\rangle \\ &= -\frac{1}{2}ln(|K_2|) + \frac{1}{2}\langle S, K_2 \rangle \ , \end{aligned}$$

where, as in (2), we omit the constant term $\frac{p}{2}ln(2\pi)$ from the calculations. In the end, we get (3):

$$\boxed{H\left(\hat{f}_n, f_{\Sigma_2}\right) \equiv \frac{1}{2}\left(\langle S, K_2 \rangle - ln(|K_2|)\right) .}$$

□

The likelihood $H\left(\hat{f}_n, f_{\Sigma_2}\right)$ follows a similar formula as the Cross Entropy between two normal distributions (2). When S defines a non degenerate normal distribution, what we actually have is $H\left(\hat{f}_n, f_{\Sigma_2}\right) = H\left(f_S, f_{\Sigma_2}\right)$. However, when $n < p$, S is singular and the density f_S is not defined. The formula (3) still holds though, and we write $H\left(S, \Sigma_2\right) := H\left(\hat{f}_n, f_{\Sigma_2}\right)$ since the formula is the same as (2) for $H\left(\Sigma_1, \Sigma_2\right)$.

Remark When the density f_S does exists, we have equality in the CE $H\left(\hat{f}_n, f_{\Sigma_2}\right) = H\left(f_S, f_{\Sigma_2}\right)$, but not in the Entropies $H\left(\hat{f}_n, \hat{f}_n\right) \neq H\left(f_S, f_S\right)$, as a consequence the KL divergences are different as well: $KL\left(\hat{f}_n, f_{\Sigma_2}\right) \neq KL\left(f_S, f_{\Sigma_2}\right)$. In practice $KL\left(f_S, f_{\Sigma_2}\right) << KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ and $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ will never reach 0, since a normal distribution will tend to be closer to another normal distribution than to an empirical one, this is particularly true with n small and Σ_2 close to S. As a result, $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$ offers a poor sense of scale, since the value 0 cannot be used as a reference. For this reason, when we represent $H\left(f_{S_{test}}, f_{\Sigma_2}\right)$ as we do in Fig. 7, we do not use it under the form of a KL with 0 as its minimum for scale reference - as we do on synthetic data in Fig. 5 - since the only KL we can compute is the mostly irrelevant $KL\left(\hat{f}_n, f_{\Sigma_2}\right)$.

7.2 Preliminary results for the model selection guarantees

To prove the controls we stated in Sections 3.2, 3.3 and 3.4, we need the two following lemmas.

Lemma 1. *Let $S^{(\lambda)} := S + \lambda I_p$. With $\widehat{K}_{\mathcal{G}} := \widehat{\Sigma}_{\mathcal{G}}^{-1}$, where $\widehat{\Sigma}_{\mathcal{G}}$ is defined as in (7), we have:*

$$\forall \mathcal{G} \in \mathcal{M}, \quad \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}} \right\rangle = p\,. \tag{17}$$

Proof: Let $\Pi_{\mathcal{G}}$ be the orthogonal projection on the edge set $E_{\mathcal{G}} \cup \{(i,i)\}_{i=1}^p$. That is to say, for any matrix $M \in \mathbb{R}^{p\times p}$, $\Pi_{\mathcal{G}}(M)_{i,j} = M_{i,j}\mathbb{1}_{(i,j)\in E_{\mathcal{G}}\cup\{(i,i)\}_{i=1}^p}$. A property of the MLE is that $\Pi_{\mathcal{G}}(\widehat{\Sigma}_{\mathcal{G}}) = \Pi_{\mathcal{G}}(S^{(\lambda)})$, i.e. the matrices have the same values on the diagonal and the edge set, see [6]. Additionally, note that, because of the sparsity of $\widehat{K}_{\mathcal{G}}$, for any matrix M, we have $\left\langle M, \widehat{K}_{\mathcal{G}} \right\rangle = \left\langle \Pi_{\mathcal{G}}(M), \widehat{K}_{\mathcal{G}} \right\rangle$. Then:

$$\begin{aligned} \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}} \right\rangle &= \left\langle \Pi_{\mathcal{G}}(S^{(\lambda)}), \widehat{K}_{\mathcal{G}} \right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}} \right\rangle &= \left\langle \Pi_{\mathcal{G}}(\widehat{\Sigma}_{\mathcal{G}}), \widehat{K}_{\mathcal{G}} \right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}} \right\rangle &= \left\langle \widehat{\Sigma}_{\mathcal{G}}, \widehat{K}_{\mathcal{G}} \right\rangle \\ \left\langle S^{(\lambda)}, \widehat{K}_{\mathcal{G}} \right\rangle &= p\,. \end{aligned}$$

□

Lemma 2. *With $\widehat{K}_{\mathcal{G}} := \widehat{\Sigma}_{\mathcal{G}}^{-1}$, where $\widehat{\Sigma}_{\mathcal{G}}$ is defined as (7), we have:*

$$\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}\,.$$

Proof: We have:

$$\begin{aligned}\left\langle S+\lambda I_p, \widehat{K}_{\mathcal{G}}\right\rangle &= p\\ \left\langle S, \widehat{K}_{\mathcal{G}}\right\rangle + \lambda tr(\widehat{K}_{\mathcal{G}}) &= p\\ tr\left(\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}}\right) + \lambda tr(\widehat{K}_{\mathcal{G}}) &= p\,.\end{aligned}$$

Since $\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}} \in S_p^+$, we have $tr\left(\widehat{K}_{\mathcal{G}}^{\frac{1}{2}} S \widehat{K}_{\mathcal{G}}^{\frac{1}{2}}\right) \geq 0$ and $\lambda tr(\widehat{K}_{\mathcal{G}}) \leq p$, i.e.

$$\left\|\widehat{K}_{\mathcal{G}}\right\|_* \leq \frac{p}{\lambda}\,.$$

□

7.3 Bounds in expectation for the CVCE solutions

We prove the results of Sections 3.2 and 3.3.

Proof of (11), (12), (13), *and* (14) *:* We want to control the expected regret $e := \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right]$. First, note that by definition of $\widehat{\mathcal{G}}^*$, we have

$$0 \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\,.$$

So the lower bound:

$$0 \leq e\,,$$

is guaranteed.
From the definition of $\widehat{\mathcal{G}}_{CV}$ (9), we get:

$$H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) \leq H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\,.$$

We have for any $\widetilde{\Sigma} \in S_p^{++}$, with $\widetilde{K} := \widetilde{\Sigma}^{-1}$:

$$H\left(S_{val}, \widetilde{\Sigma}\right) = H\left(\Sigma, \widetilde{\Sigma}\right) + \frac{1}{2}\left\langle S_{val} - \Sigma, \widetilde{K}\right\rangle\,.$$

Hence:

$$\begin{aligned}H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) \leq H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) &+ \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle\\ &- \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\,.\end{aligned} \tag{18}$$

Since $K_{\widehat{\mathcal{G}}^*}$ is defined from S_{expl} uniquely, and independently of S_{val}, we get

$$\begin{aligned}\mathbb{E}\left[\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle \middle| S_{expl}\right] &= \left\langle \mathbb{E}\left[S_{val} - \Sigma | S_{expl}\right], \widehat{K}_{\widehat{\mathcal{G}}^*}\right\rangle\\ &= 0\,.\end{aligned} \tag{19}$$

From (18) and (19) we get:

$$\begin{aligned}\mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right)\right] \leq \mathbb{E}\left[H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right]&\\ + \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right]\,.&\end{aligned}$$

Which is exactly the result of Eq. (11):

$$\boxed{e \leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right]\,.}$$

As we discussed in Section 3.3, to obtain Eq. (11), we only used the definitions of $\widehat{\mathcal{G}}_{CV}$ for the upper bound and $\widehat{\mathcal{G}}^*$ for the lower bound. Since we assume nothing on the model family $\mathcal{M}$, those bounds are somewhat optimal in terms of the available information. Additionally, (11) is actually independent of how the symmmetric positive matrices $\{\widehat{\Sigma}_{\mathcal{G}}\}_{\mathcal{G}\in\mathcal{M}}$ are defined as long as they are function only of S_{expl}. They do not need to be associated with a different graph each, or with any graph for that matter. They do not need to be solutions of the MLE problem (7) and could be for example all the solutions on the path of solution of the l_1-penalised likelihood optimisation problem (1).
To get a more explicit control on the CVCE however, we need the assumption that $\widehat{\Sigma}_{\mathcal{G}}$ is the constrained MLE defined in (7).
Let $\Sigma_\infty := \max_{i,j} |\Sigma_{ij}|$. We call $E_{\max}$ the union of the maximal edge sets in $\mathcal{M}$, $d_{\max} = |E_{max}| \leq \frac{p(p-1)}{2}$ its cardinal and Π_{max} the orthogonal projection on $E_{\max} \cup \{(i,i)\}_{i=1}^p$. We have:

$$\begin{aligned}e &\leq \frac{1}{2}\mathbb{E}\left[\left\langle \Sigma - S_{val}, \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right]\\ &= \frac{1}{2}\mathbb{E}\left[\left\langle \Pi_{\widehat{\mathcal{G}}_{CV}}\left(\Sigma - S_{val}\right), \widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\rangle\right]\\ &\leq \frac{1}{2}\mathbb{E}\left[\left\|\Pi_{\widehat{\mathcal{G}}_{CV}}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left\|\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\|_F^2\right]^{\frac{1}{2}}\\ &\leq \frac{1}{2}\mathbb{E}\left[\left\|\Pi_{max}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left\|\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right\|_*^2\right]^{\frac{1}{2}}\\ &\leq \frac{1}{2}\left(\sum_{i=1}^p \mathbb{E}\left[\left(\Sigma^{ii} - S_{val}^{ii}\right)^2\right]\right.\\ &\qquad \left. + \sum_{(i,j)\in E_{max}} \mathbb{E}\left[\left(\Sigma^{ij} - S_{val}^{ij}\right)^2\right]\right)^{\frac{1}{2}} \frac{p}{\lambda}\\ &\leq \frac{1}{2}\left(\frac{2\Sigma_\infty^2}{n_{val}}\left(p + 2d_{max}\right)\right)^{\frac{1}{2}} \frac{p}{\lambda}\,.\end{aligned}$$

From which we finally get the result of (12):

$$\boxed{e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{\left(p + 2d_{max}\right)^{\frac{1}{2}} p}{\sqrt{n_{val}}}\,.}$$

If $E_{\max}$ is dependent on the *exploration* data - because the graph family $\mathcal{M}$ was built from S_{expl} for instance - we have:

$$\begin{aligned}&\mathbb{E}\left[\left\|\Pi_{max}\left(\Sigma - S_{val}\right)\right\|_F^2\right]^{\frac{1}{2}}\\ &\quad= \left(\sum_{i=1}^p \mathbb{E}\left[\left(\Sigma^{ii} - S_{val}^{ii}\right)^2\right]\right.\\ &\qquad \left. + \mathbb{E}\left[\sum_{i,j\in E_{max}} \mathbb{E}\left[\left(\Sigma^{ij} - S_{val}^{ij}\right)^2 \middle| S_{expl}\right]\right]\right)^{\frac{1}{2}}\\ &\quad\leq \left(\frac{2\Sigma_\infty^2}{n_{val}}\left(p + 2\mathbb{E}\left[d_{max}\right]\right)\right)^{\frac{1}{2}}\,.\end{aligned}$$

We get the control (13), the same as (12) but with an additional expectation term:

$$\boxed{e \leq \frac{\Sigma_\infty}{\lambda\sqrt{2}} \frac{\left(p + 2\mathbb{E}\left[d_{max}\right]\right)^{\frac{1}{2}} p}{\sqrt{n_{val}}} .}$$

In order to prove (14), we start by showing how the regret is bounded by operator norm $\|\Sigma - S_{val}\|_2$. By tracial matrix Holder inequality:

$$\begin{aligned}\left\langle S_{val} - \Sigma, \widehat{K}_{\widehat{\mathcal{G}}_{CV}} \right\rangle &\leq \|\Sigma - S_{val}\|_2 \left\| \widehat{K}_{\widehat{\mathcal{G}}_{CV}} \right\|_* \\ &= \|\Sigma - S_{val}\|_2 \, tr\left(\widehat{K}_{\widehat{\mathcal{G}}_{CV}}\right) \\ &\leq \frac{\|\Sigma - S_{val}\|_2}{\lambda} p .\end{aligned}$$

Then, using (11), we get:

$$e \leq \mathbb{E}\left[\|\Sigma - S_{val}\|_2\right] \frac{p}{2\lambda} . \tag{20}$$

To prove (14), we first recall Theorem 4 of [34]:

Theorem 4 of [34]. *Let $X_1, X_2, ..., X_n$ be i.i.d. weakly square integrable centered random vectors in a separable Banach space with norm $\|.\|$ and Σ be their covariance operator. If X is Gaussian, then there exist an absolute constant c, independent of the problem, such that:*

$$\mathbb{E}\left[\left\|\widehat{\Sigma} - \Sigma\right\|\right] \leq c \|\Sigma\| \, max\left(\sqrt{\frac{\mathbb{E}\left[\|X\|\right]^2}{n\|\Sigma\|}}, \frac{\mathbb{E}\left[\|X\|\right]^2}{n\|\Sigma\|}\right), \tag{21}$$

where $\|.\|$ for operators denotes the operator norm associated with the vector norm $\|.\|$, that is to say:

$$\|\Sigma\| = \sup_{\|u\|=1} \|\Sigma u\| .$$

In our case, $X \sim \mathcal{N}(0_p, \Sigma)$ is a Gaussian vector in the Banach space $\mathbb{R}^p$, with the euclidean norm $\|X\|_2$, that verifies the integrability properties of the Theorem and whose covariance operator is the covariance matrix Σ. Hence the theorem can be applied. The operator norm for a symmetric positive matrix Σ associated with the euclidean norm is also called the spectral norm, since it corresponds to the highest eigenvalue: $\|\Sigma\|_2 = \lambda_{max}(\Sigma)$.
For a Gaussian vector: $Z \sim \mathcal{N}(0_p, I_p)$, we have:

$$\mathbb{E}\left[\|Z\|_2\right] \leq \sqrt{p}.$$

Since $K^{\frac{1}{2}} X \sim \mathcal{N}(0_p, I_p)$, and

$$\begin{aligned}\|X\|_2 &= \left\|\Sigma^{\frac{1}{2}} K^{\frac{1}{2}} X\right\|_2 \\ &\leq \left\|\Sigma^{\frac{1}{2}}\right\|_2 \left\|K^{\frac{1}{2}} X\right\|_2 ,\end{aligned}$$

we have:

$$\mathbb{E}\left[\|X\|_2\right] \leq \left\|\Sigma^{\frac{1}{2}}\right\|_2 \sqrt{p} .$$

Since $\|\Sigma\|_2 = \lambda_{max}(\Sigma)$, we have by definition, $\left\|\Sigma^{\frac{1}{2}}\right\|_2 = \|\Sigma\|^{\frac{1}{2}}$. In the end, when we apply (21) to our case, we get:

$$\mathbb{E}\left[\|S_{val} - \Sigma\|_2\right] \leq c\lambda_{max}(\Sigma) max\left(\sqrt{\frac{p}{n_{val}}}, \frac{p}{n_{val}}\right) . \tag{22}$$

We apply this concentration result on (20) to obtain (14):

$$\boxed{e \leq c\frac{\lambda_{max}(\Sigma)}{\lambda} p\left(\sqrt{\frac{p}{n_{val}}} \vee \frac{p}{n_{val}}\right) .}$$

□

7.4 Bounds in probability for the CVCE solutions

We prove the results of Section 3.4.

Proof of (15) *and* (16)*:* We want to lower bound the probability that the regret is small: $P := \mathbb{P}\left(\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta\right)$. The concentration dynamic driving the results comes from the convergence of random Wishart matrix S_{val} towrds its average Σ, which is made stronger by the number of observations n_{val} in the *validation* set. Since:

$$\begin{aligned}\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq & \\ &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right)\right| \\ &+ \left|H\left(S_{val}, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| ,\end{aligned}$$

then

$$\begin{aligned}&\forall \mathcal{G} \in \mathcal{M}, \ \left|H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}\right)\right| \leq \frac{\delta}{2} \\ \implies &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta .\end{aligned}$$

Since:

$$H\left(S_{val}, \widehat{\Sigma}_{\mathcal{G}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\mathcal{G}}\right) = \frac{1}{2}\left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle ,$$

then

$$\begin{aligned}&\forall \mathcal{G} \in \mathcal{M}, \ \left|\left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle\right| \leq \delta \\ \implies &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta .\end{aligned} \tag{23}$$

From the logical implication (23), we can take two path to derive two different bounds: one with a more general expression, and a more precise one taking into consideration the sparsity of the models. For the first one, note that $S_{val} = \Sigma^{\frac{1}{2}} W \Sigma^{\frac{1}{2}}$ where $n_{val} W \sim \mathcal{W}_p(I_p, n_{val})\, p$ is a standard Wishart matrix. Then we have:

$$\begin{aligned}\forall \mathcal{G}, \ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle &= \left\langle W - I_p, \Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}} \right\rangle \\ &\leq \|W - I_p\|_F \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F \\ &\leq \|W - I_p\|_F \max_{\mathcal{G} \in \mathcal{M}} \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F .\end{aligned}$$

We plug this result into (23) to obtain:

$$\begin{aligned}&\|W - I_p\|_F \max_{\mathcal{G} \in \mathcal{M}} \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F \leq \delta \\ \implies &\forall \mathcal{G} \in \mathcal{M}, \ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle \leq \delta \\ \implies &\left|H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right)\right| \leq \delta .\end{aligned}$$

We end up with the control (15) by taking the probability in the previous expression:

$$\boxed{P \geq \mathbb{P}\left(\|W - I_p\|_F \leq \frac{\delta}{\max_{\mathcal{G} \in \mathcal{M}} \left\|\Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}}\right\|_F}\right) .}$$

For the second result, let $\Pi_{\mathcal{G}}$ and Π_{max} be the orthogonal projections on the edge sets $E_{\mathcal{G}} \cup \{(i,i)\}_{i=1}^{p}$ and $E_{max} \cup \{(i,i)\}_{i=1}^{p}$ respectively. We have:

$$\begin{aligned}\forall \mathcal{G},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle &= \left\langle \Pi_{\mathcal{G}}(S_{val} - \Sigma), \widehat{K}_{\mathcal{G}} \right\rangle \\ &\leq \left\| \Pi_{\mathcal{G}}(S_{val} - \Sigma) \right\|_F \left\| \widehat{K}_{\mathcal{G}} \right\|_F \\ &\leq \left\| \Pi_{max}(S_{val} - \Sigma) \right\|_F \max_{\mathcal{G} \in \mathcal{M}} \left\| \widehat{K}_{\mathcal{G}} \right\|_F .\end{aligned}$$

Hence we get, from (23), the logical implication:

$$\begin{aligned}&\left\| \Pi_{max}(S_{val} - \Sigma) \right\|_F \max_{\mathcal{G} \in \mathcal{M}} \left\| \widehat{K}_{\mathcal{G}} \right\|_F \leq \delta \\ \Longrightarrow\ &\forall \mathcal{G},\ \left\langle S_{val} - \Sigma, \widehat{K}_{\mathcal{G}} \right\rangle \leq \delta \\ \Longrightarrow\ &\left| H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}_{CV}}\right) - H\left(\Sigma, \widehat{\Sigma}_{\widehat{\mathcal{G}}^*}\right) \right| \leq \delta .\end{aligned}$$

From which we get the control (16) by taking the probability of the events:

$$\boxed{P \geq \mathbb{P}\left(\left\| \Pi_{max}(S_{val} - \Sigma) \right\|_F \leq \frac{\delta}{\max\limits_{\mathcal{G} \in \mathcal{M}} \left\| \widehat{K}_{\mathcal{G}} \right\|_F} \right) .}$$

We underline that we obtain the two controls (15) and (16) directly from logical implications. Hence, they remain true when every probability is taken conditionally to any random variable, for instance the *exploration* data set, or the sufficient statistic built from it: S_{expl}. □

Remark Since $\forall \mathcal{G} \in \mathcal{M}$, $\left\| \widehat{K}_{\mathcal{G}} \right\|_* \leq \frac{p}{\lambda}$, both $\max\limits_{\mathcal{G} \in \mathcal{M}} \left\| \Sigma^{-\frac{1}{2}} \widehat{K}_{\mathcal{G}} \Sigma^{-\frac{1}{2}} \right\|_F$ and $\max\limits_{\mathcal{G} \in \mathcal{M}} \left\| \widehat{K}_{\mathcal{G}} \right\|_F$ are bounded random variables. They depend only on the *exploration* empirical covariance S_{expl} and can be seen as constants of the problem if working conditionally to the *exploration* set. Likewise, Π_{max} is a deterministic function conditionally to S_{expl}.

REFERENCES

[1] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.

[2] C. Giraud, S. Huet, and N. Verzelen, "Graph selection with ggmselect," *Statistical applications in genetics and molecular biology*, vol. 11, no. 3, 2012.

[3] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[4] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine learning research*, vol. 9, no. Mar, pp. 485–516, 2008.

[5] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, p. 186, 2009.

[6] A. P. Dempster, "Covariance selection," *Biometrics*, pp. 157–175, 1972.

[7] C. Giraud, "Estimation of Gaussian graphs by model selection," *Electronic Journal of Statistics*, vol. 2, pp. 542–563, 2008.

[8] M. Yuan, "High dimensional inverse covariance matrix estimation via linear programming," *Journal of Machine Learning Research*, vol. 11, no. Aug, pp. 2261–2286, 2010.

[9] T. Cai, W. Liu, and X. Luo, "A constrained l_1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 594–607, 2011.

[10] T. Sun and C.-H. Zhang, "Sparse matrix inversion with scaled lasso," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3385–3418, 2013.

[11] G. V. Rocha, P. Zhao, and B. Yu, "A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (splice)," *arXiv preprint arXiv:0807.3734*, 2008.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[13] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, "First-order methods for sparse covariance selection," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, pp. 56–66, 2008.

[14] L. Li and K.-C. Toh, "An inexact interior point method for l 1-regularized sparse covariance selection," *Mathematical Programming Computation*, vol. 2, no. 3-4, pp. 291–315, 2010.

[15] X. Yuan, "Alternating direction methods for sparse covariance selection," *preprint*, vol. 2, no. 1, 2009.

[16] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," in *Advances in neural information processing systems*, 2010, pp. 2101–2109.

[17] C. Wang, D. Sun, and K.-C. Toh, "Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2994–3013, 2010.

[18] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik, "Sparse inverse covariance matrix estimation using quadratic approximation," in *Advances in neural information processing systems*, 2011, pp. 2330–2338.

[19] J. Duchi, S. Gould, and D. Koller, "Projected subgradient methods for learning sparse gaussians," *arXiv preprint arXiv:1206.3249*, 2012.

[20] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic journal of statistics*, vol. 6, p. 2125, 2012.

[21] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu *et al.*, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[22] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of statistics*, vol. 37, no. 6B, p. 4254, 2009.

[23] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, "High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[24] J. Fan, Y. Feng, and Y. Wu, "Network exploration via the adaptive lasso and scad penalties," *The annals of applied statistics*, vol. 3, no. 2, p. 521, 2009.

[25] B. Li, H. Chun, and H. Zhao, "Sparse estimation of conditional graphical models with application to gene networks," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 152–167, 2012.

[26] Z. Ren, T. Sun, C.-H. Zhang, H. H. Zhou *et al.*, "Asymptotic normality and optimalities in estimation of large gaussian graphical models," *The Annals of Statistics*, vol. 43, no. 3, pp. 991–1026, 2015.

[27] J. Janková and S. van de Geer, "Honest confidence regions and optimality in high-dimensional precision matrix estimation," *Test*, vol. 26, no. 1, pp. 143–162, 2017.

[28] J. Jankova, S. Van De Geer *et al.*, "Confidence intervals for high-dimensional inverse covariance estimation," *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 1205–1229, 2015.

[29] J. Jankova and S. van de Geer, "Inference in high-dimensional graphical models," *arXiv preprint arXiv:1801.08512*, 2018.

[30] E. Levina, A. Rothman, and J. Zhu, "Sparse estimation of large covariance matrices via a nested lasso penalty," *The Annals of Applied Statistics*, pp. 245–263, 2008.

[31] S. Zhou, P. Rütimann, M. Xu, and P. Bühlmann, "High-dimensional covariance estimation based on Gaussian graphical models," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2975–3026, 2011.

[32] C. Uhler, "Geometry of maximum likelihood estimation in Gaussian graphical models," *The Annals of Statistics*, vol. 40, no. 1, pp. 238–261, 2012.

[33] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[34] V. Koltchinskii and K. Lounici, "Concentration inequalities and moment bounds for sample covariance operators," *arXiv preprint arXiv:1405.2468*, 2014.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[36] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans, "Enhancement of mr images using registration for signal averaging," *Journal of computer assisted tomography*, vol. 22, no. 2, pp. 324–333, 1998.
[37] J.-D. Tournier, F. Calamante, and A. Connelly, "Mrtrix: diffusion tractography in crossing fiber regions," *International Journal of Imaging Systems and Technology*, vol. 22, no. 1, pp. 53–66, 2012.

ACKNOWLEDGMENTS

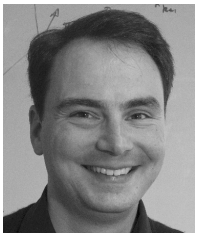
The research leading to these results has received funding from the European Research Council (ERC) under grant agreement No 678304, European Union's Horizon 2020 research and innovation program under grant agreement No 666992 (EuroPOND) and No 826421 (TVB-Cloud), and the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (IHU-A-ICM). The authors would like to thank Pascal Houillier for his insightful comments on the nephrological experiments.

Thomas Lartigue received graduate degrees from the École polytechnique and the École Normale Supérieur Paris-Saclay. He is carrying a PhD thesis funded by INRIA at the Centre de Mathématiques Appliquées of Ecole polytechnique. His research interests include Computational Statistics and Machine Learning, ranging from parameter estimation and Bayesian inference to optimisation.

Simona Bottani is a PhD student funded by INRIA at Aramis Lab team at the Brain and Spine Institute in Paris. Her PhD focuses on machine learning for differential diagnosis of neurodegenerative diseases. Bottani received a master degree on Biomedical engineer at Politecnico di Torino in December 2016.

Stéphanie Baron is a biochemist at Georges Pompidou European Hospital (Assistance Publique-Hopitaux de Paris), in Physiology Department. Her research interests include endocrinology, adrenal gland and hypertension. She received a PhD degree in Physiology from Paris Descartes University.

Olivier Colliot , PhD, is a Research Director at CNRS. He is the co-head of the ARAMIS Lab (Paris, France), a multidisciplinary laboratory dedicated to data science and machine learning applied to neurological diseases. His research interests include medical image computing, machine learning, image analysis and decision support systems. He received the PhD from Telecom ParisTech in 2003 and the Habilitation degree from University Paris-Sud in 2011. He is a member of the Editorial Board of Medical Image Analysis (Elsevier).



Stanley Durrleman is INRIA researcher, co-head of the ARAMIS Lab at the Brain Institute in Paris and founding director of the ICM Center for Neuroinformatics. He has developed statistical and computational approaches to create personalised digital brain models from multimodal patients data including image and clinical data. These models reproduce and predict the effect of a disease on brain anatomy and function in any patient. He received several awards including the MICCAI young investigator award and ERC starting grant from the European research council.

Stéphanie Allassonnière , Professor of Applied Mathematics in Paris Descartes School of medicine. She received her PhD degree in Applied Mathematics (2007), studies one year as postdoctoral fellow in the Center for Imaging Science, JHU, Baltimore. She joined the Applied Mathematics department of Ecole Polytechnique in 2008 as assistant professor and moved to Paris Descartes school of medicine in 2016 as Professor. Her researches focus on statistical analysis of medical databases in order to: understanding the common features of populations, designing classification, early prediction and decision support systems.