Stable network inference in high-dimensional graphical model using single-linkage

Emilie Devijver, Rémi Molinier, Mélina Gallopin

Stable network inference in high-dimensional graphical model using single-linkage

Emilie Devijver, Mélina Gallopin, Rémi Molinier

June 13, 2024

Abstract

Stability, akin to reproducibility, is crucial in statistical analysis. This paper examines the stability of sparse network inference in high-dimensional graphical models, where selected edges should remain consistent across different samples. Our study focuses on the Graphical Lasso and its decomposition into two steps, with the first step involving hierarchical clustering using single linkage. We provide theoretical proof that single linkage is stable, evidenced by controlled distances between two dendrograms inferred from two samples. Practical experiments further illustrate the stability of the Graphical Lasso's various steps, including dendrograms, variable clusters, and final networks. Our results, validated through both theoretical analysis and practical experiments using simulated and real datasets, demonstrate that single linkage is more stable than other methods when a modular structure is present.

1 Introduction

Interpretability, reproducibility and stability have become central challenges in statistics due to the recent advances in massive data and black-box models [36]. From a learning theory perspective, stability is crucial for generalization [7] while in practice, stability is fundamental for interpretability. This paper focuses on algorithmic stability, which pertains to the robustness of a procedure against data perturbation: does a method provide the same results on two perturbed data sets? Data perturbation techniques such as jackknife, sub-sampling or bootstrap have been extensively studied both theoretically and practically to assess the stability of statistical methods. [27] propose a framework for evaluating the stability of data set and predictive algorithms, concluding that even an inherently unstable method can appear stable over generated data if under the model. Recent methods are driven by the concept of stability [35, 34]. Stability in prediction has been explored accross various models, including random forests for interaction studies [3], bagging [30], and feature selection [26].

Network inference is a domain within statistics where stability is particularly critical. When performing network inference on two data sets derived from the same model with a small sample size using classical methods, the resulting inferred networks are often markedly different. This variability arises from the large number of parameters that need to be estimated and the complexity of the optimization task involved. However, without stability, interpretability becomes challenging, which undermines one of the major advantages of graphical models. This is especially pertinent in the context of regulatory networks derived from real omics data, where observations are typically limited [13, 19, 24]. As a result, practitioners have often criticized the developed methods, opting instead to manually select an appropriate subset of variables to focus on. However, such external knowledge is not always available and could be enhanced by a deeper understanding of the data and the application of machine learning tools.

Among various tools, graphical models are popular for network inference, and particularly valued for their interpretability. Gaussian Graphical Models (GGMs) are famous for embodying the Markov property. This property links the edges of the corresponding dependency graph between variables to the non-zero coefficients of the inverse covariance matrix. Hence, GGMs facilitate understanding the complex relationships among variables by translating statistical dependencies into a graphical representation, where each edge signifies a

direct conditional dependency between variables. The Graphical Lasso [14, 37] is a classical estimator that provides a sparse inverse covariance matrix $\Theta = \Sigma^{-1}$, solution of the following optimization problem: for a sample $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ coming from a random variable $\mathbf{Y} \sim \mathcal{N}_p(0, \Sigma)$, and its sample covariance estimate S,

$$\hat{\Theta}^{GL}(\lambda; S) = \underset{\Theta}{\operatorname{argmax}} \{\log\det\Theta - \operatorname{tr}(S\Theta) - \lambda\|\Theta\|_1\}$$

over nonnegative definite matrices Θ, and where λ is a nonnegative tuning parameter. Many algorithms have been proposed for solving the Graphical Lasso problem, but we focus here on the decomposition given in [33, 22]:

Step 1 Identify the connected components of the undirected graph with adjacency matrix A associated to the thresholded sample covariance;

Step 2 Perform Graphical Lasso with parameter $\lambda \geq 0$ on each connected component separately.

This approach has been practically used in [10, 18] to reduce the number of parameters to estimate for a fixed level of regularization, thereby improving computational efficiency. In [31], it was demonstrated that identifying the connected components in the Graphical Lasso solution (first step) is equivalent to performing single linkage hierarchical clustering based on a similarity matrix derived from the absolute values of the elements of the sample covariance matrix S. This decomposition allows flexibility in the choice of linkage in hierarchical clustering. [31] switched to average linkage, critiquing the chain effect of single linkage, and selected a model with two clusters, inferring a sparser model within each module. Conversely, [11] retained single linkage but provided non-asymptotic theoretical foundations for selecting the number of clusters.

In this paper, we argue that this decomposition into two steps enhances the stability of network inference. We experimentally illustrate this improvement and theoretically prove that single linkage is stable, whereas other classical linkages, such as average linkage, are not.

Several methods have been proposed to stabilize variable selection in GGMs, primarily based on resampling. In [1, 23], the authors suggest subsampling the observations, running a model on each sample, and retaining variables selected consistently across all or most samples. Both papers provide theoretical results that guarantee good performance asymptotically with increasing sample sizes. Building on [1], [9] evaluate the stability and accuracy of gene regulatory network inference using bootstrap aggregation. Additionally, [17], drawing from [1, 23], focuses specifically on bootstrap sampling for network inference. More recently, [6] proposed a score to measure the overall stability of the set of selected features, introducing a new calibration strategy for stability selection. In a broader context, [20] introduced `ESCV`, while [2] proposed removing the most influential observations to achieve stable networks, akin to the jackknife method.

However, these methods require substantial computation because they rely on subsampling. Furthermore, large sample sizes are necessary to ensure good performance with subsampling techniques.

Our main idea is that **estimators can be stable by construction and do not necessarily require additional steps to achieve stability**. This intrinsic stability can lead to more efficient and robust network inference.

In this paper, we make the following contributions:

- We derive theoretically the stability of the decomposition of the network into independent modules using hierarchical clustering;
- We show experimentally on simulated data and on real data sets the stability of the hierarchical clustering, of the subsequent clusters, and of the inferred network.

The remainder of the paper is organized as follows. In Section 2, we introduce the main theoretical result, about the stability of the hierarchical clustering. Section 3 investigates the numerical stability through several experiments on simulated and real dataset: study of the hierarchical clustering for several linkages, study of the considered clusters when selecting a model in the dendogram, and study of the inferred network.

2 Theoretical result for the stability of the modular decomposition

In this section, we are interested in the stability of the hierarchical clustering, in the sense that, if two samples are observed generated from the same distribution, we want to measure how close are the two dendograms provided by the hierarchical clustering. Let $(\mathbf{y}_1, \dots, \mathbf{y}_n)$ and $(\mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \tilde{\mathbf{y}}_i, \mathbf{y}_{i+1} \dots, \mathbf{y}_n)$ be two samples in $\mathbb{R}^p$ from the same multivariate normal distribution with density $\phi_p(\mathbf{0}, \Sigma)$ where $\Sigma_{j,j} = 1$ for all $j \in \{1, \dots, p\}$. We assume that observations are standardized, and we focus on empirical correlations matrices.

We start by the definition of a dendogram.

Definition 1. A *dendrogram* over $\{1, \dots, p\}$ is an application $\theta : [0, \infty) \to \mathcal{P}(\{1, \dots, p\})$, where $\mathcal{P}(\{1, \dots, p\})$ is the set of all partitions of $\{1, \dots, p\}$, such that

1. $\theta(0)$ is the partition with only singletons;
2. there exists $M > 0$ such that for all $t > M$, $\theta(t) = \{1, \dots, p\}$;
3. for every $t_1 \leq t_2$, the partition $\theta(t_1)$ is a refinement of the partition $\theta(t_2)$; and
4. for all $t_0 > 0$, there exists $\epsilon > 0$ such that for all $t \in [t_0, t_0 + \epsilon]$, $\theta(t) = \theta(t_0)$.

In other words, θ defines a nested family of partitions of $\{1, \dots, p\}$ which, according to the two other properties, starts with only singletons and ends with the whole space. The last condition ensures right-continuity, which allow the existence of some minimums (for example, in the definition of Ψ that follows). We denote Θ_p the set of all dendrograms on $\{1, \dots, p\}$.

We work here with ultrametrics, that are associated to dendrograms through a one-to-one mapping. Ultrametric spaces are metric spaces which satisfy a stronger type of triangle inequality.

Definition 2. A metric space (X, u) is called an *ultrametric space* if, for all $(x, x', x'') \in X^3$,

$$\max(u(x, x'), u(x', x'')) \geq u(x, x'').$$

For a finite set $\{1, \dots, p\}$, we denote $\mathcal{U}_p$ the set of all ultrametrics on $\{1, \dots, p\}$.

Theorem 9 in [8] gives a one to one correspondence

$$\Psi \colon \Theta_p \to \mathcal{U}_p$$

where, for $\theta \in \Theta_p$, $u = \Psi(\theta)$ is the ultrametric on $\{1, \dots, p\}$ defined for all $(x, y) \in \{1, \dots, p\}^2$ by

$$u(x, y) = \min \{t \geq 0 \mid x \text{ and } y \text{ are in the same subset in the partition } \theta(t)\}.$$

Note that $u = \Psi(\theta)$ is also, by definition, the *cophenetic* distance associated to the dendogram θ: $u(i, j)$ corresponds to the height at which stage i and j are merged together. We compare those cophenetic distances for two dendograms using the following distance.

Definition 3. The distance d_{coph} is defined by, for two dendograms $\theta_1, \theta_2 \in \Theta_p$, and their associated ultrametrics $u_1 = \Psi(\theta_1)$ and $u_2 = \Psi(\theta_2)$,

$$d_{\text{coph}}(\theta_1, \theta_2) = \max_{1 \leq i,j \leq p} |u_1(i, j) - u_2(i, j)|. \tag{2.1}$$

The inverse $\theta = \Psi^{-1}(u)$ for $u \in \mathcal{U}_p$ is given, for $t \geq 0$, by $\theta(t)$ to be the partition obtained from the equivalence relation $\sim_{u,t}$ where, for $(x, y) \in \{1, \dots, p\}^2$,

$$x \sim_{u,t} y \quad \iff \quad u(x, y) \leq t.$$

We will denote by C_p the complete simple graph with $\{1, \dots, p\}$ as set of vertices and a path in C_p with ν vertices will be encoded by a map $\eta \colon \{1, 2, \dots, \nu\} \to \{1, 2, \dots, p\}$ where, for all $k \in \{1, 2, \dots \nu\}$, $\eta(\nu)$ yields the kth vertex of the path. We introduce in the next definition the application u_A and then show that it is an ultrametric on $\{1, \dots, p\}$.

Definition 4. Let $A \in \mathbb{R}^{p\times p}$ be a symmetric matrix with positive nondiagonal entries, and zeros on the diagonal. We define the following application.

$$u_A \colon \{1,\dots,p\}^2 \longrightarrow \mathbb{R}$$
$$(i,j) \longmapsto \begin{cases} 0 \text{ if } i=j, \\ \min\limits_{\eta \text{ a path from } i \text{ to } j \text{ in } C_p} \left\{ \max\limits_k (A_{\eta(k),\eta(k+1)}) \right\} \text{ elsewhere.} \end{cases}$$

Proposition 1. *Let $A \in \mathbb{R}^{p\times p}$ be a symmetric matrix with positive nondiagonal entries, and zeros on the diagonal. The application u_A defines an ultrametric on $\{1,\dots,p\}$.*

Proof. We check all the properties of an ultrametric.

Positive definiteness: for all $(i,j) \in \{1,2,\dots,p\}^2$, $u_A(i,j) \geq 0$ and, by assumption, $u_A(i,j) = 0$ if and only if $i = j$.

Symmetric: since a is symmetric, u_A is also symmetric.

Strong triangle inequality: let us prove that for all $(i,j,k) \in \{1,2,\dots,p\}^3$,

$$u_A(i,k) \leq \max(u_A(i,j), u_A(j,k)).$$

Fix $(i,j,k) \in \{1,2,\dots,p\}^3$ and consider two paths in the complete graph C_p

$$\eta_1 \colon \{1,2,\dots,k_1\} \to \{1,2\dots,p\},$$
$$\eta_2 \colon \{1,2,\dots,k_2\} \to \{1,2,\dots,p\},$$

such that

$$u_A(i,j) = \max_{l\in\{1,\dots,k_1-1\}} \left(a_{\eta_1(l),\eta_1(l+1)}\right),$$
$$u_A(j,k) = \max_{l\in\{1,\dots,k_2-1\}} \left(a_{\eta_2(l),\eta_2(l+1)}\right).$$

Then, if we set $k = k_1 + k_2 - 1$ and $\eta_1 \cdot \eta_2$ the path η_1 followed by η_2 (which is a path from i to k in C_p), then η has k vertices and we have

$$\begin{aligned} u_A(i,k) &= \min_{\eta} \max_{l\in\{1,\dots,k-1\}} \left(a_{\eta(l),\eta(l+1)}\right) \\ &\leq \max_{l\in\{1,\dots,k-1\}} \left(a_{\eta_1\cdot\eta_2(l),\eta_1\cdot\eta_2(l+1)}\right) \qquad (\eta_1\cdot\eta_2 \text{ is a path from } i \text{ to } k) \\ &= \max\left(\max_{l\in\{1,\dots,k_1-1\}} \left(a_{\eta_1(l),\eta_1(l+1)}\right), \max_{l\in\{1,\dots,k_2-1\}} \left(a_{\eta_2(l),\eta_2(l+1)}\right) \right) \\ &= \max\left(u_A(i,j), u_A(j,k)\right). \end{aligned}$$

Hence, u_A is an ultrametric on $\{1,2,\dots,p\}$. □

For a symmetric matrix $A \in \mathbb{R}^p$ with positive nondiagonal entries and zeros on the diagonal, we denote by $\theta_A = \Psi^{-1}(u_A)$ the dendrogram associated to the ultrametric u_A.

Remark that, if the matrix A is associated to a distance d, the dendogram θ_A is exactly the one obtained by the single linkage hierarchical clustering with the distance d (see [8, Corollary 14]). One can particularly use $A = \mathbf{1} - |S^1|$, with S^1 the sample covariance matrix and $\mathbf{1}$ corresponds to the matrix with 1 for each coefficient, which is the one constructed in the first step of the Graphical Lasso [31].

The next proposition gives a control on the distance introduced in Definition 3, between dendograms induced by two different matrices.

Proposition 2. *Let $A_1, A_2 \in \mathbb{R}^{p\times p}$ be two symmetric matrices with positive entries, and zeros on the diagonal. Then*

$$d_{coph}(\theta_{A_1}, \theta_{A_2}) \leq \|A_1 - A_2\|_{\max}$$

where $||.||_{\max}$ corresponds to the maximum element of the matrix.

Proof. Set $m = \|A_1 - A_2\|_{\max}$. Let $(i,j) \in \{1,2,\dots,p\}^2$ and let $\eta_2 \colon \{1,2,\dots,K_2\} \to \{1,2,\dots p\}$ be a path such that

$$u_{A_2}(i,j) = \max_{k \in \{1,\dots,K_2-1\}} [A_2]_{\eta_2(k),\eta_2(k+1)}.$$

Then we have,

$$\begin{aligned} u_{A_1}(i,j) &\leq \max_{k \in \{1,\dots,K_2-1\}} [A_1]_{\eta_2(k),\eta_2(k+1)} && \text{(by definition of } u_{A_1}) \\ &\leq \max_{k \in \{1,\dots,K_2-1\}} \left(m + [A_2]_{\eta_2(k),\eta_2(k+1)}\right) && \text{(by definition of } m) \\ &= m + u_{A_2}(i,j) && \text{(by choice of } \eta_2). \end{aligned}$$

Hence, $u_{A_1}(i,j) - u_{A_2}(i,j) \leq m$ and symmetrically, $u_{A_2}(i,j) - u_{A_1}(i,j) \leq m$. Therefore, for all $(i,j) \in \{1,2,\dots,p\}^2$, $|u_{A_1}(i,j) - u_{A_2}(i,j)| \leq m$. Thus,

$$\max_{i,j} |u_{A_1}(i,j) - u_{A_2}(i,j)| \leq m.$$

□

Finally, we control the stability of the dendogram constructed in the first step of the Graphical Lasso in the following proposition.

Proposition 3. *Let two samples $(\mathbf{y}_1,\dots,\mathbf{y}_n)$ and $(\mathbf{y}_1,\dots,\tilde{\mathbf{y}}_i,\dots,\mathbf{y}_n)$ where $\tilde{\mathbf{y}}_i \sim \mathbf{y}_i \sim \mathbf{Y}$ and are iid, and S and $\tilde{S}$ the corresponding sample covariance matrices. Then, for $\alpha \in (0,1)$, with probability $1-\alpha$,*

$$\|S - \tilde{S}\|_\infty \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

Proof. For $1 \leq j,k \leq n$,

$$[S - \tilde{S}]_{j,k} = \frac{1}{n-1}\left(y_{i,j}y_{i,k} - \tilde{y}_{i,j}\tilde{y}_{i,k}\right).$$

From [25], we know the distribution function of Y_jY_k:

$$f_{Y_jY_k}(z;\rho) = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\left(\frac{\rho z}{1-\rho^2}\right) K_0\left(\frac{\|z\|}{1-\rho^2}\right),$$

where $\rho = \Sigma_{j,k}$ is the correlation between Y_j and Y_k. Eventually, we can compute the first two moments of $[S - \tilde{S}]_{j,k}$:

$$\begin{aligned} E([S-\tilde{S}]_{j,k}) &= 0, \\ Var([S-\tilde{S}]_{j,k}) &= Var\left(\frac{1}{n-1}\left(y_{i,j}y_{i,k} - \tilde{y}_{i,j}\tilde{y}_{i,k}\right)\right) \\ &= \frac{2}{(n-1)^2} Var\left(y_{i,j}y_{i,k}\right) \\ &= \frac{2}{(n-1)^2}(1+\Sigma_{j,k}^2), \end{aligned}$$

where the second line comes from the independence of y and $\tilde{y}$, and the last equality comes from [15], which characterizes the moments of the product of zero mean correlated normal random variables. We need to

control $\|S - \tilde{S}\|_\infty$. The union bound gives:

$$\begin{aligned}
\mathbb{P}(\|S - \tilde{S}\|_\infty \geq \eta) &= \mathbb{P}(\max_{j,k} |S_{j,k} - \tilde{S}_{j,k}| \geq \eta) \\
&\leq \sum_{j,k} \mathbb{P}(|S - \tilde{S}|_{j,k} \geq \eta) \\
&= \sum_{j,k} \mathbb{P}(\frac{1}{n-1} |y_{i,j}y_{i,k} - \tilde{y}_{i,j}\tilde{y}_{i,k}| \geq \eta) \\
&\leq \sum_{j,k} 2 \cdot \frac{1 + \Sigma^2_{k,j}}{(n-1)^2\eta^2},
\end{aligned}$$

where the last inequality comes from Tchebychev's inequality. Using that the correlations are bounded by 1, we get that

$$\mathbb{P}(\|S - \tilde{S}\|_\infty \geq \eta) \leq \frac{4p^2}{(n-1)^2\eta^2}.$$

□

This result is derived for a fixed sample size. Non-asymptotically, we have a control on the difference between the two dendrograms. Moreover, note that we used the Tchebychev inequality, but there may exist tighter concentration inequality.

Proposition 3 gives a stability result about the ultrametric induced by the empirical correlation matrix. Using the one-to-one correspondence Ψ, it can also be interpreted in terms of stability of the induced dendrogram as explained in Carlsson and Mémoli [8, Section 3.5]. It leads to the following theorem, which is our main theoretical contribution.

Theorem 1. *Let two samples* $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ *and* $(\mathbf{y}_1, \ldots, \tilde{\mathbf{y}}_i, \ldots, \mathbf{y}_n)$ *where* $\tilde{\mathbf{y}}_i \sim \mathbf{Y}$ *and are iid, and* S *and* $\tilde{S}$ *the corresponding sample covariance matrices. Then, for* $\alpha \in (0, 1)$*, with probability* $1 - \alpha$*,*

$$d_{coph}(|\theta_{\mathbf{1}-|S|}|, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \frac{2p}{(n-1)\sqrt{\alpha}}.$$

Proof. By Proposition 2 we have

$$d_{coph}(|\theta_{\mathbf{1}-|S|}|, \theta_{\mathbf{1}-|\tilde{S}|}) \leq \left\| \mathbf{1} - |S| - \left(\mathbf{1} - |\tilde{S}|\right) \right\|_{\max} = \left\| |S| - |\tilde{S}| \right\|_{\max}.$$

By the triangular inequality, the last term is less or equal to $\|S - \tilde{S}\|_{\max}$ and then Theorem 1 follows from Proposition 3. □

Asymptotically, the two collections of modules detected by the Graphical Lasso on two samples where only one observation differs, varying the regularization parameter λ, are the same. This means that the single linkage used in the first step of the Graphical Lasso is a good choice, with respect to the stability of the collection of models that is considered.

Similarly to Carlsson and Mémoli [8, Remark 17], we can show that the complete linkage and the average linkage are unstable: small perturbations of the matrix A may lead to large perturbations of the corresponding ultrametric.

3 Experiments

In this section, we evaluate in practice the stability of each step of the Graphical Lasso. First we provide the experimental design, describing the data generation process, and the two real dataset we are studying. Then, we illustrate 1/ the stability of dendograms using hierarchical clustering, varying the linkage (illustrating exactly Theorem 1), 2/ the stability of the clusters get by cutting the dendogram with some model selection criterion, 3/the stability of the network inference.

Table 1: Stability of dendograms using hierarchical clustering with several linkages on generated data, BRCA and equities. We display the normalized distance and its standard deviation in parenthesis. Best results are bolded.

coph_n	AL	CL	ML	SL	WL
Generated data	0.53 (0.12)	0.72 (0.12)	0.54 (0.12)	**0.33** (0.06)	0.72 (0.14)
BRCA	0.81 (0.08)	0.91 (0.05)	0.85 (0.06)	**0.59** (0.12)	1.01 (0.10)
Equities	0.87 (0.05)	0.97 (0.03)	0.89 (0.05)	**0.73** (0.09)	1.18 (0.27)

3.1 Experimental design

The design of the simulations is as follows. For a fixed structured covariance matrix Σ with a block diagonal structure, we simulate V samples of n observations from a p-variate normal distribution with a mean of zero and the structured covariance matrix Σ. We then compare the inferred networks pairwise, resulting in $V(V-1)/2$ comparisons. The number of variables is set to $p = 100$, the sample size to $n = 70$, and the number of samples to $V = 17$ per fixed covariance matrix. We consider $R = 5$ different covariance matrices (generated randomly, each with the same block decomposition), with the number of blocks in the diagonal matrix set to $K = 15$, and each block containing 6 or 7 variables.

Two real datasets are considered.

The BRCA dataset is a gene expression dataset for patients with breast cancer, measured with RNA-Sequencing. The data are generated by the TCGA Research Network: `http://cancergenome.nih.gov/`, and downloaded from the web portals `https://tcga-data.nci.nih.gov/tcga/` using the TCGA2STAT tool [32]. We have $n = 1212$ samples and $p = 9191$ genes, but we focus on the $p = 200$ most variable genes. We construct 17 batches of size 70, leading to 1190 observations.

Equities dataset includes stock market data available in the R package `huge` and has been studied in [31]. It contains closing prices of 452 stocks over 1258 trading days. We focus on the $p = 200$ most variables stocks' close prices and we construct 17 patches of size 70.

3.2 Stability of hierarchical clustering: which linkage method?

In this section, we validate the theoretical results obtained in Section 2. When applying hierarchical clustering, various linkage methods can be used. We compare the performance in stability of the most well-known methods: average linkage (AL), complete linkage (CL), McQuitty linkage (ML), single linkage (SL), and Ward linkage (WL). The measure used to compare dendograms is the distance introduced in Definition 3, which we normalize to facilitate the analysis: for two matrices A_1, A_2, and their associated dendograms $\theta_{A_1}, \theta_{A_2}$,

$$d^N_{\text{coph}}(\theta_{A_1}, \theta_{A_2}) = \max_{1\leq i,j\leq p} \left| \frac{u_{A_1}(i,j)}{\max(u_{A_1})} - \frac{u_{A_2}(i,j)}{\max(u_{A_2})} \right|.$$

Table 1 presents the stability of dendrograms generated using hierarchical clustering with various linkage methods across different data sets: generated data, BRCA and equities. The table displays the normalized distance d^N_{coph} along with its standard deviation in parentheses. The best results in each row are highlighted in bold.

When the covariance matrix has a block-diagonal structure, as in the generated data, single linkage (SL) is clearly the most stable method, exhibiting the lowest normalized distance with a relatively low standard deviation. This indicates that single linkage is less sensitive to small perturbations in the data, maintaining consistent clustering structures.

For the real data sets, the conclusion holds consistently. In the BRCA data, single linkage again demonstrates superior stability, compared to other methods, suggesting its robustness in clustering biological data where sample variability is often high. Similarly, in the equities data set, single linkage achieves the best stability, indicating its effectiveness in financial data clustering, which often involves high-dimensional and noisy data.

Table 2: ARI between the clusters get by hierarchical clustering using several linkages and several model selection criterion on generated data, BRCA and equities. Best results are bolded.

Data	criterion	single	average	complete	ward	mcquitty
Generated data	K	0.19	0.45	0.33	**0.65**	0.45
	2K	**0.85**	0.8	0.69	0.78	0.81
	2	0.07	0.02	0.02	**0.40**	0.02
	SH	**0.80**	0.65	0.66	0.68	0.68
	BIC	0.10	0.05	0.03	**0.47**	0.05
BRCA	2	0.09	0.10	0.02	**0.78**	0.05
	SH	**0.57**	0.49	0.32	0.37	0.44
	BIC	**0.39**	0.26	0.11	0.27	0.17
Equities	2	0.00	-0.01	0.04	**0.61**	0.00
	SH	**0.36**	0.35	0.26	0.22	0.31
	BIC	**0.21**	0.18	0.09	0.15	0.16

Analyzing the methods in order of stability across all three data sets, single linkage (SL) is the most stable, followed by average linkage (AL), McQuitty linkage (ML), complete linkage (CL), and Ward's linkage (WL). The higher values for average, McQuitty, complete, and Ward's linkages suggest these methods are more sensitive to data perturbations, resulting in less stable dendrograms. Ward's linkage, in particular, shows the highest values and standard deviations, indicating it is the least stable method in this context.

These results underscore the importance of selecting an appropriate linkage method for hierarchical clustering, especially when stability is a critical concern. Single linkage's robustness across different data types suggests it as a preferable choice for ensuring consistent clustering outcomes in practical applications.

3.3 Stability of a clustering

In this section, the dendrogram is cut to focus on clustering. Given that we are considering Gaussian Graphical Models, we can recast this task as a model selection problem. We evaluate two model selection criteria: the Bayesian Information Criterion (BIC) [29] and the slope heuristic (SH) [5, 4]. Additionally, when knowing the ground truth on generated data, we consider the model with 2 clusters, the true number of clusters K (for generated data), and twice the true number of clusters, $2K$.

Table 2 displays the Adjusted Rand Index (ARI) [28] between clusters derived from hierarchical clustering, cut according to different model selection criteria. The ARI measures the similarity between two partitions, with an ARI of 1 indicating a perfect match. We evaluate the following linkage methods: single linkage (SL), average linkage (AL), complete linkage (CL), McQuitty linkage (ML), and Ward's linkage (WL).

For the generated data, the single linkage (SL) and Ward's linkage (WL) show the best performance. Although Ward's linkage performs poorly in terms of distance between dendograms, as seen in Section 3.2, it performs well when considering clustering, likely due to high variability in early merges but more stability with larger clusters. Notably, single linkage combined with $2K$ clusters achieves the highest ARI, indicating that considering a higher number of clusters enhances stability. Among non-oracle methods, the slope heuristic (SH) combined with single linkage performs the best. Generally, SH provides stable results, whereas BIC performs poorly.

For the real data sets (BRCA and equities), the conclusions are similar: single linkage (SL) and Ward's linkage (WL) are the most competitive. Ward's linkage with 2 clusters is the most stable for both data sets, but not sparse, followed closely by single linkage combined with the slope heuristic (SH).

Table 3 displays the number of clusters selected on average by the slope heuristic and BIC for each dataset. The number of clusters selected by BIC is generally smaller, often underestimating the true number in the simulated data, while the slope heuristic tends to overestimate the number of clusters. Single linkage tends to select a higher number of clusters, contributing to its stability. As with the $2K$ scenario, the slope heuristic's tendency to overestimate the number of clusters generally contributes to greater stability.

Table 3: Number of clusters selected (in mean) for several linkages and several model selection criterion on generated data (100 variables, 15 groups), BRCA (200 variables) and equities (100 variables).

Data	criterion	single	average	complete	ward	mcquitty
simulated	SH	25	22	28	22	24
	BIC	10	2	2	5	3
equities	SH	155	100	70	60	90
	BIC	105	40	6	5	28
BRCA	SH	112	41	18	18	28
	BIC	88	14	5	7	9

These results indicate that the choice of linkage method and model selection criterion significantly impacts the stability and accuracy of clustering. Single linkage and the slope heuristic generally provide the most stable results across different data sets and scenarios.

3.4 Stability of inferred networks

In this section, we evaluate the stability of networks inferred by classical methods. Stability is assessed using the *normalized Hamming distance* between two graphs G_1 and G_2, with respective adjacency matrices A_1 and A_2. The normalized Hamming distance is defined as

$$d_H(G_1, G_2) = \frac{2\|A_1 - A_2\|_1}{\|A_1\|_1 + \|A_2\|_1}. \tag{3.1}$$

This metric provides a measure of the difference between two graphs, normalized by the total number of edges in both graphs. Additionally, we report the density of the inferred graphs, which is the proportion of nonzero coefficients in the adjacency matrix, and the CPU time required for the computations.

For the generated data, we further calibrate the estimation methods using several performance metrics: sensitivity TP/(TP+FN), specificity TN/(TN+FP), precision TP/(TP+FP), and false discovery rate (FDR) FP/(TP+FP), where TP denotes true positive, FN denotes false negative, FP denotes false positives, and TN denotes true negatives. While these metrics do not directly relate to stability, they help identify methods that infer networks close to the true structure. One prefers a high precision, recall and specificity, a density close to 0.03 on the generated data (the value on the true graph), a low normalized Hamming distance and a low CPU time.

We compare the following strategies, based on or extended from the Graphical Lasso:

- One-step Graphical Lasso methods:
  - BIC: Regularization parameter selected using Bayesian Information Criterion [29].
  - EBIC: Extended Bayesian Information Criterion with $\gamma = 0.5$ [12].
  - STARS: Stability Approach to Regularization Selection [21].
  - ESCV: Extended Stability Criterion Validation [20].
- Stabilized methods based on the one-step Graphical Lasso:
  - BoLasso (BL): Bootstrap Lasso [1], with regularization parameter fixed by cross-validation, using bootstrap over $m = 100$ samples of size n subsampled from the observations with replacement
  - Stability Selection (SS): [23], with $m = 100$ samples of size $n/2$ subsampled without replacement, and regularization parameter selected such that $\sqrt{0.8p}$ variables are chosen.
- Two-step Graphical Lasso methods:

Table 4: Performance on simulated dataset. We compare the performance in estimation (evaluated by the precision, Prec, the recall, Recall, and the specificity, Spec, and the density of the graph; the performance in stability (evaluated by the normalized hamming distance) and the computation time (evaluated by the CPU time). Best scores are bolded.

1 step	BIC	EBIC	STARS	ESCV	BL	SS
Dens	0.14	**0.03**	0.06	0.00	**0.03**	0.01
	0.04	*0.01*	*0.00*	*0.00*	*0.00*	*0.00*
Hamm	0.37	0.04	0.12	**0.00**	**0.01**	**0.00**
	0.07	*0.02*	*0.01*	*0.00*	*0.00*	*0.00*
CPU	16	16	386	235	323	1681
2 steps	SL-SH_{BIC}	SL-SH_{EBIC}	SL-SH_{STARS}	SL-SH_{ESCV}	SL-SH_{BL}	AL-2_{sparse}
Dens	0.04	0.04	0.01	0.01	0.02	0.04
	0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
Hamm	0.03	0.06	0.02	0.05	**0.01**	0.06
	0.01	*0.01*	*0.01*	*0.01*	*0.00*	*0.01*
CPU	12	8	164	102	1334	11

1 step	BIC	EBIC	STARS	ESCV	BL	SS
Prec	0.39	0.93	0.67	**1.00**	**1.00**	**1.00**
	0.08	*0.04*	*0.02*	*0.00*	*0.00*	*0.00*
Recall	**0.76**	0.60	0.70	0.26	0.61	0.44
	0.03	*0.15*	*0.01*	*0.00*	*0.01*	*0.01*
Spec	0.89	**1.00**	0.97	**1.00**	**1.00**	**1.00**
	0.04	0.00	0.00	0.00	0.00	0.00
2 steps	SL-SH_{BIC}	SL-SH_{EBIC}	SL-SH_{STARS}	SL-SH_{ESCV}	SL-SH_{BL}	AL-2_{sparse}
Prec	0.94	0.94	0.99	0.99	**1.00**	0.83
	0.05	*0.05*	*0.01*	*0.01*	*0.00*	*0.03*
Recall	0.69	0.69	0.39	0.37	0.57	0.66
	0.02	*0.02*	*0.06*	*0.04*	*0.02*	*0.02*
Spec	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99
	0.00	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*

– Single Linkage: As highlighted in theory and practice, cut with the slope heuristic, with regularization parameter selected by BIC, STARS, ESCV, and BoLasso within each module.

– CGL: Average linkage with 2 clusters, where in each module the sparser model is selected [31].

Note that Stability Selection was not run in the two-step Graphical Lasso methods due to high computational cost.

3.4.1 Results on simulated dataset

Table 4 details the performance of various methods in inferring from simulated data, focusing on metrics such as normalized Hamming distance, graph density, and CPU time.

In the simulated data context, the single linkage method paired with ESCV achieves the highest stability as indicated by the normalized Hamming distance. This method is closely followed by single linkage with BoLasso, which excels in graph density. These approaches, especially those utilizing single linkage, outperform both one-step methods and CGL in stability. Important Observation: Graphical Lasso using ESCV often appears very stable (Hamming distance is zero) but infers an empty network (density is zero), making it uninteresting.

Table 5: Performance on BRCA. We compare the density, the performance in stability (evaluated by the normalized Hamming distance Hamm) and the computation time (evaluated by the CPU time).

1 step	BIC	EBIC	STARS	ESCV	BL	SS
Dens	0.05	0.00	0.09	0.00	0.01	0.00
	0.07	*0.00*	*0.01*	*0.00*	*0.00*	*0.00*
Hamm	0.19	0.00	0.20	0.00	0.02	0.01
	0.16	*0.00*	*0.01*	*0.00*	*0.00*	*0.00*
CPU	73	73	1621	971	1071	7919
2 steps	SL-SH$_{\text{BIC}}$	SL-SH$_{\text{EBIC}}$	SL-SH$_{\text{STARS}}$	SL-SH$_{\text{ESCV}}$	SL-SH$_{\text{BL}}$	AL-2$_{\text{sparse}}$
Dens	0.02	0.02	0.01	0	0.01	0.26
	0.01	*0*	*0*	*0*	*0*	*0.03*
Hamm	0.04	0.04	0.02	0	0.01	0.68
	0.01	*0.01*	*0.01*	*0*	*0*	*0.04*
CPU	118	14	250	560	2372	2

Table 6: Performance on equities. We compare the density, the performance in stability (evaluated by the normalized Hamming distance Hamm) and the computation time (evaluated by the CPU time).

1 step	BIC	EBIC	STARS	ESCV	BL	SS
Dens	0.00	0.00	0.07	0.00	0.00	0.00
	0.00	*0.00*	*0.01*	*0.00*	*0.00*	*0.00*
Hamm	0.00	0.00	0.20	0.00	0.01	0.01
	0.00	*0.00*	*0.01*	*0.00*	*0.00*	*0.00*
CPU	67	68	1535	917	990	7690
2 steps	SL-SH$_{\text{BIC}}$	SL-SH$_{\text{EBIC}}$	SL-SH$_{\text{STARS}}$	SL-SH$_{\text{ESCV}}$	SL-SH$_{\text{BL}}$	AL-2$_{\text{sparse}}$
Dens	0.01	0.01	0	0	0.01	0.72
	0.00	*0*	*0*	*0*	*0*	*0.01*
Hamm	0.03	0.05	0.01	0.01	0.03	0.79
	0.01	*0.01*	*0.01*	*0.01*	*0.01*	*0.02*
CPU	80	6	112	71	651	187

Graphical Lasso methods using BIC and STARS do not perform well in estimation, producing overly dense graphs and lacking stability. Among the one-step methods, Graphical Lasso with EBIC, BoLasso, and Stability Selection demonstrate better performance in both estimation quality and stability, albeit with slower computation times due to bootstrap procedures.

Two-step methods generally show improved performance with better estimation accuracy and increased stability. This enhancement is attributed to the block-diagonal network structure inherent in the data generation process, which aligns well with the decomposition strategy used in these methods. While these approaches generally require more computation time, they offer superior stability, particularly evident in methods employing single linkage (SL-SH).

In summary, EBIC emerges as the standout among one-step methods, balancing density, stability, and computational efficiency. Two-step methods, particularly those leveraging single linkage, enhance stability by effectively utilizing the network structure. However, the overall choice of method should consider a trade-off between estimation quality, stability, and computational demands based on specific application requirements.

3.4.2 Real data analysis

Tables 5 and 6 present the performance on the BRCA and Equities datasets, respectively, in terms of density, normalized Hamming distance (Hamm), and computation time (CPU time in seconds).

One-step Methods: ESCV and EBIC show very high stability (Hamming distance is zero) but infer

empty networks (density is zero) on both the BRCA and Equities datasets. While stable, they lack practical utility due to this issue. STARS outperforms BIC and BoLasso in terms of network estimation but comes with significantly higher computational costs. BoLasso and Stability Selection (SS) show promise in both stability and network estimation quality. However, they are computationally intensive, especially Stability Selection.

Two-step Methods: Single Linkage with SH demonstrates notable improvements in stability and estimation performance compared to one-step methods. It effectively leverages the block-diagonal network structure present in the generated data. CGL is generally unstable, confirming theoretical expectations about the limitations of average linkage methods in this context. Overall, two-step methods improve network estimation and stability across both datasets. They mitigate the limitations observed in one-step methods, particularly in capturing the block-diagonal structure of the generated data.

In conclusion, while one-step methods like STARS show competitive performance, especially in terms of estimation accuracy, two-step methods, particularly those utilizing Single Linkage with appropriate selection criteria, offer superior stability and estimation quality, albeit at increased computational costs. These findings underscore the importance of method selection based on both performance metrics and computational feasibility in practical applications of graphical model inference.

4 Discussion and conclusion

In this paper, we propose an analysis of stability for several network inference methods, with a focus on hierarchical clustering methods using different linkages, influenced by the decomposition of the graphical lasso into two steps. Our study highlights the potential of single linkage in scenarios with a modular structure, challenging conventional wisdom and opening new avenues for stable network inference methods.

Contrary to the common advice to avoid single linkage due to its chaining property [16], our results demonstrate that single linkage is more stable than other methods when a modular structure is present. This finding is supported by both theoretical analysis and practical experiments.

While our theoretical results are robust, we were unable to provide a complete proof of stability for the full method combining single linkage with any model selection criterion, and particularly considering the slope heuristic (SL+SH). This challenge arises from the complexity involved in accounting for model selection within the stability framework. Addressing this limitation remains an open question and a promising direction for future research.

References

[1] Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pp. 33–40.

[2] Bar-Hen, A. and J. M. Poggi (2016). Influence measures and stability for graphical models. *Journal of Multivariate Analysis 147*, 145–154.

[3] Basu, S., K. Kumbier, J. B. Brown, and B. Yu (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Science 115*(8), 1943–1948.

[4] Baudry, J.-P., C. Maugis, and B. Michel (2012). Slope heuristics: overview and implementation. *Statistics and Computing 22*(2), 455–470.

[5] Birgé, L. and P. Massart (2001). Gaussian model selection. *Journal of the European Mathematical Society 3*(3), 203–268.

[6] Bodinier, B., S. Filippi, T. H. Nøst, J. Chiquet, and M. Chadeau-Hyam (2023). Automated calibration for stability selection in penalised regression and graphical models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, qlad058.

[7] Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *Journal of Machine Learning Research 2*(Mar), 499–526.

[8] Carlsson, G. and F. Mémoli (2010). Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research 11*, 1425–1470.

[9] Colby, S., R. McClure, C. Overall, R. Renslow, and J. Mcdermott (2018). Improving network inference algorithms using resampling methods. *BMC Bioinformatics 19*.

[10] Danaher, P., P. Wang, and D. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(2), 373–397.

[11] Devijver, E. and M. Gallopin (2018). Block-diagonal covariance selection for high-dimensional gaussian graphical models. *Journal of the American Statistical Association 113*(521), 306–314.

[12] Foygel, R. and M. Drton (2010). Extended bayesian information criteria for gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 604–612. Curran Associates, Inc.

[13] Frazee, A. C., B. Langmead, and J. T. Leek (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics 12*(449), 1–30.

[14] Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

[15] Gaunt, R. E. (2022). The basic distributional theory for the product of zero mean correlated normal random variables. *Statistica Neerlandica 76*(4), 450–470.

[16] Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

[17] Haury, A.-C., F. Mordelet, P. Vera-Licona, and J.-P. Vert (2012). Tigress: Trustful inference of gene regulation using stability selection. *BMC Systems Biology 6*(1), 145.

[18] Hsieh, C.-J., M. A. Sustik, I. S. Dhillon, and P. Ravikumar (2014). QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation. *Journal of Machine Learning Research 15*, 2911–2947.

[19] Krumsiek, J., K. Suhre, T. Illig, J. Adamski, and F. J. Theis (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology 5*(1), 21.

[20] Lim, C. and B. Yu (2016). Estimation Stability With Cross-Validation (ESCV). *Journal of Computational and Graphical Statistics 25*(2), 464–492.

[21] Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 1432–1440. Curran Associates, Inc.

[22] Mazumder, R. and T. Hastie (2012). Exact covariance thresholding into connected components for large-scale Graphical Lasso. *Journal of Machine Learning Research 13*, 781–794.

[23] Meinshausen, N. and P. Buhlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

[24] Michailidis, G. and F. d'Alché Buc (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical Biosciences 246*(2), 326–334.

[25] Nadarajah, S. and T. K. Pogány (2016). On the distribution of the product of correlated normal random variables. *Comptes Rendus Mathematique 354*(2), 201–204.

[26] Pfister, N., E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann (2021). Stabilizing variable selection and regression. *The Annals of Applied Statistics 15*(3), 1220 – 1246.

[27] Philipp, M., T. Rusch, K. Hornik, and C. Strobl (2017). Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics 131*.

[28] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

[29] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

[30] Soloff, J. A., R. F. Barber, and R. Willett (2024). Bagging provides assumption-free stability. *Journal of Machine Learning Research 25*(131), 1–35.

[31] Tan, K., D. Witten, and A. Shojaie (2015). The Cluster Graphical Lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis 85*, 23–36.

[32] Wan, Y. W., G. I. Allen, and Z. Liu (2015). TCGA2STAT: Simple TCGA data access for integrated statistical analysis in R. *Bioinformatics 32*(6), 952–954.

[33] Witten, D. M., J. H. Friedman, and N. Simon (2011). New insights and faster computations for the Graphical Lasso. *Journal of Computational and Graphical Statistics 20*(4), 892–900.

[34] Wu, S., A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences 113*(16), 4290–4295.

[35] Yu, B. (2013). Stability. *Bernoulli 19*(4), 1484–1500.

[36] Yu, B. and R. Barter (2024). *Veridical Data Science*.

[37] Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35.