Stability of a network inference procedure in high-dimension

Emilie Devijver [1] & Mélina Gallopin [2] & Rémi Molinier [1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France. Mail : {emilie.devijver, remi.molinier}@univ-grenoble-alpes.fr*

[2] *Université Paris-Sud, Institut de Biologie Intgrative de la Cellule, CNRS 91100 Orsay, France. Mail : melina.gallopin@u-psud.fr*

Résumé. L'inférence de réseaux permet d'évaluer et de représenter les dépendences entre des variables continues. Les modèles graphique gaussiens ont été développés pour résoudre ce problème en grande dimension sous certaines hypothèses. Ce papier porte sur la stabilité d'une procédure d'inférence appelée `shock` et introduite dans Devijver and Gallopin (2018), qui infère un réseau modulaire via une matrice de covariance diagonale par blocs. Cette structure a beaucoup d'avantages, dont la réduction de dimension, l'interprétabilité et la stabilité. Ce dernier point est explicité dans ce papier, d'un point de vue théorique (via des arguments topologiques) et d'un point de vue numérique.

Mots-clés. Inférence de réseaux, Stabilité, Classification hiérarchique.

Abstract. Network inference is widely utilized to evaluate and represent dependencies between continuous variables. Gaussian graphical models have been developed and tackle the high-dimension problem through several assumptions. This presentation deals with the stability of a procedure called `shock` and introduced in Devijver and Gallopin (2018), which infers a modular network using a block-diagonal decomposition of the covariance matrix. This structure has strong advantages, among such reducing the dimension, facilitating the interpretation and being stable. The stability of the procedure is supported by strong theoretical guarantees based on topological tools, intensive simulations and real data analysis.

Keywords. Network inference, Stability, Hierarchical clustering

1 Introduction

Evaluating the correlations between variables has become an important question, as with real data sets the independence assumption is clearly violated. Among others tools, graphical models are popular for their interpretation, and Gaussian graphical models (GGMs) are famous for the Markov property, relating the edges of the corresponding graph to the non-zero coefficients of the inverse covariance matrix. In high-dimensional contexts, this sparsity has been studied thorough an ℓ_1 penalized log-likelihood method (Friedman

et al., 2007). In Devijver and Gallopin (2018), the authors proposed a non-asymptotic model selection procedure, called `shock`, to infer a modular network, supported by theoretical guarantees ensuring that the procedure is adaptive minimax to the structure of the covariance matrix.

One main drawback of the GGM's method is the stability for finite distance, particularly for methods based on the ℓ_1 penalized log-likelihood: if we run the methods on two data sets coming from the same model with small sample size, the inferred networks will be different. Stability, as reproducibility, is very important in statistics, as described in Yu (2013). Without stability, interpretability is difficult, whereas interpretability is a major advantage of graphical models. One example of applications of graphical models are regulatory networks inferred from omics data with a limited number of observations (Akbani et al, 2014).

Some methods have been proposed to stabilize the variable selection, initially for the Lasso estimator in the general framework of the linear regression, more recently for the GGM's. In Bach (2008) and Meinshausen and Buhlmann (2010), the author propose to subsample the observations, run a model on each sample, and keep variables selected every time on every samples, or a high fraction of the samples. In both papers, theoretical results guarantee good performances asymptotically in the number of observations. For the specific context of network inference, Liu et al (2010) introduced `StARS`. However, all those methods require computation, because they are rely on subsampling. Moreover, large sample size are needed to subsample.

In this presentation, we propose to prove, theoretically and numerically, that the procedure `shock` is stable by construction, without adding any stabilization step by subsampling.

2 Model and method

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be a sample in $\mathbb{R}^p$ from a multivariate normal distribution with density $\phi_p(\mathbf{0}, \Sigma)$ where $\Sigma_{j,j} = 1$ for all $j \in \{1, \ldots, p\}$. Our goal is to study the stability of the network inference method called `shock`. The network is decomposed onto independent components by clustering the variables, using single linkage hierarchical clustering. The choice of the final model is performed using the slope heuristics.

The output of hierarchical methods can be regarded as finite ultrametric spaces, which we consider with the Gromov-Hausdorff distance. We denote by (Y, u_Y) the finite ultrametric space corresponding to the output yielded by single linkage hierarchical clustering, where similarities are computed through the empirical covariance matrix S. Roughly speaking, a *dendogram* (X, θ) over a finite set X is defined to be a nested family of partitions, usually represented graphically as a rooted tree, with $\theta : [0, \infty) \to \mathcal{P}(X)$ the parameter representing the notion of scale, and reflecting in the height of the different scales. θ must satisfy conditions, as the initial decomposition of the space is the space

itself; for large enough parameter, the partition of the space becomes trivial. Remark that from Theorem 9 in Carlsson and Mémoli (2010), dendograms and ultrametrics are equivalent through the one to one mapping denoted Ψ, such that $\Psi(\theta)$ is defining the ultrametric and $\Psi^{-1}(u)$ is denoting the dendogram. Thus, if $\mathcal{X} = \cup_n \mathcal{X}_n$ denotes the set of finite metric spaces, and $\mathcal{U} = \cup_n \mathcal{U}_n$ the set of all finite ultrametric spaces, *a hierarchical clustering method* is defined to be a map

$$\mathcal{I} : \mathcal{X} \to \mathcal{U} \text{ s.t. } (X, d) \in \mathcal{X}_n \mapsto (X, u) \in \mathcal{U}_n, n \in \mathbb{N}.$$

As an example, we consider the maximal sub-dominant ultrametric: there is a canonical construction, leading to $\mathcal{I}^* : \mathcal{X} \to \mathcal{U}$ given by $(X, d) \mapsto (X, u^*)$ where

$$u^*(x, x') := \min \left\{ \max_{i=0,\dots,k-1} d(x_i, x_{i+1}) \text{ s.t.} x = x_0, \dots, x_k = x' \right\}$$

$\mathcal{I}^*$ corresponds to the single linkage hierarchical clustering, which is stable, as proved in Carlsson and Mémoli (2010). The single linkage plays an important role here, complete linkage and average linkage not satisfying the same property.

In `shock`, the dendogram is cut using the slope heuristic, a non asymptotic model selection criterion having good guarantees in this framework (see Devijver and Gallopin, 2018). The slope heuristics is choosing a time t^{SH} such that the model selected is deduced from $(\Psi^{-1}(u^*))(t^{\text{SH}})$. Then, the last step of `shock` consists in inferring in each module a network using the graphical Lasso estimator, where the regularization parameter is selected by the BIC as the dimension is no longer high.

In this presentation, we are interested in stability for network inference, in the sense that, if two samples are generated from the same distribution, we want to measure how close are the two inferred networks. The stability will be measured through the normalized *Hamming distance*, defined by, for two graphs G_1 and G_2 with respective adjacency matrices A_1 and A_2,

$$d_H^{\text{norm}}(G_1, G_2) = \frac{\|A_1 - A_2\|_1}{\|A_1\|_1 + \|A_2\|_1}.$$

3 Theoretical results for stability

First, we use a matrix version of the Bernstein inequality to control the concentration of the empirical correlation matrix.

Lemma 1 *Let $\mathbf{y}^1 = (\mathbf{y}_1^1, \dots, \mathbf{y}_n^1)$ and $\mathbf{y}^2 = (\mathbf{y}_1^2, \dots, \mathbf{y}_n^2)$ be two samples in $\mathbb{R}^p$ from the same multivariate normal distribution with density $\phi_p(\mathbf{0}, \Sigma)$, where Σ is a correlation matrix. Denote S_1 and S_2 the two empirical correlation matrices. Then, with probability*

$1-\alpha$, $\|S_1 - S_2\| \leq q_{p,n,\alpha}$, *where*

$$q_{p,n,\alpha} = \begin{cases} 4\sqrt{\frac{2p^2}{3n}\log(2p/\alpha)} & \text{if } \alpha \geq 2p/\exp(3n/32) \\ \frac{32p}{3n}\log(2p/\alpha) & \text{if } \alpha \leq 2p/\exp(3n/32). \end{cases}$$

and where the norm corresponds to the spectral norm (the largest eigenvalue).

For a symmetric matrix $A \in \mathbb{R}^{p\times p}$ with positive entries, we define, for $(i,j) \in \{1,2,\ldots,p\}^2$, for C_p the complete simple graph over vertices set $\{1,\ldots,p\}$,

$$u_A(i,j) = \begin{cases} 0 & \text{if } i = j, \\ \min\left\{\max_k (A_{\eta(k),\eta(k+1)}) \mid \eta \text{ a path from } i \text{ to } j \text{ in } C_p\right\} & \text{elsewhere.} \end{cases}$$

This defines an ultrametric on $\{1,2,\ldots,p\}$. It is used with matrices $A = \mathbf{1} - S$ where S is an empirical correlation matrix and $\mathbf{1}$ corresponds to the matrix with 1 for each coefficient. Thus, the associated dendogram is defined by $\Psi^{-1}(u_A)$, corresponding to the one get by thresholding the matrix A.

Proposition 1 *Let* $\mathbf{y}^1 = (\mathbf{y}_1^1,\ldots,\mathbf{y}_n^1)$ *and* $\mathbf{y}^2 = (\mathbf{y}_1^2,\ldots,\mathbf{y}_n^2)$ *be two samples in* $\mathbb{R}^p$ *from the same multivariate normal distribution with density* $\phi_p(\mathbf{0},\Sigma)$*, where* Σ *is a correlation matrix. Let* S_1 *and* S_2 *be the corresponding empirical correlation matrices and set* $u_1 = u_{\mathbf{1}-|S_1|}$ *and* $u_2 = u_{\mathbf{1}-|S_2|}$.

Then, with probability $1-\alpha$, $\max_{i,j}|u_1(i,j) - u_2(i,j)| \leq \|S_1 - S_2\|_{\max} \leq q_{p,n,\alpha}$.

We are currently working on controlling the model selection step used to cut the two dendograms (based on the slope heuristic) to obtain a similar bound for the two clusterings produced by the `shock` procedure.

4 Real data analysis (Lingle et al., 2016)

In this section, we compare the performance in stability of the proposed method with other methods introduced in the introduction, on a real data set BRCA (Lingle et al., 2016). The results shown here are in whole based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`. Results are similar for simulated data and other real data sets.

We compare the following strategies for network inference: 1/**GlassoBIC:** graphical lasso with regularization parameter chosen using the BIC criterion, 2/ **shock:** partition is detected using the slope heuristic dimension jump, and the regularization parameters in each graphical lasso problem are chosen using the BIC criterion, 3/**STARS:** graphical lasso with regularization parameter chosen using the STARS criterion, 4/ **BoLasso**, 5/ **StabilitySelection**.

We observe $n = 900$ individuals, and we restrict ourselves to the $p = 200$ most variables genes. We decompose the 900 individuals into 17 subsamples of size $n = 70$, for which we run every method. We compare all the graphs inferred for each method with the normalized Hamming distance between each pair of graphs. As the computation time is really different for the several methods, we also plot the CPU time to infer a network on one data set. Results are provided in Figure 1.

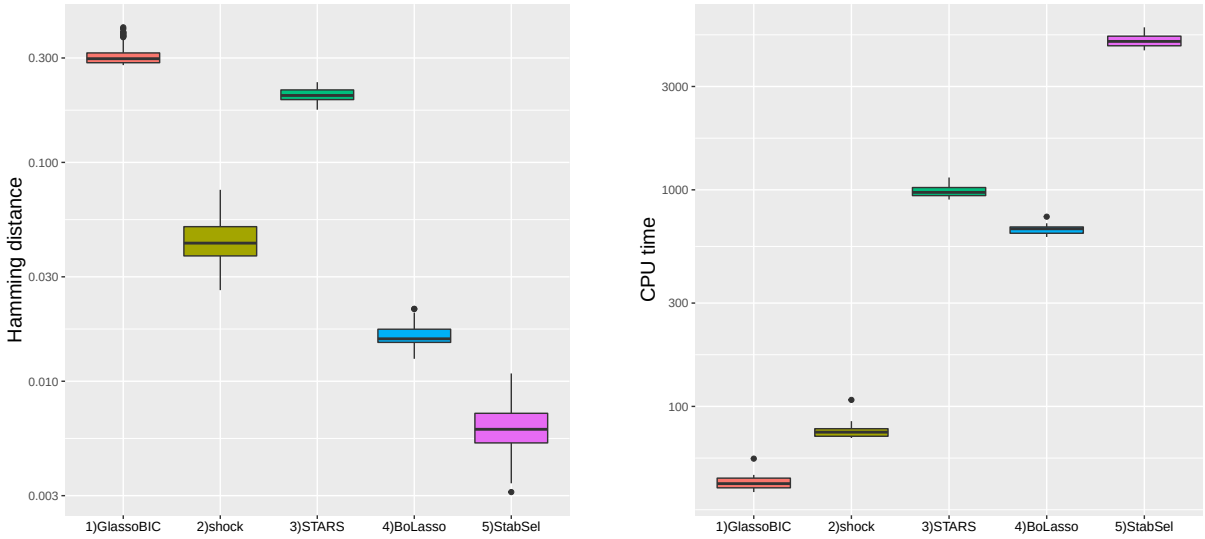

Figure 1: Results for the BRCA data set (Lingle et al., 2016). Left: normalized Hamming distance between pairs of networks inferred on subsamples. Right: CPU time for computing one network.

Bibliographie

Akbani, R. and Ng, P., Werner, H., Shahmoradgoli, M., Zhang, F. , Ju, Z. , Liu, W., Yang, J-Y , Yoshihara, K. and Li, J. et al. (2014). A pan-cancer proteomic perspective on the cancer genome atlas, *Nature Communications*, 5, pp. 3887.

Bach, F. (2008). Bolasso: model consistent lasso estimation through the bootstrap, *Machine Learning, Proceedings of the Twenty-Fifth International Conference*, pp. 33–40.

Carlsson, G. and Mémoli, F. (2010). Characterization, stability and convergence of hierarchical clustering methods, *Journal of Machine Learning Research*, 11, pp. 14251470.

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

Friedman, J., Hastie, T. and Tibshirani, T. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, 9, pp.432–441.
Liu, H. , Roeder, K.M. and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models, *Advances in Neural Information Processing Systems 23*, pages 1432–1440.
Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics*, 34, pp. 14361462.
Yu, B. (2013). Stability, *Bernoulli*, 19, pp.1484–1500.
Lingle, W., Erickson, B. J., Zuley, M. L., Jarosz, R., Bonaccio, E., Filippini, J., Gruszauskas, N. (2016). Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection.