Multi-parameters model selection for network inference

Veronica Tozzo[1] and Annalisa Barla[1]

Università degli Studi di Genova, GE, 16146, Italy
veronica.tozzo@dibris.unige.it

Abstract. Network inference is the reverse-engineering problem of infering graphs from data. With the always increasing availability of data, methods based on probability assumptions that infer multiple intertwined networks have been proposed in literature. These methods, while being extremely flexible, have the major drawback of presenting a high number of hyper-parameters that need to be tuned. The tuning of hyper-parameters, in unsupervised settings, can be performed through criteria based on likelihood or stability. Likelihood-based scores can be easily generalised to the multi hyper-parameters setting, but their computation is feasible only under certain probability assumptions. Differently, stability-based methods are of general application and, on single hyper-parameter, they have been proved to outperform likelihood-based scores. In this work we present a multi-parameters extension to stability-based methods that can be easily applied on complex models. We extensively compared this extension with likelihood-based scores on synthetic Gaussian data. Experiments show that our extension provides a better estimate of models of increasing complexity providing a valuable alternative of existing likelihood-based model selection methods.

Keywords: model selection, network inference, multi hyper-parameters

1 Introduction

Networks are present in the majority of natural phenomena [2] comprising physics [4, 28], biology [14, 43] and social sciences [6, 41]. The underlying network structure may sometimes be available, but, in general, *network inference* methods shall be used to infer it from data. Here, we focus on inference methods based on a probability assumption, *probabilistic graphical models* (PGMs), in which the connections in the network encode conditional dependencies between the nodes. These methods, typically, embed prior knowledge to guide the inference process and ease the computational burden. Usually, the models we are considering, assume sparsity of the solution, achieved by imposing an ℓ_1 penalty. This constraint reduces the original search space size of $2^{d(d-1)/2}$ (where d is the number of nodes) by forcing to zero the weaker connections [1, 13, 24, 32, 45, 46]. Other priors may be used to include more complex hypothesis such latent variables, multiple classes, multi-levels, dynamism and more [8, 11, 15, 16, 38, 48]. All priors are imposed through penalty functions, each of them regulated by a corresponding hyper-parameter. The problem of choosing the best set of hyper-parameters, also known as *model selection*, is one of the most challenging task in machine learning. Indeed, even if theoretical bounds exist for some statistical models often these do not work in practice as the estimation depends on a sample size that is usually not available ($n \ll d$). The optimal models are therefore selected by empirically evaluating the performance on data. In the context of network inference this task is particularly difficult given the unsupervised nature of the problem, which therefore relies on likelihood scores [5, 7, 9–12, 17, 35] or stability measures [22, 23, 27]. Likelihood and penalised likelihood scores (BIC [15] or AIC [11]) are typically nested in a cross-validation schema that may possibly lead to overfit [21, 40, 42]. Moreover, such scores may be conditionally applied based on the assumed probability distributions as the computation of the normalization constant of the joint distribution may be infeasible [1, 32, 46]. As an alternative, one can consider stability-based methods whose aim is to find the optimal value of the hyper-parameters that maximises stability of the inferred graph at multiple resampling of the data [22, 23]. These criteria have proved to be more effective than likelihood-based scores [22] and with some distribution assumptions are the only possible choice. Such methods were later extended to consider graphlets stability *i.e.* to verify the presence of non-isomorphic subgraphs across experimental subsampling [27, 31].

The aim of this paper is twofold, on one hand we provide a comprehensive description of the available likelihood-based scores for multi-parameters model selection; on the other, we extend stability-based methods to the multi-parameters case, also including graphlets stability in the context of sparse network infernece [22, 23, 27]. The key of our extension lies in considering the possible combinations of hyper-parameters as a unique parameter Λ defined as a tuple. Such tuple is formed in such a way that the first entry is the parameter that mostly regulates sparsity followed by the other parameters ordered, again, with respect to their impact on sparsity. We compared the general stability-based method, both with and without graphlets, with the general likelihood-based scores on three possible multi-parameters extension of the graphical lasso [13] in the context of Gaussian Graphical Models (GGMs). This is due to the fact that only with Gaussian data we can explictly compute the likelihood. Results show that our stability-based extension always overcome likelihood-based selection methods as its single hyper-parameter counterpart did in the single-network inference case. The remainder of this paper is organised as follows: in Section 2 we present the problem of network inference; in Section 3 we list our definition of generalised likelihood-based scores for model selection; Section 4 contains our main constribution as a general multi-parameters stability-based model selection method; in Section 5 we show synthetic experiments. Lastly, in Section 6 we conclude with a brief recap and we suggest possible future work in this area of research.

2 Network inference

Probability-based multiple network inference aims at estimating K graphs $G_k = (V, E_k)$ for $k = 1, \ldots, K$ where $V = \{1, \ldots, d\}$ are the nodes and $E_k \subseteq V \times V$

is the set of edges that connect such nodes in the network k. The inference of the weighted adjacency matrices of such graphs $\boldsymbol{\Theta} = (\Theta_1, \dots, \Theta_K)$ is performed from observations $\boldsymbol{X} = (X_1, \dots, X_K) \in \mathbb{R}^{n_1 \times d} \times \dots \times \mathbb{R}^{n_K \times d}$. We define a generic form for the inference problem as

$$\underset{\Theta_1,\dots,\Theta_K}{\text{minimize}} \sum_{k=1}^{K} \Big[- \ell(\Theta_k, X_k) + \alpha \|\Theta_k\|_{1,od} \Big] + \sum_{p=1}^{P} \beta_p \mathcal{P}_p(\Theta_1, \dots, \Theta_K) \qquad (1)$$

where $\|\Theta_k\|_{1,od}$ is the off-diagonal ℓ_1 norm that enforces sparsity on the off-diagonal elements of each adjacency matrix Θ_k and $\mathcal{P}_p$ is typically a sum of penalties, controlled by the hyper-parameter β_p, applied on combinations of the precision matrices. The main hyper-parameter, α, regulates the sparsity of the solution, a fundamental assumption to reduce the complexity of the problem at hand.

As previously mentioned for the rest of the paper we will use GGMs, as they allow to compute the likelihood of the model. In this case, data are assumed to be sampled from a multivariate normal distribution. Each graph k is inferred from samples $X_k \in \mathbb{R}^{n_k \times d} \sim \mathcal{N}(0, \Theta_k^{-1})$ where each *precision matrix* Θ_k is the inverse of the covariance matrix that encodes the conditional dependencies between variables, *i.e.*, is the weighted adjacency matrix of the graph G_k [13,19,24]. The related log-likelihood is defined as $\ell(\Theta_k, S_k) = \log \det(\Theta_k) + \operatorname{tr}(\Theta_k S_k)$ where $S_k = \frac{1}{n_k} X_k^\top X_k$ is the empirical covariance matrix. We instantiate the functional in Equation (1) to provide example of possibly multiple GGMs that we will later use for the analysis.

- by taking $K = 1$ and $P = 0$ Equation (1) has the same form of the standard graphical lasso problem [13];
- by taking K to be the number of classes present in the problem $P = 1$ and the related penalty $\mathcal{P}_1 = \sum_{k=1}^{K} \sum_{k' \neq k} \psi(\Theta_k - \Theta_{k'})$ we are considering the Joint Graphical Lasso problem [11,15]. Where ψ is the distance function among the precision matrices of the classes;
- by taking K as the number of time points in a time series, $P = 1$ and the related penalty $\mathcal{P}_1 = \sum_{k=1}^{K-1} \psi(\Theta_{k+1} - \Theta_k)$ we are considering the Time-Varying Graphical Lasso. Here, again, the function ψ is a distance function [16,17].

In the rest of the paper we will denote with ζ the solver for a generic network inference method that takes in input the set of matrices $\boldsymbol{X}$ and gives as output the set of adjacency matrix of the graph $\boldsymbol{\Theta}$.

3 Likelihood scores for multi-parameters model selection

Likelihood-based model selection methods rely on the possibility of computing the likelihood of the model under analysis. Therefore, as previously mentioned it is not possible to use the rest of the definition for some PGMs (*e.g.*, Ising, Poisson, Exponential [1, 45, 46]). When possible, likelihood is used as a score and inserted in a cross validation schema as K-Fold, Monte Carlo or Gaussian process-based Bayesian optimisation procedures [26, 36]. Such scores are easily extendible to the multi-parameters multi-networks case as it suffices to take the mean of the scores on the single networks. Let us consider K graphs in d variables, for which we have $\boldsymbol{X} = (X_1, \dots, X_K)$ observations each of them having n_k samples and the related empirical covariance matrices $\boldsymbol{S} = (S_1, \dots, S_K)$. We denote Λ the generic hyper-parameters tuple of the model in consideration and the inferred precision matrices inferred with the specific choice of hyper-parameters as $\boldsymbol{\Theta}_\Lambda$. Then, the generalised scores are:

- *Generalized likelihood score.* We define the generalized likelihood score as

$$\ell\ell(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}) = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \ell(\Theta_\Lambda^k, S_k) \right] \tag{2}$$

  such score was used in [17,38] to perform model selection on multi-parameter multi-network inference.
- *Generalised Bayesian Information Criterion* (BIC) [37]. It considers the *degrees of freedom* of the model in order to prevent overfitting for an increasing complexity of the model in analysis. In a graphical model selection problem the degree of freedom are the number of non-zero elements in the matrix [50]. Here, we take into account for the incremented number of degree of freedom given by the K graphs.

$$\text{BIC}(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}) = \ell\ell(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}) - \left(\sum_{k=1}^{K} \frac{\log(n_k)}{n_k} \right) \|\boldsymbol{\Theta}_\Lambda\|_{od,0} \tag{3}$$

  where $\|\boldsymbol{\Theta}_\Lambda\|_{od,0}$ is the number of non-zero elements in the off-diagonal of the matrix $\boldsymbol{\Theta}_\Lambda$. The BIC is a common score method for unsupervised problem as it leads to asymptotically consistent model selection when the number of variables d is fixed and the number of samples n_k increases. The AIC method for multi-networks [11, 33] differs from this formulation only for the penalty that, instead of being proportional to the number of samples, is simply multiplied by 2. Due to this resemblence, we do not include it in the following comparison.
- *Generalised Extended BIC* (EBIC), defined as

$$\text{EBIC}(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}, \epsilon) = \ell\ell(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}) - \sum_{k=1}^{K} \left(\frac{\log(n_k)}{n_k} + 4\epsilon \frac{\log(d)}{n_k} \right) \|\boldsymbol{\Theta}_\Lambda\|_{od,0}. \tag{4}$$

  This score proposes a trade-off between the positive selection rate and the false discovery rate based on the choice of the positive parameters ϵ, which following the literature is selected as $\epsilon = 0.5$ [5,7,9,12,35]. A further extension that furtherly penalises the degree of freedom (suitable when analysing large graphs) is defined as [10]

$$\text{EBIC}_m(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}, \epsilon) = \ell\ell(\boldsymbol{\Theta}_\Lambda, \boldsymbol{S}) - \sum_{k=1}^{K} \left(\frac{\log(n_k)}{n_k} + 4\epsilon \frac{\log(Kd(d-1)/2)}{n_k} \right) \|\boldsymbol{\Theta}_\Lambda\|_{od,0} \tag{5}$$

  where $Kd(d-1)/2$ is the total number of off-diagonal elements in the K precision matrices.

Algorithm 1 Bernoulli indicator variance (BV)

```
for $\mathit{\Lambda} \in \{\mathit{\Lambda}_1, \mathit{\Lambda}_2, \dots\}$ do
    for $m = 1, \dots, M$ do
        $\boldsymbol{X}^m = (X_1^{z1}, \dots, X_K^{zK})$
        $\boldsymbol{\Theta}_\Lambda^m = \zeta(\boldsymbol{X}^m)$
        $(\bar{\boldsymbol{\Theta}}_\Lambda^m)_{kij} = \begin{cases} 1 \text{ if } (\boldsymbol{\Theta}_\Lambda^m)_{kij} \neq 0 \\ 0 \text{ otherwise} \end{cases}$ for $k = 1, \dots, K$ and $i, j = 1, \dots, d$
    end for
    $\hat{\boldsymbol{\Theta}}_\Lambda^{\boldsymbol{z}} = \frac{1}{M} \sum_{m=1}^M \bar{\boldsymbol{\Theta}}_\Lambda^m$
    $\xi_\Lambda^{\boldsymbol{z}} = 2\hat{\boldsymbol{\Theta}}_\Lambda^{\boldsymbol{z}}(1 - \hat{\boldsymbol{\Theta}}_\Lambda^{\boldsymbol{z}})$
end for
return $[\xi_{\Lambda_1}^{\boldsymbol{z}}, \xi_{\Lambda_2}^{\boldsymbol{z}}, \dots]$
```

4 Stability based multi-parameters model selection

Model selection approaches based on stability of the result are widely used in unsupervised settings as clustering [18, 39]. In the context of graphical models they were proposed in [22, 23], the most used is called *Stability Approach to Regularisation Selection* (StARS), in which the best model is selected as the one that uses the minimum amount of regularisation still producing a sparse and stable graph under random sub-sampling of the initial dataset. StARS selects the best value for the hyper-parameter α by analysing the trend of stability as α varies. Indeed, as $\alpha \to \infty$ the inferred graph is completely sparse, *i.e.*, no edges are present. Therefore for $\alpha \to \infty$ the graph is stable under random sub-sampling of the data. On the other hand, the same holds when $\alpha = 0$ as the graph is complete and therefore there is no variation in the inferred edges. StARS selects the best α based on the possibility to order the regularisation parameters from the strongest regularisation to the weakest. The goal is to choose α^* such that the true graph E is contained in the inferred $E(\alpha^*)$, i.e. the graph is over-selected [22].

Multi-parameters relation order. In the context of multi-parameters the ordering is trickier as different parameters act on different part of the inference which may or may not impact sparsity and stability. Here, we define a single parameter $\mathit{\Lambda} = (\alpha, \beta_I, \beta_{II}, \dots, \beta_{P-th})$ as a tuple of hyper-parameters. Such hyper-parameters are not randomly positioned within the tuple but according to their impact on the sparsity of the problem. Therefore, α that directly acts on the ℓ_1 penalty is the most important hyper-parameter, followed by a certain order of the remaining hyper-parameters such that β_I acts on edges sparsity more than β_{II} and so on and so forth. Note that in this case β_I is not necessarily associated to P_1 but to the the penalty that has the most impact on sparsity.

When performing model selection, given the possible ranges for all the hyper-parameters and their order, we compute a grid of values naming each point of the grid as $\mathit{\Lambda}_i = (\alpha^i, \beta_I^i, \dots, \beta_{P-th}^i)$. We sort the tuples $\mathit{\Lambda}_i$ following the inverse lexicographic order so that α^i is the parameter that changes less frequently. In this way to the first tuple $\mathit{\Lambda}_1$ corresponds the most regularised (and therefore sparse) graph. With this choice we give more emphasis to the hyper-parameter α which governs the sparsity of the solution. Note that, the inverse lexicographic

Algorithm 2 m-StARS stability computations

```
$[\xi^{\boldsymbol{z}}_{\Lambda_1}, \xi^{\boldsymbol{z}}_{\Lambda_2}, \dots] = \mathrm{BV}(\boldsymbol{X}, \zeta, (\Lambda_1, \Lambda_2, \dots))$
for $\Lambda \in \{\Lambda_1, \Lambda_2, \dots\}$ do
    $D^{\boldsymbol{z}}_{\Lambda} = \sum_{k=1}^{K} \sum_{i<j} (\hat{\xi}^{\boldsymbol{z}}_{\Lambda})_{kij} \Big/ K \binom{d}{2}$
    $D^{\boldsymbol{z}}_{\Lambda} = \sup\limits_{\lambda \leq \Lambda} \{D_\lambda\}$
end for
return $\Lambda^* = \arg\min\limits_{\Lambda}[D^{\boldsymbol{z}}_{\Lambda} \leq 0.05]$.
```

order is arbitrary and there may be other types of ordering applied on the βs that may lead to the same stability solutions. We remark that, similarly to the choice of the search interval, this sorting criterion is guided by domain knowledge on the model in use.

Re-sampling In order to perform re-sampling it is necessary to define how many samples to randomly drawn from each dataset and how many time we should perform this procedure. We define $\boldsymbol{z} = (z_k)_{k=1}^{K}$ as a tuple containing the amount of sub-samples drawn at random without replacement from each dataset X_k. Each value $z_k \in [1, n_k]$ is taken proportionally to the original number of samples in such a way that if $n_{k'} \geq n_k$ then $z_{k'} \geq z_k$. The suggested choice for z_k is $z_k = \min(10\sqrt{n_k}, 0.9n_k)$ which allows to select a reasonable amount of sub-sample from the original dataset even when the original sample size is low [22]. Given z_k there are $M_k = \begin{pmatrix} n_k \\ z_k \end{pmatrix}$ sets of possible sub-samples without repetitions. Ideally, one would sub-sample all the possible sub-sets $M = \min(M_k)_{k=1}^{K}$, but, for computational reasons, this is often unfeasible. In the experiments we will present, the sample size is in the order of hundreds therefore we opt to sub-sample for $M = 100$ times with the guarantee of reaching the same stability results [29].

Single edge stability Given the solver ζ and the set of sorted hyper-parameters Λ_i, we perform the procedure described in Algorithm 2 that we call *multiple-StARS* (m-StARS) following the original naming. The instabilities, $D^{\boldsymbol{z}}_{\Lambda_i}$, are computed by m-StARS for each Λ_i as the mean of the how much the edges vary across the M sub-samples. The variance is computed as the Bernoulli indicator variance (see Algorithm 1), and its mean across edges gives us a global indicator of the instability of the graph. If the inference is stable the edge variance is 0, whereas when the inference is random, the variance is $\frac{1}{2}$ which is the maximum possible value. Given the instabilities for all Λs, we force the curve as Λ_i varies to be monotone and then select the best Λ^* as the one that provides the sparser graph under the accepted stability threshold of instability (0.05) [22]. Note that Λ^* depends on the block sizes $\boldsymbol{z}$ and therefore this method has some efficiency loss in low dimension [22].

Graphlet stability The m-StARS procedure is based on single edge stability which, by definition, ignore higher-order structures. It is possible to extend the

Algorithm 3 mg-StARS stability computations

```
$[\xi^{\boldsymbol{z}}_{\Lambda_1}, \xi^{\boldsymbol{z}}_{\Lambda_2}, \ldots] = \mathrm{BV}(\boldsymbol{X}, \zeta, (\Lambda_1, \Lambda_2, \ldots))$
$\Lambda^{ub} = \arg\min_{\Lambda} \left\{ \min \left[ \frac{\sum_{k=1}^{K} \sum_{i<j} 4(\xi^{\boldsymbol{z}}_{\Lambda})_{kij}(1-\left(\xi^{\boldsymbol{z}}_{\Lambda})_{kij}\right)}{K\binom{D}{2}} \right] \leq 0.05 \right\}$
$\Lambda^{lb} = \arg\min_{\Lambda} \left\{ \min \left[ \frac{\sum_{k=1}^{K} \sum_{i<j} (\xi^{\boldsymbol{z}}_{\Lambda})_{kij}}{K\binom{D}{2}} \right] \leq 0.05 \right\}$
for $\Lambda \in [\Lambda_{lb}, \Lambda_{ub}]$ do
    $(\rho^m_\Lambda)_k = GCV((\Theta^m_\Lambda)_k)$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$
    $\hat{d}^{\boldsymbol{z}}_{\Lambda} = \frac{2}{KM(M-1)} \sum_{k=1}^{K} \sum_{m>m'} \|(\rho^m_\Lambda)_k - (\rho^{m'}_\Lambda)_k\|_2$
end for
return $\Lambda^* = \operatorname{argmin}_{\Lambda} \hat{d}^{\boldsymbol{z}}_{\Lambda}$
```

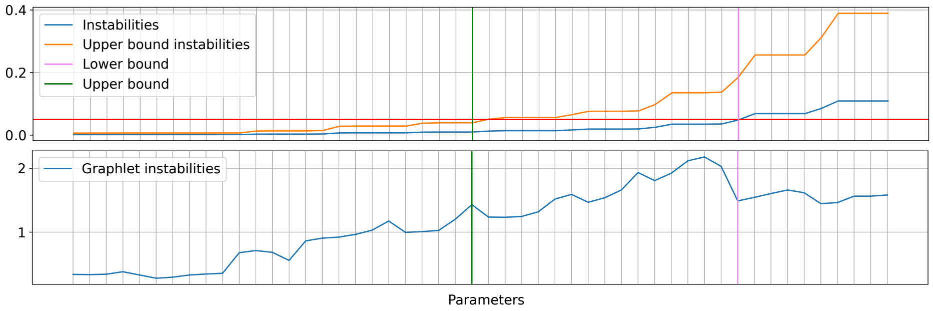

Fig. 1. An example of instabilities obtained applying our method for for a specific range of Λ values on the Joint Graphical Lasso. The top panel depicts the upper and lower bound and the interval of interested where to choose the most stable network. The bottom panel shows the graphlets instabilities curve.

concept to include *graphlets* [27] which are small (typically 4 or 5 nodes) connected non-isomorphic sub-graphs of a network widely used to characterize or compare networks [25, 30, 31]. We exploit the concept of Graphlet Correlation Vector (GCV) [34] that is used as a method to compute distances between networks.

Graphlets instability, similarly to single edge instability, is computed as the mean of the distances of the GCVs on the K graphs across all M repetitions. The main drawback of this instability, though, is that it is highly variable and it cannot easily be monotonised. Therefore, we select an interval of interest by observing the behaviour of the single edge instabilities and then we select, in this interval, the network with the lowest graphlets instability [27]. The procedure for the selection of the hyper-parameters, that we call *multi graphlets StARS* (mg-StARS), is described in Algorithm 3.

5 Experiments and results

We designed four experiments to assess the efficacy of the proposed model selection methods testing m-StARS and mg-StARS against likelihood-based scores. For the model selection with likelihood-based scores we used a 3-fold cross-validation schema training the model on a subset of data and testing it on the remaining part. We tested all model selection strategies on three GGMs model with multiple hyper-parameters, in particular the Joint Graphical Lasso (JGL) [11], the Time-varying Graphical Lasso (TGL) [17] and the Latent Graphical Lasso (LGL) [8]. For JGL we generated a random graph of 20 nodes which is the common set of edges of the three graph classes ($K = 3$) that we generated by randomly adding some edges. For TGL we devised two experiments in which we generated 10 time-evolving networks of 100 variables ($K = 10$) with two different evolution schema: smooth changes (TGL-ℓ_2) and punctual changes (TGL-ℓ_1). Both JGL and TGL have two hyper-parameters α that regulates sparsity and β that regulates the similarity of the network across classes/times, we sorted them as $\Lambda = (\alpha, \beta)$. For LGL we generated a perturbed observed network on 100 observed variables with 5 latent ($K = 1$) following the generation schema presented in [47]. LGL has two hyper-parameters α that regulates sparsity and τ that controls the amount of estimated latent variables, we ordered them as $\Lambda = (\alpha, \tau)$. For all the experiments and all the classes/times we generated $n = 100$ samples. We adapted the range of parameters to the specific case and we used 4-nodes graphlets in the mg-StARS computation. For all experiments we computed Precision-Recall (PR) and ROC curves by considering the edges of the graphs as binary classes and computing the thresholds by looking at the weights of the edges. All the code and the experiments are available for reproducibility in a general Python library for graphical models inference[1].

In Figure 1 an example of instabilities obtained applying our method for the experiment on JGL. It is noticeable that the instabilities assume a sort of step ascendance, which assesses the validity of ordering the hyper-parameters according to their impact to sparsity. We can observe that in this case the model selected with m-StARS or with mg-StARS is different. In the other experiments (not reported for space constraints) both algorithms selected the same model. The performance of m-StARS, mg-StARS, and likelihood-based scores is reported in the of Figure 2. Looking at the curves we observe that the model selected for mg-StARS performs worse than likelihood-based scores, while, if we simply use m-StARS we obtain better results.

In the remaining Figure 3, 4, and 5 we show the curves obtained for LGL, TGL-ℓ_1 and TGL-ℓ_2, respectively. In all cases m-StARS (which, in this case, is equivalent to mg-StARS) performs better than likelihood-based strategies. None of the likelihood-based scores outstands with respect to the others across experiments.

[1] https://github.com/veronicatozzo/regain/

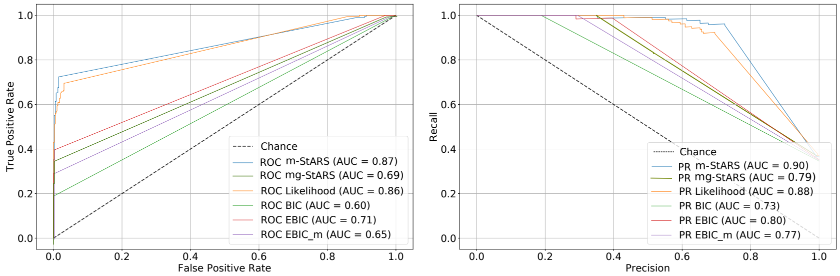

Fig. 2. ROC and Precision-Recall curves of different model selection methods for the Joint Graphical Lasso.

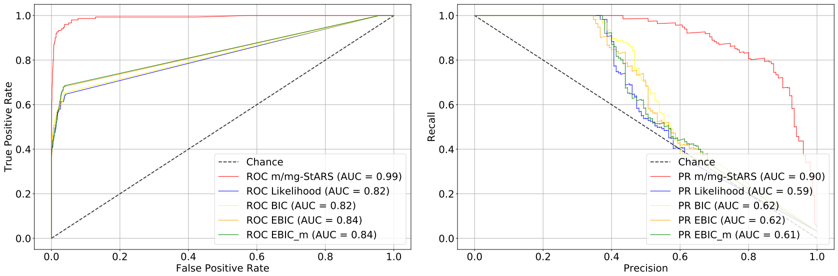

Fig. 3. ROC and Precision-Recall curves of different model selection methods for the Latent Graphical Lasso.

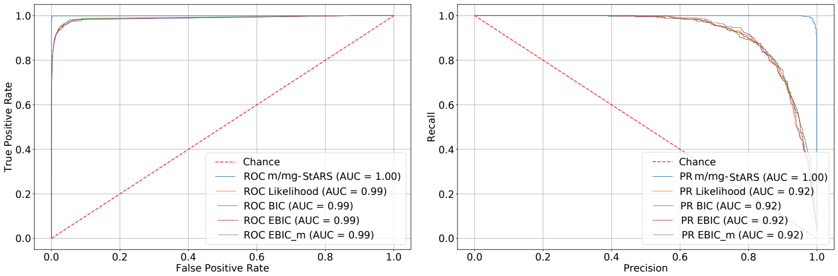

Fig. 4. ROC and Precision-Recall curves of different model selection methods for the Time-varying Graphical Lasso with ℓ_1 temporal evolution.

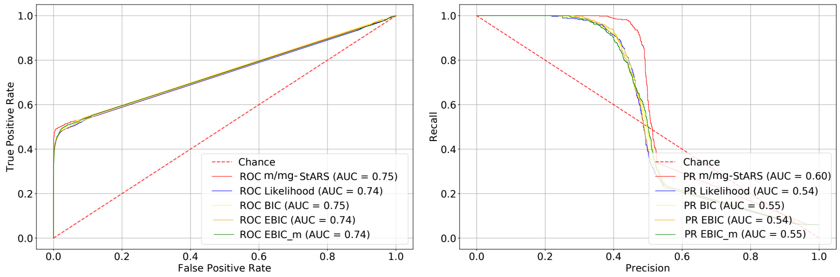

Fig. 5. ROC and Precision-Recall curves of different model selection methods for the Time-varying Graphical Lasso with ℓ_2 temporal evolution.

6 Conclusion

We presented an extension for model selection based on stability of the result for network inference methods that present more than one hyper-parameter. We showed the validity of the proposed method on Gaussian Graphical Models comparing m-StARS and mg-StARS with likelihood-based cross validation schema noticing that m-StARS always provides a better estimate of the model. We remark that, in cases of non-Gaussian data, stability-based model selection criteria are the only possible choice. Therefore, a suitable method for multi hyper-parameters selection is necessary for further exploring more complex models on other distributions [20, 44, 49]. From the methodological perspective, graphlets stability has proved to be less effective than single edge stability. For future work it would be interesting to exploit other types of stability possibly looking at topological descriptors capturing higher order relations such as persistent homology [3] and to explore other type of sorting for the hyper-paramters other than inverse lexicographic order.

References

1. Genevera I Allen and Zhandong Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE transactions on nanobioscience*, 12(3):189–198, 2013.
2. Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101, 2004.
3. Mattia G Bergomi, Massimo Ferri, Pietro Vertechi, and Lorenzo Zuffi. Beyond topological persistence: Starting from networks. *arXiv preprint arXiv:1901.08051*, 2019.
4. Martin J Blunt, Matthew D Jackson, Mohammad Piri, and Per H Valvatne. Detailed physics, predictive capabilities and macroscopic consequences for pore-network models of multiphase flow. *Advances in Water Resources*, 25(8-12):1069–1089, 2002.
5. Małgorzata Bogdan, Jayanta K Ghosh, and RW Doerge. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167(2):989–999, 2004.
6. Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.
7. Karl W Broman and Terence P Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002.
8. Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.
9. Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.
10. Lulu Cheng, Liang Shan, and Inyoung Kim. Multilevel gaussian graphical model for multilevel networks. *Journal of Statistical Planning and Inference*, 190:1–14, 2017.

11. Patrick Danaher, Pei Wang, and Daniela M Witten. 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, 2014.
12. Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.
13. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
14. Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.
15. Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.
16. David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396. ACM, 2015.
17. David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 205–213. ACM, 2017.
18. Tilman Lange, Volker Roth, Mikio L Braun, and Joachim M Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.
19. Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.
20. Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.
21. Hongzhe Li and Jiang Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2005.
22. Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.
23. Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
24. Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.
25. Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:CIN–S680, 2008.
26. Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.
27. Christian L Müller, Richard Bonneau, and Zachary Kurtz. Generalized stability approach for regularized graphical models. *arXiv preprint arXiv:1605.07072*, 2016.
28. Alessandro Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 38(33):R309, 2005.
29. Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.
30. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

31. Natasa Pržulj, Derek G Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.
32. Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.
33. Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81, 1986.
34. Anida Sarajlić, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. Graphlet-based characterization of directed networks. *Scientific reports*, 6:35098, 2016.
35. David Siegmund. Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika*, 91(4):785–800, 2004.
36. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.
37. Petre Stoica and Yngve Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.
38. Federico Tomasi, Veronica Tozzo, Saverio Salzo, and Alessandro Verri. Latent variable time-varying network inference. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2338–2346. ACM, 2018.
39. Ulrike Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.
40. Ivan Vujačić, Antonino Abbruzzo, and Ernst Wit. A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of statistical computation and simulation*, 85(18):3628–3640, 2015.
41. Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 322–331. IEEE, 2007.
42. Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.
43. Darren J Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116, 2007.
44. Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2014.
45. Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.
46. Eunho Yang, Pradeep K Ravikumar, Genevera I Allen, and Zhandong Liu. On poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.
47. Ming Yuan. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1968–1972, 2012.
48. Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.
49. Marinka Žitnik and Blaž Zupan. Gene network inference by fusing data from diverse distributions. *Bioinformatics*, 31(12):i230–i239, 2015.
50. Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.*, 35(5):2173–2192, 10 2007.