Estimation robuste sur les graphes aléatoires : les divergences comme alternative à la vraisemblance

EDMSA 2024

Cyprien Ferraris
Michel Broniatowski, Frederic Guilloux, Annick Valibouze, Mohamed Achibi

16 Mai 2024

Sorbonne Université (LPSM) – Safran Aircraft Engines (SAE)

Objective

Graph clustering has many applications: finance, genomics, social networks, industry ...

Objective

Graph clustering has many applications: finance, genomics, social networks, industry ...

A parametric model is often considered. However, in practice, data may have outliers!

Objective

Graph clustering has many applications: finance, genomics, social networks, industry ...

A parametric model is often considered. However, in practice, data may have outliers!

How to deal with it?

Outline

Weighted Graph

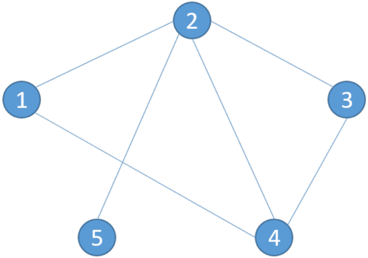

Figure: Example of a graph.

Weighted Graph

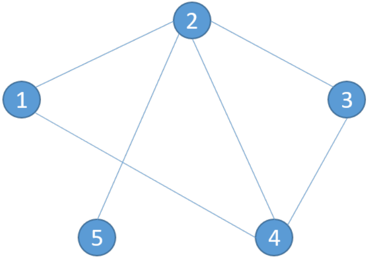

Figure: Example of a graph.

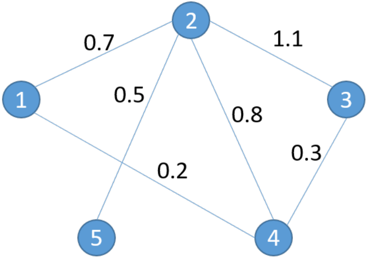

Figure: Example of a weighted graph

Weighted Graph

(Weighted) Adjacency Matrix :

$$\boldsymbol{W} = \begin{pmatrix} & 0.7 & & 0.2 & \\ 0.7 & & 1.1 & 0.8 & 0.5 \\ & 1.1 & & 0.3 & \\ 0.2 & 0.8 & 0.3 & & \\ & 0.5 & & & \end{pmatrix}$$

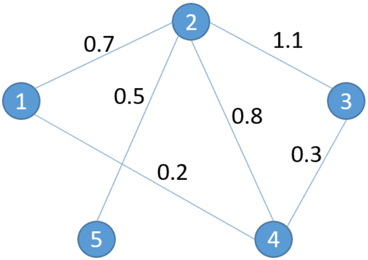

Figure: Example of a weighted graph

Which Model for the Data ?

A graph with blocks!

Which Model for the Data ?

A graph with blocks!

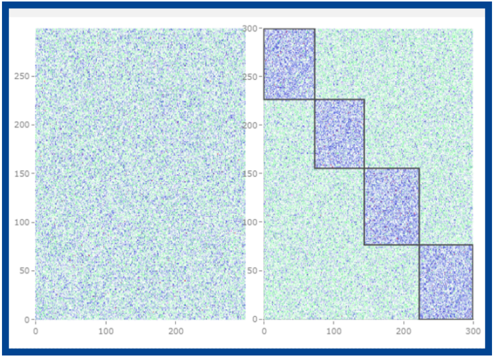

Figure: Adjacency matrix of some graph with blocks, on the left the initial graph and on the right the graph permuted according to the blocks

Stochastic Block Model (SBM)

Mathematically, for a graph with n nodes, an SBM with K blocks can be simulated from the following steps [1]

1. for each node i a latent partition c_i^* is drawn with probability $(\pi_k)_{1 \leq k \leq K}$.
2. the weighted adjacency matrix is drawn such that $W_{ij} \mid c_i^*, c_j^* \overset{i.i.d.}{\sim} P_{\theta^*_{c_i^*, c_j^*}}$, $i < j$, $W_{ij} = W_{ji}$.

Estimation of the SBM

Which estimation methods?

Estimation of the SBM

Which estimation methods? **The Likelihood**

Estimation of the SBM

Which estimation methods? **The Likelihood**

$$\sum_{\mathbf{c}\in\{1,\ldots,K\}^n} \prod_{i=1}^{n} \pi_{c_i} \prod_{ij} P_{\theta_{c_i,c_j}}(W_{ij})$$

Estimation of the SBM

Which estimation methods? **The Likelihood**

$$\sum_{\mathbf{c}\in\{1,\dots,K\}^n} \prod_{i=1}^{n} \pi_{c_i} \prod_{ij} P_{\theta_{c_i,c_j}}(W_{ij})$$

Complicated ...

Estimation of the SBM

Which estimation methods? **The Likelihood**

$$\sum_{\mathbf{c}\in\{1,\dots,K\}^n} \prod_{i=1}^{n} \pi_{c_i} \prod_{ij} P_{\theta_{c_i,c_j}}(W_{ij})$$

Complicated ...

A possibility : variational inference, which maximizes a lower bound of the likelihood which is easier to compute.

Partial Likelihood

Another alternative to the likelihood :

Partial Likelihood

Another alternative to the likelihood : **partial likelihood**.
The idea is to consider the latent partition as a parameter.

$$\mathrm{pl}(\boldsymbol{\theta}, \mathbf{c}; \boldsymbol{W}) = \prod_{ij} P_{\theta_{c_i, c_j}}(W_{ij})$$

Likelihood Lack of Robustness

However, the likelihood is known to be non-robust to some misspecifications such as outliers.

Likelihood Lack of Robustness

However, the likelihood is known to be non-robust to some misspecifications such as outliers.

In such cases, other methods may be more adapted depending on the kind of misspecifation, example **Hellinger distance**

$$H(P, Q) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ} \right)^2$$

Likelihood vs Hellinger under Misspecifications

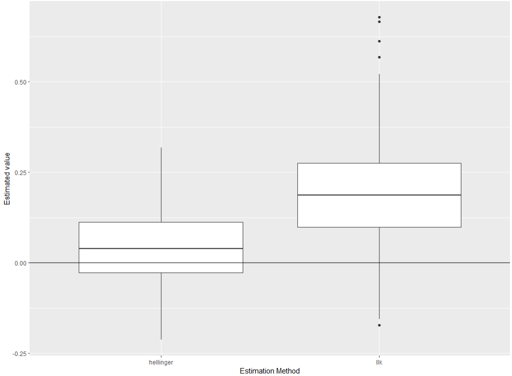

Figure: Comparison between Hellinger and likelihood estimation for N(0,1) with 5% outliers N(4,1) : 300 replications, with 100 points

General Family of Estimation Methods : φ-Divergences

To gain robustness properties we try to use what is called φ-divergence :

$$D_\varphi(P, Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ = E_Q\left[\varphi\left(\frac{dP}{dQ}\right)\right]$$

Where φ is a convex function such that $\varphi(1) = 0 \Rightarrow$

General Family of Estimation Methods : φ-Divergences

To gain robustness properties we try to use what is called φ-divergence :

$$D_\varphi(P,Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ = E_Q\left[\varphi\left(\frac{dP}{dQ}\right)\right]$$

Where φ is a convex function such that $\varphi(1) = 0 \Rightarrow$

$$D_\varphi(P,Q) = 0 \iff P = Q$$

General Family of Estimation Methods : φ-Divergences

To gain robustness properties we try to use what is called φ-divergence :

$$D_\varphi(P, Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ = E_Q\left[\varphi\left(\frac{dP}{dQ}\right)\right]$$

Where φ is a convex function such that $\varphi(1) = 0 \Rightarrow$

$$D_\varphi(P, Q) = 0 \iff P = Q$$

φ-divergence notably includes Kullback-Leibler, Hellinger, χ^2 and L_1 estimation.

Cressie-Read Divergence

A familly of φ-divergences, called Cressie-Read divergences or γ-power divergences is given by

$$\varphi(x) = \varphi_\gamma(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{1}$$

$\gamma \notin \{0, 1\}$.

Cressie-Read Divergence

A familly of φ-divergences, called Cressie-Read divergences or γ-power divergences is given by

$$\varphi(x) = \varphi_\gamma(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{1}$$

$\gamma \notin \{0, 1\}$. When $\gamma = 1/2$ the estimate is the same as the Hellinger distance.

Principle of our Approach I

Idea

Replace the likelihood by divergence

Principle of our Approach I

Idea

Replace the likelihood by divergence

To do so we reorganize the terms

Principle of our Approach I

Idea

Replace the likelihood by divergence

To do so we reorganize the terms

$$\mathrm{pll}(\boldsymbol{\theta}, \mathbf{c}; \boldsymbol{W}) = \sum_{i<j} \log\left(P_{\theta_{c_i,c_j}(W_{ij})}\right) = \sum_{1\leq k,l\leq K} \sum_{i:c_i=k,\, j:c_j=k} \log P_{\theta_{k,l}}(W_{ij})$$

$$= \sum_{k,l} n_k n_l \left(\frac{1}{n_k n_l} \sum_{i:c_i=k,\, j:c_j=k} \log P_{\theta_{k,l}}(W_{ij})\right)$$

$n_k = \sum_i I(c_i = k)$, the number of nodes in block k.

Principle of our Approach II

From $\text{pll}(\boldsymbol{\theta}, \mathbf{c}) = \sum_{k,l} n_k n_l \left(\frac{1}{n_k n_l} \sum_{i:c_i=k,\, j:c_j=k} P_{\theta_{k,l}}(W_{ij}) \right)$, we recognize the likelihood of each blocks.

Principle of our Approach II

From $\text{pll}(\boldsymbol{\theta}, \mathbf{c}) = \sum_{k,l} n_k n_l \left(\frac{1}{n_k n_l} \sum_{i:c_i=k,\, j:c_j=k} P_{\theta_{k,l}}(W_{ij}) \right)$, we recognize the likelihood of each blocks.

We replace it by a divergence

New Block Divergence Criterion [2]

$$\text{BDC}(\boldsymbol{\theta}, \mathbf{c}; \boldsymbol{W}) = \text{BDC}(\boldsymbol{\theta}, \mathbf{c}; (\hat{F}^{(k,l)})_{k,l}) = \sum_{k,l} n_k n_l D_\varphi(\hat{F}^{(k,l)}, P_{\theta_{kl}})$$

$\hat{F}^{(k,l)}$ an estimator of the distribution in the block k, l.

Why Should it Work?

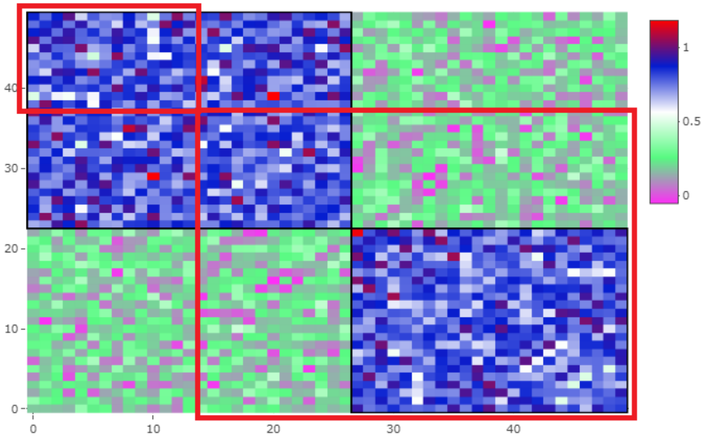

Figure: Illustration of the criterion principle

Some Remarks

We are not restricted to φ-Divergence,
Other divergences can be used or Cramèr-von Mises.

Some Remarks

We are not restricted to φ-Divergence,
Other divergences can be used or Cramèr-von Mises.
As estimator, we can use the kernel density estimator

$$\hat{f}^{(k,l)}(x) = \frac{1}{n_k n_l} \sum_{c_i=k, c_j=l} K_h(x - W_{ij})$$

where h is some band-with size.

Estimation of the Model

In SBM estimation, there are three things to estimate

- the number of blocks K,
- the parameters of the model $\boldsymbol{\theta}$,
- the latent partition $\mathbf{c}$

Estimation of the Model

In SBM estimation, there are three things to estimate

- the number of blocks K,
- the parameters of the model $\boldsymbol{\theta}$,
- the latent partition $\mathbf{c}$

As the model may be misspecified, so $K, \boldsymbol{\theta}, \mathbf{c}$ may not be defined. We consider then

$$\boldsymbol{\theta}^*, \mathbf{c}^* \in \text{argmin}_{\boldsymbol{\theta}, \mathbf{c}} \, \text{BDC}(\boldsymbol{\theta}, \mathbf{c}, (F^{(k,l)})_{k,l})$$

$F^{(k,l)}$ the true distribution in the block k, l.

Parameters Estimation

We assume K fixed, we consider

$$\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{c}} \in \operatorname{argmin}_{\boldsymbol{\theta},\mathbf{c}} \mathrm{BDC}(\boldsymbol{\theta}, \mathbf{c}, (\hat{F}^{(k,l)})_{k,l})$$

Parameters Estimation

We assume K fixed, we consider

$$\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{c}} \in \operatorname{argmin}_{\boldsymbol{\theta},\mathbf{c}} \operatorname{BDC}(\boldsymbol{\theta}, \mathbf{c}, (\hat{F}^{(k,l)})_{k,l})$$

To ensure the partition recovery, we consider

$$\|\mathbf{e} - \mathbf{c}\|_{P_n} = \min_{\sigma} \sum_{i=1}^{n} |\sigma(e_i) - c_i|$$

σ a permutation of $\{1, \ldots, K\}$.

Consistency of Estimates

Under the hypothesis

- Mixtures of edges distributions does not correspond to a member of P_θ.
- φ is γ-Holder for $\gamma \in (0, 1]$,
- $\frac{n}{Klog(K)} \to \infty$.

Consistency of Estimates

Under the hypothesis

- Mixtures of edges distributions does not correspond to a member of P_θ.
- φ is γ-Holder for $\gamma \in (0, 1]$,
- $\frac{n}{Klog(K)} \to \infty$.

Result

We have for any ϵ,

$$\mathbb{P}\left(\frac{1}{n}\|\tilde{\mathbf{c}} - \mathbf{c}^*\|_{P_n} > \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0$$

Optimization of the Criterion

Optimization on the space of partition $\rightarrow$ NP-hard problem.

Optimization of the Criterion

Optimization on the space of partition → NP-hard problem. Thus, we consider an alternative problem with a relaxed version of a partition,

$$\boldsymbol{c} \in [0,1]^{n\times K} \; : \; \sum_{k=1}^{K} c_{ik} = 1 \quad .$$

Optimization of the Criterion

Optimization on the space of partition $\rightarrow$ NP-hard problem. Thus, we consider an alternative problem with a relaxed version of a partition,

$$\boldsymbol{c} \in [0,1]^{n \times K} \ : \ \sum_{k=1}^{K} c_{ik} = 1 \quad .$$

c_{ik} can be interpreted as the probability that the node i belongs to the class k.

Optimization of the Criterion

Optimization on the space of partition $\rightarrow$ NP-hard problem. Thus, we consider an alternative problem with a relaxed version of a partition,

$$\mathbf{c} \in [0,1]^{n \times K} \; : \; \sum_{k=1}^{K} c_{ik} = 1 \quad .$$

c_{ik} can be interpreted as the probability that the node i belongs to the class k.
We then use $n_k = \sum_{i=1}^{n} c_{ik}$

Optimization of the Criterion

Optimization on the space of partition $\rightarrow$ NP-hard problem. Thus, we consider an alternative problem with a relaxed version of a partition,

$$\boldsymbol{c} \in [0,1]^{n\times K} \ : \ \sum_{k=1}^{K} c_{ik} = 1 \quad .$$

c_{ik} can be interpreted as the probability that the node i belongs to the class k.
We then use $n_k = \sum_{i=1}^{n} c_{ik}$

$\hat{f}^{(k,l)}(x) = \sum_{ij} \frac{c_{ik}}{n_k} \frac{c_{jl}}{n_l} K_h(x - W_{ij})$.

Estimation using Projected Gradient Descent

Algorithm 1 Optimization of criterion BDC

Require: A weighted matrix $\boldsymbol{W}$ of size $n \times n$, a number of blocks K, gradient step η, initial partition $\mathbf{c}^{(0)}$ and number of iterations T.

Ensure: **Local minimum** $\boldsymbol{\theta}^{(T)}$ and $\mathbf{c}^{(T)}$ of BDC.

for $t = 1$ to T **do**

Compute $\boldsymbol{\theta}^{(t)} = \text{argmin}_{\boldsymbol{\theta}} \text{BDC}(\boldsymbol{\theta}, c^{(t-1)}, \boldsymbol{W})$, for $1 \leq k, l \leq K$.

Compute $\nabla^{(t)} = \nabla_{\mathbf{c}} \text{BDC}(\boldsymbol{\theta}^{(t)}, \mathbf{c}^{(t-1)}, \boldsymbol{W})$

$\mathbf{c}_{i.}^{(t)} = \text{Proj}(\mathbf{c}_{i.}^{(t-1)} - \eta \nabla_{i.}^{(t)})$, for $1 \leq i \leq n$.

Where Proj is the l_2 projection on the probability simplex.

end for

Model Selection with Divergence

In the Divergence literature: Divergence Information Criterion (DIC) [3].
The idea is to look at an asymptotic unbiased version of the criterion, i.e.

$$E_{\boldsymbol{W}}[\mathrm{BDC}(\tilde{\theta}(\boldsymbol{W}), \tilde{\boldsymbol{C}}(\boldsymbol{W}), \boldsymbol{W}) + \mathrm{Penalty}(K)] \to 0 \tag{2}$$

Model Selection : Choosing K

Problem here : often high biased estimators of the divergence.

Model Selection : Choosing K

Problem here : often high biased estimators of the divergence.

Model Selection under an SBM

With some regularity assumption, for power divergence with $\gamma \in (0, 1]$, the penalty term

$$log(n) \sum_{k,l} \frac{n_k n_l}{n^2} (n_k n_l)^{-2\gamma/3}$$

can be used as asymptotic model selection criterion.

Edges Missepecification

Our motivation to use divergences is their robustness properties!
Under an SBM, a classical example is

$$W_{ij} \mid c_i^*, c_j^* \sim (1-\epsilon) P_{\theta_{c_i^*, c_j^*}} + \epsilon P_{ij} \tag{M}$$

where

$\epsilon \in [0, 1]$ represent the proportion of misspecifications,

the $(P_{ij})_{ij}$ are the misspecified probability distributions.

KL-Divergence and Robustness

Possible to analyze influence function

KL-Divergence and Robustness

Possible to analyze influence function

[4] demonstrate that for two distributions P and Q on the same sample space Ω, for any event $\mathrm{E} \subset \Omega$,

$$Q(\mathrm{E}) \leq \max\left\{\frac{2KL(Q,P)}{1 - log(P(E))}, e\sqrt{2P(\mathrm{E})}\right\} \quad .$$

KL-Divergence and Robustness

Possible to analyze influence function

[4] demonstrate that for two distributions P and Q on the same sample space Ω, for any event $\mathrm{E} \subset \Omega$,

$$Q(\mathrm{E}) \leq \max\left\{\frac{2KL(Q,P)}{1-log(P(E))}, e\sqrt{2P(\mathrm{E})}\right\} \quad .$$

Applying the BDC to SBM is asymptotically consistent if $o(\frac{n}{log(n)})$ edges are misspecified.

Example for K=2

2 models with both $K = 2$, $\pi_1 = \pi_2 = 0.5$.

Example for K=2

2 models with both $K = 2$, $\pi_1 = \pi_2 = 0.5$.

$$W_{ij} \mid \mathbf{c} \sim \text{Binomial}(10, 0.2\delta_{c_i,c_j} + 0.6) \qquad \text{(M1)}$$

Example for K=2

2 models with both $K = 2$, $\pi_1 = \pi_2 = 0.5$.

$$W_{ij} \mid \mathbf{c} \sim \text{Binomial}(10, 0.2\delta_{c_i,c_j} + 0.6) \tag{M1}$$

and

$$W_{ij} \mid \mathbf{c} \sim 0.5\text{Binomial}(10, 0.2\delta_{c_i,c_j} + 0.6) + 0.5\delta_{\{0\}}. \tag{M2}$$

Hellinger vs Likelihood under the Model (M1)

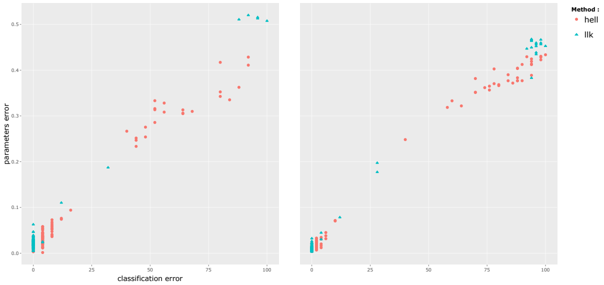

Figure: Comparison of estimation error of estimates based on Hellinger and the partial likelihood under model M1 for $n = 50$ (left) and $n = 100$ (right)

Hellinger vs Likelihood under the Model (M2)

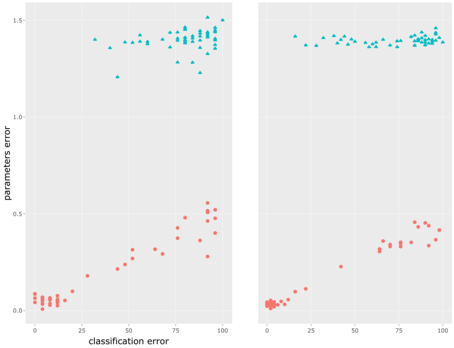

Figure: Comparison of estimation error of estimates based on Hellinger and the partial likelihood under model M2 for $n = 50$ (left) and $n = 100$ (right)

Conclusion

Summary

We provide new criteria for parametric estimation in SBM, which has the following properties,

- The estimated parameters is asymptotically consistent
- The estimated parameters inherit robustness properties to outliers by using φ-divergence.

Conclusion

Summary

We provide new criteria for parametric estimation in SBM, which has the following properties,

- The estimated parameters is asymptotically consistent
- The estimated parameters inherit robustness properties to outliers by using φ-divergence.

Perspectives

More studies of the robustness properties of model selection outside the SBM

Thank You

Bibliography

- Holland, P., Laskey, K. & Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*. **5**, 109-137 (1983,6)
- Ferraris, C. Phi Divergence and Consistent Estimation for Stochastic Block Model. *Data Analysis And Related Applications 3 Theory And Practice – New Approaches*. (2024,3)
- Mattheou, K., Lee, S. & Karagrigoriou, A. A model selection criterion based on the BHHJ measure of divergence. *Journal Of Statistical Planning And Inference*. **139**, 228-235 (2009,2)
- Makarychev, K., Makarychev, Y. & Vijayaraghavan, A. Learning Communities in the Presence of Errors. *Conference On Learning Theory*. pp. 1258-1291 (2016,6),

Work between Safran Aircraft Engines (SAE) and the LPSM (PhD thesis under CIFRE contract)

Work under the supervision of

- Michel Broniatowski (LPSM)
- Frédéric Guilloux (LPSM)
- Annick Valibouze (LIP6-LPSM)
- Mohamed Achibi (SAE)

Abstract

Les modèles de graphes aléatoires par blocs sont étudiés depuis quelques années maintenant et trouvent beaucoup d'applications dans des domaines variés : finance, génomique, réseaux sociaux etc. La plupart des critères d'estimation de la littérature sur ces modèles reposent sur la vraisemblance des données. Or le maximum de vraisemblance peut être fortement influencé par la présence de valeurs aberrantes ou d'autres déviations au modèle de graphe théorique considéré. Afin d'obtenir des estimateurs plus robustes, nous avons développé de nouveaux critères basés sur l'utilisation de divergences entre graphes, i.e., entre les lois générant ces graphes. Nous démontrons la consistance des estimateurs associés à ces critères. De plus nous proposons un critère de sélection de modèle afin de choisir le nombre de blocs. Enfin, nous illustrons l'intérêt de ces méthodes par rapport à la vraisemblance en présence de mauvaises spécifications du modèle.