Université de Montréal

# Apprentissage statistique des modèles de graphes aléatoires exponentiels : théorie et méthodes

par

**Gabriel Fortin-Leblanc**

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

26 août 2024

© Gabriel Fortin-Leblanc, 2024

# Université de Montréal

Faculté des arts et des sciences


Ce mémoire intitulé

## Apprentissage statistique des modèles de graphes aléatoires exponentiels : théorie et méthodes

présenté par


## Gabriel Fortin-Leblanc


a été évalué par un jury composé des personnes suivantes :

*Alejandro Murua*
(président-rapporteur)

*Florian Maire*
(directeur de recherche)

*Lucas Benigni*
(membre du jury)

# Résumé

Les modèles de graphes aléatoires de la famille exponentielle sont flexibles et permettent d'analyser des relations entre des objets. Malheureusement, leur souplesse vient avec un lot de difficultés. Échantillonner des graphes provenant de ces modèles est le plus souvent infaisable, ce qui oblige d'utiliser des techniques comme Métropolis-Hastings. Approximer l'estimateur de vraisemblance maximale est ardu, et l'utilisation de techniques souvent peu connues est nécessaire. En statistique bayésienne, une approximation par à une loi normale de la loi à postériori est tout ce qu'on peut espérer. Dans cette monographie, on commence par montrer comment échantillonner selon ces modèles, puisque cette tâche revient tout au long du document. Il s'en suit une étude détaillée sur l'estimateur de vraisemblance maximale, et puis finalement, on montre comment obtenir une approximation de la loi à postériori dans le cadre bayésien.

Un grand effort est apporté à l'approximation par Monte Carlo par chaînes de Markov de l'estimateur de vraisemblance maximale. On y voit les conditions nécessaires et suffisantes à son existence et à son unicité, mais aussi à l'existence et l'unicité de son approximation. On apporte des améliorations aux algorithmes déjà existants pour mieux prendre en compte la théorie et assurer leur robustesse. On termine en démontrant le comportement asymptotique de l'approximation. Les précédentes études ont simplement appliqué la méthode pour ce type de modèle sans s'assurer de la validité de l'application, ce qu'on rectifie.

**Mots-clés :** Statistique computationelle, Statistique bayésienne, Modèles de graphes aléatoires exponentiels, Estimateur de vraisemblance maximale par chaîne de Markov, Inférence variationelle

# Abstract

Exponential random graph models are flexible and allow analyzing relations between any type of objects. Unfortunately, they come with a large number of nuisances. Sampling random graphs is impossible. It left no choice except to use methods such as Metropolis-Hastings. Approximating the maximum likelihood estimator is not trivial, and employing not well-known tools are needed. In Bayesian statistics, a normal law approximation is the closest we can get from the posterior law. In this monograph, we first show how to sample random graphs since it is required for all other techniques. After that, a well-developed study on the maximum likelihood estimator is done. Finally, we show how to approximate the posterior law.

A great effort has been made to clarify the Markov chain Monte Carlo approximation of the maximum likelihood estimator. We exhaustively enumerate the conditions for the existence and the uniqueness of the estimator and its approximation. We then bring improvements to existing algorithms for better reflecting the theory and strengthen robustness. Lastly, we show the asymptotic behaviour of this approximation. Previous studies simply applied the technique without verifying all needed conditions. We pass through all those conditions, and apply modifications when needed.

**Keywords :** Computational statistics, Bayesian statistics, Exponential random graph models, Markov chain maximum likelihood estimator, Variational inference

# Table des matières

**Résumé** . . . . . v

**Abstract** . . . . . vii

**Liste des tableaux** . . . . . xi

**Table des figures** . . . . . xiii

**Liste des sigles et des abréviations** . . . . . xv

**Remerciements** . . . . . xvii

**Introduction** . . . . . 1

Définitions . . . . . 2

Littérature . . . . . 3

Organisation et contributions . . . . . 4

**Chapitre 1. Échantillonnage** . . . . . 7

1.1. *Tie-no-tie* . . . . . 7

1.2. Borne de convergence . . . . . 10

**Chapitre 2. Estimateur de vraisemblance maximale** . . . . . 13

2.1. Existence et unicité de l'estimateur . . . . . 13

2.2. Pseudo-vraisemblance . . . . . 21

2.3. Maximum de vraisemblance par Monte Carlo par chaînes de Markov . . . . . 23

2.3.1. Adaptation pour les modèles de graphes aléatoires de la famille exponentielle 28

2.3.2. Convergence de l'estimateur . . . . . 33

2.3.3. Comportement asymptotique du maximum de vraisemblance par Monte Carlo par chaînes de Markov . . . . . 33

2.3.4. Algorithmes en deux étapes : approximation de l'estimateur de vraisemblance maximale . . . 38
2.3.4.1. Pseudo-observation . . . 39
2.3.4.2. Unicité du maximum . . . 41
2.3.4.3. Algorithme modifié . . . 41
2.3.5. Travaux futurs . . . 47

**Chapitre 3. Inférence de la loi à postériori . . . 49**

3.1. Méthodes classiques en inférence variationnelle . . . 50
3.1.1. Inférence variationnelle pour les modèles de graphes aléatoires . . . 51
3.1.2. Travaux futurs . . . 54

**Références . . . 55**

# Liste des tableaux

2.1 Comparaison entre l'Algorithme 2.3.20 et l'Algorithme 2.3.22 pour le Modèle 1 appliqué au graphe Karate Club Network [**69**]. Pour chaque paramètre d'expérience, 50 expériences ont été produites. On suppose qu'une expérience n'a pas convergé si le nombre d'itérations dépasse 100. . . . . 43

2.2 Comparaison entre l'Algorithme 2.3.20 et l'Algorithme 2.3.22 pour le Modèle 2 appliqué au graphe Karate Club Network [**69**]. Pour chaque paramètre d'expérience, 50 expériences ont été produites. On suppose qu'une expérience n'a pas convergé si le nombre d'itérations dépasse 100. . . . . 44

# Table des figures


0.1 Exemples de graphes . . . . . 2

0.2 Exemples de statistiques exhaustives . . . . . 4

0.3 Graphes simulés avec deux statistiques exhaustives : le nombre d'arêtes et le nombre de 2-étoiles (chemin de longueur 2) . . . . . 5

1.1 Simulation de l'algorithme de Métropolis pour le modèle avec deux statistiques exhaustives, le nombre d'arêtes et le degré du poids géométrique, pour des graphes à 4 sommets et selon le paramètre $\theta = (0.5, -0.2)$ . . . . . 9

2.1 Graphe composé de clubs de karaté illustrant les interactions à l'extérieur de la pratique du sport [**69**] . . . . . 45

2.2 Exécution de l'Algorithme 2.3.22 avec le nombre d'arêtes et le nombre d'étoiles à trois branches comme statistiques exhaustives. Le graphe observé est en haut à gauche, les statistiques exhaustives pour chacun des graphes possibles sont représentées en haut à droite, où la couleur représente la fréquence d'un même point, la deuxième ligne représente la fonction $\phi_{\omega_0}^{(K)}$ à chacune des itérations, et la troisième ligne représente les statistiques exhaustives des graphes échantillonnés. 46

2.3 100 réalisations MCMCMLE pour le Modèle 1 pour le *Karate Club Network* . . . . . 47

3.1 Approximation de la loi à postériori par une loi normale en minimisant la divergence Kullback-Leibler renversée pour le modèle composé du nombre d'arêtes et du nombre de chemins de longueur deux . . . . . 53

# Liste des sigles et des abréviations

| | |
|---|---|
| ERGM | Modèle de graphes aléatoires de la famille exponentielle, de l'anglais *Exponential Random Graph Model* |
| MLE | Estimateur de vraisemblance maximale, de l'anglais *Maximum Likelihood Estimator* |
| MPLE | Estimateur de pseudo-vraisemblance maximale, de l'anglais *Maximum PseudoLikelihood Estimator* |
| MCMC | Monte Carlo par Chaînes de Markov, de l'anglais *Markov Chain Monte Carlo* |
| MCMCMLE | Estimateur de vraisemblance maximale par Monte Carlo par Chaînes de Markov, de l'anglais *Markov Chain Monte Carlo Maximum Likelihood Estimator* |
| VI | Inférence variationelle, de l'anglais *Variational Inference* |
| RKL | (divergence) Kullback-Leibler renversée, de l'anglais *Reverse Kullback-Leibler (divergence)* |

| | |
|---|---|
| FKL | (divergence) Kullback-Leibler avant, de l'anglais *Forward Kullback-Leibler (divergence)* |
| ELBO | borne inférieure de l'évidence, de l'anglais *Evidence Lower BOund* |

# Remerciements

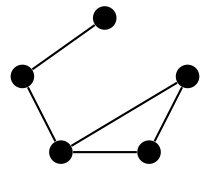
Je ne peux pas imaginer un meilleur superviseur que le professeur Florian Maire. Il a toujours été présent, et m'a encouragé à retourner toutes les pierres sur mon chemin. J'ai énormément appris à ses côtés. Pour tout son temps et son soutien, je le remercie de tout cœur.

Je dois remercier le Département de mathématiques et de statistique de l'Université de Montréal, la Fondation Desjardins ainsi que les Fonds de recherche du Québec pour leurs généreuses bourses qui m'ont permis de me concentrer sur mes études sans me soucier du côté financier.

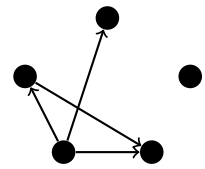
# Introduction

Soit un ensemble $V$. On définit un graphe $G$ comme le couple $(V, E)$, où $E \subset V^2$. On nomme les éléments de $V$ des sommets, et les éléments de $E$ des arêtes. Soit une paire de sommets $u, v \in V$. On dit que $u$ et $v$ sont adjacents si $(u, v) \in E$. On remarque que $E$ peut être vu comme une relation sur $V$. Si $E$ est symétrique, c'est-à-dire que si $(u, v) \in E$ alors $(v, u) \in E$, on dit que $G$ est non dirigé. Dans ce cas, on identifie les arêtes $(u, v)$ et $(v, u)$ comme une seule arête $\{u, v\}$. Sinon, on nomme souvent les arêtes comme des arcs, et souvent la relation est notée $A$. Si $E$ est non réflexive, c'est-à-dire que si pour tout $u \in V$ on a $(u, u) \notin E$ et que $G$ est non dirigé, on dit que $G$ est simple. Il existe d'autres types de graphes, comme les multigraphes, où plusieurs instances d'une même arête sont permises, ou les graphes pondérés, où chaque arête est associée à un poids, mais on se concentrera sur les graphes simples. Si $V$ est fini, on note $n = |V|$, et on peut représenter $G$ par une matrice $A \in \{0, 1\}^{n \times n}$, qu'on appelle une matrice d'adjacence, où $A_{ij} = 1$ si $\{i, j\} \in E$ et 0 sinon. Des exemples de graphes sont représentés dans la Figure 0.1. Généralement, un graphe est une structure mathématique qui permet d'encoder des relations. Elle est composée de sommets qui peuvent représenter des objets quelconques ainsi que d'arêtes qui relient des paires de sommets entre eux. On peut penser à des échanges économiques entre des pays, ou encore à des contacts entre des personnes.

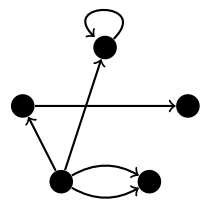
Les modèles de graphes aléatoires de la famille exponentielle, *Exponential Random Graph Models*, (ERGM) sont grandement utilisés en sociologie [**34, 39, 21**], en science politique [**13, 54**] et en économie [**22, 26, 37**], mais d'autres disciples commencent aussi à les utiliser, comme la biologie [**55**] et la médecine [**68**]. Dans l'unique but de simplifier la démarche, plusieurs résultats qui seront rencontrés dans le document tiendront pour acquis que le graphe est simple, mais ils peuvent aisément être généralisés à des graphes orientés, ou même avec des statistiques associées aux sommets et aux arêtes.



(a) Un graphe simple



(b) Un graphe orienté



(c) Un multigraphe

**Figure 0.1** – Exemples de graphes

# Définitions

Soit $n \in \mathbb{Z}_+$ un nombre de sommets. On définit $\mathcal{Y}_n$ l'ensemble de tous les graphes simples à $n$ sommets. On note l'ensemble des sommets $V_n = \{1, \ldots, n\}$ et la fonction $E_n : \mathcal{Y}_n \to \binom{V_n}{2}$, qui prend un graphe et retourne l'ensemble de ses arêtes. On définit l'espace paramétrique comme un sous-ensemble d'un espace euclidien $\Gamma \subset \mathbb{R}^d$. Dans le but de simplifier l'écriture, on a défini les sommets comme les nombres naturels $1, \ldots, n$, mais ils peuvent être des représentants d'objets plus complexes. On remarque aussi que les graphes de $\mathcal{Y}_n$ partagent le même ensemble de sommets, mais ont des ensembles d'arêtes différents. Même si la théorie peut être généralisée à plusieurs graphes, on supposera qu'une seule réalisation est observée. On définit la fonction $s : \mathcal{Y}_n \to \mathbb{R}^d$ qui pour tout graphe simple retourne un vecteur de statistiques exhaustives, et on pose l'ensemble de toutes les statistiques exhaustives $\mathcal{S} = \{s(y) : y \in \mathcal{Y}_n\}$. Soit $Y$ une variable aléatoire qui prend ses réalisations dans $\mathcal{Y}_n$. On définit la fonction de masse de $Y$, la fonction de vraisemblance, par

$$p_\theta(y) = \frac{\exp(s(y)^{\mathrm{T}}\theta)}{\sum_{y' \in \mathcal{Y}_n} \exp(s(y')^T\theta)} \quad \text{pour tout } y \in \mathcal{Y}_n, \theta \in \Gamma. \tag{0.0.1}$$

On écrira la constante de normalisation, ou la fonction de partition, par

$$z(\theta) = \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta). \tag{0.0.2}$$

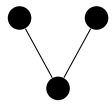
On observe en réécrivant ainsi

$$p_\theta(y) = \exp\left(s(y)^T\theta - \log z(\theta)\right),$$

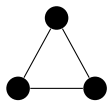
que cette loi fait partie de la famille exponentielle, d'où le nom. Tout au long du présent document, on supposera que $n$ est bien défini, et de même pour la fonction des statistiques exhaustives $s$. Il faut cependant garder en tête que la fonction $s$ est définie par l'utilisateur et qu'aucune contrainte n'est posée sur cette dernière. On observe que l'étude asymptotique des ERGM peut être faite de deux manières : soit en faisant tendre $n$, le nombre de sommets, vers l'infini, ou soit en faisant tendre le nombre de graphes observés vers l'infini. Lors de ces études, le cadre sera bien spécifié.

## Littérature

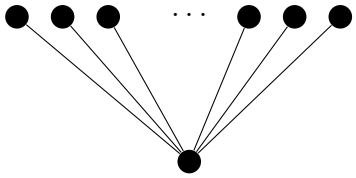
Ces premiers modèles ont vu le jour en 1959 dans les travaux d'Erdős et Rényi [**14**] (republié en 2022). Chaque arête était alors indépendante les unes des autres, et le modèle pouvait s'exprimer comme une variable de loi binomiale en ne considérant que le nombre d'arêtes. Ces modèles se sont complexifiés depuis en ajoutant de la dépendance entre les arêtes, et les statistiques exhaustives se sont diversifiées [**15, 56, 51, 43, 28, 65**]. Des exemples de statistiques intéressantes sur les graphes sont représentés dans la Figure 0.2. Une étude du commerce mondial du blé a été faite à l'aide de ces modèles [**22**]. Une statistique alors intéressante pour cet exemple était le nombre de pays qui avaient un certain nombre de partenaires d'exportation. D'autres variantes de ces modèles se sont ajoutées : en différenciant les sommets entre eux [**62, 66, 9, 10**], en considérant une collection structurée ou non de graphes [**49, 64**]. Pour ces différents modèles, plusieurs techniques de choix de modèles ont été adaptées à ceux-ci [**29, 7, 6**]. Dû à l'incapacité de calculer explicitement la fonction de vraisemblance pour des graphes de tailles moyennes et plus, beaucoup de chercheurs se sont concentrés sur des méthodes d'approximation du MLE, ou encore simplement sur des méthodes d'estimation de la fonction de vraisemblance [**59, 19, 30, 8, 27, 31, 25, 6, 41, 53**]. Si cette fonction est difficile à calculer, la loi à postériori l'est tout autant. Même des techniques modernes telles que l'inférence variationnelle mènent à des approximations acceptables [**61, 70**], mais sans justifications. Par ailleurs, toutes ces méthodes demandent d'échantillonner selon la fonction de vraisemblance. La méthode par défaut utilise Metropolis [**57**] même si le temps de convergence peut être exponentiel selon le nombre de sommets [**4**]. Des exemples de réalisations de graphes aléatoires peuvent être observés dans la Figure



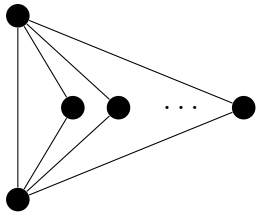
(a) Le nombre de deux-étoiles (*two-stars*)



(b) Le nombre de triangle



(c) Le degré du poids géométrique (*geometrically weighted degree*)



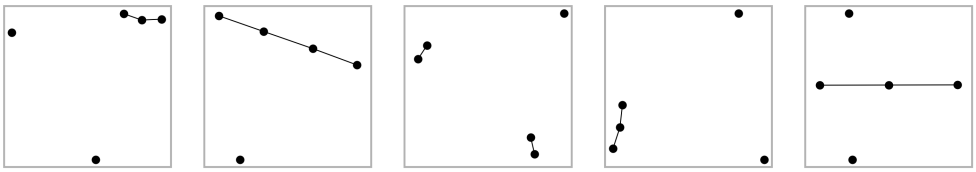
(d) Le degré du poids géométrique de partenaires communs (*geometrically weighted edgewise shared partner*)

**Figure 0.2** – Exemples de statistiques exhaustives

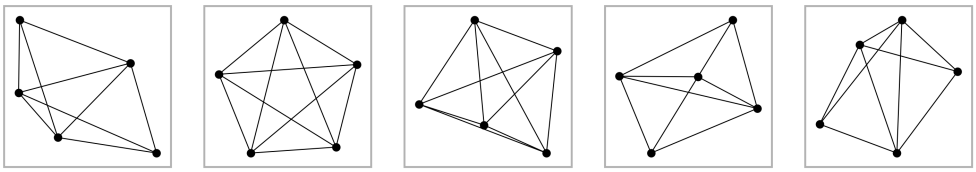
0.3. Malgré des améliorations apportées à la technique [**58**], une grande partie de l'espace paramétrique concentre presque toute la probabilité sur quelques graphes de l'espace : soit des graphes presque complets, soit des graphes presque vides [**23, 11, 12**]. Dû à leur grande expressivité, et à l'aisance de leur interprétation, les ERGM sont très utiles en pratiques, même s'ils sont très instables : une petite variation du paramètre entraîne une dégénérescence de la loi, un phénomène connu en physique statistique sous le nom de régime à changement de phases. Certains ont essayé de créer des modèles restreints pour éviter de tomber sur des lois dégénérées [**33**].

# Organisation et contributions

Une méthode pour simuler des graphes aléatoires provenant d'un ERGM est présentée au Chapitre 1. On y voit non seulement une méthode pour construire une chaîne de Markov uniformément ergodique dans la Section 1.1, mais aussi quelques propriétés de convergence de cette chaîne dans la Section 1.2. Une grande partie des efforts ont été accordés à l'étude



(a) Graphes simulés selon le paramètre $\theta = (0.5, -1)$



(b) Graphes simulés selon le paramètre $\theta = (0.5, 0.5)$

**Figure 0.3** – Graphes simulés avec deux statistiques exhaustives : le nombre d'arêtes et le nombre de 2-étoiles (chemin de longueur 2)

de l'estimateur de vraisemblance maximale. Une méthode pour approcher cet estimateur est décrite, et une analyse de celui-ci est faite au Chapitre 2. Plus précisément, on pose les conditions nécessaires et suffisantes à l'existence et à l'unicité de l'estimateur dans la Section 2.1. On propose ensuite deux approximations de cet estimateur dans les Sections 2.2 et 2.3, l'estimateur de pseudo-vraisemblance maximale et l'estimateur de vraisemblance maximale par Monte Carlo par chaînes de Markov. L'accent est plutôt porté sur ce dernier. On y voit la méthode générale, puis dans la Section 2.3.1, comment l'adapter au ERGM. Sous certaines hypothèses, on montre que cet estimateur converge vers l'estimateur de vraisemblance maximale dans la Section 2.3.2, on identifie sa loi asymptotique dans la Section 2.3.3. On présente l'algorithme communément utilisé pour trouver cet estimateur dans la Section 2.3.4, et les modifications à apporter pour prendre en compte toutes les particularités des ERGM dans les Sections 2.3.4.1, 2.3.4.2 et 2.3.4.3. On termine ce chapitre en proposant des pistes pour de futurs travaux sur cet estimateur dans la Section 2.3.5. Finalement, l'inférence de la loi à postériori dans le cadre bayésien est abordée au Chapitre 3. On y présente la méthode d'inférence variationnelle et comment l'appliquer au ERGM dans les Sections 3.1 et 3.1.1.

Dans la plupart des précédents travaux, on caractérisait mal l'existence de l'estimateur de vraisemblance maximale, et lorsqu'il existait on supposait qu'il était toujours unique. Or, ce n'est pas le cas, et on caractérise son existence et son unicité. En supposant son existence, il se peut que l'estimateur de vraisemblance maximale par Monte Carlo par chaînes de Markov n'existe pas toujours, et lorsqu'il existe, n'est pas toujours unique. Les conditions d'existence

et d'unicité de cette approximation ont trop longtemps été glissées sous le tapis, et ce pourtant même si l'unicité est une condition à la convergence de cet estimateur. On redéfinit cet estimateur afin de s'assurer de son existence et de son unicité. Étant donné l'ignorance de ces conditions, l'algorithme fréquemment utilisé pouvait échouer ou même encore mener à différentes estimations pour différentes exécutions. En apportant des modifications, on renforce leur robustesse, ce qui est aussi appuyé empiriquement.

# Chapitre 1

# Échantillonnage

Simuler une variable aléatoire provenant d'un ERGM est une tâche courante qui apparaît dans presque toutes les méthodes qui impliquent de tels modèles. Dans des cas particuliers, il est possible d'échantillonner parfaitement des graphes, comme pour le modèle Erdős-Rényi [**14**], le modèle où toutes les arêtes sont indépendantes les unes des autres. Dans le cas où le nombre de sommets est petit, la méthode de *coupling from the past* [**48**] peut être utilisée. Cependant, pour la plupart des ERGM, la méthode par défaut nécessite d'échantillonner sur une chaîne de Markov ergodique avec une certaine loi stationnaire. L'algorithme de Metropolis-Hastings [**24**] a été adapté par Snijders [**57**] pour construire une chaîne de Markov sur laquelle on se permet d'échantillonner des graphes. Cet algorithme est nommé *tie-no-tie* dû à la méthode de changement d'état.

## 1.1. *Tie-no-tie*

On rappelle d'abord la méthode de Hastings, qui étend la méthode de Metropolis [**42**]. Elle permet de créer une chaîne de Markov avec une distribution stationnaire choisie à l'avance, disons $\pi$. Puisqu'elle n'impose pas de loi de transition, l'utilisateur doit toujours vérifier que la chaîne résultante est irréductible. Pour la suite, l'espace d'état sera supposé fini, comme ce sera le cas avec les ERGM. Les états seront identifiés par $1, \ldots, N$, et on notera $X(t)$ l'état qu'occupe notre variable aléatoire au temps $t \in \mathbb{Z}_+$. Dans le cas des ERGM, la chaîne en est une de graphes. Soit $Q$ une matrice des transitions quelconque sur le même espace d'état. Pour toute paire d'états $i, j$, on pose

$$\alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi_i Q_{ij}}{\pi_j Q_{ji}}}, \tag{1.1.1}$$

où $\pi_i$ est la probabilité d'être à l'état $i$ sous la distribution stationnaire souhaitée et où $s_{ij}$ est une fonction symétrique, comme par exemple

$$s_{ij} = \begin{cases} 1 + \frac{\pi_i Q_{ij}}{\pi_j Q_{ji}} & \text{si } \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}} \geq 1; \\ 1 + \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}} & \text{sinon.} \end{cases} \tag{1.1.2}$$

Si $X(t) = i$ et que l'état proposé au temps $t+1$ est $j$, alors $X(t+1) = j$ avec probabilité $\alpha_{ij}$ et $X(t+1) = i$ avec probabilité $1 - \alpha_{ij}$. L'état initial peut être choisi au hasard ou non, car cela n'influence pas la distribution stationnaire. En substituant l'Équation 1.1.2 dans l'Équation 1.1.1 et en posant $Q_{ij} = Q_{ji}$, on obtient la probabilité d'acceptation obtenue par la méthode de Metropolis, soit

$$\alpha_{ij}^{\mathrm{M}} = \min\left\{1, \frac{\pi_j}{\pi_i}\right\}. \tag{1.1.3}$$

Pour le cas d'un ERGM, on pose certaines notations. Pour une variable aléatoire $Y$ et pour une paire de sommets $\{u, v\}$, on note $Y_{uv}^{+} = (V_n, E(Y) \cup \{\{u, v\}\})$, $Y_{uv}^{-} = (V_n, E(Y) \backslash \{\{u, v\}\})$, et

$$Y_{uv}^{c} = \begin{cases} Y_{uv}^{+} & \text{si } \{u, v\} \notin E(Y); \\ Y_{uv}^{-} & \text{si } \{u, v\} \in E(Y); \end{cases}$$

qui est simplement le graphe $Y$, mais en ayant modifié l'état de l'arête $\{u, v\}$. Soit $Y(t)$ le graphe au temps $t \in \mathbb{N}$. La fonction de transition « choisit » en fait une paire de sommets, disons $\{u, v\}$ et $Y(t)_{uv}^{c}$ est le prochain état proposé. Le choix de la paire de sommets peut être fait aléatoirement ou même simplement en passant à travers chaque paire un à un comme l'échantillonneur de Gibbs. On peut retrouver ce résultat en partant de la méthode de Hastings. Si on pose $Q_{ij}$ comme $n(n-1)/2$ si les graphes des états $i$ et $j$ partagent toutes sauf une arête, et 0 sinon, alors la probabilité par la méthode de Metropolis est

$$\alpha_{ij} = \min\left\{1, \exp\left(\theta^{\mathrm{T}}(s(Y(t)_{uv}^{c}) - s(Y(t)_{uv}))\right)\right\},$$

où $\{u, v\}$ est l'arête que ne partagent pas les graphes des états $i$ et $j$.

La Figure 1.1 montre le résultat de 15 itérations de l'algorithme de Métropolis-Hastings (Métropolis « classique ») pour des graphes de quatre sommets utilisant le modèle composé du nombre d'arêtes et du degré du poids géométrique comme statistiques exhaustives. Les états sont tous les graphes possibles et sont disposés horizontalement selon l'ordre partiel du nombre d'arêtes. On remarque que le graphe vide est l'état initial et on observe les pas effectués par la chaîne de Markov.

Plusieurs améliorations peuvent être apportées aux techniques d'échantillonnage des ERGM. Dans un premier temps, l'algorithme de Métropolis-Hastings nécessite régulièrement de calculer des statistiques exhaustives de graphes, et plusieurs de ces graphes sont isomorphes. L'utilisation d'une fonction de hachage qui est invariante à des isomorphes permettrait d'accélérer l'exécution en enregistrant dans une table de hachage les statistiques

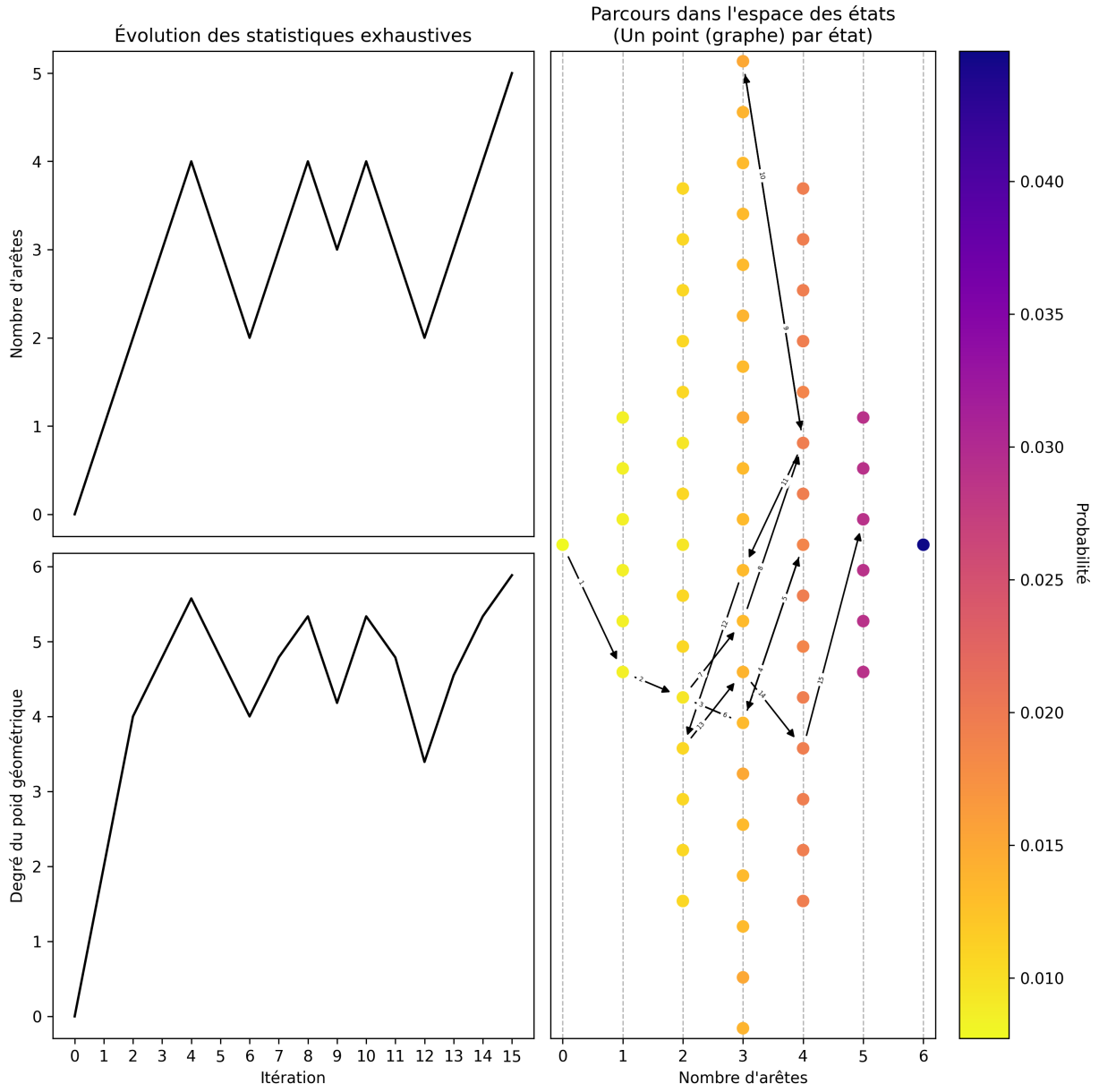



**Figure 1.1** – Simulation de l'algorithme de Métropolis pour le modèle avec deux statistiques exhaustives, le nombre d'arêtes et le degré du poids géométrique, pour des graphes à 4 sommets et selon le paramètre $\theta = (0.5, -0.2)$

exhaustives déjà calculées. Sous une autre optique, aucune borne dépendant du nombre de sommets sur le temps de mélange n'a été trouvée. Une tentative a été faite dans ce sens [**4**], mais le cas général reste encore ouvert.

## 1.2. Borne de convergence

Malgré qu'une borne selon le nombre de sommets n'est pas encore trouvée, pour un choix de statistiques exhaustives et un nombre de sommets fixes, le temps de mélange est acceptable. Pour simplifier l'écriture, on pose le noyau de transition en $k \in \mathbb{Z}_+$ transitions pour une certaine chaîne de Markov $\{X_t\}_{t\in\mathbb{Z}_+}$ sur un espace d'état $S$ avec $x \in S$ comme état initial,

$$\mathbb{P}^k(A|x) = \mathbb{P}(X_k \in A|X_0 = x) \quad \text{pour tout ensemble mesurable } A \subset S.$$

**Définition 1.2.1.** Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov sur l'espace d'états $S$. La chaîne satisfait la **condition de minorité**, *minorization condition*, sur un ensemble $C \subset S$ s'il existe une mesure de probabilité $\rho$ sur $S$, un nombre entier positif $t_0 \in \mathbb{Z}_+$, et un nombre réel $\varepsilon > 0$ tels que pour tout état $x \in C$ et pour tout ensemble mesurable $A \subset S$

$$\mathbb{P}^{t_0}(A|x) \geq \varepsilon\rho(A).$$

On dit que l'ensemble $C$ est **petit**. Si tout l'espace d'état $S$ est petit, alors la chaîne est **uniformément ergodique**.

Pour résumé en d'autres mots, un ensemble d'états $C$ est petit, si à partir de n'importe quel état de cet ensemble, on peut se rendre vers n'importe quel ensemble d'états avec une probabilité bornée inférieurement qui ne dépend pas de cet état.

**Remarque 1.2.2.** Il est aisé de voir qu'une chaîne de Markov irréductible sur un espace d'état fini est uniformément ergodique. Ainsi, lorsqu'on échantillonne des graphes aléatoires provenant d'un ERGM par une chaîne de Markov en utilisant l'algorithme *tie-no-tie*, on obtient une chaîne qui est uniformément ergodique.

**Proposition 1.2.3.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov sur l'espace d'état $S$ qui admet une loi stationnaire $\pi$. Si la chaîne est uniformément ergodique, c'est-à-dire qu'il existe une mesure de probabilité $\rho$ sur $S$, un nombre entier positif $t_0 \in \mathbb{Z}_+$, et un nombre réel $\varepsilon > 0$ tels que pour tout état initial $x \in S$ et pour tout ensemble mesurable $A \subset S$*

$$\mathbb{P}^{t_0}(A|x) \geq \varepsilon\rho(A),$$

*alors, pour tout état $x \in S$*

$$\left\|\mathbb{P}^t(\,\cdot\,|x) - \pi(\cdot)\right\|_{\mathrm{TV}} \leq (1-\varepsilon)^{\lfloor t/t_0 \rfloor}. \tag{1.2.1}$$

*(Théorème 6.15 de* [**47**]*)*

Une première observation provenant de la Proposition 1.2.3 est que conditionnellement à un certain ERGM, le temps de mélange se produit exponentiellement vite. Comme pour

le fléau de la dimensionnalité en régression, la constante $\varepsilon$ dépend souvent du nombre de sommets, mais dû à la grande flexibilité des ERGM, la dépendance de $\varepsilon$ en ce paramètre n'est pas connue dans le cas général.

# Chapitre 2

# Estimateur de vraisemblance maximale

Lors de l'étude de l'estimateur de vraisemblance maximale, *Maximum likelihood estimator*, (MLE) du modèle ERGM, on supposera que l'espace paramétrique est $\Gamma = \mathbb{R}^d$, ce qui amène son lot de difficultés qui seront à discuter au cours du présent chapitre. Dans le cas où l'espace paramétrique $\Gamma$ est compact, alors une grande partie des difficultés disparaissent, car dans ce cas, le MLE existe toujours et les critères nécessaires pour approximer l'estimateur sont satisfaits. On rappelle aussi que $\mathcal{S} = \{s(y) : y \in \mathcal{Y}_n\}$ est l'ensemble des statistiques exhaustives.

On commence le chapitre en définissant le MLE et sous quelles conditions il existe et est unique. L'estimateur de pseudo-vraisemblance maximale (MPLE), qui souhaite se substituer au MLE, est présenté à la Section 2.2 puisqu'il peut être utile dans certains contextes. L'estimateur de vraisemblance maximale par Monte Carlo par chaînes de Markov, *Markov chains Monte Carlo maximum likelihood estimator*, (MCMCMLE) est présenté à la Section 2.3.

## 2.1. Existence et unicité de l'estimateur

On définit l'estimateur de vraisemblance maximale, qu'on notera $\hat{\theta}_{\mathrm{ML}}$ comme

$$\hat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta \in \Gamma} p_\theta(Y). \tag{2.1.1}$$

Cet estimateur n'est pas bien défini pour toute réalisation de $Y$, et lorsqu'il existe, il peut ne pas être unique. En effet, en se basant sur la théorie de Barndorff-Nielsen [**2**] (réimprimé en 2014), Handcock [**23**] pose des conditions nécessaires et suffisantes à son existence. Pour ce faire, on doit définir l'enveloppe affine, l'enveloppe convexe et l'intérieur convexe relatif d'un ensemble.

**Définition 2.1.1.** Soit $C = \{c_1, \ldots, c_n\}$ un ensemble de $n$ points d'un espace euclidien $\mathbb{R}^d$. On définit l'**enveloppe affine** (*affine hull*) l'ensemble des points $x \in \mathbb{R}^d$ tel qu'il existe des

nombres réels $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ avec $\sum_{i=1}^n \lambda_i = 1$ où

$$x = \sum_{i=1}^n \lambda_i c_i.$$

On note l'ensemble de ces points $\mathrm{aff}(C)$.

On notera $V_C$ l'espace vectoriel induit par l'enveloppe affine de $C$, c'est-à-dire pour $u \in \mathrm{aff}(C)$, un élément pivot, on définit cet espace comme

$$V_C = \{v - u : v \in \mathrm{aff}(C)\}.$$

Par extension, on pose que la dimension de l'enveloppe affine est la dimension de l'espace vectoriel induit,

$$\dim \mathrm{aff}(C) := \dim V_C.$$

**Remarque 2.1.2.** En effet, $V_C \subseteq \mathbb{R}^d$ est un sous-espace vectoriel. Pour deux vecteurs $v_1, v_2 \in V_C$, il existe des nombres réels $\lambda_1^{(1)}, \ldots, \lambda_n^{(1)}, \lambda_1^{(2)}, \ldots, \lambda_n^{(2)}, \lambda_1^{(u)}, \ldots, \lambda_n^{(u)}$, avec $\sum_{i=1}^n \lambda_i^{(1)} = \sum_{i=1}^n \lambda_i^{(2)} = \sum_{i=1}^n \lambda_i^{(u)} = 1$, tels que $v_1 = \sum_{i=1}^n \lambda_i^{(1)} c_i - \sum_{i=1}^n \lambda_i^{(u)} c_i$ et $v_2 = \sum_{i=1}^n \lambda_i^{(2)} c_i - \sum_{i=1}^n \lambda_i^{(u)} c_i$, et alors pour des nombres réels $a_1, a_2 \in \mathbb{R}$, on trouve que $a_1 v_1 + a_2 v_2 \in V_C$, puisque

$$\begin{aligned}
a_1 v_1 + a_2 v_2 &= a_1 \sum_{i=1}^n \left(\lambda_i^{(1)} - \lambda_i^{(u)}\right) c_i + a_2 \sum_{i=1}^n \left(\lambda_i^{(2)} - \lambda_i^{(u)}\right) c_i \\
&= \sum_{i=1}^n \left(a_1 \lambda_i^{(1)} + a_2 \lambda_i^{(2)} - (a_1 + a_2)\lambda_i^{(u)}\right) c_i \\
&= \sum_{i=1}^n \left(a_1 \lambda_i^{(1)} + a_2 \lambda_i^{(2)} - (a_1 + a_2 - 1)\lambda_i^{(u)}\right) c_i - \sum_{i=1}^n \lambda_i^{(u)} c_i,
\end{aligned}$$

et en remarquant que,

$$\sum_{i=1}^n \left(a_1 \lambda_i^{(1)} + a_2 \lambda_i^{(2)} - (a_1 + a_2 - 1)\lambda_i^{(u)}\right) = a_1 + a_2 - (a_1 + a_2 - 1) = 1.$$

**Définition 2.1.3.** Soit $C = \{c_1, \ldots, c_n\}$ un ensemble de $n$ points d'un espace euclidien $\mathbb{R}^d$. On définit l'**enveloppe convexe** (*convex hull*) l'ensemble des points $x \in \mathbb{R}^d$ tel qu'il existe des nombres réels $\lambda_1, \ldots, \lambda_n \geq 0$ avec $\sum_{i=1}^n \lambda_i = 1$ où

$$x = \sum_{i=1}^n \lambda_i c_i.$$

On note l'enveloppe convexe $\mathrm{conv}(C)$. On remarque que

$$\mathrm{conv}(C) \subset \mathrm{aff}(C).$$

**Définition 2.1.4.** Soit $C = \{c_1, \ldots, c_n\}$ un ensemble de $n$ points d'un espace euclidien $\mathbb{R}^d$. On définit l'**intérieur convexe relatif** l'ensemble des points $x \in \operatorname{aff}(C)$ tel qu'il existe un nombre réel $\delta > 0$

$$B(x, \delta) \cap \operatorname{aff}(C) \subset \operatorname{conv}(C),$$

où $B(x, \delta) = \{y \in \mathbb{R}^d : \|x - y\| < \delta\}$ est la boule ouverte de centre $x$ et de rayon $\delta$ . On note l'ensemble de ces points $\operatorname{rcint}(C)$.

On appelle la **frontière convexe relative** l'ensemble des points $x \in \operatorname{aff}(C)$ qui sont dans l'enveloppe convexe de $C$, mais pas dans son intérieur convexe relatif. On note l'ensemble de ces points $\operatorname{rcbd}(C)$.

On remarque que l'intérieur convexe relatif est l'intérieur de l'enveloppe convexe relativement à la dimension de cette enveloppe convexe, mais ce n'est pas l'intérieur de l'enveloppe convexe ni l'intérieur relatif de $C$. Pour démontrer cette différence, prenons $C = \{(-1, 1, 0), (0, 0, 0), (1, 1, 0)\}$. Dans ce cas-ci, l'intérieur de l'enveloppe convexe est vide, car pour tout nombre réel $\delta > 0$, et pour tout point $x \in \operatorname{conv}(C)$, $x + (0, 0, \delta) \notin \operatorname{conv}(C)$. L'intérieur relatif quant à lui est aussi vide, car pour tout point $x \in C$, et pour tout nombre réel $\delta > 0$, il existe toujours un vecteur $\bar{\delta} \in V_C$, $\|\bar{\delta}\| = \delta$, tel que $x + \bar{\delta} \notin C$. La définition de l'intérieur convexe relatif est comme une combinaison de ces deux définitions. Pour cet ensemble en particulier l'intérieur convexe relatif est tous les points $(x, y, z) \in \mathbb{R}^3$ tels que

$$y > |x|;$$
$$y < 1, \text{ et};$$
$$z = 0.$$

**Proposition 2.1.5.** *Soit $C = \{c_1, \ldots, c_n\}$ un ensemble de $n$ points d'un espace euclidien $\mathbb{R}^d$. Si le point $x \in \mathbb{R}^d$ est dans l'intérieur convexe relatif de $C$, alors pour tout vecteur $v \neq 0_d$ de $V_C$, l'espace vectoriel induit par l'enveloppe affine de $C$ (Remarque 2.1.2), il existe un point $c_v \in C$ tel que*

$$(c_v - x)^{\mathrm{T}} v > 0.$$

*Démonstration.* Si $C$ est un singleton, alors la proposition est vraie par vacuité, car $V_C = \{0\}$. On assume à partir de maintenant que $|C| \geq 2$. On montre par contradiction l'implication. On suppose alors qu'il existe un vecteur $v \neq 0_d$ de $V_C$ tel que pour tout point $c \in C$,

$$(c - x)^{\mathrm{T}} v \leq 0.$$

Puisque $x \in \operatorname{rcint}(C)$, il existe un nombre réel $\delta > 0$ tel que

$$B(x, \delta) \cap \operatorname{aff}(C) \subseteq \operatorname{conv}(C).$$

En posant

$$r = \frac{\delta}{\|v\|} v,$$

on a que $x + r \in \operatorname{conv}(C)$, et donc il existe une combinaison convexe $\lambda_1, \ldots, \lambda_n \geq 0$ telle que

$$\begin{aligned} x + r &= \sum_{i=1}^{n} \lambda_i c_i \\ \iff r &= \sum_{i=1}^{n} \lambda_i (c_i - x). \end{aligned}$$

Mais alors

$$\begin{aligned} r^{\mathrm{T}} v &= \sum_{i=1}^{n} \lambda_i (c_i - x)^{\mathrm{T}} v \\ \iff \frac{\delta}{\|v\|} v^{\mathrm{T}} v &= \sum_{i=1}^{n} \underbrace{\lambda_i}_{\geq 0} \underbrace{(c_i - x)^{\mathrm{T}} v}_{\leq 0} \leq 0, \end{aligned}$$

implique que $v = 0_d$, ce qui contredit notre hypothèse. □

**Proposition 2.1.6.** *Soit $C = \{c_1, \ldots, c_n\}$ un ensemble de $n$ points d'un espace euclidien $\mathbb{R}^d$. Si le point $x \in \operatorname{aff}(C)$ s'exprime par une combinaison convexe strictement positive des éléments de $C$, alors le point $x$ est un élément de l'intérieur convexe relatif de $C$.*

*Démonstration.* On montre l'implication par contradiction. On suppose alors qu'il existe des nombres réels $\lambda_1, \ldots, \lambda_n > 0$, avec $\sum_{i=1}^{n} \lambda_i = 1$, tel que

$$x = \sum_{i=1}^{n} \lambda_i c_i,$$

mais que $x \notin \operatorname{rcint}(C)$. Il doit cependant se trouver sur la frontière convexe relative de $C$, car il respecte la Définition 2.1.3. Il existe donc un point $y \in \operatorname{aff}(C)$, $y \neq x$, tel que pour tout $\delta \in (0, 1)$, $\delta x + (1 - \delta) y \notin \operatorname{conv}(C)$, sinon $x \in \operatorname{rcint}(C)$. Puisque $y \in \operatorname{aff}(C)$, il existe des nombres réels $\gamma_1, \ldots, \gamma_n$, avec $\sum_{i=1}^{n} \gamma_i = 1$, tels que $y = \sum_{i=1}^{n} \gamma_i c_i$. On note alors

$$\begin{aligned} z(\delta) &= \delta x + (1 - \delta) y \\ &= \delta \sum_{i=1}^{n} \lambda_i c_i + (1 - \delta) \sum_{i=1}^{n} \gamma_i c_i \\ &= \sum_{i=1}^{n} \left[ \delta \lambda_i + (1 - \delta) \gamma_i \right] c_i. \end{aligned}$$

Pour tout $i \in \{1, \ldots, n\}$, on rappelle que $\lambda_i > 0$, mais que $\gamma_i$ prend une valeur réelle quelconque. On remarque aussi qu'il existe toujours un $\delta \in (0,1)$ tel quel

$$\begin{aligned} &\delta\lambda_i + (1-\delta)\gamma_i \geq 0 \\ \iff &\frac{\delta}{(1-\delta)}\lambda_i \geq -\gamma_i, \end{aligned}$$

puisque la fonction $\delta \mapsto \delta/(1-\delta)$ projette sur tout $\mathbb{R}_+$. Prenons un certain $\delta^\star \in (0,1)$ tel que pour tout $i \in \{1, \ldots, n\}$, $\delta^\star\lambda_i + (1-\delta^\star)\gamma_i \geq 0$. De plus, on a

$$\delta^\star \sum_{i=1}^n \lambda_i + (1-\delta_\star) \sum_{i=1}^n \gamma_i = \delta^\star + 1 - \delta^\star = 1.$$

En posant $\theta_i = \delta^\star\lambda_i + (1-\delta^\star)\gamma_i$, on trouve que $z(\delta^\star) = \sum_{i=1}^n \theta_i c_i$ s'exprime comme une combinaison convexe des éléments de $C$, ce qui contredit notre hypothèse que pour tout $\delta \in (0,1)$, $z(\delta) \notin \operatorname{conv}(C)$. Ainsi, nécessairement $x \in \operatorname{rcint}(C)$. □

**Proposition 2.1.7.** *La fonction log-vraisemblance,*

$$\ell(\theta) = \log p_\theta(Y) = s(Y)^{\mathrm{T}}\theta - \log z(\theta), \quad \theta \in \Gamma$$

*est*

*(1) concave sur son domaine, et ;*

*(2) strictement concave sur son domaine si, et seulement si,* $\dim \operatorname{aff}(\mathcal{S}) = d$.

Si on a $\dim \operatorname{aff}(\mathcal{S}) < d$, alors il y a une « surparamétrisation » du modèle. On comprend que l'information d'au moins une statistique exhaustive est comprise par les autres statistiques. Ce problème de multicolinéarité des statistiques exhaustives se retrouve aussi en régression linéaire où l'estimateur des moindres carrées n'est plus unique lorsque les variables explicatives souffrent de multicolinéaritées.

*Démonstration.* On s'inspire de la preuve de Bayly et coll. [**3**] qui est simple et claire. Par la forme de la log-vraisemblance, si on montre que $\log z(\theta)$ est convexe, alors on aura montré que la fonction $\ell(\theta)$ est concave et strictement concave si $\log z(\theta)$ est strictement convexe.

On commence par montrer le premier point, c'est-à-dire que la log-vraisemblance $\ell(\theta)$ est concave pour tout $\theta \in \Gamma$. Avant de continuer, on rappelle l'inégalité de Hölder pour le cas discret. Pour une paire de nombres $p, q \in [1, \infty]$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, alors pour tout $x, y \in \mathbb{R}^d$,

$$\sum_{i=1}^d |x_i y_i| \leq \left(\sum_{i=1}^d |x_i|^p\right)^{1/p} \left(\sum_{i=1}^d |y_i|^q\right)^{1/q},$$

avec égalité si et seulement s'il existe un nombre réel $c$ tel que $|x_i|^p = c|y_i|^q$ pour tout $i = 1, \ldots, d$. Pour toute paire $\theta_1, \theta_2 \in \Gamma$ et pour tout $\lambda \in [0, 1]$, on pose les fonctions

$$f : \mathcal{Y}_n \to \mathbb{R}_+ : y \mapsto \exp(s(y)^{\mathrm{T}}\theta_1\lambda),$$

et

$$g : \mathcal{Y}_n \to \mathbb{R}_+ : y \mapsto \exp(s(y)^{\mathrm{T}}\theta_2(1-\lambda)).$$

En utilisant l'inégalité de Hölder avec $p = \frac{1}{\lambda}$ et $q = \frac{1}{1-\lambda}$, on obtient

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}_n} f(y)g(y) \leq \left(\sum_{y\in\mathcal{Y}_n} f(y)^p\right)^{1/p} \left(\sum_{y\in\mathcal{Y}_n} g(y)^q\right)^{1/q} \\
\iff &\sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta_1\lambda + s(y)^{\mathrm{T}}\theta_2(1-\lambda)) \\
&\leq \left(\sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta_1\lambda/\lambda)\right)^{\lambda} \left(\sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta_1(1-\lambda)/(1-\lambda))\right)^{(1-\lambda)} \\
&\iff z(\theta_1\lambda + \theta_2(1-\lambda)) \leq z(\theta_1)^{\lambda} z(\theta_2)^{(1-\lambda)} \\
&\iff \log(z(\theta_1\lambda + \theta_2(1-\lambda))) \leq \lambda\log(z(\theta_1)) + (1-\lambda)\log(z(\theta_2)).
\end{aligned}$$

On a donc montré que $\log z(\theta)$ est convexe, et donc que $\ell(\theta)$ est concave.

On continue en montrant le deuxième point, c'est-à-dire que la log-vraisemblance est strictement concave si, et seulement si, $\dim \operatorname{aff}(\mathcal{S}) = d$. Pour établir ce résultat, on va plutôt montrer qu'il existe un nombre réel $c$ tel que pour tout graphe $y \in \mathcal{Y}$, $f(y)^p = cg(y)^q$ si, et seulement si, $\dim \operatorname{aff}(\mathcal{S}) = d$. On a qu'un tel nombre réel $c$ existe si, et seulement si, pour toute paire de graphes $y_1, y_2 \in \mathcal{Y}$,

$$\frac{f(y_1)^p}{g(y_1)^q} = \exp(s(y_1)^{\mathrm{T}}(\theta_1 - \theta_2)) = \exp(s(y_2)^{\mathrm{T}}(\theta_1 - \theta_2)) = \frac{f(y_2)^p}{g(y_2)^q}.$$

On sait que $s(y_1) - s(y_2) = s(y_1) - u - s(y_2) + u \in V_{\mathcal{S}}$ pour un certain élément pivot $u \in \operatorname{aff}(\mathcal{S})$. Si $\dim V_{\mathcal{S}} = \dim \operatorname{aff}(\mathcal{S}) < d$, il existe alors une paire de points distincts $\theta_1, \theta_2 \in \Gamma$ qui sont telle que $\theta_1 - \theta_2 \in V_{\mathcal{S}}^{\perp}$. Pour cette paire de points, on obtient $(s(y_1) - s(y_2))^{\mathrm{T}}(\theta_1 - \theta_2) = 0$, ce qui montre que sous cette condition, la log-vraisemblance n'est pas strictement concave. Maintenant si $\dim V_{\mathcal{S}} = \dim \operatorname{aff}(\mathcal{S}) = d$, alors pour toute paire de points distincts $\theta_1, \theta_2 \in \Gamma$, par construction de l'espace $V_{\mathcal{S}}$, il existe toujours une paire de graphes $y_1, y_2 \in \mathcal{Y}$ telle que $|(s(y_1) - s(y_2))^{\mathrm{T}}(\theta_1 - \theta_2)| > 0$, ce qui montre que dans ce cas, la fonction de log-vraisemblance est strictement concave. □

**Remarque 2.1.8.** On se place dans le cas où $s(Y) \in \operatorname{rcint}(\mathcal{S})$. Si $\dim \operatorname{aff}(\mathcal{S}) < d$, alors il peut exister une infinité de $\theta \in \Gamma$ qui maximise la vraisemblance. On peut déduire par la

Proposition 2.1.7 quelles sont ces autres solutions. Soit un point $\theta \in \mathbb{R}^d$, et soit $v \in V_{\mathcal{S}}^{\perp}$ et $u \in \operatorname{aff}(\mathcal{S})$. Puisque pour tout graphe $y \in \mathcal{Y}_n$, $s(y) - u$ se trouve dans $V_{\mathcal{S}}$, on a

$$\begin{aligned}
\ell(\theta + v) &= s(Y)^{\mathrm{T}}(\theta + v) - \log z(\theta + v) \\
&= (s(Y) - u)^{\mathrm{T}}(\theta + v) + u^{\mathrm{T}}(\theta + v) - \log \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\theta + v)) \\
&= (s(Y) - u)^{\mathrm{T}}(\theta + v) + \log \exp(u^{\mathrm{T}}(\theta + v)) - \log \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\theta + v)) \\
&= (s(Y) - u)^{\mathrm{T}}(\theta + v) - \log \sum_{y \in \mathcal{Y}_n} \exp((s(y) - u)^{\mathrm{T}}(\theta + v)) \\
&= (s(Y) - u)^{\mathrm{T}}\theta - \log \sum_{y \in \mathcal{Y}_n} \exp((s(y) - u)^{\mathrm{T}}\theta) \\
&= (s(Y) - u)^{\mathrm{T}}\theta - \log \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta) + u^{\mathrm{T}}\theta \\
&= \ell(\theta).
\end{aligned}$$

**Proposition 2.1.9.** *L'estimateur de vraisemblance maximale existe si et seulement si $s(Y)$, le vecteur de statistiques exhaustives de $Y$, est dans l'intérieur convexe relatif de $\mathcal{S}$, c'est-à-dire si, et seulement si,*

$$s(Y) \in \operatorname{rcint}(\mathcal{S}).$$

*Démonstration.* On s'inspire de la preuve de Bayly et coll. [**3**]. Si $|\mathcal{S}| = 1$, alors nécessairement $s(Y) \in \operatorname{rcint}(\mathcal{S})$ et tous les points de $\Gamma$ est un MLE par la Proposition 2.1.7. On assume à partir de maintenant que $|\mathcal{S}| > 1$. On en profite pour rappeler qu'on suppose ici que l'espace paramétrique $\Gamma$ est tout l'espace $\mathbb{R}^d$.

On remarque que la fonction de log-vraisemblance $\ell$ est dérivable pour tout point $\theta \in \mathbb{R}^d$. On suppose qu'il existe un $\theta^\star \in \Gamma$ tel que $[\nabla \ell](\theta^\star) = 0$. Par la Proposition 2.1.7, on sait que $\ell(\theta^\star) = \sup_{\theta \in \Gamma} \ell(\theta)$. Le gradient de la log-vraisemblance est

$$\begin{aligned}
\nabla \ell(\theta) &= s(Y) - \frac{1}{z(\theta)} \sum_{y \in \mathcal{Y}_n} s(y) \exp(s(y)^{\mathrm{T}}\theta) \\
&= s(Y) - \sum_{y \in \mathcal{Y}_n} s(y) p_\theta(y).
\end{aligned}$$

et lorsque $\theta = \theta^\star$, on trouve

$$\begin{aligned}
0 &= s(Y) - \sum_{y \in \mathcal{Y}_n} s(y) p_{\theta^\star}(y) \\
\iff s(Y) &= \sum_{y \in \mathcal{Y}_n} s(y) p_{\theta^\star}(y).
\end{aligned}$$

Puisque pour tout $y \in \mathcal{Y}_n$, $p_{\theta^\star}(y) > 0$, on a que $s(Y)$ est dans l'intérieur convexe relatif de $\mathcal{S}$ par la Proposition 2.1.6.

En supposant que $s(Y)$ est dans l'intérieur convexe relatif de $\mathcal{S}$, on montre maintenant par contradiction qu'il existe un $\theta^\star$ tel que le supremum de la log-vraisemblance est atteint. Donc, on suppose que la vraisemblance n'est pas bornée, ainsi pour tout $K \in \mathbb{R}$, il existe un $\theta \in \Gamma$ tel que

$$K < \ell(\theta). \tag{2.1.2}$$

En particulier, il existe un point $\theta$ respectant l'Équation 2.1.2 et qui se trouve seulement dans $V_{\mathcal{S}}$ par la Remarque 2.1.8. Il existe donc une suite $(\theta_n)_{n\in\mathbb{N}}$ dans $V_{\mathcal{S}}$ telle que $\|\theta_n\| \to \infty$ et $\ell(\theta_n) \to \infty$ en rappelant que $\ell$ est continue. On pose la suite des termes normalisés ainsi $\underline{\theta}_n = \frac{\theta_n}{\|\theta_n\|}$, et on remarque que la suite $(\underline{\theta}_n)_{n\in\mathbb{N}}$ est contenue dans l'hypersphère de rayon 1, qui est compacte. Par le théorème de Bolzano-Weierstrass, il existe une sous-suite $\left(\underline{\theta}_{n_k}\right)_{k\in\mathbb{N}}$ qui converge dans l'hypersphère, et notons $\underline{\theta}$ cette limite. Nécessairement, cette limite se trouve aussi dans $V_{\mathcal{S}}$. Pour cette limite, il existe nécessairement un graphe $\tilde{y} \in \mathcal{Y}_n$ et un nombre réel $\delta > 0$, et en particulier un inférieur à 1, tel que $(s(\tilde{y}) - s(Y))^{\mathrm{T}}\underline{\theta} > \delta$, sinon $s(Y)$ ne serait pas dans l'intérieur convexe relatif de $\mathcal{S}$ par la Proposition 2.1.5. Puisqu'il existe un tel graphe $\tilde{y}$ et un tel nombre réel $\delta$, l'ensemble ouvert

$$O = \left\{u \in \mathbb{R}^d : \|u\| = 1, \underline{\theta}^{\mathrm{T}} u > \delta/2\right\}$$

sur l'hypersphère qui contient $\underline{\theta}$ est tel que pour tout vecteur $v \in O$, on a que $(s(\tilde{y}) - s(Y))^{\mathrm{T}} v > \delta/2 > 0$. Puisque $\left(\underline{\theta}_{n_k}\right)_{k\in\mathbb{N}}$ converge vers $\underline{\theta}$, qui est compris dans l'ouvert $O$, alors il existe un nombre naturel $M \in \mathbb{N}$, tel que pour tout $k \geq M$, on a que $\underline{\theta}_{n_k} \in O$. Pour un tel nombre $k \geq M$, on a

$$\begin{aligned}
\ell(\theta_{n_k}) = s(Y)^{\mathrm{T}}\theta_{n_k} - \log \sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}\theta_{n_k}) &\leq s(Y)^{\mathrm{T}}\theta_{n_k} - \log \exp(s(\tilde{y})^{\mathrm{T}}\theta_{n_k}) \\
&= s(Y)^{\mathrm{T}}\theta_{n_k} - s(\tilde{y})^{\mathrm{T}}\theta_{n_k} \\
&= (s(Y) - s(\tilde{y}))^{\mathrm{T}}\theta_{n_k} \\
&= \|\theta_{n_k}\|(s(Y) - s(\tilde{y}))^{\mathrm{T}}\underline{\theta}_{n_k} \\
&\leq -\|\theta_{n_k}\|\frac{\delta}{2},
\end{aligned}$$

mais $\|\theta_{n_k}\| \to \infty$, ce qui contredit que $\ell(\theta_n) \to \infty$. On a montré que si $s(Y)$ est dans l'intérieur convexe relatif de $\mathcal{S}$, et donc que le MLE existe. $\square$

Un intérêt pourrait être porté afin de trouver une technique qui identifie si $s(Y)$ est dans l'intérieur convexe relatif de $\mathcal{S}$, et aussi si $\dim \operatorname{aff}(\mathcal{S}) = d$, et ce indépendamment du choix de statistiques exhaustives.

La cardinalité de $\mathcal{Y}_n$ est de $2^{\frac{n(n-1)}{2}}$, ce qui ne permet pas de passer à travers l'ensemble dès que $n \geq 8$, car le nombre de graphes augmente exponentiellement selon le nombre de sommets. Donc, la constante de normalisation qui nécessite de sommer à travers tous les graphes n'est pas calculable, et ainsi le MLE d'une manière classique est impossible. Plusieurs techniques sont utilisées pour surmonter ce problème, mais aucune ne permet d'obtenir l'estimateur exact. Une première technique est de supposer que chacune des paires de sommets est indépendante étant données les autres paires. Cette technique est expliquée dans la Section 2.2. Une autre technique est d'estimer la vraisemblance par Monte Carlo par chaînes de Markov, *Markov Chain Monte Carlo*, (MCMC), qui est expliquée dans la Section 2.3.

## 2.2. Pseudo-vraisemblance

Dans cette section, pour toute paire de nombres $i, j \in \{1, \ldots, n\}$, on note $Y_{ij}$ la variable aléatoire qui prend ses valeurs dans $\{0, 1\}$ et qui retourne 1 si $Y$ contient l'arête $\{i, j\}$ et 0 sinon. On note aussi $Y_{-ij}$ la variable aléatoire qui prend ces valeurs dans $\{0, 1\}^{\frac{n(n-1)}{2}-1}$ et qui représente toutes les paires du graphe sauf la paire $\{i, j\}$. Finalement, on définit les variables aléatoires $Y_{ij}^+ = (V_n, E(Y) \cup \{i, j\})$, $Y_{ij}^- = (V_n, E(Y) \backslash \{i, j\})$, et on pose $\Delta_{ij}(Y) = s(Y_{ij}^+) - s(Y_{ij}^-)$, le vecteur de changement de statistiques exhaustives lorsqu'on change le statut de l'arête $\{i, j\}$.

**Définition 2.2.1.** On définit la **pseudo-vraisemblance**, comme la fonction de vraisemblance conditionnelle de $Y_{ij}$ selon toutes les autres arêtes, $Y_{-ij}$, c'est-à-dire pour tout $\theta \in \Gamma$,

$$L_{pl}(\theta) = \prod_{1 \leq i < j \leq n} p_\theta(Y_{ij} | Y_{-ij}),$$

avec

$$p_\theta(Y_{ij} | Y_{-ij}) = \mathbb{P}_\theta(Y_{ij} = 1 | Y_{-ij})^{Y_{ij}} \mathbb{P}_\theta(Y_{ij} = 0 | Y_{-ij})^{1-Y_{ij}}.$$

On pose alors l'estimateur de pseudo-vraisemblance maximale, *Maximum PseudoLikelihood Estimator*, (MPLE) comme

$$\hat{\theta}_{\text{MPL}} \in \arg\max_{\theta \in \Gamma} L_{pl}(\theta)$$

L'idée d'utiliser le MPLE comme substitut à l'usuel MLE est proposée par Strauss et Ikeda [**59**]. Contrairement, à la vraisemblance, la pseudo-vraisemblance est facilement calculable. On a pour toute paire de sommets $\{i, j\}$ et pour tout $\theta \in \mathbb{R}^d$,

$$\mathbb{P}_\theta(Y_{ij} = 1 | Y_{-ij}) = \frac{\mathbb{P}_\theta(Y_{ij}^+)}{\mathbb{P}_\theta(Y_{ij}^+) + \mathbb{P}_\theta(Y_{ij}^-)},$$

et donc pour tout $\theta \in \Gamma$,

$$\begin{aligned}
L_{pl}(\theta) &= \prod_{1\leq i<j\leq n} \mathbb{P}_\theta(Y_{ij}=1|Y_{-ij})^{Y_{ij}}\mathbb{P}_\theta(Y_{ij}=0|Y_{-ij})^{1-Y_{ij}} \\
&= \prod_{i\leq i<j\leq n} \left(\frac{\mathbb{P}_\theta(Y_{ij}^+)}{\mathbb{P}_\theta(Y_{ij}^-)}\right)^{Y_{ij}} \frac{\mathbb{P}_\theta(Y_{ij}^-)}{\mathbb{P}_\theta(Y_{ij}^+)+\mathbb{P}_\theta(Y_{ij}^-)} \\
&= \prod_{i\leq i<j\leq n} \left(\frac{\exp(\theta^{\mathrm{T}}s(Y_{ij}^+))}{\exp(\theta^{\mathrm{T}}s(Y_{ij}^-))}\right)^{Y_{ij}} \frac{\exp(\theta^{\mathrm{T}}s(Y_{ij}^-))}{\exp(\theta^{\mathrm{T}}s(Y_{ij}^+))+\exp(\theta^{\mathrm{T}}s(Y_{ij}^-))} \\
&= \prod_{1\leq i<j\leq n} \frac{\exp(\theta^{\mathrm{T}}(s(Y_{ij}^+)-s(Y_{ij}^-)))^{Y_{ij}}}{\exp(\theta^{\mathrm{T}}(s(Y_{ij}^+)-s(Y_{ij}^-)))+1}.
\end{aligned}$$

Strauss et Ikeda ont montré que le MPLE est bien défini, et que trouver cet estimateur revient à trouver le MLE d'un modèle de régression logistique [**59**]. En effet, pour toute paire de sommets $\{i,j\}$ et pour tout $\theta \in \Gamma$, on a

$$\begin{aligned}
\mathrm{logit}(\mathbb{P}_\theta(Y_{ij}=1|Y_{-ij})) &= \log \frac{\mathbb{P}_\theta(Y_{ij}=1|Y_{-ij})}{\mathbb{P}_\theta(Y_{ij}=0|Y_{-ij})} \\
&= \log \frac{\mathbb{P}_\theta(Y_{ij}^+)}{\mathbb{P}_\theta(Y_{ij}^-)} \\
&= \theta^{\mathrm{T}}\Delta_{ij}(Y).
\end{aligned}$$

Ainsi, on peut créer préalablement un jeu de données

$$\left\{(Y_{ij}, \Delta_{ij}(Y)) : \{i,j\} \in V_n^2\right\},$$

où $Y_{ij}$ joue le rôle de variable réponse et $\Delta_{ij}(Y)$, le rôle de variable explicative, et on peut utiliser une technique de régression logistique pour trouver le MPLE. Le modèle de régression logistique est bien connu. Albert et Anderson nomment une condition nécessaire et suffisante à l'existence du MLE pour ce type de modèle [**1**].

**Définition 2.2.2.** Pour un échantillon de $n$ points $X_1, \ldots, X_n$ dans $\mathbb{R}^d$ associés à deux classes, disons $\{0,1\}$, tel que pour $i = 1, \ldots, n$, $Y_i = 0$ si $X_i$ est associé à la classe 0 et $Y_i = 1$ sinon. On dit que les données se **chevauchent** si pour tout vecteur $u \in \mathbb{R}^d$, il existe un $i \in \{1, \ldots, n\}$ tel que

$$(2Y_i - 1)u^{\mathrm{T}}X_i < 0. \tag{2.2.1}$$

À priori, il n'est pas aisé de voir ce que l'Équation 2.2.1 signifie réellement. En d'autres termes, on dit que les points d'un jeu de données défini comme à la Définition 2.2.2 se chevauchent s'il n'existe pas d'hyperplan qui sépare parfaitement les points selon leur classe d'assignation. Dans le cas où

$$(2Y_i - 1)u^{\mathrm{T}}X_i > 0,$$

il existe un plan qui sépare parfaitement les points selon leur classe. On dit alors que les points sont séparables. Il faut noter que les points peuvent aussi être quasi-séparables dans le cas où des points de classes différentes se trouve exactement sur le plan qui les sépare. Supposons qu'on a un jeu de données défini comme à la Définition 2.2.2, et supposons qu'il existe un hyperplan $H$ qui sépare parfaitement les points selon leur classe. Posons $u$ un vecteur normal à cet hyperplan $H$. Pour un certain point $i \in \{1, \ldots, n\}$ disons de classe 1 sans perdre de généralité, on a

$$(2Y_i - 1)u^{\mathrm{T}} X_i = u^{\mathrm{T}} X_i.$$

Nécessairement, soit $u^{\mathrm{T}} X_i > 0$, ou pour $\bar{u} = -u$, qui est aussi un vecteur normal de l'hyperplan $H$, $\bar{u}^{\mathrm{T}} X_i > 0$.

**Proposition 2.2.3.** *Pour des données $(X_1, Y_1), \ldots, (X_n, Y_n)$, où pour tout $i = 1, \ldots, n$, $X_i \in \mathbb{R}^d$ et $Y_i \in \{0, 1\}$, le MLE pour un modèle de régression logistique existe si et seulement si les données se chevauchent. (Théorème 3 de* [**1**]*)*

De ce résultat, Handcock a pu noter le prochain corollaire [**23**].

**Corollaire 2.2.4.** *Pour un ERGM, le MPLE existe si et seulement si pour tout $u \in \mathbb{R}^d$, il existe une paire de sommets $(i, j)$ telle que*

$$(2Y_{ij} - 1)u^{\mathrm{T}} \Delta_{ij}(Y) < 0.$$

En utilisant la programmation linéaire, Albert et Anderson [**1**] et Santner et Duffy [**52**] proposent une méthode simple afin de vérifier si les données se chevauchent.

Étant donné l'absence de la fonction de partition, cet estimateur est rapide et facile à trouver, mais il suppose l'indépendance conditionnelle des paires de sommets, connaissant le reste du graphe. Cette condition est trop forte pour les situations où le modèle ERGM est utilisé.

## 2.3. Maximum de vraisemblance par Monte Carlo par chaînes de Markov

On présente d'abord une méthode introduite par Geyer dans sa thèse [**17**] qui permet d'approximer le MLE par des méthodes MCMC. Cette méthode donne suite à de nombreuses publications [**16, 19, 18, 60**] qui détailleront la méthode dans différents cas. On va d'abord décrire sa méthode de manière générale avec les résultats importants [**18**], qu'on utilisera ensuite pour approximer le MLE pour dans le cadre des ERGM.

**Définition 2.3.1.** Pour un espace paramétrique $\Omega$, soit $\{h_\omega : \omega \in \Omega\}$ une famille de fonctions non négatives qui prennent place sur un espace mesurable et qui sont toutes intégrables et non nulles presque partout sous une certaine mesure $\mu$. En notant la fonction de normalisation

$$c(\omega) = \int h_\omega(x) d\mu(x) \in \mathbb{R}^+, \quad \text{pour tout } \omega \in \Omega,$$

on définit la **famille des densités normalisées**

$$\left\{ f_\omega(x) = \frac{h_\omega(x)}{c(\omega)} : \omega \in \Omega \right\}.$$

Soit $\omega^\star \in \Omega$ un certain paramètre et soit $X$ une variable aléatoire de loi $f_{\omega^\star}$. On cherche le MLE sous ce modèle, c'est-à-dire qu'on cherche un $\hat{\omega} \in \Omega$ qui permet d'atteindre $\sup_{\omega \in \Omega} f_\omega(X)$. On suppose aussi que la fonction de normalisation soit difficilement évaluable. Dans ce cas, trouver un tel estimateur devient difficile. Pour un certain paramètre $\omega_0 \in \Omega$ connu, on pose

$$\psi_{\omega_0}(\omega) = \log f_\omega(X) - \log f_{\omega_0}(X), \quad \text{pour tout } \omega \in \Omega, \tag{2.3.1}$$

la différence de deux log-vraisemblances. On note que la seconde fonction ne dépend pas de $\omega$. Ainsi, maximiser $\psi_{\omega_0}(\omega)$ est équivalent à maximiser $f_\omega(X)$ sous $\omega$. En observant que

$$\begin{aligned}
\frac{c(\omega)}{c(\omega_0)} &= \int \frac{h_\omega(x)}{c(\omega_0)} d\mu \\
&= \int \frac{h_\omega(x)}{h_{\omega_0}(x)} \frac{h_{\omega_0}(x)}{c(\omega_0)} d\mu(x) \\
&= \int \frac{h_\omega(x)}{h_{\omega_0}(x)} f_{\omega_0}(x) d\mu(x),
\end{aligned}$$

qui est une espérance prise sous $f_{\omega_0}$, on peut réécrire $\psi_{\omega_0}$ ainsi

$$\begin{aligned}
\psi_{\omega_0}(\omega) &= \log \frac{h_\omega(X)}{h_{\omega_0}(X)} - \log \frac{c(\omega)}{c(\omega_0)} \\
&= \log \frac{h_\omega(X)}{h_{\omega_0}(X)} - \log \int \frac{h_\omega(x)}{h_{\omega_0}(x)} f_{\omega_0}(x) d\mu(x).
\end{aligned}$$

Soit $X_1, X_2, \ldots, X_K$ $K$ variables aléatoires de loi $f_{\omega_0}$ ou d'une chaîne de Markov ergodique avec $f_{\omega_0}$ comme distribution stationnaire indépendante de la variable $X$. On pose l'approximation

$$\psi_{\omega_0}^{(K)}(\omega) = \log \frac{h_\omega(X)}{h_{\omega_0}(X)} - \log \frac{1}{K} \sum_{k=1}^{K} \frac{h_\omega(X_k)}{h_{\omega_0}(X_k)}, \tag{2.3.2}$$

et de celle-ci, on pose l'estimateur de vraisemblance maximale par Monte Carlo par chaînes de Markov, *Markov chains Monte Carlo maximum likelihood estimator*, (MCMCMLE) comme

$$\hat{\omega}_K \in \arg\max_{\omega \in \Omega} \psi_{\omega_0}^{(K)}(\omega) \quad \text{si un maximum existe.} \tag{2.3.3}$$

**Théorème 2.3.2.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov ergodique sur l'espace d'état $S$ avec $\pi$ comme loi stationnaire. Alors, pour toute fonction mesurable $f : S \to R$ telle que $\int |f(x)|\pi(dx) < \infty$,*

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i) \xrightarrow{\text{p.s.}} \int f(x)\pi(dx).$$

*(Théorème 3 de* [**63**]*)*

**Remarque 2.3.3.** On remarque assez facilement que le Théorème 2.3.2 se généralise pour les fonctions à valeurs dans $\mathbb{R}^d$.

Le Théorème 2.3.2 montre que $\psi_{\omega_0}^{(K)}(\omega)$ converge presque sûrement vers $\psi_{\omega_0}(\omega)$, et ce pour tout $\omega \in \Omega$ et tout $\omega_0 \in \Omega$. Avant de présenter le principal résultat de Geyer [**18**], on doit présenter les trois définitions suivantes.

**Définition 2.3.4.** Soit une suite de fonctions $(f_n)_{n\in\mathbb{N}}$ d'un ensemble quelconque $\mathcal{X}$ vers $\mathbb{R}$. On dit que la suite $(x_n)_{n\in\mathbb{N}}$ est un $\varepsilon_n$**-maximum** pour $(f_n)_{n\in\mathbb{N}}$ si pour tout $n \in \mathbb{N}$

$$f_n(x_n) \geq \sup_{x\in\mathcal{X}} f_n(x) - \varepsilon_n$$

pour une certaine suite positive $(\varepsilon_n)_{n\in\mathbb{N}}$ qui converge vers 0.

**Définition 2.3.5.** Soit une suite de fonctions $(f_n)_{n\in\mathbb{N}}$ d'un ensemble quelconque $\mathcal{X}$ vers $\mathbb{R}$. On dit que $(f_n)_{n\in\mathbb{N}}$ **hypoconverge** vers une fonction $f : \mathcal{X} \to \mathbb{R}$, noté $f_n \xrightarrow{h} f$, si :

— pour tout $x \in \mathcal{X}$ et pour toute suite $(x_n)_{n\in\mathbb{N}}$ de $\mathcal{X}$ qui converge vers $x$

$$\limsup_{n\in\mathbb{N}} f_n(x_n) \leq f(x), \text{et ;}$$

— pour tout $x \in \mathcal{X}$ il existe une suite $(x_n)_{n\in\mathbb{N}}$ qui converge vers $x$ tel que

$$\liminf_{n\in\mathbb{N}} f_n(x_n) \geq f(x).$$

**Définition 2.3.6.** Soit une fonction $f : \mathcal{X} \to \mathbb{R}$. La fonction $f$ est **semi-continue inférieurement** (**semi-continue supérieurement**) en $x_0 \in \mathcal{X}$ si pour tout nombre $y \in \mathbb{R}$ tel que $y < f(x_0)$ ($f(x_0) < y$), il existe un voisinage $U$ de $x_0$ tel que pour tout $x \in U$, $y < f(x)$ ($f(x) < y$).

**Théorème 2.3.7.** *Soit une famille de densités normalisées induite par la famille de fonction positive $\{h_\omega : \omega \in \Omega\}$ posé à la Définition 2.3.1 où $\Omega$ est équipée d'une base, et soit un paramètre connu $\omega_0 \in \Omega$. Pour $X_1, X_2, \ldots$ une suite de variables aléatoires provenant d'une chaîne de Markov ergodique avec $f_{\omega_0}$ comme distribution stationnaire et une variable aléatoire observable $X$, on pose $\psi_{\omega_0}$ la fonction définie comme à l'Équation 2.3.1 ainsi que son estimation par la moyenne $\psi_{\omega_0}^{(K)}$ défini comme à l'Équation 2.3.2.*

*Si*

*(1) pour tout $x$, la fonction $\omega \mapsto h_\omega(x)$ est $f_{\omega_0}$ presque sûrement semi-continue inférieurement, et ;*

*(2) pour la variable observée $X_{obs}$, la fonction $\omega \mapsto h_\omega(X_{obs})$ est $f_{\omega_0}$ presque sûrement semi-continue supérieurement ;*

*alors la suite de fonction $\left(\psi_{\omega_0}^{(K)}\right)_{K\in\mathbb{N}}$ hypoconverge $f_{\omega_0}$ presque sûrement vers $\psi_{\omega_0}$. (Théorème 1 de* [**18**]*)*

À priori, le Théorème 2.3.7 nous permet seulement de dire que sous certaines conditions, une suite de fonctions va hypoconverger vers une certaine fonction, c'est-à-dire que pour toute suite convergente, la convergence de cette suite appliquée à la suite de fonctions « se maintient ». Or, une proposition du même ouvrage permet d'entrevoir l'utilité du précédent théorème.

**Proposition 2.3.8.** *Soit une suite de fonction $(f_n)_{n\in\mathbb{N}}$ de $\mathcal{X}$ vers $\mathbb{R}$ qui hypoconverge vers une fonction $f : \mathcal{X} \to \mathbb{R}$ et soit $(x_n)_{n\in\mathbb{N}}$ une suite $\varepsilon_n$-maximum pour la suite $(f_n)_{n\in\mathbb{N}}$ qui converge vers une certaine valeur $x \in \mathcal{X}$. Alors,*

$$\lim_{n\to\infty} f_n(x_n) = \sup_{x'\in\mathcal{X}} f(x') = f(x).$$

*(Proposition 1 de* [**18**]*)*

La Proposition 2.3.8 nous permet de dire que sous certaines conditions, si la suite des $\varepsilon_n$-maximum de $\left(\psi_{\omega_0}^{(K)}\right)_{K\in\mathbb{N}}$ converge, alors elle converge vers le maximum de $\psi_{\omega_0}$. Avant de poser le prochain corollaire, qui complète le résultat, on pose un lemme.

**Lemme 2.3.9.** *Soit une suite $(x_n)_{n\in\mathbb{N}}$ et un point $x$ tous deux contenus dans un ensemble compact $X \subset \mathbb{R}^d$ tel que toute sous-suite convergente de la suite $(x_n)_{n\in\mathbb{N}}$, converge vers le point $x$. Alors, la suite $(x_n)_{n\in\mathbb{N}}$ converge vers le point $x$.*

*Démonstration.* On montre la contraposée de l'implication. On suppose qu'il existe un nombre réel $\varepsilon > 0$, tel que pour tout nombre $N \in \mathbb{N}$, il existe un nombre naturel $n \geq N$ tel que

$$\|x_n - x\| > \varepsilon.$$

À partir de ça, on construit une sous-suite $(x_{n_k})_{k\in\mathbb{N}}$ ainsi

$$n_1 = \inf\{n \in \mathbb{N} : \|x_n - x\| > \varepsilon\}, \text{ et ;}$$
$$n_k = \inf\{n \in \mathbb{N} : n > n_{k-1}, \|x_n - x\| > \varepsilon\}, \quad \text{pour tout } k \geq 2.$$

Cette sous-suite est bien sûr aussi contenue dans l'ensemble compact $X$, et par le théorème de Bolzano-Weierstrass, la sous-suite $(x_{n_k})_{k\in\mathbb{N}}$ contient une sous-suite $(x_{n_{k_j}})_{j\in\mathbb{N}}$ (une sous sous-suite) qui converge. Par construction, elle ne peut pas converger vers le point $x$, et puisqu'une sous sous-suite est aussi une sous-suite, alors il existe une sous-suite de $(x_n)_{n\in\mathbb{N}}$ qui ne converge pas vers le point $x$. □

**Corollaire 2.3.10.** *Soit une famille de densités normalisées induite par la famille de fonctions positives $\{h_\omega : \omega \in \Omega\}$ posée à la Définition 2.3.1 où $\Omega$ est équipée d'une base dénombrable (second-countable topological space), et soit un paramètre connu $\omega_0 \in \Omega$. Pour $X_1, X_2, \ldots$ une suite de variables aléatoires provenant d'une chaîne de Markov ergodique avec $f_{\omega_0}$ comme distribution stationnaire et une variable aléatoire observable $X$, on pose $\psi_{\omega_0}$ la fonction définie comme à l'Équation 2.3.1 ainsi que son estimation par la moyenne $\psi_{\omega_0}^{(K)}$ définie comme à l'Équation 2.3.2. On assume que*

- *la fonction $\omega \mapsto h_\omega(X)$ est semi-continue inférieurement, sauf sous un ensemble de mesure nulle, et semi-continue supérieurement conditionnellement à la variable $X$ ;*
- *la fonction $\psi_{\omega_0}$ possède un unique maximum, et ;*
- *il existe une suite $(\omega_K)_{K\in\mathbb{N}}$ d'$\varepsilon_K$-maximum pour la suite $\left(\psi_{\omega_0}^{(K)}\right)_{K\in\mathbb{N}}$ qui est presque sûrement compris dans un ensemble compact ;*

*alors la suite $(\omega_K)_{K\in\mathbb{N}}$ converge presque sûrement vers le maximum de $\psi_{\omega_0}$. (Corollaire 1 de* [**18**]*)*

Malgré que la « preuve » est présente dans l'article, elle contient certaines zones d'ombre, et donc on présente une preuve plus détaillée.

*Démonstration.* Soit $(Q, \mathcal{F})$ l'espace mesurable induite par le problème, et soit $Q_0 \in \mathcal{F}$ tel que $\mathbb{P}(Q_0) = 1$, et tel que pour toute réalisation $q \in Q_0$, sous cette réalisation, la suite $(\omega_K)_{K\in\mathbb{N}}$ est comprise dans un ensemble compact $C_q \subset \Omega$. Le reste de la preuve est faite sous une telle réalisation $q \in Q_0$.

Par le théorème de Bolzano-Weierstrass, la suite $(\omega_K)_{K\in\mathbb{N}}$ contient une sous-suite convergente $(\omega_{K_j})_{j\in\mathbb{N}}$. Cette sous-suite est $\varepsilon_{K_j}$-maximum pour la suite $\left(\psi_{\omega_0}^{(K_j)}\right)_{j\in\mathbb{N}}$, qui hypoconverge vers $\psi_{\omega_0}$. Par la Proposition 2.3.8, la sous-suite $(\omega_{K_j})_{j\in\mathbb{N}}$ converge vers l'unique maximum $\hat{\omega}$ de $\psi_{\omega_0}$. Ainsi, puisque toute sous-suite convergente de $(\omega_K)_{K\in\mathbb{N}}$ converge vers le point $\hat{\omega}$, par le Lemme 2.3.9, la suite $(\omega_K)_{K\in\mathbb{N}}$ converge vers le point $\hat{\omega}$. □

### 2.3.1. Adaptation pour les modèles de graphes aléatoires de la famille exponentielle

Pour des ERGM, les résultats de Geyer ne peuvent pas simplement s'appliquer. On remarque que

$$\{\exp(s(\cdot)^{\mathrm{T}}\theta) : \theta \in \Gamma\}$$

forme une famille de fonctions non négatives et que la famille de fonctions de masse induite

$$\{p_\theta(\cdot) : \theta \in \Gamma\}$$

est continue selon $\theta$. Soit $\theta_0 \in \Gamma$, un vecteur connu quelconque de l'espace paramétrique. Le problème d'optimisation pour trouver le MLE est équivalent à

$$\begin{aligned}
\arg\max_{\theta\in\Gamma} \ell(\theta) &= \arg\max_{\theta\in\Gamma} \{\ell(\theta) - \ell(\theta_0)\} \\
&= \arg\max_{\theta\in\Gamma} \left\{ s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \frac{z(\theta)}{z(\theta_0)} \right\} \\
&= \arg\max_{\theta\in\Gamma} \left\{ s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \sum_{y\in\mathcal{Y}_n} \frac{\exp(s(y)^{\mathrm{T}}\theta)}{z(\theta_0)} \right\} \\
&= \arg\max_{\theta\in\Gamma} \left\{ s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \sum_{y\in\mathcal{Y}_n} \frac{\exp(s(y)^{\mathrm{T}}(\theta-\theta_0))\exp(s(y)^{\mathrm{T}}\theta_0)}{z(\theta_0)} \right\} \\
&= \arg\max_{\theta\in\Gamma} \left\{ s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\theta-\theta_0)) p_{\theta_0}(y) \right\}.
\end{aligned}$$

Donc, trouver le maximum de vraisemblance revient à trouver le maximum de

$$\phi_{\theta_0}(\theta) = s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \sum_{y\in\mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\theta-\theta_0)) p_{\theta_0}(y) \qquad \text{pour tout } \theta \in \Gamma. \tag{2.3.4}$$

Si on est capable de simuler des graphes provenant de la loi de fonction de masse $p_{\theta_0}$, alors on peut approximer la somme, qui est une espérance, par la moyenne. Pour un échantillon de $K$ graphes, $Y_1, \ldots, Y_K$, provenant de cette loi, on peut approximer la fonction objective par

$$\phi_{\theta_0}^{(K)}(\theta) = s(Y)^{\mathrm{T}}(\theta-\theta_0) - \log \frac{1}{K} \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta-\theta_0)) \quad \text{pour tout } \theta \in \Gamma. \tag{2.3.5}$$

Comme il est vu au Chapitre 1, les graphes aléatoires provenant de ERGM sont simulés par chaîne de Markov ergodique avec la loi stationnaire voulue. La Remarque 1.2.2 et le Théorème 2.3.2 nous permet de conclure que la chaîne de Markov issue de l'algorithme *tie-no-tie* admet une loi des grands nombres pour toute fonction intégrable, ce qui implique que pour tout $\theta \in \Gamma$ et pour tout $\theta_0 \in \Gamma$, $\phi_{\omega_0}^{(K)}(\theta)$ converge presque sûrement vers $\phi_{\omega_0}(\theta)$. On observe en

fait que les fonctions des Équations 2.3.4 et 2.3.5 sont les fonctions correspondantes aux fonctions des Équations 2.3.1 et 2.3.2 dans le cadre des ERGM. Contrairement au cas de Geyer, la fonction $\phi_{\theta_0}^{(K)}$ de l'Équation 2.3.5 n'admet pas toujours un maximum, et lorsqu'il existe il peut ne pas être unique. Similairement aux Propositions 2.1.7 et 2.1.9, on obtient les Propositions 2.3.12 et 2.3.11. Pour un échantillon de $K$ graphes provenant de la loi $p_{\theta_0}$ avec le paramètre connu $\theta_0 \in \mathbb{R}^d$, on pose d'abord

$$\hat{\mathcal{S}}_{\theta_0}^{(K)} = \{s(Y_1), \ldots, s(Y_K)\}. \tag{2.3.6}$$

**Proposition 2.3.11.** *Le maximum de la fonction $\phi_{\omega_0}^{(K)}$ (Équation 2.3.5) existe si et seulement si $s(Y)$ est dans l'intérieur convexe relatif de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$.*

*Démonstration.* La preuve est exactement la même que celle de la Proposition 2.1.9, mais en remplaçant $\mathcal{S}$ par $\hat{\mathcal{S}}_{\theta_0}^{(K)}$. □

La Proposition 2.3.11 nous donne une condition nécessaire et suffisante pour s'assurer de l'existence d'un maximum pour la fonction $\phi_{\omega_0}^{(K)}$. Mais, le Corollaire 2.3.10 nécessite aussi que chaque maximum soit unique pour assurer la convergence.

**Proposition 2.3.12.** *La fonction de l'Équation 2.3.5 est*

*(1) concave sur son domaine, et ;*

*(2) strictement concave sur son domaine si et seulement si* $\dim \operatorname{aff}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right) = d$.

*Démonstration.* Malgré que la preuve est la même que celle de la Proposition 2.1.7, mais en altérant certains aspects, on préfère l'écrire par souci de transparence. On pose d'abord un échantillon de $K$ graphes $Y_1, \ldots, Y_K$, et un certain paramètre connu $\theta_0 \in \Gamma$. La fonction de l'Équation 2.3.5 est

$$\begin{aligned}\phi_{\theta_0}^{(K)}(\theta) &= s(Y)^{\mathrm{T}}(\theta - \theta_0) - \log \frac{1}{K} \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0)) \\ &= s(Y)^{\mathrm{T}}(\theta - \theta_0) - \log \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0)) + \log K.\end{aligned}$$

Elle est (strictement) concave si et seulement si

$$\log \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0))$$

est (strictement) convexe. On commence par montrer la concavité. Pour $\theta_1, \theta_2 \in \mathbb{R}^d$ et $\lambda \in [0, 1]$, on pose les fonctions

$$f : \mathcal{Y}_n \to \mathbb{R}_+ : y \mapsto \exp(s(y)^{\mathrm{T}}(\theta_1 - \theta_0)\lambda),$$

et

$$g : \mathcal{Y}_n \to \mathbb{R}_+ : y \mapsto \exp(s(y)^{\mathrm{T}}(\theta_2 - \theta_0)(1-\lambda)).$$

En utilisant l'inégalité de Hölder avec $p = \frac{1}{\lambda}$ et $q = \frac{1}{1-\lambda}$, on obtient

$$\begin{aligned}
\sum_{i=1}^{K} f(Y_i)g(Y_i) &\leq \left(\sum_{i=1}^{K} f(Y_i)^p\right)^{1/p} \left(\sum_{i=1}^{K} g(Y_i)^q\right)^{1/q} \\
\iff \sum_{i=1}^{K} &\exp(s(Y_i)^{\mathrm{T}}(\lambda\theta_1 + (1-\lambda)\theta_2 - \theta_0)) \\
&\leq \left(\sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta_1 - \theta_0))\right)^{\lambda} \left(\sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta_2 - \theta_0))\right)^{1-\lambda} \\
\iff \log \sum_{i=1}^{K} &\exp(s(Y_i)^{\mathrm{T}}(\lambda\theta_1 + (1-\lambda)\theta_2 - \theta_0)) \\
&\leq \lambda \log \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta_1 - \theta_0)) + (1-\lambda) \log \sum_{i=1}^{K} \exp(s(Y_i)^{\mathrm{T}}(\theta_2 - \theta_0)).
\end{aligned}$$

La fonction de l'Équation 2.3.4 est donc concave pour tout $\theta \in \mathbb{R}^d$.

On continue en montrant la stricte concavité si, et seulement si, $\dim \operatorname{aff}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right) = d$. On suppose que $\dim \operatorname{aff}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right) < d$. Il existe alors une paire de points distincts $\theta_1, \theta_2 \in \mathbb{R}^d$ tels que $\theta_1 - \theta_2 \in V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}}^{\perp}$. Pour toute paire de graphes $Y_1', Y_2' \in \{Y_1, \ldots, Y_K\}$, on a que $s(Y_1') - s(Y_2') \in V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}}$, et donc

$$\begin{aligned}
(s(Y_1') - s(Y_2'))^{\mathrm{T}}(\theta_1 - \theta_2) &= 0 \\
\iff s(Y_1')^{\mathrm{T}}(\theta_1 - \theta_2) &= s(Y_2')^{\mathrm{T}}(\theta_1 - \theta_2).
\end{aligned}$$

Ainsi, on a

$$\frac{f(Y_1')^p}{g(Y_1')^q} = \exp(s(Y_1')^{\mathrm{T}}\theta_1 - s(Y_1')^{\mathrm{T}}\theta_2) = \exp(s(Y_2')^{\mathrm{T}}\theta_1 - s(Y_2')^{\mathrm{T}}\theta_2) = \frac{f(Y_2')^p}{g(Y_2')^q},$$

ce qui implique qu'il existe un nombre réel $c$ tel que $f(Y') = cg(Y')$ pour tout $Y' \in \{Y_1, \ldots, Y_K\}$.

On suppose maintenant que $\dim \operatorname{aff}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right) = d$. Pour toute paire de points distincts $\theta_1, \theta_2 \in \Gamma$, on a que $\theta_1 - \theta_2 \in V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}}$. Par construction de l'espace $V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}}$, il existe une paire de graphes $Y_1', Y_2' \in \{Y_1, \ldots, Y_n\}$ tels que $|(s(Y_1') - s(Y_2')^{\mathrm{T}}(\theta_1 - \theta_2))| > 0$. Ainsi, dans ce cas, les fonctions $f$ et $g$ sont linéairement indépendantes, ce qui implique que l'égalité de Hölder n'est jamais rencontrée. En d'autres mots, la fonction de l'Équation 2.3.5 est strictement concave. □

Il se peut que le MCMCMLE n'existe pas comme il a été montré dans les Propositions 2.3.11, et ce même si le MLE existe. Mais encore, la Proposition 2.3.12 montre que, lorsqu'il existe, il peut ne pas être unique. Cette difficulté peut être résolue en se concentrant sur les maximums de plus petites normes, car il existe un unique maximum de plus petite norme. On se replace dans le contexte de la Remarque 2.1.8. On est dans le cas où le MLE existe, c'est-à-dire que $s(Y) \in \operatorname{rcint}(\mathcal{S})$, mais aussi où le MLE n'est pas unique, c'est-à-dire que le cas où $\dim \operatorname{aff}(\mathcal{S})$. On a alors vu que si $\theta^\star$ est un MLE possible, alors pour tout vecteur $u \in V_{\mathcal{S}}^\perp$, $\theta^\star + v$ est aussi un MLE. Notons par $\theta^\star_{V_{\mathcal{S}}}$ la projection orthogonale de $\theta^\star$ sur l'espace vectoriel $V_{\mathcal{S}}$, et par $\theta^\star_{V_{\mathcal{S}}^\perp}$ la projection orthogonale de $\theta^\star$ sur $V_{\mathcal{S}}^\perp$. On remarque alors

$$\begin{aligned}\|\theta^\star + v\| &= \|\theta^\star_{V_{\mathcal{S}}} + \theta^\star_{V_{\mathcal{S}}^\perp} + v\| \\ &= \|\theta^\star_{V_{\mathcal{S}}}\| + \|\theta^\star_{V_{\mathcal{S}}^\perp} + v\|.\end{aligned}$$

Le MLE avec la plus petite norme est celui qui se trouve uniquement dans $V_{\mathcal{S}}$. On doit alors ajuster la définition du MCMCMLE pour les ERGM. Le MCMCMLE est alors

$$\hat{\theta}_K \in \begin{cases} \theta_0 & \text{si } s(Y) \notin \operatorname{rcint}\left(\hat{\mathcal{S}}^{(K)}_{\theta_0}\right); \\ \arg\max_{\theta \in V_{\hat{\mathcal{S}}^{(K)}_{\theta_0}}} \phi^{(K)}_{\omega_0}(\theta) & \text{sinon,} \end{cases} \tag{2.3.7}$$

où $V_{\hat{\mathcal{S}}^{(K)}_{\theta_0}}$ est l'espace vectorielle induite par l'enveloppe affine de $\hat{\mathcal{S}}^{(K)}_{\theta_0}$. Bien sûr que choisir $\theta_0$ lorsque $\phi^{(K)}_{\omega_0}$ n'admet pas de maximum est arbitraire. Cette valeur est aussi bonne que $\theta_0$ est près du MLE.

**Lemme 2.3.13.** *Soit $S$ un espace fini d'états, et soit $X_1, X_2, \ldots$ une chaîne de Markov à temps discret irréductible. Pour $n \in \mathbb{Z}_+$, on note $S_n = \{X_1, \ldots, X_n\}$ les $n$ premiers états visités. Alors,*

$$\mathbb{P}\left(\liminf_{n\to\infty}\{S_n = S\}\right) = 1.$$

*Démonstration.* Notons $E_n = \{S_n = S\}$ l'événement où tous les états ont été visités par les $n$ premières variables de la chaîne. On remarque que si les $n$ premières variables ont visité tous les états, alors il en est de même pour les $n+1$ premières variables, $E_n \subset E_{n+1}$. On a que

$$\left(\liminf_{n\in\mathbb{N}} E_n\right)^c = \limsup_{n\in\mathbb{N}} E_n^c = \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} E_n^c.$$

Puisque la suite $(E_n)_{n\in\mathbb{N}}$ est croissante, alors la suite $(E_n^c)_{n\in\mathbb{N}}$ est décroissante. Ainsi, pour tout nombre entier positif $m$,

$$\bigcup_{n\geq m} E_n^c = E_m^c,$$

et

$$\bigcap_{m=1}^{\infty} E_m^c = \lim_{m\to\infty} E_m^c.$$

On a les inclusions suivantes

$$\begin{aligned}\lim_{m\to\infty} E_m^C &= \{\exists j \in S : X_1 \neq j, X_2 \neq j, \dots\} \\ &\subset \bigcup_{j\in S} \{X_1 \neq j, X_2 \neq j, \dots\} \\ &= \bigcup_{j\in S} \bigcup_{i\in S\setminus\{j\}} \{X_1 = i, X_2 \neq j, X_3 \neq j, \dots\},\end{aligned}$$

et puisque la chaîne est irréductible pour toute paire d'états $i, j \in S$

$$\mathbb{P}(X_2 \neq j, X_3 \neq j, \dots | X_1 = i) = 0,$$

alors

$$\begin{aligned}\mathbb{P}\left(\lim_{m\to\infty} E_m^c\right) &\leq \sum_{j\in S} \sum_{i\in S\setminus\{j\}} \mathbb{P}\left(X_1 = i, X_2 \neq j, X_3 \neq j, \dots\right) \\ &= \sum_{j\in S} \sum_{i\in S\setminus\{j\}} \mathbb{P}\left(X_2 \neq j, X_3 \neq j, \dots | X_1 = i\right) \mathbb{P}(X_1 = i) \\ &\leq 0.\end{aligned}$$

Donc, $\mathbb{P}(\limsup_{n\in\mathbb{N}} E_n^c) = 0$ et $\mathbb{P}(\liminf_{n\in\mathbb{N}} E_n) = 1$. □

Le Lemme 2.3.13 dit simplement que pour une chaîne de Markov irréductible, il existe presque sûrement un temps $T$ tel qu'à partir de ce temps, tous les états auront été visités au moins une fois. Sous la condition d'existence du MLE, c'est-à-dire que $s(Y) \in \operatorname{rcint}(\mathcal{S})$, on a l'inclusion suivante

$$\left\{\hat{\mathcal{S}}_{\theta_0}^{(K)} = \mathcal{S}\right\} \subset \left\{s(Y) \in \operatorname{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)\right\}, \tag{2.3.8}$$

ce qui implique qu'il existe presque sûrement un moment $T \in \mathbb{N}$ tel que pour tout nombre naturel $K \geq T$ la fonction $\phi_{\theta_0}^{(K)}$ de l'Équation 2.3.5 possède un maximum. Ainsi, notre estimateur est celui de Geyer pour un $K$ suffisamment grand.

Pour une chaîne de Markov $\{X_t\}_{t\in\mathbb{Z}_+}$ sur l'espace $S$, on pose le critère d'arrêt de couverture

$$\tau_{\text{cov}} = \inf\{t \in \mathbb{Z}_+ : \{X_0, \dots, X_t\} = S\},$$

et le temps de couverture

$$t_{\text{cov}} = \max_{x\in S} \mathbb{E}[\tau_{\text{cov}} | X_0 = x].$$

On pose aussi le temps d'atteinte, *hitting time*,

$$t_{\text{hit}} = \max_{x,y\in S} \mathbb{E}[\tau_y | X_0 = x],$$

où

$$\tau_y = \inf\{t \in \mathbb{Z}_+ : X_t = y\}.$$

**Proposition 2.3.14.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov irréductible sur l'espace d'état fini $S$. Pour $|S| = n > 1$ états, on a l'inégalité suivante*

$$t_{\text{hit}} \leq t_{\text{cov}} \leq t_{\text{hit}} \sum_{i=1}^{n-1} \frac{1}{i}.$$

*(Commentaire et Théorème 11.2 de* [**36**]*)*

La Proposition 2.3.14 donne une idée en moyenne sur l'ordre d'atteinte du moment où les maximums des fonctions $\left(\phi_{\theta_0}^{(K)}\right)$ existeront toujours. Malheureusement, le temps $t_{\text{hit}}$ est difficile à calculer.

### 2.3.2. Convergence de l'estimateur

Il y a quatre conditions à respecter afin d'appliquer le Théorème 2.3.7, la Proposition 2.3.8 et le Corollaire 2.3.10 pour notre estimateur :

— l'espace paramétrique doit contenir une **base dénombrable** ;
— la fonction $\theta \mapsto \exp(s(Y)^{\mathrm{T}}\theta)$ doit être **semi-continue** à certains points ;
— l'**existence** presque sûre d'un **unique** maximum pour la fonction $\phi_{\theta_0}^{(K)}$, et ;
— l'existence presque sûre d'un **ensemble compact** $C \subset \Gamma$ qui contient la suite des maximums $\hat{\theta}_K$ pour la suite de fonction $\left(\phi_{\omega_0}^{(K)}\right)_{K\in\mathbb{N}}$.

L'espace paramétrique $\Gamma \subset \mathbb{R}^d$ (avec la topologie standard) est muni d'une base dénombrable induite par l'espace $\mathbb{R}^d$. On rappelle que pour un espace métrique, l'espace est muni d'une base dénombrable si, et seulement si, l'espace est séparable. La fonction $\theta \mapsto \exp(s(Y)^{\mathrm{T}}\theta)$ est continue pour tout $\theta \in \Gamma$. Pour un $K$ suffisamment grand, il existe un unique maximum pour $\phi_{\omega_0}^{(K)}$. Finalement, la condition la plus contraignante est celle de la compacité. Malgré que cette condition ne soit que rarement respectée, cette technique est celle par défaut pour approximer le MLE.

### 2.3.3. Comportement asymptotique du maximum de vraisemblance par Monte Carlo par chaînes de Markov

Dans ses travaux, Geyer s'est aussi intéressé à la loi asymptotique du MCMCMLE. Il se trouve, comme le Théorème central limite, que la loi asymptotique du MCMCMLE est aussi une normale [**18**]. Pour énoncer son théorème de manière générale, on se replace sous la notation de la Section 2.3. On a un espace paramétrique $\Omega$, qui « engendre » une famille

de densités normalisées $\{f_\omega : \omega \in \Omega\}$, et pour un certain paramètre $\omega^\star$ soit une variable aléatoire $X$ de loi $f_{\omega^\star}$. On suppose ici que $\Omega \subset \mathbb{R}^d$.

**Théorème 2.3.15.** *Soit $\hat{\omega}_K$ un MCMCMLE et soit $\hat{\omega}$ un MLE. Sous les conditions suivantes :*

*(1) Le MLE $\hat{\omega}$ est unique et il existe un voisinage $U$ de $\hat{\omega}$ qui est contenu dans $\Omega$ ;*

*(2) Le MCMCMLE $\hat{\omega}_K$ converge en probabilité vers le MLE $\hat{\omega}$ ;*

*(3) La fonction $\omega \mapsto h_\omega(X)$ est doublement dérivable et ses dérivées sont intégrables sous $\mu$ ;*

*(4) $\sqrt{K}\left[\nabla\psi_{\omega_0}^{(K)}\right](\hat{\omega}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, A)$ pour une certaine matrice de covariance $A$ ;*

*(5) $B = -[\nabla^2\psi](\hat{\omega})$ est une matrice définie positive, et ;*

*(6) toutes les dérivées partielles triples de $\psi_{\omega_0}^{(K)}$ sont uniformément bornées en probabilité dans un voisinage du MLE $\hat{\omega}$ ;*

*alors*

$$\sqrt{K}\,(\hat{\omega}_K - \hat{\omega}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, B^{-1}AB^{-1})$$

*(Théorème 3 de* [**18**]*)*

Dans le cas des ERGM, la Proposition 2.1.7 et la Proposition 2.1.9 donnent les conditions nécessaires et suffisantes à l'existence et l'unicité du MLE, ce qui satisfait la Condition 1 et la Condition 5. En assumant la condition de compacité nommée dans la Section 2.3, le Corollaire 2.3.10 implique la Condition 2. La Condition 3 est remplie par la mesure finie utilisée des ERGM et par la dérivabilitée de la fonction exponentielle. De plus, par cette dernière fonction, la Condition 6 est aussi respectée. Pour vérifier la Condition 4, on doit d'abord poser quelques résultats.

**Théorème 2.3.16.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov uniformément ergodique sur l'espace d'état $S$ avec $\pi$ comme loi stationnaire. Alors, pour toute fonction mesurable $f : S \to \mathbb{R}$ telle que $\int f^2(x)\pi(dx) < \infty$, il existe un nombre réel positif $\sigma_f^2$ où*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \int f(x)\pi(dx)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_f^2).$$

*(Théorème 5 de* [**63**]*)*

**Corollaire 2.3.17.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov uniformément ergodique sur l'espace d'état $S$ avec $\pi$ comme loi stationnaire. Alors, pour toute fonction mesurable $f : S \to \mathbb{R}^d$ telle que pour toute paire d'indices $1 \le i, j \le d$, $\int f_i(x)f_j(x)\pi(dx) < \infty$, il existe une matrice*

*de covariance $\Sigma_f$ où*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \int f(x)\pi(dx)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma_f).$$

*Démonstration.* Pour tout vecteur $u \in \mathbb{R}^d$, la fonction $g_u : S \to \mathbb{R} : x \mapsto u^{\mathrm{T}} f(x)$ respecte les hypothèses du Théorème 2.3.16. En effet, on a

$$\begin{aligned}\int g_u^2(x)\pi(dx) &= \int u^{\mathrm{T}} f(x) f^{\mathrm{T}}(x) u \pi(dx) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} u_i \left[\int f_i(x) f_j(x)\pi(dx)\right] u_j < \infty.\end{aligned}$$

Par ce théorème, on a le résultat suivant

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} g_u(X_i) - \int g_u(x)\pi(dx)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{g_u}^2). \tag{2.3.9}$$

Soit une suite de variables aléatoires $(U_n)_{n\in\mathbb{N}}$ avec les fonctions caractéristiques $(\varphi_{U_n})_{n\in\mathbb{N}}$ et soit une variable aléatoire $V$ de fonction caractéristique $\varphi_V$, tous dans $\mathbb{R}^d$. La suite $(U_n)_{n\in\mathbb{N}}$ converge en loi vers la variable $V$ si, et seulement si, pour tout $t \in \mathbb{R}^d$, la suite $(\varphi_{U_n}(t))_{n\in\mathbb{N}}$ converge vers $\varphi_V(t)$. On note

$$W_N = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \int f(x)\pi(dx)\right), \quad \text{et ;}$$

$W$ une variable aléatoire de loi $\mathcal{N}_d(0, \Sigma_f)$. Pour tout vecteur $t \in \mathbb{R}^d$, on a

$$\mathbb{E}\left[e^{it^{\mathrm{T}}W_n}\right] = \mathbb{E}\left[e^{i(t^{\mathrm{T}}W_N)}\right] \to \mathbb{E}\left[e^{i(t^{\mathrm{T}}W)}\right] = \mathbb{E}\left[e^{it^{\mathrm{T}}W}\right],$$

par le résultat de l'Équation 2.3.9, ce qui complète la preuve. □

**Corollaire 2.3.18.** *Soit $\{X_t\}_{t\in\mathbb{Z}_+}$ une chaîne de Markov uniformément ergodique sur l'espace euclidien $\mathbb{R}^d$ avec $\pi$ comme loi stationnaire qui est telle que pour toute paire d'indices $1 \le i, j \le d$, $\int |x_i x_j|\pi(dx) < \infty$, et soit $g : \mathbb{R}^d \to \mathbb{R}^p$ une fonction différentiable sur tout son domaine. Alors, pour une certaine variance $A \in \mathbb{R}^{p\times p}$, on obtient*

$$\sqrt{N}\left(g\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) - g\left(\int x\pi(dx)\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0_p, A).$$

*Démonstration.* Pour simplifier l'écriture, on note $1/N\sum_{i=1}^{N} X_i = \bar{X}_N$ et $\mathbb{E}[X] = \int x\pi(dx)$. Par le théorème de Taylor, il existe un point $\xi$ entre $\bar{X}_N$ et $\mathbb{E}[X]$, c'est-à-dire que pour tout $1 \le i \le d$,

$$\min\left((\bar{X}_N)_i, (\mathbb{E}[X])_i\right) \le \xi_i \le \max\left((\bar{X}_N)_i, (\mathbb{E}[X])_i\right),$$

on a

$$g\left(\bar{X}_N\right) = g\left(\mathbb{E}[X]\right) + [\nabla g]\left(\xi\right)^{\mathrm{T}}\left(\bar{X}_N - \mathbb{E}[X]\right),$$

où $\nabla g$ est la matrice jacobienne de $g$. Par le Théorème 2.3.2, on trouve que $\bar{X}_N \to \mathbb{E}[X]$ presque sûrement, ce qui implique que $\xi \to \mathbb{E}[X]$ presque sûrement, et par le Corollaire 2.3.17, on a

$$\sqrt{N}\left(\bar{X}_N - \mathbb{E}[X]\right) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0_d, \Sigma)$$

pour une certaine matrice de covariance $\Sigma$. Donc, par le théorème de Slutsky, on obtient le résultat

$$\sqrt{N}\left(g\left(\bar{X}_N\right) - g\left(\mathbb{E}[X]\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, A),$$

où plus particulièrement $A = [\nabla g](\mathbb{E}[X])^{\mathrm{T}}\Sigma[\nabla g](\mathbb{E}[X])$. □

La prochaine proposition montre que la Condition 4 est respectée.

**Proposition 2.3.19.** *Soit $\nabla\phi_{\theta_0}^{(K)}$, le gradient de l'Équation 2.3.5. On assume que le MLE $\hat{\theta} \in \mathbb{R}^d$ existe. Alors,*

$$\sqrt{K}\left[\nabla\phi_{\theta_0}^{(K)}\right](\hat{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, A),$$

*où $A$ est une certaine matrice de covariance.*

*Démonstration.* Dans le Remarque 1.2.2, on note que la chaîne de Markov obtenue par l'algorithme du Chapitre 1 possède une certaine loi stationnaire $p_\theta$, mais est aussi uniformément ergodique. Le gradient de l'Équation 2.3.5 est

$$\begin{aligned}
\nabla\phi_{\theta_0}^{(K)}(\theta) &= s(Y) - \sum_{i=1}^{K} \frac{\exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0))}{\sum_{j=1}^{K}\exp(s(Y_j)^{\mathrm{T}}(\theta - \theta_0))} s(Y_i) \\
&= \sum_{i=1}^{K} \frac{\exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0))(s(Y) - s(Y_i))}{\sum_{j=1}^{K}\exp(s(Y_j)^{\mathrm{T}}(\theta - \theta_0))} \\
&= \frac{\frac{1}{K}\sum_{i=1}^{K}\exp(s(Y_i)^{\mathrm{T}}(\theta - \theta_0))(s(Y) - s(Y_i))}{\frac{1}{K}\sum_{j=1}^{K}\exp(s(Y_j)^{\mathrm{T}}(\theta - \theta_0))}.
\end{aligned}$$

On remarque que pour tout graphe $y \in \mathcal{Y}_n$,

$$\exp(s(y)^{\mathrm{T}}(\hat{\theta} - \theta_0))p_{\theta_0}(y) = \frac{e^{s(y)^{\mathrm{T}}\hat{\theta}}}{z(\theta_0)}.$$

On pose les fonctions mesurables suivantes

$$U : \mathcal{Y}_n \to \mathbb{R}^d : y \mapsto \exp(s(y)^{\mathrm{T}}(\hat{\theta} - \theta_0))(s(Y) - s(y)), \quad \text{et}$$

$$L : \mathcal{Y}_n \to \mathbb{R} : y \mapsto \exp(s(y)^{\mathrm{T}}(\hat{\theta} - \theta_0)),$$

et pour simplifier l'écriture, on définit $U_i = U(Y_i)$ et $L_i = L(Y_i)$ pour tout $i = 1, \ldots, K$. Leur espérance, conditionnellement à $Y$, est

$$\mathbb{E}[U] = \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\hat{\theta} - \theta_0))(s(Y) - s(y))p_{\theta_0}(y) = \frac{z(\hat{\theta})}{z(\theta_0)} \sum_{y \in \mathcal{Y}_n} (s(Y) - s(y))p_{\hat{\theta}}(y), \quad \text{et}$$

$$\mathbb{E}[L] = \sum_{y \in \mathcal{Y}_n} \exp(s(y)^{\mathrm{T}}(\hat{\theta} - \theta_0))p_{\theta_0}(y) = \frac{z(\hat{\theta})}{z(\theta_0)}.$$

En posant la fonction $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d : x, y \mapsto x/y$, on retrouve

$$g\left(\frac{1}{K}\sum_{k=1}^{K} U_k, \frac{1}{K}\sum_{k=1}^{K} L_k\right) = \nabla \phi_{\theta_0}^{(K)}(\hat{\theta}), \quad \text{et}$$

$$g(\mathbb{E}[U], \mathbb{E}[L]) = \sum_{y \in \mathcal{Y}_n} (s(Y) - s(y))p_{\hat{\theta}}(y),$$

ce qui est le gradient de la log-vraisemblance appliqué au MLE. Ce terme est donc nul. Par le Corollaire 2.3.18, qui se trouve être la méthode Delta appliquée au chaîne de Markov, on conclue que

$$\sqrt{K}\left[\nabla \phi_{\theta_0}^{(K)}\right]\left(\hat{\theta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0_d, A),$$

pour une certaine matrice de covariance $A \in \mathbb{R}^{d \times d}$. □

Ainsi, sous certaines conditions de régularité, le résultat du Théorème 2.3.15 est valide pour les ERGM. Comme il est proposé par Hummel et coll., on peut estimer la variance du MCMCMLE par

$$\mathrm{Var}\left(\hat{\theta}_K | Y\right) \approx \frac{1}{K}\left[\hat{I}\left(\hat{\theta}_K\right)\right]^{-1} \hat{V}\left(\hat{\theta}_K\right)\left[\hat{I}\left(\hat{\theta}_K\right)\right]^{-1},$$

où

$$\hat{I}(\theta) = \sum_{k=1}^{K} w_k s(Y_k)s(Y_k)^{\mathrm{T}} - \left[\sum_{k=1}^{K} w_k s(Y_k)\right]\left[\sum_{k=1}^{K} w_k s(Y_k)\right]^{\mathrm{T}},$$

avec les poids

$$w_k = \frac{\exp((\theta - \theta_0)^{\mathrm{T}} s(Y_k))}{\sum_{l=1}^{K} \exp((\theta - \theta_0)^{\mathrm{T}} s(Y_l))},$$

est l'approximation de la matrice $B$, et où

$$\hat{V}(\theta) = \frac{1}{K^2}\left\{\sum_{k=1}^{K} \exp((\theta - \theta_0)^{\mathrm{T}} s(Y_k))\right\}^2$$

est l'approximation de la matrice $A$.

### 2.3.4. Algorithmes en deux étapes : approximation de l'estimateur de vraisemblance maximale

Un point important à comprendre ici est qu'il est impossible de simplement utiliser le résultat de Geyer pour trouver le MCMCMLE dû aux particularités des ERGM. Selon l'espace paramétrique défini, il se peut que certaines fonctions de la suite $\left(\phi_{\theta_0}^{(K)}\right)_{K\in\mathbb{Z}_+}$ n'admettent pas de maximum. Il faut

(1) Trouver un $\theta_0 \in \Gamma$ suffisamment près du MLE, et ;

(2) Approximer le MLE en utilisant les résultats de Geyer.

Malheureusement, un $\theta_0$ suffisamment près du MLE n'assure pas que chaque fonction de la suite ait un maximum, elle augmente simplement la probabilité d'un tel événement. Un algorithme en deux étapes a été pensé par Hunter et Handcock [**30**], et a été revue et améliorée par Hummel et coll. [**27**].

**Algorithme 2.3.20.** Algorithme pour trouver un point $\theta_0 \in \Gamma$ suffisamment près du MLE (Algorithme de la page 7 de [**27**])

---

1: Soit $t = 0$ le nombre d'itérations, $K \in \mathbb{Z}_+$ la taille d'échantillon utilisé, $\theta^{(t)} = \theta^{(0)}$ le paramètre initial, et $Y$ le graphe observé.

2: Utiliser l'algorithme proposé au Chapitre 1 pour générer $K$ graphes $Y_1, Y_2, \ldots, Y_K$ qui suivent la loi $p_{\theta_0}$.

3: Calculer $\bar{\xi}^{(t)}$, la moyenne des statistiques suffisantes de l'échantillon.

4: Trouver le plus grand $\gamma^{(t)} \in (0, 1]$ tel que

$$\gamma^{(t)} s(Y) + (1 - \gamma^{(t)})\bar{\xi}^{(t)}$$

est dans l'enveloppe convexe de $s(Y_1), s(Y_2), \ldots, s(Y_K)$. Ensuite, poser

$$\hat{\xi}^{(t)} = \gamma^{(t)} s(Y) + (1 - \gamma^{(t)})\bar{\xi}^{(t)}$$

comme pseudo-observation.

5: Trouver le prochain paramètre $\theta^{(t+1)}$ tel que

$$\theta^{(t+1)} = \arg\max_{\theta\in\Gamma} \phi_{\theta^{(t)},\hat{\xi}^{(t)}}^{(K)}(\theta)$$

où $\phi_{\theta^{(t)},\hat{\xi}^{(t)}}^{(K)}(\theta)$ est la même fonction que $\phi_{\theta^{(t)}}^{(K)}(\theta)$ sauf que $s(Y)$ a été remplacé par $\hat{\xi}^{(t)}$.

6: Incrémenter $t$. Retourner à l'étape 2 jusqu'à ce que deux itérations consécutives avec un $\gamma^{(t)} = 1$ aient été complétées.

---

Une fois que l'Algorithme 2.3.20 eut été complété, l'optimisation de l'étape 5 avec une valeur de $K$ grande permet d'approximer le MCMCMLE. Les auteurs proposent, afin de garder une certaine stabilité, d'effectuer l'optimisation de l'étape 5 en cherchant le meilleur $\theta$ dans une boule (ou un cube) autour de $\theta^{(t)}$. Un algorithme quasi-Newton qui accepte des bornes, comme L-BFGS-B, qui est une généralisation de la méthode de Zhu et coll. [**71**], est un choix acceptable pour effectuer cette tâche. Cet algorithme comporte plusieurs failles :

— Malgré qu'on peut avoir l'impression que $\theta^{(t)}$ se rapproche du MLE, rien n'est moins certain. Du meilleur de nos connaissances, il n'existe pas de preuve de convergence de cet algorithme.

— Le choix de pseudo-observation n'est pas optimal. Si $s(Y) \notin \operatorname{aff}(\hat{\mathcal{S}}_{\theta_0}^{(K)})$, alors la meilleure pseudo-observation sera simplement la moyenne des statistiques exhaustives de l'échantillon $\bar{\xi}^{(t)}$, ce qui fait « perdre » une itération. Même sans ce cas, la définition de cette pseudo-observation n'est pas optimale. Comme il est vu au Chapitre 1, il peut être coûteux de générer des graphes aléatoires alors que chaque échantillon doit être utilisé à bon escient.

— Les auteurs supposent que le l'estimateur est unique, ce qui n'est pas garanti. Or, il est nécessaire d'opter pour une réécriture pour s'assurer de la convergence comme il a été vu à la Section 2.3.

— Obtenir une bonne valeur de $\theta_0$ ne garantie pas l'existence du MCMCMLE pour une certaine valeur de $K$ choisi.

**2.3.4.1. Pseudo-observation.** Sauf pour Bayly et coll. [**3**], la plupart des autres auteurs ont supposé que l'enveloppe affine de $\mathcal{S}$ était de pleine dimension, et de ce fait, lorsque le MLE (ou le MCMCMLE) existait, il était unique. Or, si ce n'est pas le cas, c'est-à-dire que si la dimension de l'enveloppe affine de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$ n'est pas de pleine dimension et que $s(Y)$ n'est pas dans cette enveloppe affine, alors le meilleur $\gamma$ qu'on peut choisir à l'étape 4 est $\gamma = 0$. Ce dernier cas se produit assez souvent lorsque $\theta_0$ est loin du MLE. De plus, la pseudo-observation trouvée avec cette méthode est un point dans l'intérieur convexe relatif de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$ qui est à mi-chemin entre $s(Y)$ et le point milieu de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$. Il serait pourtant préférable de trouver le point dans l'intérieur convexe relatif de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$ le plus près de $s(Y)$. Cet algorithme peut donc être modifié pour mieux refléter les Propositions 2.3.12 et 2.3.11.

On redéfinit d'abord la pseudo-observation $\hat{\xi}$,

$$\hat{\xi} \in \arg \min_{\xi \in \operatorname{rcint}(\hat{\mathcal{S}}_{\theta_0}^{(K)})} \|s(Y) - \xi\|^2. \tag{2.3.10}$$

La solution n'existe pas dès que $s(Y) \notin \operatorname{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)$, car $\operatorname{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)$ est un ensemble ouvert et la solution se trouverait sur la fermeture de cet ensemble. Dans ce cas, on pourrait tout de même se satisfaire d'une solution valide, c'est-à-dire d'un point qui se trouve dans

l'intérieur convexe relatif, mais qui soit arbitrairement près du point critique se trouvant sur la fermeture. Le problème de l'Équation 2.3.10 peut être réécrit sous une forme plus facile à gérer. Soit $Y_1, \ldots, Y_K$ l'échantillon de $K$ graphes qui suivent la loi $p_{\theta_0}$. On pose $C \in \mathbb{R}^{d \times K}$ la matrice composée des $K$ vecteurs $s(Y_1), \ldots, s(Y_K)$ comme colonnes. Le nouveau problème d'optimisation est

$$\begin{aligned} \hat{\lambda} \in \arg\min_{\lambda \in \mathbb{R}^K} \|s(Y) - C\lambda\|^2 \\ \text{sous les contraintes} \quad 0 < \lambda_i, \quad \text{pour tout } i = 1, \ldots, K \\ \sum_{i=1}^{K} \lambda_i = 1. \end{aligned}$$

Les contraintes obligent à chercher parmi les combinaisons convexes, mais l'inégalité stricte des $K$ premières contraintes obligent que $\hat{\xi} = C\lambda$ soit un point de l'intérieur convexe relatif de $\hat{\mathcal{S}}_{\theta_0}^{(K)}$. Ce problème de programmation quadratique peut se résoudre avec une méthode de point intérieur présentée par Gondzio [**20**]. Cependant, une simple reparamétrisation permet d'utiliser une descente de gradient sans se soucier des contraintes.

**Définition 2.3.21.** On définit la fonction **logit multivariée** (ou encore *softmax*) $\Psi : \mathbb{R}^{K-1} \to \mathbb{R}^K$ tel que pour tout $\gamma \in \mathbb{R}^{K-1}$,

$$\begin{aligned} \Psi_j(\gamma) &= \frac{\exp(\gamma_j)}{1 + \sum_{k=1}^{K-1} \exp(\gamma_k)} \quad \text{pour } j = 1, \ldots, K-1 \\ \Psi_K(\gamma) &= \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\gamma_k)}. \end{aligned}$$

De plus, sa matrice jacobienne $J_\Psi \in \mathbb{R}^{K \times (K-1)}$ est, pour tout $\gamma \in \mathbb{R}^{K-1}$, $i = 1, \ldots, K$ et $j = 1, \ldots, K-1$,

$$\begin{aligned} [J_\Psi]_{ij}(\gamma) &= \Psi_i(\gamma)(1 - \Psi_i(\gamma)) \quad \text{si } i = j \\ [J_\Psi]_{ij}(\gamma) &= -\Psi_i(\gamma)\Psi_j(\gamma) \quad \text{sinon.} \end{aligned}$$

On remarque que quelque soit $\gamma \in \mathbb{R}^{K-1}$, la fonction logit multivariée est telle qu'on a que $\Psi_i(\gamma) > 0$, pour tout $i = 1, \ldots, K$, et $\sum_{i=1}^{K} \Psi_i(\gamma) = 1$. En utilisant la reparamétrisation $\lambda = \Psi(\gamma)$, on peut réécrire finalement notre problème d'optimisation ainsi

$$\hat{\gamma} \in \arg\min_{\gamma \in \mathbb{R}^{K-1}} \|s(Y) - C\Psi(\gamma)\|^2.$$

Ce problème peut être résolu avec n'importe quel algorithme de descente de gradient, comme ADAM [**35**]. On note qu'il peut exister une infinité de solutions, car souvent une infinité de combinaisons convexes peut représenter un même point. Ainsi, des algorithmes qui font intervenir la matrice hessienne, comme Newton-Raphson, peuvent ne pas fonctionner.

Dû à cet aspect, le critère d'arrêt ne doit pas tenir compte du pas effectué entre deux itérations, mais plutôt à une variation de la fonction objective. Dans le cas où la solution se trouve plutôt dans la fermeture de l'intérieur convexe relatif, en commençant avec un point valide, comme pour $\gamma = 0_{K-1}$, qui représente la moyenne des statistiques exhaustives, plus la tolérance, c'est-à-dire la borne sur la variation de la fonction objective entre deux itérations avant l'arrêt, est basse, plus l'algorithme va se rapprocher de la frontière de l'enveloppe convexe, mais tout en restant dans l'intérieur convexe relatif. Il peut être alors judicieux de prendre une tolérance pas trop près de 0, comme $10^{-3}$ par exemple, pour s'assurer que la pseudo-observation est à bonne distance de la frontière. Finalement, on peut conclure que l'événement $s(Y) \in \operatorname{rcint}(\hat{\mathcal{S}}_{\theta_0}^{(K)})$ est rencontré lorsque $\|s(Y) - C\Psi(\hat{\gamma})\|^2 = 0$. Comme pour l'Algorithme 2.3.20, on supposera qu'une valeur $\theta_0 \in \Gamma$ suffisamment près du MLE est trouvée lorsque deux itérations sont complétées en ayant rencontré cet événement.

**2.3.4.2. Unicité du maximum.** Comme expliqué précédemment, il faut que les points maximums soient uniques, afin de s'assurer de la convergence du MCMCMLE. Il a déjà été noté qu'on peut se concentrer sur les points de plus petite norme qui se trouvent tous dans l'espace $V_{\mathcal{S}}$. Or, bien sûr, lorsqu'on résout le problème d'optimisation de l'étape 5, on ne sait pas si cet algorithme est de pleine dimension. On peut cependant se restreindre à l'espace

$$\hat{V}_{\hat{\mathcal{S}}_{\theta_0}^{(K)}} = \operatorname{vect}\left(s(Y_1) - s(Y), \ldots, s(Y_K) - s(Y)\right).$$

On remarque que

$$V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}} \subset \hat{V}_{\hat{\mathcal{S}}_{\theta_0}^{(K)}} \subset V_{\mathcal{S}},$$

et que $V_{\hat{\mathcal{S}}_{\theta_0}^{(K)}}$ converge presque sûrement vers $V_{\mathcal{S}}$ lorsque $K$ converge vers l'infini. Plus spécifiquement, ces deux espaces sont égaux au plus tard lorsque la chaîne $s(Y_1), \ldots, s(Y_K)$ a couvert l'espace $\mathcal{S}$. On pose alors $B$ comme la matrice composée des vecteurs $s(Y_1) - s(Y), \ldots, s(Y_K) - s(Y)$ comme colonne. Le rang est bien sûr supérieurement borné par $\min(K, d)$, alors pour extraire une base linéairement indépendante, on pose $\underline{B}$ la matrice composée des vecteurs propres de $B$ associé à des valeurs propres strictement positives. On pose $\underline{B} \in \mathbb{R}^{d \times p}$. On peut alors changer le problème d'optimisation de l'étape 5 ainsi

$$a^{(t+1)} = \arg\max_{a \in \mathbb{R}^p} \phi_{\theta^{(t)}, \hat{\xi}^{(t)}}(\underline{B}a),$$

et poser $\theta^{(t+1)} = \underline{B}a^{(t+1)}$. Par construction, il existe un unique $a \in \mathbb{R}^p$ qui maximise la fonction.

**2.3.4.3. Algorithme modifié.** En prenant en compte nos précédents résultats, on obtient l'Algorithme 2.3.22.

**Algorithme 2.3.22.** Algorithme itératif modifié en deux étapes pour trouver le MCMCMLE

---

1: Soit
   — $t = 0$ le nombre d'itérations ;
   — $\theta^{(t)} = \theta^{(0)}$ le paramètre initial ;
   — $K \in \mathbb{Z}_+$ la taille d'échantillon utilisé, et ;
   — $Y$ le graphe observé.

2: Utiliser l'algorithme proposé au Chapitre 1 pour générer $K$ graphes $Y_1, Y_2, \ldots, Y_K$ qui suivent la loi $p_{\theta_0}$.

3: Poser $\hat{\mathcal{S}}_{\theta_0}^{(K)} = \{s(Y_i) : i = 1, \ldots, K\}$ et $\underline{B} = \text{base}(s(Y_i) - s(Y) : i = 1, \ldots, K)$ où base$(A)$ retourne une matrice composée d'une base linéairement indépendante de vect$(A)$ comme colonne.

4: Si $s(Y) \in \text{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)$, poser $\hat{\xi} = s(Y)$, sinon trouver la pseudo-observation

$$\hat{\xi} \in \arg \min_{\xi \in \text{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)} \|s(Y) - \xi\|^2,$$

de la même manière qu'il est expliqué dans la Section 2.3.4.1.

5: Trouver combinaison $\hat{a}$ telle que

$$\hat{a} \in \arg \max_{a \in \mathbb{R}^p} \phi_{\theta^{(t)}, \hat{\xi}}^{(K)} (\underline{B}a)$$

où $\phi_{\theta^{(t)}, \hat{\xi}}^{(K)}(\theta)$ est la même fonction que $\phi_{\theta_0}^{(K)}(\theta)$ de l'Équation 2.3.5 où $\theta_0 = \theta^{(t)}$, sauf que $s(Y)$ a été remplacé par $\hat{\xi}$.

6: $\theta^{(t+1)} = \underline{B}\hat{a}$.

7: Incrémenter $t$. Retourner à l'étape 2 jusqu'à ce que deux itérations consécutives se fassent avec $s(Y) \in \text{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)$.

---

Des éléments restent encore à améliorer :

— La distance utilisée à l'Étape 4 est la distance $L_2$ et elle ne prend pas en compte la géométrie intrinsèque de l'ensemble $\hat{\mathcal{S}}_{\theta_0}^{(K)}$, et ;

— Aucune preuve formelle n'a été faite pour s'assurer que la suite $(\theta^{(t)})_{t \in \mathbb{N}}$ s'approche d'un MLE.

Pour mettre à l'épreuve l'amélioration apportée, on effectue une expérience. Pour deux modèles différents, on compare la moyenne du nombre d'itérations nécessaire pour atteindre un critère d'arrêt. Le graphe observé sera celui provenant du jeu de données *Karate Club Network* [**69**]. Il est composé de 34 sommets et de 78 arêtes. Chaque sommet représente un club de karaté, et pour une paire de clubs, il existe une arête si des membres de chacun

des clubs se côtoient à l'extérieur des séances de karaté. Le graphe est représenté dans la Figure 2.1. Pour le choix des modèles, on s'inspire de modèles déjà existants d'une ancienne publication [**61**] :

(1) Le modèle est composé du nombre d'arêtes et le degré du poids géométrique de partenaires communs avec une pénalité de 0,2.

(2) Le modèle est composé du nombre d'arêtes, le degré du poids géométrique de partenaires communs avec une pénalité de 0,2, et le degré du poids géométrique avec une pénalité de 0,8.

Le critère d'arrêt est changé dans les deux algorithmes pour s'assurer que la comparaison représentée est fiable. Le critère d'arrêt est la vérification explicite que $s(Y) \in \operatorname{rcint}\left(\hat{\mathcal{S}}_{\theta_0}^{(K)}\right)$. Pour chacun des modèles, on teste les algorithmes pour différentes tailles d'échantillon $K$, soit 50, 100 et 200. Un temps de brûlure de 500 et un facteur d'amincissement de 25 est utilisé pour chaque chaîne de Markov. Lors de l'optimisation pour trouver le prochain paramètre, on restreint notre domaine dans un carré de 0,2 unité de large centré autour du précédent paramètre. Le paramètre initial est toujours le MPLE. Pour chacun des cas, 50 exécutions sont effectuées. On impose cependant un nombre maximal de 100 itérations avant de supposer qu'une exécution n'a pas convergé. Ce nombre a été choisi par des observations empiriques pour ces modèles appliqués à ce graphe. Les expériences qui n'ont pas convergé à l'intérieur de 100 itérations ne semblent jamais converger. Par exemple, pour l'exécution de l'algorithme « MCMCMLE classique » avec un échantillon de taille 200, l'algorithme a convergé 45 fois sur 50, et parmi ces 45 fois, le critère d'arrêt a été satisfait après 58,42 itérations en moyenne et avec un écart-type de 19,44 itérations.

| **Grandeur d'échantillon** ($K$) | **Algorithmes** (Nombre d'exécution avec convergence sur 50) Moyenne ± Écart-type | |
|---|---|---|
| | **MCMCMLE classique** | **MCMCMLE amélioré** |
| **50** | (50) 45,86 ± 16,29 | (50) 34,44 ± 9,85 |
| **100** | (50) 42,62 ± 13,41 | (50) 21,74 ± 3,57 |
| **200** | (45) 58,42 ± 19,44 | (50) 16,72 ± 2,13 |

**Tableau 2.1** – Comparaison entre l'Algorithme 2.3.20 et l'Algorithme 2.3.22 pour le Modèle 1 appliqué au graphe Karate Club Network [**69**]. Pour chaque paramètre d'expérience, 50 expériences ont été produites. On suppose qu'une expérience n'a pas convergé si le nombre d'itérations dépasse 100.

| **Grandeur d'échantillon** ($K$) | **Algorithmes** (Nombre d'exécution avec convergence sur 50) Moyenne ± Écart-type | |
|---|---|---|
| | **MCMCMLE classique** | **MCMCMLE amélioré** |
| **50** | (5) 65,20 ± 19,05 | (21) 73,43 ± 18,78 |
| **100** | (14) 66,29 ± 22,88 | (50) 53,32 ± 18,07 |
| **200** | (9) 75,44 ± 15,38 | (50) 30,54 ± 9,29 |

**Tableau 2.2** – Comparaison entre l'Algorithme 2.3.20 et l'Algorithme 2.3.22 pour le Modèle 2 appliqué au graphe Karate Club Network [**69**]. Pour chaque paramètre d'expérience, 50 expériences ont été produites. On suppose qu'une expérience n'a pas convergé si le nombre d'itérations dépasse 100.

Pour le Modèle 1, on remarque que l'Algorithme 2.3.22 donne toujours de meilleurs résultats et est aussi bien plus constant en termes d'écart-type que l'Algorithme 2.3.20. Quant au Modèle 2, pour la plupart des exécutions, l'Algorithme 2.3.20 n'a pas convergé. Encore ici, l'Algorithme 2.3.22 semble mieux se comporter.

Dans la Figure 2.2, on a effectué une simulation de l'Algorithme 2.3.22. On a pris le modèle composé du nombre d'arêtes et du nombre d'étoiles à trois branches comme statistiques exhaustives appliqué à des graphes de six sommets. On a ensuite pris un certain paramètre ($\theta = (1, -0.3)$) pour générer le graphe observé. On a ensuite exécuté l'algorithme, qui a pris cinq itérations, comprenant l'itération finale pour converger. On observe l'évolution de l'approximation de la log-vraisemblance qui converge tranquillement vers la log-vraisemblance au fur et à mesure que les graphes simulés se retrouvent dans la même région que notre graphe observé. L'itération finale permet de trouver $\hat{\theta} = (1.19, -0.36)$ comme MCMCMLE.

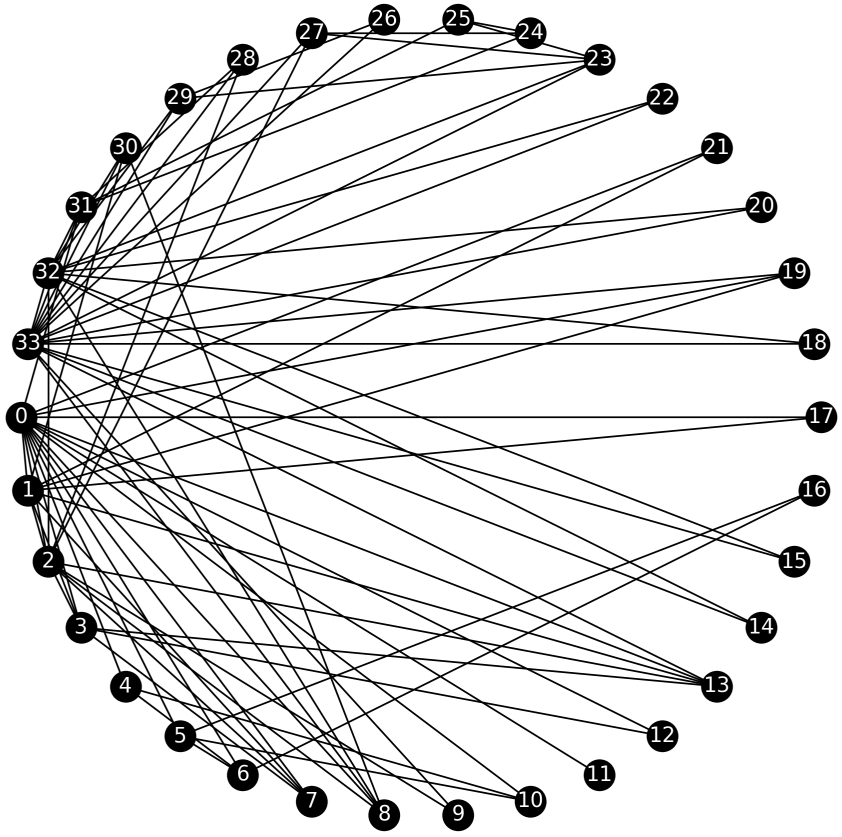



**Figure 2.1** – Graphe composé de clubs de karaté illustrant les interactions à l'extérieur de la pratique du sport [**69**]

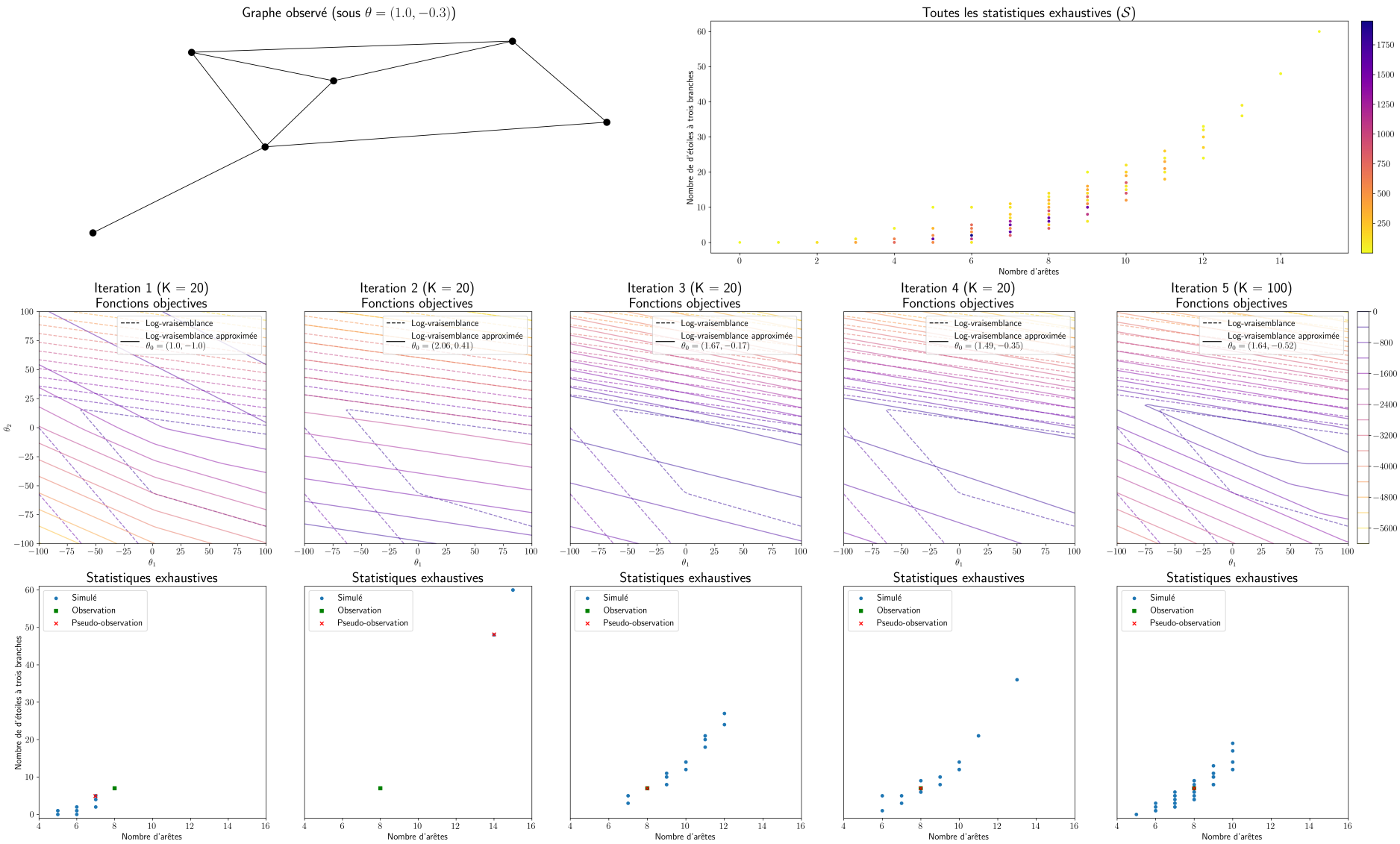



**Figure 2.2** – Exécution de l'Algorithme 2.3.22 avec le nombre d'arêtes et le nombre d'étoiles à trois branches comme statistiques exhaustives. Le graphe observé est en haut à gauche, les statistiques exhaustives pour chacun des graphes possibles sont représentées en haut à droite, où la couleur représente la fréquence d'un même point, la deuxième ligne représente la fonction $\phi_{\omega_0}^{(K)}$ à chacune des itérations, et la troisième ligne représente les statistiques exhaustives des graphes échantillonnés.

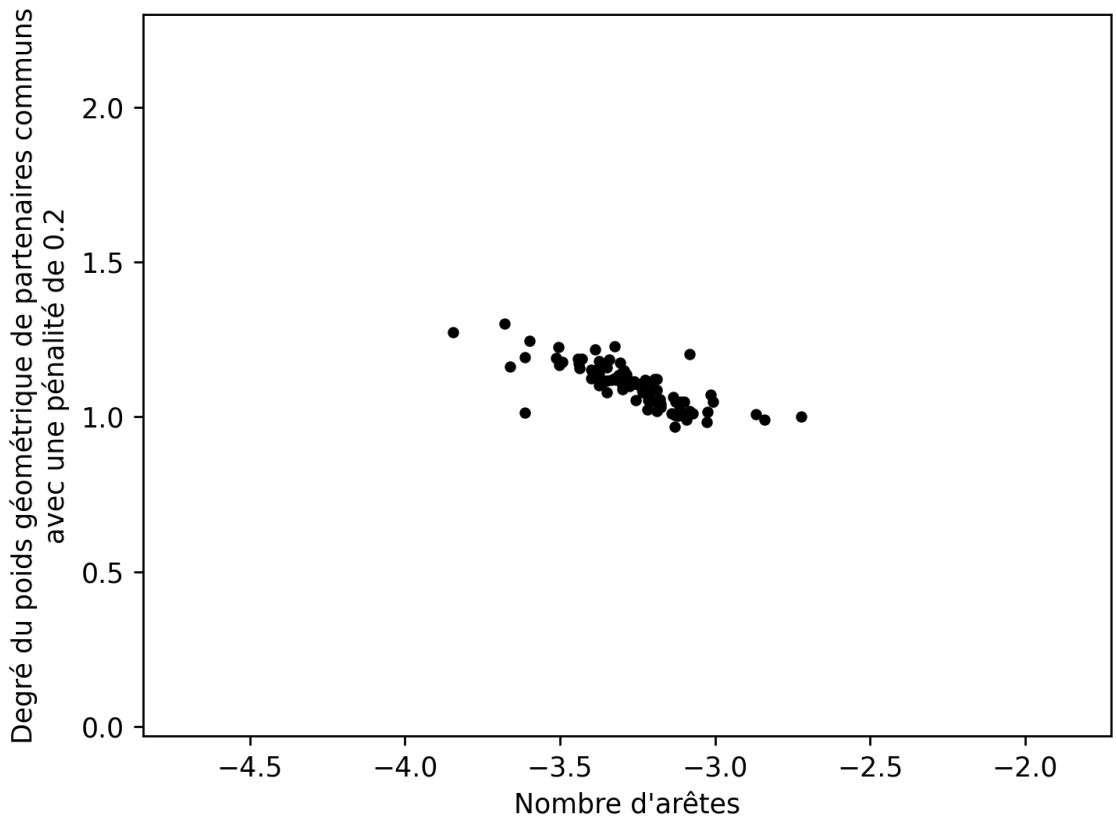



**Figure 2.3** – 100 réalisations MCMCMLE pour le Modèle 1 pour le *Karate Club Network*

On illustre maintenant le comportement asymptotique avec un exemple. On reprend le même graphe, *Karate Club Network*, on exécute l'Algorithme 2.3.22 100 fois pour trouver le MCMCMLE pour le Modèle 1. Les 100 réalisations sont représentées dans la Figure 2.3.

### 2.3.5. Travaux futurs

Par l'inclusion de l'Équation 2.3.8, on sait que, quel que soit le choix de $\theta_0 \in \Gamma$, il existe presque sûrement un moment où pour toutes les fonctions subséquentes de la suite $\left(\phi_{\theta_0}^{(K)}\right)_{K\in\mathbb{Z}_+}$ le maximum existe. Cette inclusion est cependant très conservatrice. La fonction $\phi_{\theta_0}^{(K)}$ est une approximation du logarithme du rapport de vraisemblances et comme il a été remarqué par Geyer et Thompson [**19**], l'approximation est aussi bonne que $\theta_0$ est près du maximum de vraisemblance. Trouver une borne moins conservatrice permettrait de choisir une valeur de $K$ qui améliorerait les chances de trouver un maximum. D'autre part, si on était capable de bornée supérieurement au temps de couverture, la Proposition 2.3.14 proposerait alors un ordre de grandeur pour $K$.

Comme il a été remarqué dans la Section 2.3.4, tant pour l'Algorithme 2.3.20 que pour l'Algorithme 2.3.22, aucune preuve ne permet de conclure que les itérations permettent d'obtenir une « bonne » valeur pour $\theta_0$, et même avec une bonne valeur de $\theta_0$, pour une valeur de $K$ choisi, le MCMCMLE n'est pas garanti d'exister. Ce sont deux problèmes majeurs qui demandent une sérieuse attention.

# Chapitre 3

# Inférence de la loi à postériori

Tout au long du présent chapitre, on supposera que l'espace paramétrique est $\Gamma = \mathbb{R}^d$. Ce chapitre prend place dans un contexte bayésien, ainsi on suppose que le paramètre est plutôt une variable aléatoire $\Theta$ qui prend ses valeurs dans l'espace paramétrique $\Gamma$. La loi à priori $\pi(\cdot)$ grandement adoptée par communauté scientifique est la loi normale avec les paramètres $\mu_0 = 0_d$ et $\Sigma_0 = \sigma_0^2 I_d$, et où $\sigma_0^2 \gg 0$. La fonction de vraisemblance $p_\theta$ est maintenant la fonction de masse conditionnelle $p(y|\theta) = p_\theta(y)$ qui est définie pour tout graphe $y \in \mathcal{Y}$ et pour toute valeur de paramètre $\theta \in \Gamma$. On remarque que la loi à postériori

$$\pi(\theta|y) = \frac{p(\theta|y)\pi(\theta)}{\int_{\mathbb{R}^d} p(\theta|y)\pi(\theta)d\theta}, \quad \text{pour tout } y \in \mathcal{Y}, \theta \in \Gamma,$$

est incalculable dû à la fonction de vraisemblance qui s'y trouve. Étant donné qu'on la trouve au numérateur ainsi qu'au dénominateur, cette densité est souvent étiquetée « doublement incalculable ». Il faut alors faire preuve d'ingéniosité pour inférer sur le paramètre.

Une première technique est d'échantillonner des réalisations de la loi à postériori en utilisant des variables auxiliaires [**45**]. Cette méthode a mené à l'algorithme d'*exchange* [**44**]. Malheureusement, il est nécessaire d'échantillonner parfaitement selon la vraisemblance, ce qui n'est pas possible dans le cadre des ERGM. De nouvelles améliorations ont été apportées à cette technique pour permettre d'échantillonner sur une chaîne de Markov [**38**], mais l'application aux ERGM n'est pas faite ici. Il est important de mentionner aussi que ces techniques présentent un coût en temps-machine élevé si les paramètres de l'algorithme ne sont pas habilement choisis.

Une autre technique est l'utilisation d'une densité de substitut, une approximation « suffisamment » près de la vraie loi. La technique classique d'inférence variationnelle est présentée à la Section 3.1. Ensuite, à la Section 3.1.1, l'inférence variationnelle est appliquée aux ERGM.

## 3.1. Méthodes classiques en inférence variationnelle

L'inférence variationnelle, *Variational Inference*, (VI) [**5**] est une technique pour approximer une densité (ou une fonction de masse) par une autre densité qui est plus facile à manipuler. L'idée est de trouver la fonction parmi un ensemble de fonctions, celle qui est « la plus proche » de la fonction à approximer. On appelle la famille variationnelle, qu'on note $\mathcal{G}$, l'ensemble des fonctions qui contient nos possibles approximations. Si l'ensemble est paramétré par un paramètre, qu'on note $\lambda$, il est appelé le paramètre variationnel. Dans notre cas, la famille variationnelle sera une sous-variété différentiable contenue dans l'ensemble des densités intégrables, et dont la fonction de transition est infiniment différentiable. Une « distance » doit être utilisée pour trouver un bon candidat dans la famille variationnelle. En inférence variationnelle, on utilise le plus souvent la divergence Kullback-Leibler

$$D_{\mathrm{KL}}[f,g] = \int f(x) \log \frac{f(x)}{g(x)} dx, \quad \text{pour deux densités } f, g. \tag{3.1.1}$$

Il existe plusieurs autres divergences, mais on se concentrera sur cette dernière. On remarque que pour toute paire de densité $f, g$ on a toujours

$$\begin{aligned} &D_{\mathrm{KL}}[f,g] \geq 0, \text{ et ;} \\ &D_{\mathrm{KL}}[f,g] = 0 \text{ si, et seulement si, } f = g. \end{aligned}$$

On remarque aussi que ce n'est pas une distance. Elle n'est pas symétrique et ne respecte pas l'inégalité du triangle. Les questions d'existence et d'unicité de l'optimum ne sont généralement (et malheureusement) pas abordées. On suppose simplement qu'il existe au moins un optimum et on se satisfait d'un local. Bien sûr, les fonctions de $\mathcal{G}$ ainsi que la fonction à approximer doivent posséder le même support, sinon en posant $c/0 = \infty$ pour une constante $c > 0$, on trouvera que $D_{\mathrm{KL}}[f,g] = \infty$ pour toute paire de densités sur des supports (presque sûrement) différents. On observe que cette divergence n'est pas symétrique et l'ordre a un impact sur l'approximation obtenue. Ainsi, si on souhaite approximer la densité $f$ par une densité de la famille $\mathcal{G}$, notre approximation peut être la solution au problème d'optimisation

$$\hat{f} \in \arg\min_{g \in \mathcal{G}} D_{\mathrm{KL}}[g, f], \tag{3.1.2}$$

ou encore la solution au problème d'optimisation

$$\hat{f} \in \arg\min_{g \in \mathcal{G}} D_{\mathrm{KL}}[f, g]. \tag{3.1.3}$$

Pour le problème de l'Équation 3.1.2, on dit qu'on minimise la divergence Kullback-Leibler renversée, *Reverse Kullback-Leibler divergence*, (divergence RKL), et pour le problème de l'Équation 3.1.3, on dit qu'on minimise la divergence Kullback-Leibler avant, *Forward Kullback-Leibler divergence*, (divergence FKL).

On observe maintenant le cas où la densité à approximer en est une à postériori. Soit $\Omega \subset \mathbb{R}^d$ un espace paramétrique où la variable aléatoire $\eta$ prend ses valeurs selon une certaine loi définie par une densité à priori $\pi(\omega)$, et soit $\mathcal{X}$ un espace dans lequel la variable aléatoire $X$ prend ses valeurs selon une loi définie par une densité $f(x)$. Les deux variables aléatoires $X$ et $\eta$ sont liées par la densité jointe $f(x, \omega)$, et la fonction de vraisemblance $f(x|\omega)$. On note la loi à postériori $\pi(\omega|x)$. Soit $\mathcal{G}$ une famille variationnelle pour $\Omega$. Le problème d'optimisation pourrait soit être posé comme à l'Équation 3.1.2, ou soit comme à l'Équation 3.1.3. Le premier cas sera le plus souvent utilisé pour des raisons de faisabilité. On remarque que dans le deuxième cas une intégration doit être faite par rapport à $\pi(\cdot|X)$, et si on utilise une méthode VI, c'est en partie parce qu'on n'est pas capable de la calculer. De plus, pour toute densité $g \in \mathcal{G}$, on a la relation suivante

$$\log f(X) = \int_{\mathbb{R}^d} g(\eta) \log \frac{f(\eta, X)}{g(\eta)} d\eta + \int_{\mathbb{R}^d} g(\eta) \log \frac{g(\eta)}{\pi(\eta|X)} d\eta.$$

Pour une certaine réalisation, la partie de gauche de l'égalité est fixe, et quant à la partie de droite, on trouve deux termes : le premier est appelé la borne inférieure de l'évidence, *evidence lower bound*, (ELBO) et le deuxième terme est la divergence Kullback-Leibler $D_{\mathrm{KL}}(g, \pi(\cdot|X))$. On remarque ainsi que minimiser la divergence Kullback-Leibler revient à maximiser la ELBO, et il est souvent plus aisé de maximiser la ELBO que de minimiser la divergence. On peut aussi imposer des structures particulières sur la famille variationnelle afin de faciliter le problème. Par exemple, on peut supposer que pour toute densité $g \in \mathcal{G}$, il existe une décomposition en plusieurs densités distinctes ainsi

$$g(\omega) = g_1(\omega_1) \dots g_d(\omega_d), \quad \text{pour tout } \omega \in \Omega.$$

Cette sous-famille est nommée *Mean Field*. Comme il peut être vu dans [**5**], trouver la fonction de cette famille qui minimise la divergence RKL est très facile. Cependant, se restreindre aux densités qui supposent l'indépendance des paramètres mène nécessairement à un plus grand biais.

### 3.1.1. Inférence variationnelle pour les modèles de graphes aléatoires

La famille variationnelle utilisée pour les ERGM est l'ensemble des densités normales, et avec une telle famille, les paramètres variationnels sont $\mu \in \mathbb{R}^d$, le paramètre de moyenne, et $\Sigma \in \mathbb{R}^{d \times d}$, le paramètre de variance. Tan et Friel ont proposé une manière de résoudre ce problème [**61**] en utilisant une méthode de descente de gradient stochastique [**50**]. Comme précédemment expliqué, notre problème pour minimiser la divergence RKL peut se ramener

au problème suivant

$$\hat{\pi} \in \arg\max_{g \in \mathcal{G}} \int_{\mathbb{R}^d} g(\theta) \log \frac{p(Y,\theta)}{g(\theta)} d\theta. \tag{3.1.4}$$

Notre fonction objective est

$$J(g) = \int_{\mathbb{R}^d} g(\theta) \log \frac{p(Y,\theta)}{g(\theta)} d\theta, \quad \text{pour toute fonction } g \in \mathcal{G},$$

ce qui peut s'écrire de manière équivalente $J(g) \equiv J(\mu, \Sigma)$ dans notre cas. Pour tout paramètre de variance $\Sigma$, il existe une unique décomposition de Cholesky $\Sigma = CC^{\mathrm{T}}$, et si on pose $\mathrm{vech} : \mathbb{R}^{d \times d} \to \mathbb{R}^{\frac{d(d+1)}{2}}$ la fonction qui vectorise une matrice triangulaire inférieure, nos mises à jour de paramètres sont

$$\begin{aligned}
\mu^{(t+1)} &= \mu^{(t)} + \rho_t [\nabla_\mu J](\mu^{(t)}, \Sigma^{(t)}), \\
C^{(t+1)} &= C^{(t+1)} + \rho_t \mathrm{vech}^{-1} \left( [\nabla_{\mathrm{vech}(C)} J](\mu^{(t)}, \Sigma^{(t)}) \right),
\end{aligned}$$

où $\rho_t$ sont les pas de gradient qui respectent les conditions nécessaires à la convergence de l'algorithme. Certaines difficultés restent à éclaircir.

On remarque d'abord que notre fonction $J$ est en fait une espérance prise sous une autre mesure, la mesure où la loi de $\Theta$ est induite par la densité $g$. Sous cette mesure, on peut poser $\Theta = CX + \mu$ où $X$ est une variable aléatoire sous la loi normale standard. On peut donc voir $\Theta$ comme une fonction de $X$. Ainsi, notre fonction objective peut s'écrire

$$J(\mu, \Sigma) = \int_{\mathbb{R}^d} \phi(x) \log \frac{p(Y, \theta(x))}{g(\theta(x))} dx = \int_{\mathbb{R}^d} \phi(x) \log p(Y, \theta(x)) dx - \int_{\mathbb{R}^d} \phi(x) \log g(\theta(x)),$$

où $\phi$ est la densité d'une normale standard. En écrivant la fonction objective de cette manière, on retrouve les paramètres dans la fonction $g$, mais aussi dans $\theta$. Sous certaines conditions de régularité, en utilisant la dérivation en chaine, on trouve

$$\begin{aligned}
[\nabla_\mu J](\mu, \Sigma) &= \int_{\mathbb{R}^d} \phi(x) \left\{ \nabla_\theta \log p(Y, \theta(x)) - \nabla_\theta \log g(\theta(x)) - \nabla_\mu \log g(\theta(x)) \right\} dx, \\
&[\nabla_{\mathrm{vech}(C)} J](\mu, \Sigma) \\
&\qquad = \int_{\mathbb{R}^d} \phi(x) \left\{ \mathrm{vech} \left( \nabla_\theta \log p(Y, \theta(x)) - \nabla_\theta \log g(\theta(x)) \right) - \nabla_{\mathrm{vech}(C)} \log g(\theta(x)) \right\} dx,
\end{aligned}$$

avec les gradients suivants

$$\begin{aligned}
\nabla_\theta \log p(Y, \theta(x)) &= s(Y) - \sum_{y \in \mathcal{Y}_n} p(y|\theta(x)) s(y) - \Sigma_0^{-1}(\theta(x) - \mu_0), \\
\nabla_\theta \log g(\theta(x)) &= -\nabla_\mu \log g(\theta(x)) = -C^{-\mathrm{T}} x, \\
\nabla_{\mathrm{vech}(C)} \log g(\theta(x)) &= \mathrm{vech}(C^{-\mathrm{T}}(xx^{\mathrm{T}} - I_d)).
\end{aligned}$$

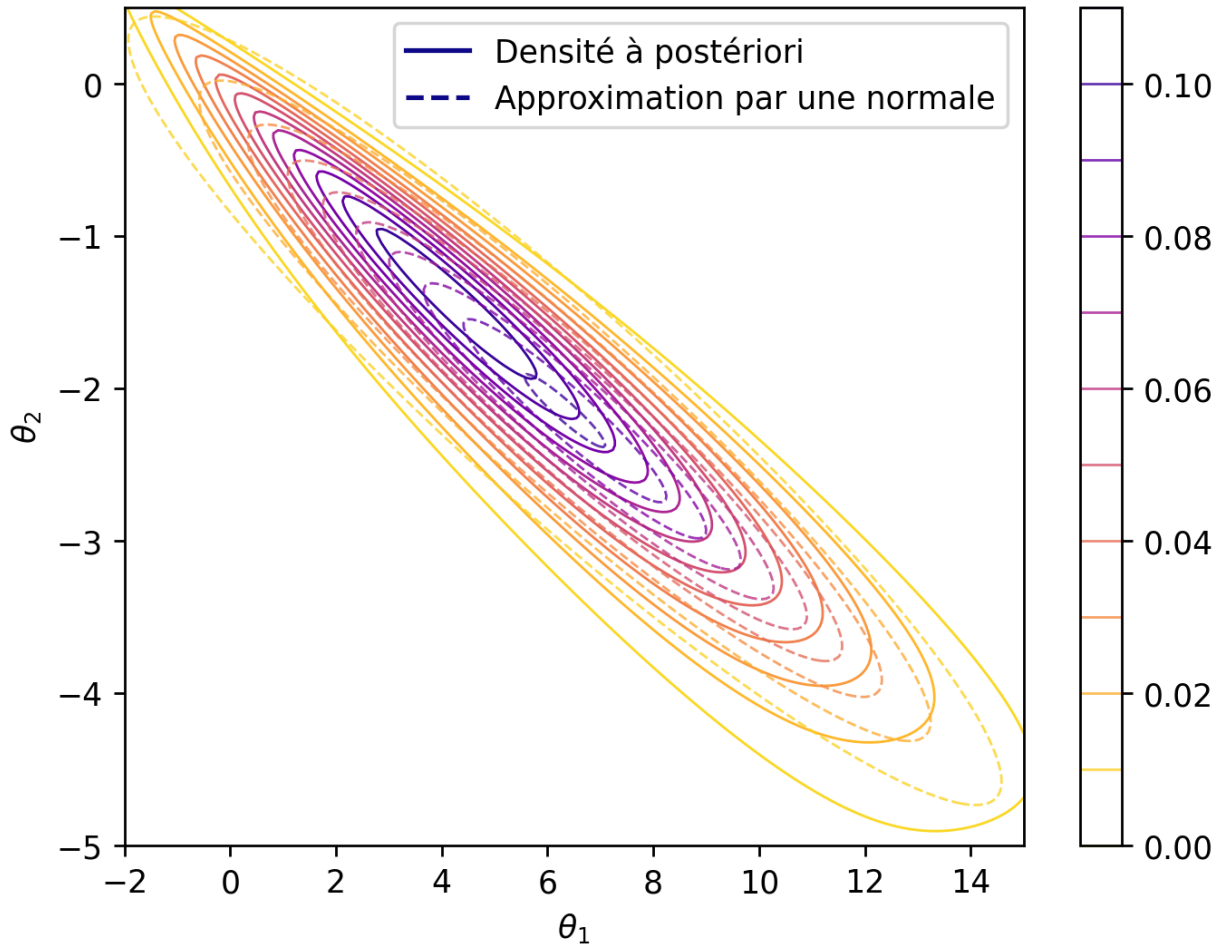



**Figure 3.1** – Approximation de la loi à postériori par une loi normale en minimisant la divergence Kullback-Leibler renversée pour le modèle composé du nombre d'arêtes et du nombre de chemins de longueur deux

L'espérance $\sum_{y \in \mathcal{Y}_n} p(y|\theta(x))s(y)$ peut être approximée par une méthode MCMC ou par *self-normalized importance sampling*. De plus, les deux principaux gradients sont des espérances qu'on peut facilement estimer par Monte Carlo en échantillonnant sur $X \sim \mathcal{N}(0_d, I_d)$.

Pour illustrer l'approximation obtenue, on prend le modèle composé du nombre d'arêtes et du nombre de chemin de longueur deux, et on échantillonne un graphe à cinq sommets suivant ce modèle. Étant donné que le nombre de sommets est faible, on peut calculer la loi à postériori pour tous les points et la comparer avec l'approximation obtenue par inférence variationnelle. Les deux densités sont représentées dans la Figure 3.1.

### 3.1.2. Travaux futurs

Même si l'approximation semble acceptable, on ignore si c'est le minimum global ou simplement un optimum local, ou encore s'il existe un unique minimum global. Dans le précédent exemple, on a pu vérifier empiriquement que la densité à postériori est unimodale, mais rien ne garantit que ce soit toujours le cas. Si elle était composée de plusieurs modes, de récentes études [**67, 32**] montrent que minimiser la divergence RKL mènerait à des approximations avec des ailes aplaties, alors que la minimiser la divergence FKL mènerait à des approximations avec des ailes plus épaisses. Or, si le but est de calculer une espérance en utilisant une méthode comme *self-nomalized importance sampling*, alors la variance de l'estimateur serait très élevée [**40**]. La méthode *Markovian score climbing* [**46**] permet de minimiser la divergence FKL dans le cas où seule la constante de normalisation de la densité à postériori est incalculable, mais aux meilleures de nos connaissances, aucune technique ne permet d'attaquer le problème lorsque la densité est doublement incalculable. Ce sont tous des avenues de travaux possibles qui restent à explorer.

# Références


[1] A. Albert et J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika **71** (1984), no. 1, 1–10.

[2] Ole Barndorff-Nielsen, *Information and exponential families*, John Wiley & Sons, Ltd, 2014.

[3] Henry Bayly, Aditya Khanna, et Kathryn Lindsey, *Existence of maximum likelihood estimates in exponential random graph models*, 2023.

[4] Shankar Bhamidi, Guy Bresler, et Allan Sly, *Mixing time of exponential random graphs*, 2008 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE.

[5] David M. Blei, Alp Kucukelbir, et Jon D. Mcauliffe, *Variational inference : A review for statisticians*, Journal of the American Statistical Association **112** (2017), no. 518, 859–877.

[6] Lampros Bouranis, Nial Friel, et Florian Maire, *Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods*, Journal of Computational and Graphical Statistics **27** (2018), no. 3, 516–528, doi : 10.1080/10618600.2018.1448832.

[7] A. Caimo et N. Friel, *Bayesian model selection for exponential random graph models*, Social Networks **35** (2013), no. 1, 11–24.

[8] Alberto Caimo et Nial Friel, *Bayesian inference for exponential random graph models*, Social Networks **33** (2011), no. 1, 41–55.

[9] Alberto Caimo et Isabella Gollini, *A multilayer exponential random graph modelling approach for weighted networks*, Computational Statistics & Data Analysis **142** (2020), 106825.

[10] _____, *Recent advances in exponential random graph modelling*, Mathematical Proceedings of the Royal Irish Academy **123A** (2023), no. 1, 1–12.

[11] Sourav Chatterjee et Persi Diaconis, *Estimating and understanding exponential random graph models*, The Annals of Statistics **41** (2013), no. 5.

[12] Ryan DeMuse, Danielle Larcomb, et Mei Yin, *Phase transitions in edge-weighted exponential random graphs : Near-degeneracy and universality*, Journal of Statistical Physics **171** (2018), no. 1, 127–144.

[13] Bruce A. Desmarais et Skyler J. Cranmer, *Micro-level interpretation of exponential random graph models with application to estuary networks*, Policy Studies Journal **40** (2012), no. 3, 402–434.

[14] P. Erdős et A. Rényi, *On random graphs. i.*, Publicationes Mathematicae Debrecen **6** (2022), no. 3-4, 290–297.

[15] Ove Frank et David Strauss, *Markov graphs*, Journal of the American Statistical Association **81** (1986), no. 395, 832–842, doi : 10.1080/01621459.1986.10478342.

[16] Charles J. Geyer, *Markov chain monte carlo maximum likelihood.*

[17] _____, *Likelihood and exponential families*, Thesis, 1990.

[18] _____, *On the convergence of monte carlo maximum likelihood calculations*, Journal of the Royal Statistical Society : Series B (Methodological) **56** (1994), no. 1, 261–274.

[19] Charles J. Geyer et Elizabeth A. Thompson, *Constrained monte carlo maximum likelihood for dependent data*, Journal of the Royal Statistical Society : Series B (Methodological) **54** (1992), no. 3, 657–683.

[20] Jacek Gondzio, *Interior point methods 25 years later*, European Journal of Operational Research **218** (2012), no. 3, 587–601.

[21] Steven M. Goodreau, *Advances in exponential random graph (p*) models applied to a large social network*, Social Networks **29** (2007), no. 2, 231–248.

[22] Ester Gutiérrez-Moya, Sebastián Lozano, et Belarmino Adenso-Díaz, *Analysing the structure of the global wheat trade network : An ergm approach*, Agronomy **10** (2020), no. 12.

[23] Mark S. Handcock, *Assessing degeneracy in statistical models of social networks*, 2003.

[24] W. K. Hastings, *Monte carlo sampling methods using markov chains and their applications*, Biometrika **57** (1970), no. 1, 97–109.

[25] Ran He et Tian Zheng, *Estimation of exponential random graph models for large social networks via graph limits*, Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM.

[26] Langong Hou, Ye Liu, et Xiaoqin He, *Research on the mechanism of regional innovation network in western china based on ergm : A case study of chengdu-chongqing shuangcheng economic circle*, Sustainability **15** (2023), no. 10, 7993.

[27] Ruth M. Hummel, David R. Hunter, et Mark S. Handcock, *Improving simulation-based algorithms for fitting ergms*, Journal of Computational and Graphical Statistics **21** (2012), no. 4, 920–939.

[28] David R. Hunter, *Curved exponential family models for social networks*, Social Networks **29** (2007), no. 2, 216–230.

[29] David R. Hunter, Steven M. Goodreau, et Mark S. Handcock, *Goodness of fit of social network models*, Journal of the American Statistical Association **103** (2012), no. 481, 248–258.

[30] David R. Hunter et Mark S. Handcock, *Inference in curved exponential family models for networks*, Journal of Computational and Graphical Statistics **15** (2006), no. 3, 565–583, doi : 10.1198/106186006X133069.

[31] David R. Hunter, Pavel N. Krivitsky, et Michael Schweinberger, *Computational statistical methods for social network models*, Journal of Computational and Graphical Statistics **21** (2012), no. 4, 856–882.

[32] Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A. Heller, Yian Ma, et Michael I. Jordan, *Variational refinement for importance sampling using the forward kullback-leibler divergence*, 2021, pp. 1819–1829.

[33] Vishesh Karwa, Sonja Petrović, et Denis Bajić, *Dergms : Degeneracy-restricted exponential family random graph models*, Network Science **10** (2022), no. 1, 82–110.

[34] Jae Hun Kim et Han Woo Park, *Identifying networked patterns in memecoin twitter accounts using exponential random graph modeling*, IT Professional **25** (2023), no. 6, 82–89.

[35] Diederik P. Kingma et Jimmy Ba, *Adam : A method for stochastic optimization*, 2017.

[36] D. A. Levin, Y. Peres, et Elizabeth L. Wilmer, *Markov chains and mixing times*, vol. 58, AMS, 2009.

[37] Juan Li, Keyin Liu, Zixin Yang, et Yi Qu, *Evolution and impacting factors of global renewable energy products trade network : An empirical investigation based on ergm model*, Sustainability **15** (2023), no. 11, 8701.

[38] Faming Liang, Ick Hoon Jin, Qifan Song, et Jun S. Liu, *An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants*, Journal of the American Statistical Association **111** (2016), no. 513, 377–393.

[39] Jingfang Liu et Yu Zeng, *Exploring user interaction patterns in an online physician interactive community based on exponential random graph models*, Humanities and Social Sciences Communications **11** (2024), no. 1.

[40] Jun S. Liu, *Monte carlo strategies in scientific computing*, Springer, 2004.

[41] Angelo Mele et Lingjiong Zhu, *Approximate variational estimation for a model of network formation*, The Review of Economics and Statistics **105** (2023), no. 1, 113–124.

[42] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, et Edward Teller, *Equation of state calculations by fast computing machines*, The Journal of Chemical Physics **21** (1953), no. 6, 1087–1092.

[43] Sumit Mukherjee et Yuanzhe Xu, *Statistics of the two star ergm*, Bernoulli **29** (2023), no. 1, 24–51, 28.

[44] Iain Murray, Zoubin Ghahramani, et David J. C. MacKay, *Mcmc for doubly-intractable distributions*, 2006, p. 359–366.

[45] J. Møller, A. N. Pettitt, R. Reeves, et K. K. Berthelsen, *An efficient markov chain monte carlo method for distributions with intractable normalising constants*, Biometrika **93** (2006), no. 2, 451–458.

[46] Christian A. Naesseth, Fredrik Lindsten, et David Blei, *Markovian score climbing : variational inference with kl(p||q)*, 2020, p. Article 1300.

[47] Esa Nummelin, *General irreducible markov chains and non-negative operators*, Cambridge University Press, Cambridge, 1984.

[48] James Gary Propp et David Bruce Wilson, *Exact sampling with coupled markov chains and applications to statistical mechanics*, Random Structures & Algorithms **9** (1996), no. 1-2, 223–252.

[49] Sa Ren, Xue Wang, Peng Liu, et Jian Zhang, *Bayesian nonparametric mixtures of exponential random graph models for ensembles of networks*, Social Networks **74** (2023), 156–165.

[50] Herbert Robbins et Sutton Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics **22** (1951), no. 3, 400–407, 8.

[51] Garry Robins, Pip Pattison, et Peng Wang, *Closure, connectivity and degree distributions : Exponential random graph (p*) models for directed social networks*, Social Networks **31** (2009), no. 2, 105–117.

[52] Thomas J. Santner et Diane E. Duffy, *A note on a. albert and j. a. anderson's conditions for the existence of maximum likelihood estimates in logistic regression models*, Biometrika **73** (1986), no. 3, 755–758.

[53] Christian S. Schmid et David R. Hunter, *Improving ergm starting values using simulated annealing*, Social Networks **76** (2024), 209–214.

[54] Tyler A. Scott, *Analyzing policy networks using valued exponential random graph models : Do government-sponsored collaborative groups enhance organizational networks ?*, Policy Studies Journal **44** (2016), no. 2, 215–244.

[55] Matthew J. Silk et David N. Fisher, *Understanding animal social structure : exponential random graph models in animal behaviour research*, Animal Behaviour **132** (2017), 137–146.

[56] Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, et Mark S. Handcock, *New specifications for exponential random graph models*, Sociological Methodology **36** (2006), no. 1, 99–153.

[57] Tom A.B. Snijders, *Markov chain monte carlo estimation of exponential random graph models*, Journal of Social Structure **3** (2002), 1–40.

[58] Alex D. Stivala, Johan H. Koskinen, David A. Rolls, Peng Wang, et Garry L. Robins, *Snowball sampling for estimating exponential random graph models for large networks*, Social Networks **47** (2016), 167–188.

[59] David Strauss et Michael Ikeda, *Pseudolikelihood estimation for social networks*, Journal of the American Statistical Association **85** (1990), no. 409, 204–212.

[60] Yun Ju Sung et Charles J. Geyer, *Monte carlo likelihood inference for missing data models*, The Annals of Statistics **35** (2007), no. 3, 990–1011.

[61] Linda S. L. Tan et Nial Friel, *Bayesian variational inference for exponential random graph models*, Journal of Computational and Graphical Statistics **29** (2020), no. 4, 910–928.

[62] S. Thiemichen, N. Friel, A. Caimo, et G. Kauermann, *Bayesian exponential random graph models with nodal random effects*, Social Networks **46** (2016), 11–28.

[63] Luke Tierney, *Markov chains for exploring posterior distributions*, The Annals of Statistics **22** (1994), no. 4, 1701–1728.

[64] Petro Tolochko et Hajo G. Boomgaarden, *Same but different : A comparison of estimation approaches for exponential random graph models for multiple networks*, Social Networks **76** (2024), 1–11.

[65] George G. Vega Yon, Andrew Slaughter, et Kayla de la Haye, *Exponential random graph models for little networks*, Social Networks **64** (2021), 225–238.

[66] Peng Wang, Garry Robins, Philippa Pattison, et Emmanuel Lazega, *Exponential random graph models for multilevel networks*, Social Networks **35** (2013), no. 1, 96–115.

[67] Yuling Yao, Aki Vehtari, Daniel Simpson, et Andrew Gelman, *Yes, but did it work ? : Evaluating variational inference*, 2018, pp. 5581–5590.

[68] Fei Yu, Helal M. El-Zaatari, Michael R. Kosorok, Andrea Carnegie, et Gaurav Dave, *The application of exponential random graph models to collaboration networks in biomedical and health sciences : a review*, Network Modeling Analysis in Health Informatics and Bioinformatics **13** (2024), no. 1.

[69] Wayne W. Zachary, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research **33** (1977), no. 4, 452–473.

[70] Qian Zhang et Faming Liang, *Bayesian analysis of exponential random graph models using stochastic gradient markov chain monte carlo*, Bayesian Analysis **-1** (2023), no. -1.

[71] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, et Jorge Nocedal, *Algorithm 778 : L-bfgs-b*, ACM Transactions on Mathematical Software **23** (1997), no. 4, 550–560.