Bachelor Project Report

Author:
Salah-Eddine Al Khiati
Supervisor:
Charles Dufour

École Polytechnique Fédérale de Lausanne

Report on Graphon and Signal Estimation

Salah-Eddine Al Khiati
salah.alkhiati@epfl.ch
École Polytechnique Fédérale de Lausanne, EPFL

December 30, 2024

Abstract

This paper presents an overview of the key theoretical results in the graphon estimation problem, with a particular focus on the stochastic block model framework. We extend this exploration to related challenges in signal estimation and the joint graphon-signal estimation task, providing analogous theoretical insights. Additionally, we adapt and evaluate a few existing methods to address the graphon-signal estimation problem, and apply them to real world data.

1 Introduction

In the last few years, graphons, a notion of graph limits [45], have emerged as a powerful unifying framework to understand graph behavior across several disciplines including biology and epidemiology [25], machine learning with graph neural networks [61], game theory [55] and more [74]. Most importantly, in network analysis [12], they have been used to generalize a number of random graph models including the stochastic block model [66], the latent space model [34] and more [33; 6; 44]. Naturally, the problem of estimating the graphon that parametrizes a random graph has recently gained a lot of momentum [75; 3; 17; 29]. Numerous studies have developed numerical methods [3; 19; 17] and derived optimal rates [29; 21; 39] for estimating graphons, yielding significant insights and practical tools in the field of network analysis [53; 80; 4]. However research on estimating the graphon-signal, an analogue to the graphon for random graph-signals, is still scarce. This problem is especially relevant in applications where noisy node signals/features/attributes are observed alongside the network structure, but also serves to study how these elements relate within the network.

This paper serves as a very brief overview of the research being done in the fields of graphon and signal estimation. We will first focus on giving the theoretical results and guarantees such as optimal rates and consistency. Then, we will propose practical methods adapted from previously tested state of the art approaches, and test them on synthetic and real data.

2 Preliminaries

A sequence of random variables $X = (X_1, X_2, \dots)$ is said to be exchangeable when for any permutation of the indices σ, the joint distribution of the sequence $X^\sigma = (X_{\sigma(1)}, \dots)$

stays invariant. De Finetti's theorem states, in brief, that for any infinite sequence of exchangeable random variables X, there exists some family of distributions $\{P_\theta\}_{\theta \in T}$ indexed by some parameter space T and a distribution on said space ν, such that: $P(X \in \cdot) = \int_T P_\theta^\infty(\cdot)\nu(d\theta)$. Where P_θ^∞ is the distribution of an infinite sequence following P_θ. This can be interpreted as first sampling some parameter $\Theta = (\theta_1, \theta_2, \dots) \sim \nu$ and then sampling $X_1, X_2, \dots | \Theta \overset{i.i.d}{\sim} P_\Theta$. If we take everything to be on $\mathbb{R}$ we can write F to be the "quantile" function of ν, so sampling $\Theta \sim \nu$ is the same as sampling $U_1, U_2, \dots \overset{i.i.d}{\sim} \mathrm{Unif}([0,1])$ and then taking $\Theta = (F(U_1), F(U_2), \dots)$. Therefore, one can characterize such exchangeable data by some underlying function with support on $[0,1]$ and with values on the parameter space. Such a result can be generalized to exchangeable arrays of two dimension with the Aldous-Hoover theorem, and even higher dimensional arrays (Kallenberg (2005)). For two dimensional symmetric binary arrays such a model describes exchangeable graphs with an underlying symmetric function called a graphon $w : [0,1]^2 \longrightarrow [0,1]$ with values in the parameter space of the Bernoulli family $\{\mathrm{Bern(p)}\}_{p \in [0,1]}$. In one dimension if we take our distribution family to be the gaussian location family $\{\mathcal{N}(\mu, 1)\}_{\mu \in \mathbb{E}}$, we call the underlying function giving values in the parameter space the signal $f : [0,1] \longrightarrow \mathbb{R}$. The signal appellation is due to our later consideration of this function to be a value function of graph nodes. This brief presentation lacks in rigor and failed to describe other subtleties, we refer to Orbanz and Roy (2015) for an excellent overview of the subject. On a parenthetic note, one can construct even more general latent variable models than the graphon (Jaeger et al. (2024); Wu et al. (2025)). But in this paper, we do not concern ourselves in studying such probabilistic symmetries and latent models, instead we directly describe our problems in the most convenient formulation in the next section.

2.1 Problem formulation

In the following, we will formulate each of the problems we will treat separately:

Problem 1. First we state the standard graphon estimation problem. Consider a simple, undirected graph G with n nodes, represented by its adjacency matrix A. Each node i is associated with a latent variable $\xi_i \overset{i.i.d}{\sim} \mathrm{Unif}[0,1]$, and the edges $\{A_{ij}\}$ are sampled as follows:

$$A_{ij} \mid \xi_i, \xi_j \overset{i.i.d}{\sim} \mathrm{Bern}(w(\xi_i, \xi_j)) \ , \ i < j = 1, \dots, n.$$

Where $w : [0,1]^2 \to [0,1]$ is a symmetric measurable function called a graphon. We call the set of such function $\mathcal{W}$. We aim to estimate w.

Problem 2. Now we state the signal estimation problem. Consider a sequence of n random variables $X_1, \dots, X_n$ each associated to a random variable $\xi_i \overset{i.i.d}{\sim} \mathrm{Unif}[0,1]$ and sampled as follows:

$$X_i \mid \xi_i \overset{i.i.d}{\sim} \mathcal{N}(f(\xi_i), 1) \ , \ i = 1, \dots, n.$$

Where $f : [0,1] \to \mathbb{R}$ is a measurable bounded function. We call the set of such functions $\mathcal{F}$, For some $r > 0$ we also write $\mathcal{F}_r = \{f \in \mathcal{F} : |f(x)| \leq r, \forall x \in [0,1]\}$. We aim to estimate f.

Problem 3. Now we describe what we will refer to as the joint estimation problem. Consider again a simple, undirected graph G with n nodes, represented by its adjacency matrix A, along with a set of node signals $X = (X_1, X_2, \ldots, X_n)$. Each node i is associated with a latent variable $\xi_i \overset{i.i.d}{\sim} \text{Unif}[0,1]$. The edges $\{A_{ij}\}$ and node signals $\{X_i\}$ are independent from each other and are sampled as follows:

$$A_{ij} \mid \xi_i, \xi_j \overset{i.i.d}{\sim} \text{Bern}(w(\xi_i, \xi_j)) \ , \ X_i \mid \xi_i \overset{i.i.d}{\sim} \mathcal{N}(f(\xi_i), 1) \ , \ i < j = 1, \ldots, n.$$

Where $w \in \mathcal{W}$ and $f \in \mathcal{F}$. The pair (w, f) is called a graphon-signal. We call the set of such pairs $\mathcal{WF}$ and again for $r > 0$ we write $\mathcal{WF}_r = \{(w, f) \in \mathcal{WF} : f \in \mathcal{F}_r\}$. We aim to estimate (w, f).

In the previous problem formulations, when it's stated that we aim to estimate w, f or (w, f) it is not very clear what it means as we first need to define a notion of distance to what we wish to estimate. The next section presents the various norms and distances we will be using and further clarifies our goal.

2.2 Norms and Distances

First we define the norms for the graphon.

Definition 2.1. *For any $w \in \mathcal{W}$:*

$$\begin{aligned} \|w\|_1 &= \int_{[0,1]^2} |w(x,y)| dxdy, \\ \|w\|_2 &= \left(\int_{[0,1]^2} w(x,y)^2 dxdy \right)^{\frac{1}{2}}, \\ \|w\|_\square &= \sup_{S,T \subset [0,1]} \left| \int_{S \times T} w(x,y) dxdy \right|. \end{aligned} \tag{1}$$

We have :

$$\begin{aligned} &\|w\|_\square \leq \|w\|_1 \leq \|w\|_2, \\ &\|w\|_2 \leq \sqrt{\|w\|_1}. \ \textit{(due to } w \leq 1\textit{)} \end{aligned}$$

Aside from the standard L_1 and L_2 norms we also defined the cut norm (1). This norm emerged in graph limit theory as the right choice (Borgs et al. (2007)). One can see what it means in such a context by considering w to be a "graph" and S, T to be subsets of nodes. Then the integral in (1) counts the edges between the nodes in the sets S and T.

As can be seen in the problem description above, two point-wise different graphons can parametrize the same random graph. Specifically, two graphon only differing by a measure preserving rearrangement. It's what motivated the use of the following distances.

Definition 2.2. *For any $w_1, w_2 \in \mathcal{W}$ and any norm N:*

$$\delta_N(w_1, w_2) = \inf_{\sigma \in \mathcal{M}} \|w_1 - w_2^\sigma\|_N,$$

where $\mathcal{M}$ is the set of measure preserving bijections from $[0,1]$ to $[0,1]$.
We have :

$$\begin{aligned}\delta_{\square}(w_1,w_2) &\leq \delta_1(w_1,w_2) \leq \delta_2(w_1,w_2), \qquad (2)\\ \delta_2(w_1,w_2) &\leq \sqrt{\delta_1(w_1,w_2)}.\end{aligned}$$

As was discussed in Janson (2011) and Orbanz and Roy (2015) we can use the distances defined above to define a quotient space of $\mathcal{W}$ where two graphons are identified if their δ distance is null or, equivalently, if they parametrize the same random graph.

We take an analogue approach for the signal and define our norms:

Definition 2.3. *For any $f \in \mathcal{F}$:*

$$\begin{aligned}\|f\|_1 &= \int_{[0,1]} |f(x)|dx,\\ \|f\|_2 &= \left(\int_{[0,1]} f(x)^2 dx\right)^{\frac{1}{2}},\\ \|f\|_{\square} &= \sup_{S\subset[0,1]} \left|\int_S f(x)dx\right|.\end{aligned}$$

We have :

$$\|f\|_{\square} \leq \|f\|_1 \leq \|f\|_2.$$

If $f \in \mathcal{F}_r$:

$$\|f\|_2 \leq \sqrt{r}\sqrt{\|f\|_1}.$$

Proposition 2.1 (Levie (2023))**.** *If $f \in \mathcal{F}_r$:*

$$\|f\|_{\square} \leq \|f\|_1 \leq 2\|f\|_{\square} \qquad (3)$$

As for the graphon a measure preserving rearrangement of f parametrizes the same random sequence. So we define our distances similarly:

Definition 2.4. *For any $f_1, f_2 \in \mathcal{F}$ and any norm N:*

$$\begin{aligned}\delta'_N(f_1,f_2) &= \inf_{\sigma\in\mathcal{M}} \|f_1 - f_2^{\sigma}\|_N,\\ &= \inf_{\sigma,\phi\in\mathcal{M}} \|f_1^{\phi} - f_2^{\sigma}\|_N.\end{aligned}$$

(by a simple change of measure, see Bogachev (2007) theorem 3.6.1)
We have :

$$\begin{aligned}\delta'_{\square}(f_1,f_2) &\leq \delta'_1(f_1,f_2) \leq \delta'_2(f_1,f_2), \qquad (4)\\ \delta'_{\square}(f_1,f_2) &\leq \delta'_1(f_1,f_2) \leq 2\delta'_{\square}(f_1,f_2). \qquad (5)\end{aligned}$$

If $f_1, f_2 \in \mathcal{F}_r$:

$$\delta'_2(f_1,f_2) \leq \sqrt{2r}\sqrt{\delta'_1(f_1,f_2)}.$$

Finally for our joint norms and distances:

Definition 2.5. *For any $(w, f) \in \mathcal{WF}$ and for any norm N:*

$$\|(w, f)\|_N = \|w\|_N + \|f\|_N.$$

Again, we see that applying a measure preserving bijection jointly to (w, f) will parametrize the same random graph-signal:

Definition 2.6. *For any $(w_1, f_1), (w_2, f_2) \in \mathcal{WF}$ and norm N:*

$$\delta''_N((w_1, f_1), (w_2, f_2)) = \inf_{\sigma \in \mathcal{M}} \|(w_1 - w_2^\sigma, f_1 - f_2^\sigma)\|_N.$$

We have:

$$\delta''_\square((w_1, f_1), (w_2, f_2)) \leq \delta''_1((w_1, f_1), (w_2, f_2)) \leq \delta''_2((w_1, f_1), (w_2, f_2)).$$

2.2.1 Step Functions and Regularity Lemmas

In the rest of this paper we will mainly work with step-functions as defined in what follows:

Definition 2.7. *We say $f \in \mathcal{F}$ is a k-step function if there exists a partition $\{I_a\}_{a \leq k}$ of $[0, 1]$ and $M = (M_1, \ldots, M_k) \in \mathbb{R}$ such that:*

$$f(x) = \sum_{a \leq K} M_a \mathbb{I}\{x \in I_a\}.$$

We note the space of such function $\mathcal{F}_k$ and similarly write $\mathcal{F}_{k,r} = \mathcal{F}_k \cap \mathcal{F}_r$.

Definition 2.8. *We call $w \in \mathcal{W}$ a k-step graphon if there exists a partition $\{I_a\}_{a \leq k}$ of $[0, 1]$ and $Q \in [0, 1]^{k \times k}_{sym}$ such that:*

$$w(x, y) = \sum_{a, b \leq k} Q_{ab} \mathbb{I}\{x \in I_a \times I_b\}.$$

We call the set of such functions $\mathcal{W}_k$

Definition 2.9. *We call $(w, f) \in \mathcal{WF}$ a k-step graphon-signal if there exists a partition $\{I_a\}_{a \leq k}$ of $[0, 1]$ and $Q \in [0, 1]^{k \times k}_{sym}$ and $M_1, \ldots, M_k \in \mathbb{R}$ such that:*

$$w(x, y) = \sum_{a, b \leq k} Q_{ab} \mathbb{I}\{x \in I_a \times I_b\} \ , \ f(x) = \sum_{a \leq K} M_a \mathbb{I}\{x \in I_a\}.$$

We call the set of such pairs $\mathcal{WF}[k]$ and similarly $\mathcal{WF}_r[k] = \mathcal{WF}_r \cap \mathcal{WF}[k]$.

To motivate our restriction to step functions, we cite the following regularity lemmas:

Theorem 2.10 (Lovász and Szegedy (2007))**.** *For every graphon $w \in \mathcal{W}$ and $\epsilon > 0$ there exists $w' \in \mathcal{W}_k$ with $k \leq \lceil 2^{\frac{2}{\epsilon^2}} \rceil$ such that:*

$$\delta_\square(w, w') \leq \epsilon.$$

This result is an analytic equivalent to Szemerédi's regularity lemma which states, in simple terms, that for every very large graph there is a small weighed graph that summarizes most of its structure. In analytic terms, it means that for any graphon we can find a step graphon which describes most of it's structure up to any error depending on the number of blocks. There is a similar result for graphon-signals.

We first give this basic result on signals alone:

Proposition 2.2. *For any $f \in \mathcal{F}_r$ and $\epsilon > 0$ there exists $f' \in \mathcal{F}_{k,r}$ with $k = \lceil \frac{r}{\epsilon} \rceil$ such that:*

$$\delta'_{\square}(f, f') \leq \epsilon.$$

Which was used in Levie (2023) to get:

Theorem 2.11 (Levie (2023))**.** *For any sufficiently small $\epsilon > 0$ and $(w, f) \in \mathcal{WF}_r$ there exists $(w', f') \in \mathcal{WF}_r[k]$ with $k \leq \lceil 2^{\frac{18}{4\epsilon^2}} \rceil$ such that:*

$$\delta''_{\square}((w, f), (w', f')) \leq \epsilon.$$

This can be interpreted in the same way as the regularity lemma for graphons. The structure of a graphon-signal can be summarized by a step graphon-signal up to any error with a sufficient number of blocks. The importance of these results is that any graphon/graphon-signal can be approximated by a stochastic block model which provides far more structure and tools for inference.

2.2.2 Stochastic Block Model

. The stochastic block model SBM with k blocks we will consider is equivalent to the model described in *Problem 1* with a k-step graphon and where we use a community assignment function $\mathcal{Z}_{n,k} \ni z : [n] \longrightarrow [k]$ to model the latent ξ_i's. We can write: $z(i) = a$ if and only if $\xi_i \in I_a$. Note that there are many other variants and extensions of the SBM (De Nicola et al. (2022)). We define the SBM parameter space:

$$\Theta_k = \{\{\theta_{ij}\} \in \mathbb{R}^{n\times n} | \theta_{ii} = 0,\ \theta_{ij} = Q_{ab} = Q_{ba} \text{ for some } \{Q_{ab}\} \in [0,1]^{k\times k} \text{ and } z \in \mathcal{Z}_{n,k}\}.$$

For the signal alone we define an analogue model with the parameter space:

$$\mathcal{U}_k = \{\{\mu_i\} \in \mathbb{R}^n | \mu_i = M_{z(i)} \text{ for some } \{M_a\} \in \mathbb{R}^k \text{ and } z \in \mathcal{Z}_{n,k}\}.$$

And for the graphon-signal joint case, known in literature as contextual SBM (CSBM):

$$\begin{aligned}\Theta\mathcal{U}[k] = \{(\{\theta_{ij}\}, \{\mu_i\}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n | \theta_{ii} = 0,\ \theta_{ij} = Q_{ab} = Q_{ba} \text{ for some } \{Q_{ab}\} \in [0,1]^{k\times k},\\ \mu_i = M_{z(i)} \text{ for some } \{M_a\} \in \mathbb{R}^k \text{ and } z \in \mathcal{Z}_{n,k}\}.\end{aligned}$$

In what follows we will also consider mixture weights in our model and write them as $\pi_a = |I_a|$. In a sens we are considering random assignments with $P(z(i) = a \mid \pi) = \pi_a$. This mixture model will prove useful for a joint estimation method we propose later. But it will be most relevant for the study of the signal problem alone. We introduce in the following section some definitions and results to that end.

2.2.3 Mixing distributions and Wasserstein distance

Definition 2.12. *We call a k-component mixing distribution any measure on $\mathbb{R}$ of the form :*

$$G = \sum_{i=1}^{k} \pi_i \delta_{q_i},$$

where δ_x is the Dirac measure centered at x and the weights π_i and centers q_i are such that $\forall i : \pi_i \geq 0, q_i \in \mathbb{R}$ and $\sum_i \pi_i = 1$. We call the set of such distributions $\mathcal{G}_k$. If we take the centers to be bounded by $M > 0$, we write $\mathcal{G}_{k,M}$.

The p-Wasserstein distance between two distributions G, G' is defined as:

$$W_p(G, G') = \inf_{\Pi} \left[\int_{\mathbb{R}^2} |x - y|^p d\Pi(x, y) \right]^{\frac{1}{p}},$$

where the infimum is taken over probability measures on $\mathbb{R}^2$ with marginals G and G' (Villani (2009)). We will focus on the $p = 1$ case and we will mostly use this more direct formulation:

$$W_p^p(G, G') = \int_{[0,1]} |F_G^{-1}(t) - F_{G'}^{-1}(t)|^p dt,$$

where F_G is the cumulative distribution function CDF of G and F_G^{-1} is the generalized inverse CDF or quantile function of G.

To better understand this distance, if we take two mixing distributions G, G' with the same weights equal to $\frac{1}{k}$. Then we have:

$$W_1(G, G') = \min_{\sigma \in S_k} \frac{1}{k} \sum_i |q_i - q'_{\sigma(i)}|.$$

The following proposition bridges mixture model theory with our problem:

Proposition 2.3. *For any f_1, f_2 step functions, we have that:*

$$\delta'_1(f_1, f_2) = W_1(G_{f_1}, G_{f_2}),$$
$$\delta'_2(f_1, f_2) = W_2(G_{f_1}, G_{f_2}),$$

where $G_f \in \mathcal{G}_k$ is the mixing distributions with weights corresponding to the measure of the blocks of f and as centers the corresponding block means, for any $f \in \mathcal{F}_k$

There exists a similar relation with the graphon δ distances, we refer to Janson (2011). An application of that relation was done in Xu et al. (2020) and Ponti (2024).

This proposition entails that our problem of estimating f is equivalent to the problem of estimating the parameters of a mixture model. It can also be seen as a de-convolution problem since we are only considering Gaussian means as parameters, and we can write:

$$\{X_i\} \overset{iid}{\sim} p_{\mathcal{N}(0,1)} * G_f,$$

where $p_{\mathcal{N}(0,1)}$ is the density of the standard normal. We can consistently estimate $p_{\mathcal{N}(0,1)} * G_f$ by some strong law of large numbers and with a parametric rate (see inequality in Massart (1990)). But de-convolving it to recover G_f is where the difficulty lies. A lot of proofs in mixture model theory revolve around that approach (see Heinrich and Kahn (2018) and their other work).

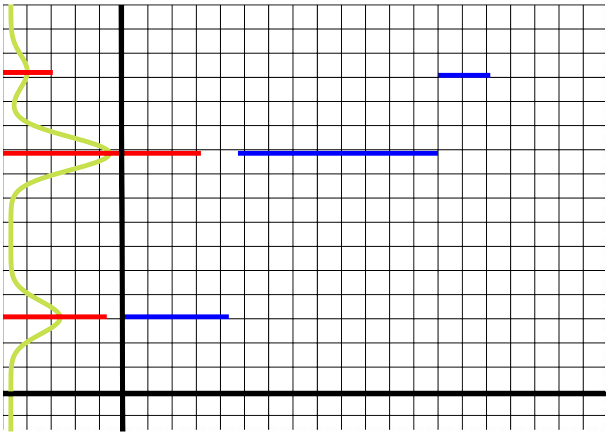

Figure 1: The step function f in blue, the mixing distributions G_f in red and the observed convolution with the gaussian location family in yellow

2.3 Notation

We state here some notation for the reader to fall back to if it is unfamiliar or unclear from context : We write $\|\cdot\|_2$ for the L_2 norm, when used with a matrix, it's the Frobenius norm. When writing a constant C that constant doesn't depend on the sample size n, we may add a subscript C_p when the constant depends on some parameter p. The symbol $a \asymp b$ means $\exists C, C' > 0 : Cb \leq a \leq C'b$, where the constants C, C' do not depend on the sample size. If we write $\asymp_k$ those constants depend on some parameter k. The big O, small o and Ω notation are with respect to (w.r.t.) sample size. When we write $\mathbb{E}_\theta$ then we are taking the expectation conditional on some θ. a.e means almost everywhere, $\xrightarrow{a.s}$ is convergence almost surely, $\xrightarrow{d}$ is convergence in distribution. we use $p(\cdot)$ to denote densities in general. $|E|$ is the measure of the set E w.r.t. Lebesgue outer measure. All integrals are taken with Lebesgue outer measure.

3 Optimal rates and consistency

In this section we present the optimal rates w.r.t. sample size for our estimation problems, along with some other results. We will constrain ourselves to step functions, i.e the SBM. There exists results for Holder smooth graphons (Gao et al. (2015)) and recently for graphons with spectral decay (Chen and Lei (2024)), but we see the step function context to be the more fundamental and interpretable one as discussed previously. Furthermore, the literature is very scarce concerning results in a smooth context (or else) for the signal case and are not as easily generalizable from the SBM as for the graphon, we will see later why that is.

There are generally two formulations for the graphon estimation problem. The first is to estimate the probability matrix θ, this is usually done with respect to the error function $\frac{1}{n^2}\|\theta - \hat{\theta}\|_2^2$ for some estimator $\hat{\theta}$. The second is to estimate the equivalence class of the graphon, that is construct some estimator $\hat{w}$ as to minimize the various δ-distances defined previously. For the signal case we will proceed analogously. We will first study the problem of estimating the means vector μ w.r.t the error $\frac{1}{n}\|\mu - \hat{\mu}\|_2^2$ for some estimator $\hat{\mu}$. Then we will tackle the problem of estimating f in the equivalence class. Afterward, we will discuss and explore how to tackle the joint problem.

3.1 Graphon Estimation Alone

In this section we put ourselves in the context of *Problem 1* with the SBM. We will list some of the theoretical results on graphon estimation.

First we would like to briefly define the estimator that yielded most of the following optimal rates. For any (Q, z) as in the SBM, define the loss function:

$$\mathcal{L}(Q, z) = \sum_{i,j} |A_{ij} - Q_{z(i)z(j)}|^2.$$

Take:

$$(\hat{Q}, \hat{z}) = arg \min_{(Q,z)} \mathcal{L}(Q, z).$$

Then our estimator is:

$$\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}.$$

Starting with optimal rates for the probability matrix, we have the following result:

Theorem 3.1 (Gao et al. (2015)). *For $1 \leq k \leq n$, we have:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k} \mathbb{E}_\theta \left[\frac{1}{n} \|\hat{\theta} - \theta\|_2^2 \right] \asymp \frac{k^2}{n^2} + \frac{\log(k)}{n}.$$

The upper-bound is verified for the least square estimator defined previously.

Now we give result for the step graphons using the δ distances. These results are largely drawn from Klopp and Verzelen (2019) and Klopp et al. (2017), simplified for our case. The least square graphon estimator is defined from the same previous estimator as:

$$\hat{w}(x, y) = \hat{Q}_{\hat{z}(\lceil nx \rceil)\hat{z}(\lceil ny \rceil)}.$$

Theorem 3.2 ((Klopp and Verzelen, 2019; Klopp et al., 2017)). *For $2 \leq k \leq n$, we have for the cut distance:*

$$\inf_{\hat{w}} \sup_{w \in \mathcal{W}_k} \mathbb{E}_w \left[\delta_\square(\hat{w}, w) \right] \asymp \sqrt{\frac{1}{n}} + \sqrt{\frac{k}{n \log(k)}}.$$

The upper-bound is achieved by the empirical graphon. For the δ_2 distance:

$$\inf_{\hat{w}} \sup_{w \in \mathcal{W}_k} \mathbb{E}_w \left[\delta_2(\hat{w}, w) \right] \asymp \frac{k}{n} + \sqrt{\frac{\log(k)}{n}} + \left(\frac{k}{n} \right)^{\frac{1}{4}}.$$

The upper-bound is achieved by the graphon least square estimator. And finally for the δ_1 distance:

$$\inf_{\hat{w}} \sup_{w \in \mathcal{W}_k} \mathbb{E}_w \left[\delta_1(\hat{w}, w) \right] \geq C_1 \left(\frac{k}{n} + \sqrt{\frac{1}{n}} + \sqrt{\frac{k}{n}} \right).$$

$$\inf_{\hat{w}} \sup_{w \in \mathcal{W}_k} \mathbb{E}_w \left[\delta_1(\hat{w}, w) \right] \leq C_2 \left(\frac{k}{n} + \sqrt{\frac{\log(k)}{n}} + \sqrt{\frac{k}{n}} \right).$$

This upper-bound is also achieved by the least square estimator.

The optimal rates for the δ_2 and δ_1 distances are the same as the ones we see for the discrete MSE (with roots taken) but with an added "agnostic error" which is the error between the true graphon and the graphon constructed from the sampled probability matrix. The only difference is that the agnostic error $\left(\frac{k}{n}\right)^{\frac{1}{4}}$ for the δ_2 norm is smaller than $\sqrt{\frac{k}{n}}$ for the δ_1 error, which is explained by (2). The optimal rate for the $\delta_\square$ distance is achieved by the empirical graphon estimator, which makes sens as this distance was used in graph limit theory where the graphon is the limit of the random graph it parametrizes. For more on the differenc between the L_1/L_2 distance and the cut distance we refer to Cai et al. (2015) appendix A.

The least square estimator described above is not usable in practice. But it can be shown to be equivalent to the maximum likelihood estimator (Wolfe and Olhede (2013)), which can be well estimated using variational expectation maximization methods (Celisse et al. (2012)). And indeed, Gaucher and Klopp (2021) proved that the variational estimator is minimax optimal and achieves the rates presented above. We will discuss this method in later sections.

3.2 Signal Estimation Alone

We put ourselves in the context of *Problem 2*. First we start with estimating the means vector. It is equivalent to a very popular problem known under different names. It is treated in Tsybakov (2009) as the gaussian sequence model and in Wasserman (2006) as the normal means problem. By letting the variance term vary with n, this model can be shown to be equivalent to a lot of non-parametric problems as discussed in the aforementioned books, making for very fundamental and informative literature. We refer to the introduction in Neykov (2022) and Biscarri (2019) for more details. In our case, we fix the variance to be 1. In this context the minimax rate turns out to be constant, as discussed in Gao et al. (2015) section 3.2.

Theorem 3.3. *Take our parameter space to be $\{q_1, q_2\}^n \in \mathbb{R}^n$, if $q_1 \neq q_2$ we have :*

$$\inf_{\hat{\mu}} \sup_{\mu \in \{q_1,q_2\}^n} \mathbb{E}_\xi \left[\frac{1}{n}\|\hat{\mu} - \mu\|^2\right] \asymp 1.$$

Where we are taking the infimum over all estimators (all measurable functions with values in $\mathbb{R}$).

From the above, one easily gets, simply by observing that the two constants parameter space is smaller than $\mathcal{U}_k$:

Corollary 3.4.

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{U}_k} \mathbb{E}_\mu \left[\frac{1}{n}\|\hat{\mu} - \mu\|^2\right] \asymp 1.$$

Remark 3.5. *We could derive much better minimax bounds dependent on various parameters such as parameter space and variance (if it wasn't fixed). We refer again to Neykov (2022).*

We can also interpret this problem as sampling a single observation of a n-multivariate gaussian, in that case the optimal estimator would be the empirical mean $\hat{\mu} = X$.

In our context, this constant error is likely due to clustering. In the graphon case the dimensionality of the problem smooths out this error which does not happen in this case. To see that the clustering error is indeed constant we give a result from Lu and Zhou (2016). First we define the parameter space:

$$\begin{aligned}\mathcal{U}_{\Delta,\alpha,k} = \{(M,z) | M = (M_1,\dots,M_k) \in \mathbb{R}^k,\ z:[n]\to[k],\\ \Delta \le \min_{a\neq b}|M_a - M_b|,\ |z^{-1}(a)| \ge \alpha n, \forall a \in [k]\}.\end{aligned}$$

And we define the error function:

$$l(z',z) = \inf_{\sigma\in S_k} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\sigma(z'(i)) \neq z(i)\}.$$

Theorem 3.6 (Lu and Zhou (2016))**.** *With k fixed, if $\alpha \asymp \frac{1}{k}$ and $\Delta \to \infty$ as n grows. Then we have:*

$$\inf_{\hat z} \sup_{(z,M)\in\mathcal{U}_{\Delta,\alpha,k}} \mathbb{E}_{(z,M)}\left(l(\hat z, z)\right) \asymp \exp\left(-(1+o(1))\frac{\Delta^2}{8}\right).$$

If instead $\Delta = O(1)$, then we have:

$$\inf_{\hat z} \sup_{(z,M)\in\mathcal{U}_{\Delta,\alpha,k}} \mathbb{E}_{(z,M)}\left(l(\hat z, z)\right) \ge c,$$

for some constant $c > 0$. And where we take the infimum over all estimators $\hat z$ (measurable functions taking values in $[k]^n \subset \mathbb{R}^n$).

Remark 3.7. *The upper-bound for this method is achieved by $\hat z$ acquired through Lloyd's algorithm under certain conditions.*

We see from this result that one way to get consistency is to take a separation between our block means Δ that diverges with n. This assumption, however, is not very realistic for our intended applications. We must note that this minimax loss formulation is pessimistic and that strong recovery of the labels is possible under other similar conditions, we refer to Awasthi and Sheffet (2012) for more on clustering with gaussian mixtures.

We show how the clustering error contributes:

$$\frac{1}{n}\|\hat\mu - \mu\|^2 = \frac{1}{n}\sum_i \mathbb{I}\{\hat z(i) = z(i)\}(\hat M_{z(i)} - M_{z(i)}) + \frac{1}{n}\sum_i \mathbb{I}\{\widehat{z(i)} \neq z(i)\}(\hat M_{\hat z(i)} - M_{z(i)}).$$

As we see, even if we get an optimal estimator for the block means, if they are separated, the clustering error contributes a constant term. In the graphon case the clustering error term is certainly also constant but gets multiplied by a dimension normalizing factor $\frac{1}{n}$ making it decrease at an optimal rate. This corresponds to the $\frac{\log(k)}{n}$ term seen previously. Such an analysis is supported by Choi et al. (2012), Massoulié (2014) and Mossel et al. (2016). We refer to Gao and Ma (2021) for an overview of the subject.

All this entails that any methods that would estimate the signal through an estimation of the means vector μ are probably to be avoided if we wish to prove consistency w.r.t the δ' distances. Sadly, most of the literature for graphon estimation does exactly that as we saw previously, and so we cannot adapt it directly.

But we can take a different approach as the signal problem is different from the graphon problem in a major way, that is the equivalence between the least-cut distance and the Wasserstein distance as seen in proposition 2.3. This enables us to translate our δ' error into an error between mixing distributions that generate our data through a convolution with the gaussian location family.

The following results have been taken from Wu and Yang (2019) (and its supplementary material) and from Doss et al. (2021) which we adapt using proposition 2.3. We get three rates, one for the general case with fixed k, one for the fully separated case, and another for a growing k (which would be useful for non-parametric smooth estimation).

First we define the parameter space for the separable case:

$$\mathcal{F}_{k,r,\Delta,\alpha} = \{f \in \mathcal{F}_{k,r} | \min_{i \neq j} |M_i - M_j| \geq \Delta, |I_i| \geq \alpha \ \forall i \in [k]\}.$$

Theorem 3.8. *Take* $2 \leq k \leq n$*:*

$$\inf_{\hat{f}} \sup_{f \in \in \mathcal{F}_{k,r}} \mathbb{E}_f \delta'_1(f, \hat{f}) \asymp_k \left(\frac{1}{n}\right)^{\frac{1}{4k-2}}.$$

The optimal rate is achieved by a k-step function estimator derived through the de-noised moments method DMM.

For another DMM k-step function estimator, if $k = O(\frac{\log n}{\log \log n})$ *and* n *is large enough w.r.t to* Δ *and* α *with probability at least* $1 - \epsilon$ *for any* $\epsilon > 0$*:*

$$\sup_{f \in \mathcal{F}_{k,r,\Delta,\alpha}} \delta'_1(\hat{f}, f) \leq C_k r \Delta^{-(2k-2)} \sqrt{\frac{\log(\frac{k}{\epsilon})}{n}},$$

where $C_k > 0$ *is dependent only on* k*.*

Now if $k \geq \Omega(\frac{\log n}{\log \log n}))$*, then there is some step function estimator* $\hat{f}$ *derived through DMM with step number in the order of* $\min(k, O(\frac{\log n}{\log \log n}))$ *such that we have with probability* $1 - \epsilon$ *for any* $\epsilon > 0$*:*

$$\sup_{f \in \mathcal{F}_{k,r}} \delta'_1(\hat{f}, f) \leq C' r \left(\frac{\log n}{\log \log n} + \sqrt{\frac{\log(\frac{1}{\epsilon})}{n^{1-c}}}\right),$$

for some absolute constant $C' > 0$ *and some constant* $c < 1$*.*
Furthermore all these rates are optimal in the minimax sens.

Remark 3.9. *A similar result to the first one can be given if we only assume* $k = O(\frac{\log n}{\log \log n})$ *instead of assuming it to be fixed.*

The first rate can be interpreted as only knowing an upper-bound k of the number of fully separated blocks. The higher the number of misspecified blocks the worse our estimates are expected to be. The parametric $\sqrt{n}$ rate is only possible when we know the exact number of fully separated blocks. We see that in the general case our rates are far worse than those of graphon estimation and are expected to be even worse for non-parametric estimation.

The estimators that achieve the optimal rates in the theorem above are derived through a de-noised method of moments, which is not evident to adapt to our goal of joint estimation. There are also rates given for the method of least distance which is somewhat analogue to the least squares method used in graphon estimation, we refer to Heinrich and Kahn (2018) for details.

So to motivate the later use of the EM algorithm, we give the following result about the maximum likelihood estimator in the separable case drawn from Redner and Walker (1984). First define:

$$\mathcal{F}_{k,r,\beta,sep} = \left\{f \in \mathcal{F}_{k,r} : |I_a| \geq \beta,\ M_a \neq M_b,\ a \neq b = 1, \dots, k\right\}.$$

Theorem 3.10. *Take the true $f \in \mathcal{F}_{k,r,\beta,sep}$ and a fixed k. Then for n large enough there exists a unique consistent maximum likelihood estimator (MLE) $\hat{f}$ such that:*

$$\delta_2'(f, \hat{f}) \xrightarrow{a.s} 0.$$

Remark 3.11. *The MLE for the parameters vector is, in fact, asymptotically normal, which suggests an asymptotic $\sqrt{n}$ rate. See the proof of the theorem for further details.*

This was expected as under separability conditions and a fixed k our problem is a regular parametric estimation problem, while the mixture case requires slightly different conditions than those of Wald (1949). However, the issue here is that we require k to be fixed, this gives no guarantee for non-parametric signal estimation where k usually grows with n. We can describe a smooth signal f by a smooth mixing distribution G_f which can be consistently estimated with the MLE (Chen (2023) chapter 2). The result from Wu and Yang (2019) we gave for a growing k should also serve to reassure us.

To conclude this section which was mostly mixture model theory we refer to Chen (2023) for a good overview of the field. For the discrete problem also known as Gaussian de-noising and how it relates to mixture models, we refer to Saha and Guntuboyina (2020). In the next section we will discuss the problem of joint estimation.

3.3 Joint Estimation

From the previous analysis of the two problems, we see that they are very different. In the case of the graphon problem the δ_1 and δ_2 loss is equivalent to the discrete loss except for an agnostic error. This is due to the additional structure of the graphon making those integral losses weaker. That explains why in graphon estimation literature the problem of estimating the graphon and that of estimating the probability matrix θ are considered equivalent. However, an exception must be made for the $\delta_\Box$ distance, where the graphon is simply a graph limit and the optimal rate is achieved by the empirical graphon. For the signal problem the δ' loss is much stronger and gives very different results from the discrete loss as it gets rid of the clustering error. This makes the problem of estimating the signal different from the problem of estimating the means vector μ, although they are closely related.

Now that we understand each problem separately we can tackle the joint estimation of a graphon-signal and put ourselves in the context of *Problem 3*. First we will briefly discuss how the additional data would change the separate estimation problems.

Regarding the estimation μ the addition of the data from the graphon doesn't fundamentally change the structure of our parameter space or, more precisely, it doesn't change our space of distributions indexed by the parameter space, thus we can use the same previous argument to get a constant minimax rate. For the estimation of θ as we already have consistency, it is not evident wether the rate would improve. However as the only leverage point of additional signal data is clustering, which is not important for graphon estimation, we believe the rate won't be improved. If we consider a joint error, we get as we expect:

Proposition 3.1.

$$\inf_{\hat{\theta},\hat{\mu}} \sup_{(\theta,\mu)\in\Theta\mathcal{U}[k]} \mathbb{E}\left[\frac{1}{n^2}\|\hat{\theta}-\theta\|^2 + \frac{1}{n}\|\hat{\mu}-\mu\|^2\right] \asymp 1.$$

For the δ distance, it is the same as for θ. But for the δ' rate, it is far less evident, as the slow rate is an issue of separation of blocks, and graph data might help in that regard but only if we have some separation assumption on our graphon blocks as well. As discussed in Dreveton et al. (2023), there is quantifiable improvement when clustering with node features when looking at phase transitions (threshold at which exact clustering with high probability becomes possible). See Abbe et al. (2022) and Braun et al. (2022) for such results. In these papers a notion of joint separation distance Δ was given with a rate similar to what we saw with clustering in the signal context, also requiring the separation to diverge with n for consistency. This would suggest that in joint estimation graphon and signal would compensate for separation showing an improvement. Giving such a result for the δ' rates is technically very challenging and is beyond the scope of this report.

Now we look to estimate the graphon-signal using the δ'' distances in $\mathcal{WF}[k]$. We expect that using estimators that would yield consistent estimates for each of the separate problem, would yield a consistent estimate for the joint problem. However, that is only the case if our estimators are "aligned" the same way our true graphon and signal are "aligned". Which means that the measure preserving bijection that minimizes distance between the signal and its estimator is the same one that minimizes distance between the graphon and its estimator and we would have $\delta'' \simeq \delta + \delta'$ which suggests that we have consistency, and a possible upper-bound rate would be a sum of the optimal rates for each separate problem. This happens naturally when we apply a joint estimation method, for example if our method entails estimating a single community assignment function for both data. Writing down a proper result would be very technical and doesn't seem very pertinent for the theory of joint estimation. We have found no direct or indirect results in literature that would suggest an improved rate (especially for the signal) w.r.t. δ'' when doing joint estimation, only results pertaining to better clustering, as we said previously, which are not evidently applicable to the integral loss. Instead we will directly propose methods for joint estimation in the next section.

But first, we need to discuss how we can model a meaningful graphon-signal relationship justifying a joint estimation. For the joint SBM case we do have a meaningful relationship in clustering, as nodes that belong to the same block have the same signal values. We'd like to generalize this concept to smooth graphon-signals. To this end, the concept of Holder smoothness is most pertinent. It has been established in multiple papers (Gao et al. (2015), Wolfe and Olhede (2013)) that the "optimal" number of blocks to fit a smooth graphon to an SBM depends on it's Holder smoothness. One can also see how approximating an α-Holder function by a block function with an error ϵ can be done with blocks in the order of $(\frac{1}{\epsilon})^{\frac{1}{\alpha}}$. A natural conclusion would be that our graphon and signal should have the same holder smoothness so as to best approximate the graphon-signal pair with a joint SBM. We propose a more precise definition with the following:

Definition 3.12. *We say that a graphon-signal $(w, f) \in \mathcal{WF}$ is jointly smooth if each function is piece-wise Holder smooth for some exponent and for any open I:*

$$w(\cdot, y) \in \mathcal{H}^{\alpha}(I), \text{ for a.e } y \in [0,1] \Longleftrightarrow f \in \mathcal{H}^{\alpha}(I),$$

for any Hölder exponent $\alpha > 0$.

Remark 3.13. *Notice how $\mathcal{WF}[k] \subset \mathcal{WF}[\mathcal{H}]$, with a Hölder exponent $\alpha > 1$.*

Of course, we mean to consider the highest Hölder exponent for each open I. With this condition, it makes sens to fit a joint SBM to such a graphon-signal, as we expect the number of blocks we need to fit a portion of the graphon will be the same number of blocks we need to fit a portion of the signal. This implies that added signal data should improve our estimation of the assignment function and/or the block weights. The reason we took this partition approach instead of considering the smallest holder exponent each function verifies is that in exploratory analysis of network data it would be best to fit an SBM with a minimal number of blocks, we would then expect portions with high smoothness to be fitted with a fewer number of blocks than low smoothness portions. This makes it easier to rearrange our estimator for better interpretability, but might also increase the loss.

For practical use, we would like to find a way to generate meaningfully related pairs as we defined previously. We will first use diffused signals to do so. We say that a signal is diffused if it's of the form $X = \sum_{l=0}^{d} y_l A^l X_0$. To define the analog version for the graphon we will first need to define the graphon integral operator (Ruiz et al. (2021)):

Definition 3.14. *Every graphon w induces a graphon integral operator T_ω
$T_w : \mathcal{F} \to \mathcal{F}$ given by :*

$$(T_w f)(v) = \int_0^1 w(u, v) f(u) du,$$

for any $f \in \mathcal{F}$

Now we define diffused signals:

Definition 3.15. *We say f is a diffused signal for some graphon $w \in \mathcal{W}$ if for some constant $f_0 \in \mathbb{R}$ and $d \in \mathbb{N}$ and sequence $\{y_l\}_{l \in \mathbb{N}} \subset \mathbb{R}$ we have :*

$$f = \sum_{l=0}^{d} y_l T_\omega^{(l)} f_0.$$

A diffused signal can be used to interpret a lot of real world phenomena like spread of information in networks. So it is grounded in reality to generate our synthetic signal in such a way. We can now give the following result:

Proposition 3.2. *Take $w \in \mathcal{W}$ such that w. If f is a diffused signal for w then we have for any open $I \in [0, 1]$:*

$$w(\cdot, y) \in \mathcal{H}^{\alpha}(I), \text{ for a.e } y \in [0,1] \Longrightarrow f \in \mathcal{H}^{\alpha}(I).$$

Corollary 3.16. *If $w \in \mathcal{W}_k$ and f is a diffused signal of w, then $(w, f) \in \mathcal{WF}[k]$.*

We can also use a different relation:

Definition 3.17. *Let $f \in \mathcal{F}$ we say that a graphon w is induced by f:*

$$w(x, y) = g(|f(x) - f(y)|)$$

Where g is Lipschitz on the image of $(x, y) \mapsto |f(x) - f(y)|$ and takes values in $[0, 1]$.

We can see that if we take g to be increasing/decreasing this relation can describe homophily/heterophily. One can easily see that:

Proposition 3.3. *If $w \in \mathcal{W}$ is induced by $f \in \mathcal{F}$ then for any open $I \in [0, 1]$:*

$$f \in \mathcal{H}^\alpha(I) \Longrightarrow w(\cdot, y) \in \mathcal{H}^\alpha(I), \text{ for a.e } y \in [0, 1].$$

4 Joint Estimation Methods

In this section we will present a few methods to estimate our graphon-signal. So again, we put ourselves in the context of *Problem 3*. What we propose is mainly adaptations of already existing methods. We will present two methods: one for SBM estimation and another for non-parametric (including smooth) estimation. We will be using and testing both for non-parametric estimation with step graphon-signals and smooth ones.

4.1 Stochastic Block Model Estimation

In this section we propose an estimation method for the parameters of the joint stochastic block model or CSBM. The method proposed here, and even the goal, are somewhat different from what non-parametric graphon estimation is, where the SBM is used as a tool to approximate the non-parametric graphon, and its estimation is not the goal of the analysis. However, we saw it fit to address this problem of SBM estimation as it is a direct application of most of the theory presented earlier. In this context we assume our true model to be the SBM and mainly aim to estimate the block parameters that define our $(w, f) \in \mathcal{WF}[k]$.

In the same way as the separate problems, our joint problem can be formulated as an incomplete data problem as was presented in Dempster et al. (1977). This suggests the use of the expectation maximization (EM) algorithm as the most natural way to proceed. Under incomplete data interpretation, z is treated as unknown data and not as a parameter, in this sens the complete data would be A, X, z.

First we introduce some notation: write our true parameters as $\Phi^\star = (Q^\star, M^\star, \pi^\star = (|I_1|, \dots, |I|_k))$, write $b(x|p) = (1-p)^{1-x}p^x$ and $g(x|\eta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\eta)^2}{2}}$ and let $\mathcal{X} = (X, A)$ be our incomplete data . For some parameter Φ, our complete data log-likelihood is:

$$\begin{aligned} L(\mathcal{X}, z) =& L(z) + L(\mathcal{X}|z) \\ =& \sum_{i \le n} \pi_{z(i)} + \sum_{i<j\le n} \log b(A_{ij}|Q_{z(i)z(j)}) + \sum_{i \le n} \log g(X_i|M_{z(i)}) \end{aligned}$$

From the expression above we can already tell that joint estimation will not worsen our approximation as the fisher information matrix FIM for each problem is positive semi-definite contributing to a better Cramer-Rameo lower bound for the variance if our estimators are unbiased. As we saw in the proof of theorem 3.10, under separability conditions our FIM for mixture parameters for signal data alone is positive definite, in that case our estimates should be strictly better in variance. This should serve to reassure us that under joint SBM (or something like (3.12) for the smooth case) we shouldn't see our estimates worsen if we use signal data.

The incomplete data likelihood would require summing over all possible z's making it's direct maximization not tractable. The EM algorithm is best suited for such problems. It works as follows:

First notice the relation between the three densities: $p(\mathcal{X}, z|\Phi) = p(z|\mathcal{X}, \Phi)p(\mathcal{X}|\Phi)$. Then we have for any Φ' :

$$\begin{aligned} L(\mathcal{X}) =& \mathbb{E}[L(\mathcal{X}, z)|z \sim p(z|\mathcal{X}, \Phi')] - \mathbb{E}[\log(p(z|\mathcal{X}, \Phi))|z \sim p(z|\mathcal{X}, \Phi')] \\ =& Q(\Phi|\Phi') - H(\Phi|\Phi') \end{aligned}$$

For some guess $\Phi^{(t)}$ at iteration t, determine $Q(\Phi|\Phi^{(t)})$ (E-step) and choose $\Phi^{(t+1)} = arg\max_\Phi Q(\Phi|\Phi^{(t)})$ (M-step). Dempster et al. (1977) proved that the likelihood $L(\mathcal{X})$ does not decrease at each iteration. For the signal estimation alone, one can derive closed form expressions for the parameters at each iteration (Redner and Walker (1984)). However, the computation of the density $p(z|\mathcal{X}, \Phi)$ is not tractable when working with the adjacency data A, making a direct application of the EM algorithm impossible in our case. No notions of neighborhood can be used in this framework either due to the structure of the graph ($p(z|A) \neq p(z|A_i)$, for more detail see Daudin et al. (2008)).

To estimate SBM parameters of graph data, Daudin et al. (2008) proposed the variational EM approach VEM. Similar to the EM approach, we maximize a lower-bound of the incomplete data log-likelihood given by:

$$\mathcal{J}(R_\mathcal{X}) = L(\mathcal{X}) - KL(R_\mathcal{X}(\cdot) \mid p(\cdot|\mathcal{X}, \Phi))$$

The KL term factorizes out the untractable density. $R_\mathcal{X}(\cdot)$ is a distribution on z that approximates the problematic density $p(z|\mathcal{X}, \Phi)$. We will do as in Daudin et al. (2008) and take the mean filed approximation:

$$R_\mathcal{X}(z) = \prod_{i \leq n} h(z_i|\tau_i)$$

Where h is the multinomial distribution and $\tau_i = (\tau_{i1}, \ldots, \tau_{ik})$ for each $i = 1, \ldots, n$.

With this approach, given some initial guess $\Phi^{(t)}$ in the E-step we first find the τ that best approximates the incomplete data log-likelihood by taking $\hat{\tau} = arg\max_\tau \mathcal{J}(R_\mathcal{X})$.

Proposition 4.1. *Given some initial parameter Φ the τ that maximizes $\mathcal{J}(R_\mathcal{X})$ satisfies a fixed point relation of the form:*

$$\hat{\tau}_{ia} \propto g(X_i|M_a)\pi_a \prod_{j \neq i} \prod_{b \leq k} b(A_{ij}|Q_{ab})^{\hat{\tau}_{ia}}$$

For all $i \leq n, a \leq k$. Where the proportionality coefficient is a normalizing term.

Given $\hat{\tau}$, in the M-step we find the parameters $\Phi^{(t+1)}$ that maximize $\mathcal{J}(R_\mathcal{X})$.

Proposition 4.2. *Given the variational parameter τ the value of the parameter Φ that maximizes $\mathcal{J}(R_\mathcal{X})$ is:*

$$\hat{\pi}_a = \frac{1}{n}\sum_{i \leq n} \tau_{ia}, \ \hat{Q}_{ab} = \frac{\sum_{i \neq j} \tau_{ia}\tau_{jb}A_{ij}}{\sum_{i \neq j} \tau_{ia}\tau_{jb}}, \ \hat{M}_a = \frac{\sum_i \tau_{ia}X_i}{\sum_i \tau_{ia}}$$

A direct result following from these two propositions is that at each iteration $\mathcal{J}(R_{\mathcal{X}})$ does not decrease. As an extra step to estimate the community assignment function, we set:

$$\hat{z}(i) = arg \max_{a \leq k} \hat{\tau}_{ia}$$

With it, we can build estimates for:

$$\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}, \quad \hat{\mu}_{\hat{z}(i)\hat{z}(j)} = \hat{M}_{\hat{z}(i)}, \quad i \neq j = 1, \ldots, n$$

Gaucher and Klopp (2021) proved the strong recovery of the maximum likelihood clustering up to permutations using this method with adjacency data. The minimax optimality of the $\hat{\theta}$ estimator with the same rate as in Gao et al. (2015) was also proved in the same paper. Following the discussion with the likelihood above, we expect the same guarantees to hold, as these estimators approximate well the maximum likelihood estimators. For the same reason, since the parameter μ is highly dependent on clustering, we expect a good estimate for f w.r.t the δ' distance.

To determine the number of blocks we use the integrated classification likelihood (ICL) method first proposed by Biernacki et al. (2000) for the gaussian mixture model and adapted by Daudin et al. (2008) for the graph SBM. We denote by m_k the joint SBM with k components. The aim is to maximize the complete data integrated likelihood:

$$p(\mathcal{X}, z \mid m_k) = \int p(\mathcal{X}, z \mid m_k, \Phi) p(\Phi \mid m_k) d\Phi$$

The ICL is derived by approximating the above term, we refer to Biernacki et al. (2000) for details.

Proposition 4.3. *The ICL for our problem is given by:*

$$ICL(m_k) = \max_{\Phi} L(\mathcal{X}, z \mid m_k, \Phi) - \frac{1}{2}\frac{k(k+3)}{2} \log \frac{n(n+1)}{2} - \frac{k-1}{2} \log n$$

In practice we take z to be some estimate $\tilde{z}$. We maximize over k given some lower-bound of the true block number. The reason a lower-bound is better (and not an upper-bound) is due to the Stirling approximation for the gamma function we use in the proof which assumes large cardinality of the groups. If we initially guessed a bigger k than the true one, we expect our algorithms to output a group of very low cardinality. And so ICL tends to miss important structure for small sample sizes as discussed in Biernacki et al. (2010). That asymptotic approximation cannot be avoided either as otherwise the computation of the likelihood is not tractable due to adjacency data. To solve this problem a variational Bayesian criterion was proposed in Latouche et al. (2010), using appropriate prior distributions instead of asymptotic approximations. However, as we are looking to use this method for non-parametric (smooth) cases, it does not seem very pertinent to pursue the problem any further and we stick to the ICL criterion, as the goal is mainly to compare it to heuristic choices for k.

We would like to mention that an EM based method similar to what we described here was implemented in Sischka and Kauermann (2022) for smooth graphon estimation and even more general models (Sischka and Kauermann (2024)). The method of iterative refinement with least squares IR-LS with spectral initialization introduced in Braun et al. (2022) will serve as comparison in the numerical experiments section. For more spectral based methods for clustering with node attributes we refer to Binkiewicz et al. (2017); Mele et al. (2021); Abbe (2023). Inference on graphs with valued nodes using an EM approach was also done in Stanley et al. (2018) but only using node features on the E-step allowing a direct use of the EM algorithm.

4.2 Non-parametric Estimation

As was discussed previously even for (piece-wise) smooth graphon-signals if a condition such as (3.12) holds, we expect to be able to directly apply the VEM method and get satisfactory results (**??**). Nevertheless, we will discuss a method that was specifically tailored for non-parametric graphon estimation, in particular we will use the neighborhood smoothing method (NBS) first introduced in Zhang et al. (2017). It works by defining a distance between nodes:

$$d_G^2(i,j) = \int_{[0,1]} |w(\xi_i, y) - w(\xi_j, y)|^2 dy$$

With which to define a neighborhood. It can be shown (Zhang et al. (2017)) by a heuristic argument that the distance can be well approximated by:

$$\hat{d}_G^2(i,j) = \max_{l \neq i,j} |\langle A_i - A_j, A_k \rangle| /n$$

To accommodate for node features, Su et al. (2020) proposed using a node feature distance:

$$\hat{d}_s^2(i,j) = \max_{l \neq i,j} |\langle X_i - X_j, X_k \rangle| /p$$

Where p is the dimension of the node features ($p = 1$ in our case). Then adding it to define a refined distance between nodes:

$$\hat{d}^2(i,j) = \hat{d}_G^2(i,j) + \lambda \hat{d}_s^2(i,j)$$

For some $\lambda \geq 0$ to be chosen appropriately. Then the neighborhood of some node i is taken to be:

$$N_i = \{j \neq i \mid \hat{d}(i,j) \leq q_i(h)\}$$

Where $q_i(h)$ is the h-th sample quantile of the set $\{\hat{d}(i,j) \mid j \neq i\}$. Where $h = C_0\sqrt{\frac{\log n}{n}}$, with C_0 a constant usually chosen as 1. With graph data only ($\lambda = 0$) Zhang et al. (2017) had proved, under reasonable conditions, that this method is consistent w.r.t the discrete loss for graphons, with a $\sqrt{\frac{\log n}{n}}$ rate. Su et al. (2020) extended this result for an arbitrary λ and under similar conditions. And it turns out that when $\lambda > 0$ the method is not consistent under general conditions depending on the dimension of node features and the level of noise. So the use of node features might even worsen our estimate. It is why a cross validation method was introduced in the same paper to chose an appropriate λ. Nevertheless, it was observed with the use of real world data, that the addition of node features does better the estimate. Such an analysis in the context of clustering with node features was also done by Zhang et al. (2016), pointing to better performance with finite samples, and mentioning a lack of analytical tools to give results on the matter. In our case, where we want to estimate the signal, the lack of explicit fitting of an SBM/finite mixture with this method gives no guarantees on consistency. Nonetheless, a natural way to proceed would be to use the estimated neighborhoods to give an estimator for μ. So we take our estimators to be:

$$\hat{\theta}_{ij} = \frac{1}{2}\left(\frac{\sum_{i' \in N_i} A_{i'j}}{|N_i|} + \frac{\sum_{j' \in N_j} A_{ij'}}{|N_j|}\right), \quad \hat{\mu}_i = \frac{\sum_{i' \in N_i} X_{i'}}{|N_i|}$$

4.3 Numerical experiments

The loss we will be using is the discrete loss for the graphon, the signal, and the joint case (taking the sum of both losses). For the signal loss a good indicator is when the error is far below 1, which is the expected error if we use the empirical estimator.

We will be testing our methods with both the SBM and the smooth case. We initialize IR-LS with the spectral method described in Braun et al. (2022). VEM is initialized with a random τ drawn from a Dirichlet distribution. Due to randomness and instability we also run it a few times and take the output with the best complete data likelihood. For the λ parameter in the FANS method, we ran it with multiple λ's from a $[0, 1]$ grid and surprisingly found $\lambda = 0$ to give the best performance in all smooth cases and especially when we considered the signal error alone, so we will take it to be 0 for all the experiments. Finally in the non-parametric (smooth and unspecified k) case we will also test VEM+ICL where the ICL only tests up to $k = 20$. For the VEM and IR-LS we will take the number of blocks to be $\lceil\sqrt{n}\rceil$.

Figure 2 compares the VEM with the IR-LS on a random SBM with specified number of blocks. They both consistently perform similarly on high samples and small block sizes with the VEM generally doing better w.r.t. signal error and on smaller sample sizes. Instabilities sometimes arise for both with high sample sizes as can be seen in the figure, but it is not recurrent. The FANS algorithm performs in general much worse when it comes to SBM's with specified k, so it won't be shown on the figure. Although, we must note that in the SBM case setting $\lambda > 0$ improved the performance very slightly.

We present the four graphons on which we will be testing smooth estimation on figure 3. The first is the gradient graphon $\frac{x+y}{2}$ and the signal is the scaled and translated diagonal. The second graphon is the latent distance model graphon $|x-y|$, the signal is a diffused signal on the graphon with the sequence $y_l = 10l$, initial constant 2 and diffused 200 times. For the third one, it's the graphon tested in Zhang et al. (2017) as an example for a graphon without strictly monotone marginals, and the signal is also diffused with sequence $y_l = l$ initial constant 2 and diffused 200 times. The fourth one is a graphon resulting from $w(x, y) = g(|f(x) - f(y)|)$ with $f = 2\cos(10x) + 16$ with g only a scalar serving to have $w \in [0, 1]$. In all cases we try to keep the range of the signal around 2-5, so that the noise is consequent.

Table 1 summarizes the results for the smooth graphons and two random SBM's (k=5,10) tested on 300 nodes. The errors are presented in triplets (Graphon error, Signal error, Joint error).

Pair tested	IR-LS	VEM	FANS	VEM+ICL
pair 1	(0.00518,0.43450,0.43968)	(0.00490,0.27002,0.27492)	(**0.00407**,**0.09563**,**0.09970**)	(0.00539,0.20323,0.20863)
pair 2	(0.00367,0.29594,0.29961)	(**0.00309**,0.31158,0.3146)	(0.00413,**0.16919**,**0.17333**)	(0.00318,0.30935,0.31253)
pair 3	(0.00501,0.29905,0.30407)	(0.00643,0.2886,0.29504)	(**0.00329,0.27042,0.27371**)	(0.00605,0.28119,0.28724)
pair 4	(0.00329,0.20740,0.21070)	(**0.0029**,0.09433,0.09725)	(0.00365,**0.03840,0.04206**)	(0.00333,0.11981,0.123)
SBM k=5	(0.00216,0.38659,0.38876)	(0.00267,0.17446,0.17714)	(0.00399,0.15917,0.16316)	(**0.00198,0.09384,0.09583)**
SBM k=10	(**0.00083**,0.32568,0.32651)	(0.00707,0.23003,0.23711)	(0.0070,0.99899,1.00603)	(0.00133,**0.1456,0.14699**)

Table 1: Error Table

We note how the FANS with $\lambda = 0$ is outperforming all the other methods when it comes to smooth signal estimation (fig 4). For the SBM case, the SBM methods were the most accurate as expected. The addition of the ICL criterion made no noticeable difference and mostly guessed block numbers close to the bound specified, showing that the penalty term is indeed not appropriate for small sample sizes. The better performance

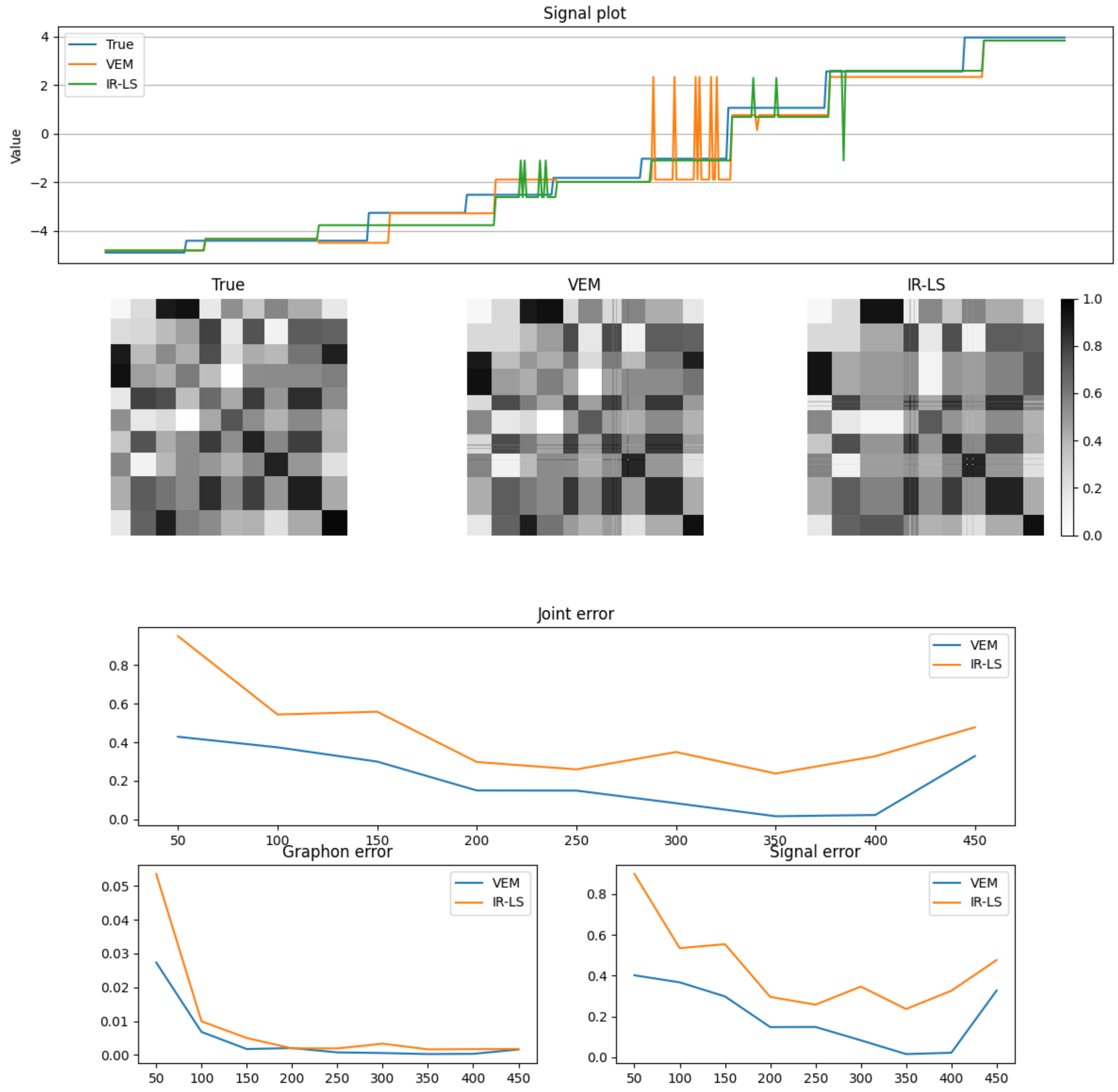

Figure 2: Plots of the estimates for IR-LS and VEM aligned using the latents (500 nodes, 10 specified blocks) and a plot of the different errors against sample size (10 specified blocks)

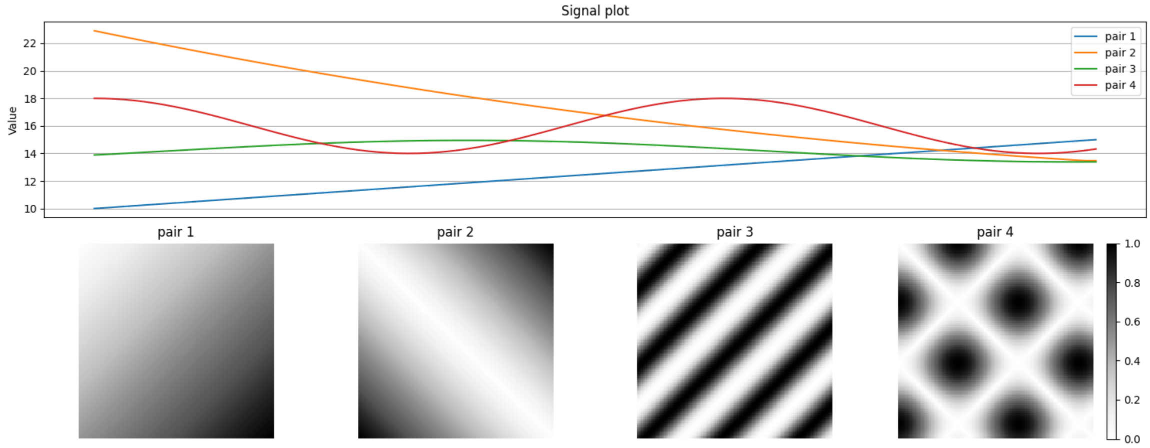

Figure 3: The four smooth graphon-signal pairs we test our methods on.

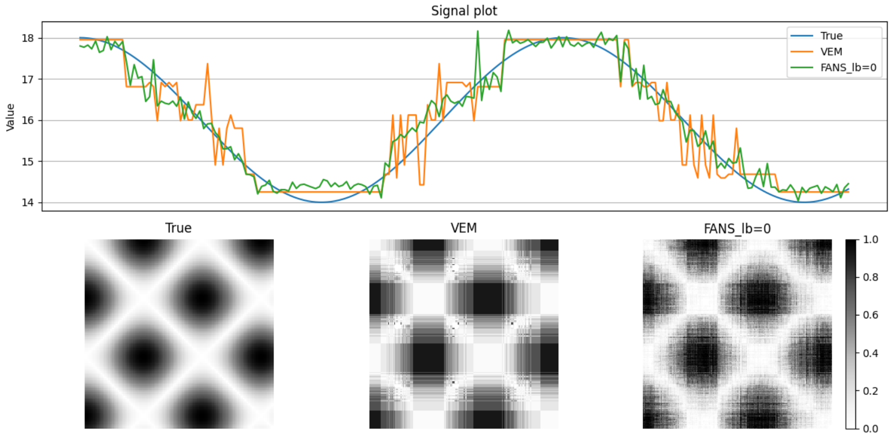

Figure 4: The VEM and the FANS output on pair 4. with 200 nodes.

with SBM's is due to a specified number of blocks closer to the true one. We must say that, with the high number of variables, these tests are certainly not fully descriptive of the behavior of these methods and implementation might also make a difference. We refer the reader to the code repository (`https://github.com/Salah4848/PDB.git`) for more figures of the tests.

4.4 Real World Data Examples

In this section we use our methods with two real world data examples. One of which serves more as a non-example and shows a different use of our algorithms. We will not delve too deep in analyzing the networks. This will serve more as an example of exploratory use of the methods.

4.4.1 Collaboration network

We use the General Relativity and Quantum Cosmology collaboration network dataset (Leskovec et al. (2007)). Available on SNAP (`snap.stanford.edu/data/ca-GrQc.html`). An edge is present between two authors if they have co-authored a paper. We take as node features the degree of each node, i.e all its collaborators. We filtered out nodes with

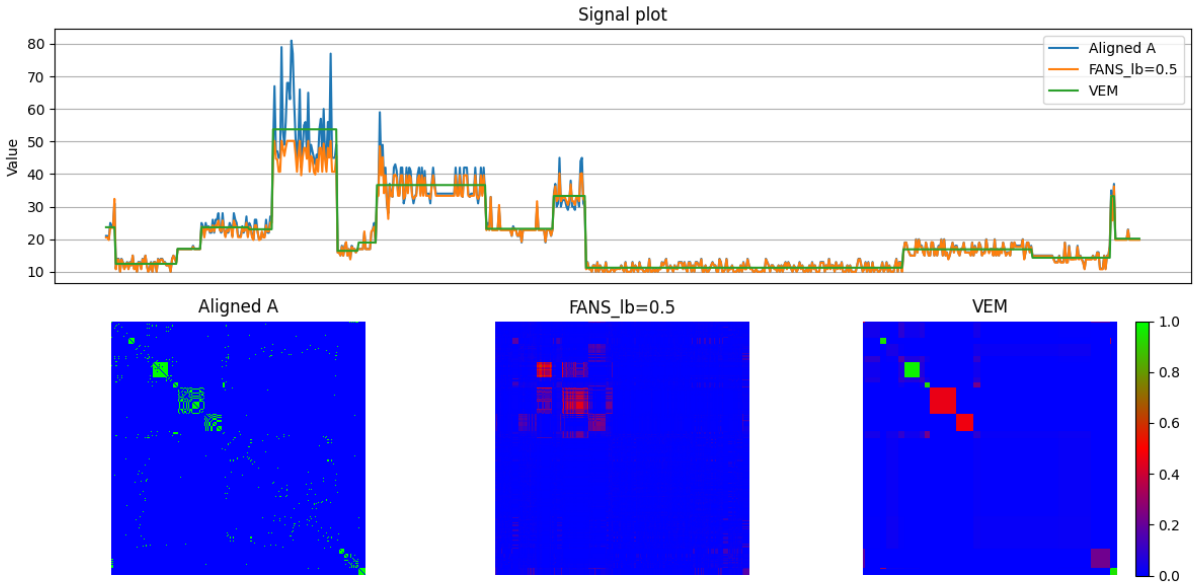

Figure 5: Results for the Collaboration dataset.

less than 10 collaborators, leaving out 737 nodes. The ICL criterion identified 16 to 20 blocks with a given bound of 50 blocks. We run both the VEM with 16 blocks and FANS algorithm choosing $\lambda = 0.5$. For other choices of $\lambda > 0$ we got similar outputs. For $\lambda = 0$ the algorithm performed poorly due to the sparsity of the network.

Figure 5 displays both the outputs of the VEM and FANS algorithms along with the aligned adjacency matrix. The alignment is done using the clusters identified by the VEM algorithm. The highly collaborative authors with an average of around 55 co-authored papers densely cluster together. Inter-cluster collaboration is mainly between small clusters. We also spot a few isolated dense clusters. Despite the sparseness, the VEM performed quit well, while the FANS did not do as well and mostly used node features giving quite faulty results when compared to the adjacency matrix.

4.4.2 Twitch Gamers Network

We use the Twitch Gamers dataset (Rozemberczki and Sarkar (2021)) also available on SNAP (`snap.stanford.edu/data/twitch_gamers.html`). Nodes are accounts on the live streaming platform Twitch. As the platform has a "follow" feature, edges represent accounts that mutually follow each other. Node features are total view count on the account. We filtered our data to keep only english language accounts with views exceeding 5 million. We end up with a network of size 580. The ICL identified 15 blocks. We run the VEM once with node features included (fig 7) and another time without node features (fig 6).

When running the VEM with node features, while we do identify a faint cluster containing nodes with the highest views, there were no real densely connected communities identified. A much better job in clustering was done when running the VEM with no node features, and visible communities were identified. We also see that under this clustering there is no real pattern seen in the views counts, showing that such a statistic does not greatly influence connectivity in the network. This highlights how such methods can be used to identify feature influence on networks.

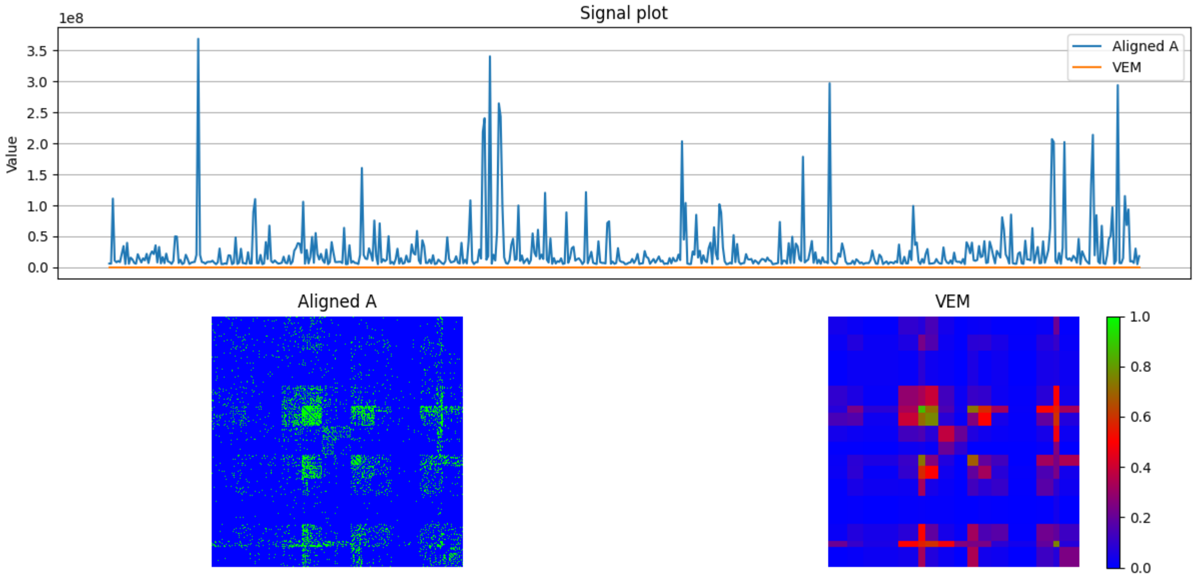

Figure 6: The VEM output when using no node features.

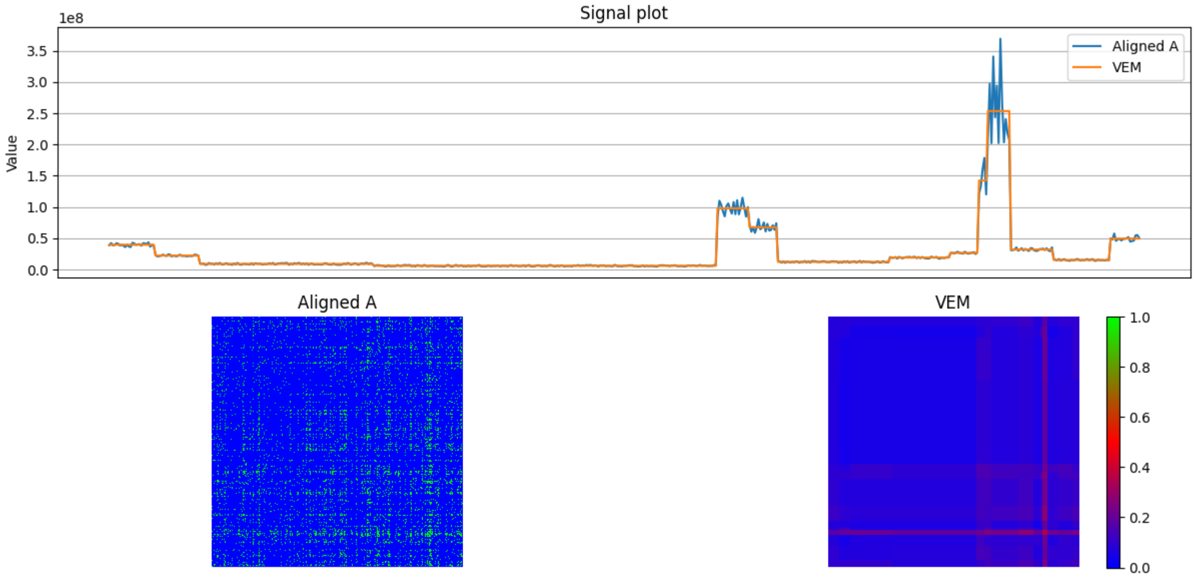

Figure 7: The VEM output when we use features data

5 Discussion

We conclude this report with a short discussion on what we've learned, possible improvements, and some subjects of interest that were out of scope for this paper. The most insightful take-away would be the stark difference in nature of the graphon and signal estimation problems. A most interesting result would to be to see a change in the problematic integral error rate of the signal when using graph data. Indeed, it seems to be mostly a matter of separation and, as we saw with clustering, additional graph data does help in distinguishing between clusters. Generalizing the results for signal estimation to smooth signals would also be of great values. But it poses a great challenge as literature for non-parametric mixtures is very scarce and the use of our metric of interest (the Wasserstein metric) is only recent. We also saw the importance of optimal transport results to bridge between our problem and mixture parameter estimation, generalizing those results for smooth signals and graphon-signals is a must, to leverage both theories of graphon estimation and mixture estimation. Concerning the discrete joint estimation problem, we saw how some recent results support an improvement in clustering under joint estimation, but the generalization of such results to the non-parametric context is hard to see. Moving on to the computational methods we proposed and particularly the adapted VEM. While we did give an argument with FIM's and the good approximation of the MLE, there was a lack of technical justification which has also proved to be beyond our reach for this project. We also believe some of the spectral methods, such as the USVT (Chatterjee (2015)), can be adapted for our use. We would also like to point out that the VEM method can be adapted for a multivariate gaussian with unknown covariance matrix. However, some of the theoretical support might not be easily adapted. Finally, we would like to point out the lack of accessible datasets relevant to our context, we do believe we can better explore the capabilities of our methods given more suitable data.

Words of gratitude to my supervisor Charles for his weekly, involved guidance for my project. I would also like to thank Prof. Sofia Olhede for the opportunity to undertake this project.

Proofs

Proof of Theorem 3.3

Proof. For any i, by LeCam's two point testing lemma (Lafferty et al., 2008):

$$\inf_{\hat{\mu}_i} \sup_{\mu_i \in \{q_1, q_2\}} \mathbb{E}\left[|\hat{\mu}_i - \mu_i|^2\right] = \inf_{\hat{\mu}_i} \sup_{P_i \in \{\mathcal{N}(q_1,1), \mathcal{N}(q_1,1)\}} \mathbb{E}\left[|\hat{\mu}_i - \mu(P_i)|^2\right] \geq \frac{|q_1 - q_2|^2}{8} e^{-KL(\mathcal{N}(q_1,1), \mathcal{N}(q_2,1))}$$

The Kullback–Leibler divergence for our distributions is :

$$KL(\mathcal{N}(q_1, 1), \mathcal{N}(q_2, 1)) = \frac{1}{2}|q_1 - q_2|^2$$

So we have for all i:

$$\inf_{\hat{\mu}_i} \sup_{\mu_i \in \{q_1, q_2\}} \mathbb{E}\left[|\hat{\mu}_i - \mu_i|^2\right] \geq c > 0$$

Where c is positive and independent of i and n.
Furthermore, since we can choose each μ_i independently (this comes from the definition

of the parameter space), we can write for any estimator $\hat{\mu}$:

$$\sup_{\mu \in \{q_1,q_2\}^n} \mathbb{E}\left[\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] = \sum_i \sup_{\mu_i \in \{q_1,q_2\}} \mathbb{E}\left[|\hat{\mu}_i-\mu_i|^2\right]$$

And by using the super-linearity of the infimum operator we have:

$$\inf_{\hat{\mu}} \sup_{\mu \in \{q_1,q_2\}^n} \mathbb{E}\left[\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] \geq \frac{1}{n}\sum_i \inf_{\hat{\mu}_i} \sup_{\mu_i \in \{q_1,q_2\}} \mathbb{E}\left[|\hat{\mu}_i-\mu_i|^2\right] \geq \frac{1}{n}nc = c$$

For the upper-bound we can take $\hat{\mu}_i = X_i$, since we have variance 1, we conclude. □

Proof of proposition 2.3

Proof. Let f_1, f_2 be step functions. We will first show that the least-L_1 distance is reached when both functions are rearranged to be increasing (= non decreasing), following the same method in the proof of the Hardy-Littlewood inequality.

Without loss of generality we assume both function to be non negative, else we simply add $\max(\sup_x f_1, \sup_x f_2)$ to both functions which conserves their L_1 distance.

First we observe that:

$$\begin{aligned}
\delta_1'(f_1,f_2) &= \inf_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} |f_1(\phi(x)) - f_2(\sigma(x))|dx \\
&= \inf_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} f_1(\phi(x)) + f_2(\sigma(x)) - 2\min(f_1(\phi(x)), f_2(\sigma(x)))dx \\
&= \int_{[0,1]} f_1(x)dx + \int_{[0,1]} f_2(x)dx - \sup_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} \min(f_1(\phi(x)), f_2(\sigma(x)))dx
\end{aligned}$$

For the removal of the measure preserving transformation under the integral see Janson (2011) lemma 5.5.
Using layer cake representation:

$$\begin{aligned}
\int_{[0,1]} \min(f_1(x), f_2(x))dx &= \int_0^\infty \int_{[0,1]} \mathbb{I}\{y \leq \min(f_1(x), f_2(x))\}dxdy \\
&= \int_0^\infty |\{x : \ y \leq f_1(x) \ \cap \ y \leq f_2(x)\}|dy
\end{aligned}$$

We can see that the measure of the intersection $|\{x : \ y \leq f_1(x) \ \cap \ y \leq f_2(x)\}|$ is maximized when both functions are rearranged to be increasing, Indeed, when f_1 is increasing $\{x : \ y \leq f_1(x)\}$ is an interval ending in 1 and the same for f_2, so their intersection is maximized when both are increasing as the smaller set in included in the bigger set. If we call $f_1^\star, f_2^\star$ the increasing rearrangements, we have:

$$\int_{[0,1]} \min(f_1(\phi(x)), f_2(\sigma(x))) \leq \int_{[0,1]} \min(f_1^\star(x), f_2^\star(x))dx$$

Since those rearrangements can be obtained through measure preserving bijections (see Levie (2023) lemma B.2):

$$\delta_1'(f_1,f_2) = \int_{[0,1]} |f_1^\star(x) - f_2^\star(x)|dx$$

Now for any $f \in \mathcal{F}_k$ we have that $F_{G_f}^{-1}$ is $f^\star$ a.e. Denote by $M_i^\star$ the distinct means of $f^\star$ in increasing order and denote by $I_i^\star$ the corresponding interval. If $t \in]0, |I_1^\star|[$ then $F_{G_f}^{-1} = \inf\{x | F_{G_f}(x) \geq t\} = M_1^\star$ (if $x < M_1^\star : F_{G_f}(x) = 0$), and if $t \in]|I_1^\star|, |I_1^\star| + |I_2|^\star[$: $F_{G_f}^{-1}(t) = M_2^\star$ and so on. So, $F_{G_f}^{-1} = f^\star$ almost everywhere and we have:

$$\delta_1'(f_1, f_2) = \int_{[0,1]} |f_1^\star(x) - f_2^\star(x)| dx = \int_{[0,1]} |F_{G_{f_1}}^{-1}(x) - F_{G_{f_2}}^{-1}(x)| dx = W_1(G_{f_1}, G_{f_2})$$

Now for the δ_2' distance:

$$\begin{aligned}
\delta_2'^2(f_1, f_2) &= \inf_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} |f_1(\phi(x)) - f_2(\sigma(x))|^2 dx \\
&= \inf_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} f_1^2(\phi(x)) + f_2^2(\sigma(x)) - 2 f_1(\phi(x)) f_2(\sigma(x)) dx \\
&= \int_{[0,1]} f_1^2(x) dx + \int_{[0,1]} f_2^2(x) dx - \sup_{\sigma,\phi \in \mathcal{M}} \int_{[0,1]} f_1(\phi(x)) f_2(\sigma(x)) dx
\end{aligned}$$

Using layer cake representation again:

$$\begin{aligned}
\int_{[0,1]} f_1(x) f_2(x) dx &= \int_0^\infty \int_0^\infty \int_{[0,1]} \mathbb{I}\{y_1 \leq f_1(x)\} \mathbb{I}\{y_2 \leq f_2(x)\} dx dy_1 dy_2 \\
&= \int_0^\infty \int_0^\infty |\{x : \ y_1 \leq f_1(x) \ \cap \ y_2 \leq f_2(x)\}| dy_1 dy_2
\end{aligned}$$

Using the same argument the intersection is maximized when both functions are increasing. So we conclude. □

Proof of theorem 3.10

Proof. We start from a result in a survey by Redner and Walker (1984) of which we explain the proof (which was not given in the paper) and verify the conditions. This gives us everything we need regarding convergence of parameters. We then translate those results in Wasserstein distance.

First take $f \in \mathcal{F}_{k,r,\beta,sep}$. We denote by $\Phi^\star = (M_1, \ldots, M_K, |I_1|, \ldots, |I_{k-1}|)$ the true parameter. $k = K$ is assumed to be fixed (w.r.t n). We easily get our gaussian mixture density function for any parameter $\Phi = (q_1, \ldots, q_K, \pi_1, \ldots, \pi_{K-1})$:

$$p(x|\Phi) = \sum_{a \leq K} \frac{\pi_a}{\sqrt{2\pi}} e^{-\frac{|x - q_a|^2}{2}} \ , \ x \in \mathbb{R}$$

And the joint log-likelihood function:

$$L(X_1, \ldots, X_n | \Phi) = \sum_{i \leq n} \log(\sum_{a \leq K} \frac{\pi_a}{\sqrt{2\pi}} e^{-\frac{|X_i - q_a|^2}{2}})$$

We recall that since we confine ourselves to $\mathcal{F}_{sep,\beta,M}$ our parameter space for means and weights is compact, let's call it C. For convenience, we will write $\Phi = (\alpha_1, \ldots, \alpha_l)$ where $l = 2K - 1$. We will list the conditions needed for the theorem to be true and see they are verified for our case :

Condition (i) : For all $\Phi \in C$, for all $i, j, k = 1, \ldots, l$ and for all $x \in \mathbb{R}$ the following exist:

$$\frac{\partial}{\partial \alpha_i} \log(p(x|\Phi)),\ \frac{\partial}{\partial \alpha_i \partial \alpha_j} \log(p(x|\Phi)),\ \frac{\partial}{\partial \alpha_i \partial \alpha_j \partial \alpha_k} \log(p(x|\Phi))$$

Condition (ii) : For all $\Phi \in C$, for all $i, j, k = 1, \ldots, l$ we have :

$$\frac{\partial}{\partial \alpha_i} p(x|\Phi) \leq f_i(x),\ \frac{\partial}{\partial \alpha_i \partial \alpha_j} p(x|\Phi) \leq f_{ij}(x),\ \frac{\partial}{\partial \alpha_i \partial \alpha_j \partial \alpha_k} p(x|\Phi) \leq f_{ijk}(x)$$

Where our dominant functions are independent of Φ and are integrable.
We also need :

$$\int_{\mathbb{R}} f_{ijk}(x) p(x|\Phi^\star) dx < \infty$$

Condition (iii) : The fisher information matrix :

$$I(\phi) = \int_{\mathbb{R}} [\nabla_\Phi \log(p(x|\phi))][\nabla_\Phi \log(p(x|\phi))]^T p(x|\Phi) dx$$

Is positive definite and well defined at $\Phi^\star$.
Condition (iv) : For all $\Phi \in C$, for all $i, j = 1, \ldots, l$:

$$\frac{\partial}{\partial \alpha_i} \log(p(.|\Phi)),\ \frac{\partial}{\partial \alpha_i \partial \alpha_j} \log(p(.|\Phi))$$

Are also continuous.

We mean by solutions to the likelihood equations any $\Phi \in C$ that satisfies :

$$\nabla_\Phi L(X_1, \ldots, X_n|\Phi) = 0$$

Chanda (1954) proved that under conditions (i),(ii) and (iii) for n sufficiently large there exists aleast one solution to the likelihood equations which is a consistent estimate to the true parameter $\Phi^\star$ and proved it's asymptotically normal. Tarone and Gruenhage (1975) proved that if condition (iv) is added then from all solutions of the likelihood equations one and only one is consistent. Peters and Walker (1978) proved (appendix A) that under our conditions this unique consistent solution is in fact strongly consistent and locally maximizes the likelihood function.

Conditions (i),(ii),(iv) are trivially verified for our density function inside a compact set. We verify the more interesting condition (iii) which is a form of identifiably. For any non zero $u \in \mathbb{R}^{2K-1}$:

$$\begin{aligned} u^T I(\Phi^\star) v &= E_{\Phi^\star}(u^T [\nabla_\Phi \log(p(x|\Phi^\star))][\nabla_\Phi \log(p(x|\Phi^\star))]^T u) \\ &= E_{\Phi^\star}([\nabla_\Phi \log(p(x|\Phi^\star))]^T u)^2 \geq 0 \end{aligned}$$

The last quantity is null if and only if :

$$\begin{aligned} &\forall x \in \mathbb{R}\ : [\nabla_\Phi \log(p(x|\Phi^\star))]^T u = 0 \text{ (for all x is due to continuity)} \\ \Leftrightarrow &\exists 0 \neq \lambda \in \mathbb{R}^{2K}\ : \sum_{i=1}^{K} \lambda_i (x - M_i) e^{-\frac{(x-M_i)^2}{2}} + \sum_{i=1}^{k-1} \lambda_{i+K} e^{-\frac{(x-M_i)^2}{2}} = 0 \end{aligned}$$

We got rid of the weights by taking $\lambda_i = \frac{u_i}{|I_i|}$ for the appropriate i's. This can be done since we took non null weights and got rid of the linearly independent $|I_K|$ in our definition. Now we we can take x big enough so that we may divide both side of our equation by $(x - M_j)e^{-\frac{(x-M_i)^2}{2}}$ where M_j is our unique maximum (we took distinct means). We get by taking $x \to +\infty$ that $\lambda_j = 0$. Now we do the same but with $e^{-\frac{(x-M_i)^2}{2}}$ which gives us $\lambda_{j+K} = 0$. We continue doing this proving that $\lambda = 0$. And so our Fisher information matrix is positive definite.

Since our distribution function verifies all the conditions within any compact, we conclude that we have a unique strongly consistent MLE $\hat{\Phi}$ in C:

$$\hat{\Phi} \xrightarrow{a.s} \Phi^\star$$

And

$$\sqrt{n}(\hat{\Phi} - \Phi^\star) \xrightarrow{d} \mathcal{N}(0, I(\Phi^\star)^{-1})$$

Where $I(\Phi^\star)$ is the fisher information matrix evaluated at $\Phi^\star$.

By the continuous mapping theorem:

$$\|\hat{\Phi} - \Phi^\star\|_2 \xrightarrow{a.s} 0$$

The last step is to bound the Wasserstein distance by the parameter norms. This will require the use of optimal transport concepts, so we refer to Peyré and Cuturi (2020) for details on the theory. For $G, G' \in \mathcal{G}_{k,M}$ we have:

$$W_p^p(G, G')= \min\{\sum_{i,j} T_{ij}|q_i - q'_j|^p : \sum_j T_{ij} = \pi_i, \sum_i T_{ij} = \pi'_j\}$$

$\{T_{ij}\}$ is a matrix called the transport plan. It tells how much mass from the point q_i (which has total mass π_i) to transport to the point q'_j (where we need to fill up π'_j worth of mass), $|q_i - q'_j|^p$ is the transport cost to move mass from q_i to q'_j. Once we understand this, it becomes easy to upper-bound the wasserstein distance by the parameter norm. Indeed since the minimum is taken over all transport plans, we can choose one in particular to upper-bound the distance. We procede as follows:

- For the point q_1, if $\pi_1 \leq \pi'_1$, we move all the mass to q'_1. If $\pi_1 > \pi'_1$ we move as much as we can to q'_1 (i.e move π'_1), but it will leave us excess mass $|\pi_i - \pi'_i|$.
- We do the same for every i. And since we want to upper-bound we can assume we will always have leftover mass $|\pi_i - \pi'_i|$.
- That leftover mass will need to be moved somewhere at some cost. But again, since we only want to upper-boudn, we assume we move it at the highest cost possible $(2r)^p$.

This gives us an upper-bound:

$$W_p^p(G, G') \leq \sum_{i \leq K} \min(\pi_i, \pi'_i)|q_i - q'_i|^p + (2r)^p \sum_{i \leq K} |\pi_i - \pi'_i|$$

We can remove the redundant weight by writing in terms of l_1 distance between the other weights. We also use equivalence of norms in $\mathbb{R}^{K-1}$ to switch to l_2 distance for the weights. And so there is some constant $C_{K,r}$ such that:

$$W_2^2(G, G') \leq C_{K,r}(\sum_{i \leq K} |q_i - q'_i|^2 + \sum_{i \leq K-1} |\pi_i - \pi'_i|^2) = C_{K,r}\|\Phi - \Phi'\|_2^2$$

Combining our results we conclude. □

Proof of proposition 3.1

Proof. We have :

$$\mathbb{E}\left[\frac{1}{n^2}\|\hat{\theta}-\theta\|^2+\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] \geq \mathbb{E}\left[\frac{1}{n}\|\hat{\mu}-\mu\|^2\right]$$

Since:

$$\sup_{\mu\in\mathcal{U}_k}\mathbb{E}\left[\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] = \sup_{(\theta,\mu)\in\Theta\mathcal{U}_k}\mathbb{E}\left[\frac{1}{n}\|\hat{\mu}-\mu\|^2\right]$$

Therefore:

$$\inf_{\hat{\theta},\hat{\mu}}\sup_{(\theta,\mu)\in\Theta\mathcal{U}_k}\mathbb{E}\left[\frac{1}{n^2}\|\hat{\theta}-\theta\|^2+\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] \geq cte > 0$$

For the upper bound, by choosing good enough estimators, like the one given in Gao et al. (2015) for $\hat{\theta}$ and $\hat{\mu}_i = X_i$, we get :

$$\inf_{\hat{\theta},\hat{\mu}}\sup_{(\theta,\mu)\in\Theta\mathcal{U}_k}\mathbb{E}\left[\frac{1}{n^2}\|\hat{\theta}-\theta\|^2+\frac{1}{n}\|\hat{\mu}-\mu\|^2\right] \leq O(1)$$

□

Proof of proposition 3.2

Proof. For any $l \in \mathbb{N}$ and $x, x' \in I$:

$$T^{(l)}f_0(x) - T^{(l)}f_0(x') = \int_{[0,1]}\cdots\int_{[0,1]}(w(x,x_1)-w(x',x_1))w(x_1,x_2)\cdots w(x_{l-1},x_l)f_0dx_1\cdots dx_l$$

Which implies:

$$\left|T^{(l)}f_0(x) - T^{(l)}f_0(x')\right| \leq Cf_0|x-x'|^\alpha$$

So we conclude:

$$|f(x)-f(x')| \leq Cf_0\sum_{l\leq d} y_l|x-x'|^\alpha$$

□

Proof of proposition 4.1

Proof. We follow the proof in Daudin et al. (2008).

$$\begin{aligned}\mathcal{J}(R_\mathcal{X}) &= \sum_z R_\mathcal{X}(z)\log(p(\mathcal{X},z|\Phi)) - \sum_z R_\mathcal{X}(z)\log(R_\mathcal{X}(z))\\ &= \mathbb{E}[L(\mathcal{X},z) \mid z\sim R_\mathcal{X}] - \mathbb{E}[\log(R_\mathcal{X}(z)) \mid z\sim R_\mathcal{X}]\\ &= \sum_{i\leq n}\sum_{a\leq k}\tau_{ia}\log(\pi_a) + \sum_{i<j\leq n}\sum_{a,b\leq k}\tau_{ia}\tau_{jb}\log(b(A_{ij}|Q_{ab}))\\ &+ \sum_{i\leq n}\sum_{a\leq k}\tau_{ia}\log(g(X_i|M_a)) - \sum_{i\leq n}\sum_{a\leq k}\tau_{ia}\log(\tau_{ia})\end{aligned}$$

We maximize $\mathcal{J}(R_{\mathcal{X}})$ w.r.t τ subject to $\sum_a \tau_{ia} = 1$, $\forall i$ by maximizing $\mathcal{J}(R_{\mathcal{X}})+\sum_i \lambda_i(\sum_a \tau_{ia}-1)$ where λ_i are the Lagrange multipliers. Differentiating w.r.t τ_{ia} we get:

$$\log(\pi_a) + \sum_{j\neq i}\sum_{b\leq k} \tau_{ib}\log(b(A_{ij}|Q_{ab})) + \log(g(X_i, M_a)) - \log(\tau_{ia}) + 1 + \lambda_i$$

The derivative is null if and only if τ_{ia} satisfies the relation given in the proposition, $\exp(\lambda_i + 1)$ being the normalizing constant. □

Proof of proposition 4.2

Proof. For π_a and Q_{ab} we can directly use the results from Daudin et al. (2008). We only differentiate for the block means M_a:

$$\sum_i \tau_{ia}\frac{-(M_a - X_i)}{g(X_i|M_a)} g(X_i|M_a) = 0$$
$$\implies M_a = \frac{\sum_i \tau_{ia}X_i}{\sum_i \tau_{ia}}$$

□

proof of proposition 4.3

Proof. To derive the ICL we assume $p(\Phi \mid m_k) = p(Q, M \mid m_k)p(\pi \mid m_K)$. A lemma from Biernacki et al. (2000) gives:

$$L(\mathcal{X}, z \mid m_k) = L(z \mid m_k) + L(\mathcal{X} \mid z, m_k)$$

For $L(z \mid m_k)$ we do as in Biernacki et al. (2000) and Daudin et al. (2008) and consider a non-informative Jeffreys prior and use a Stirling approximation for the Γ function. Getting at the end:

$$L(z \mid m_k) = \max_{\pi} L(z \mid \pi, m_k) - \frac{k-1}{2}\log n$$

For the second term we do, again, as in both papers and use a Bayesian information criterion (BIC) approximation:

$$L(\mathcal{X} \mid z, m_k) \simeq \max_{Q,M} L(\mathcal{X} \mid z, Q, M, m_k) - \frac{1}{2}\left(\frac{k(k+1)}{2} + k\right)\log\left(\frac{n(n-1)}{2} + n\right)$$

Summing both gives the desired expression. □

References

[1] Abbe, E. (2023). Community Detection and Stochastic Block Models. arXiv:1703.10146.

[2] Abbe, E., Fan, J., and Wang, K. (2022). An ℓ_p theory of PCA and spectral clustering. arXiv:2006.14062 [math].

[3] Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. arXiv:1311.1731 [stat].

[4] Avella-Medina, M., Parise, F., Schaub, M. T., and Segarra, S. (2020). Centrality Measures for Graphons: Accounting for Uncertainty in Networks. *IEEE Transactions on Network Science and Engineering*, 7(1):520–537. Conference Name: IEEE Transactions on Network Science and Engineering.

[5] Awasthi, P. and Sheffet, O. (2012). Improved Spectral-Norm Bounds for Clustering. arXiv:1206.3204 [cs].

[6] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073. Publisher: Proceedings of the National Academy of Sciences.

[7] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Biernacki, C., Celeux, G., and Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

[9] Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377. arXiv:1411.2158 [stat].

[10] Biscarri, W. (2019). *Statistical methods for binomial and Gaussian sequences.* text, University of Illinois at Urbana-Champaign.

[11] Bogachev, V. I. (2007). *Measure Theory.* Springer, Berlin, Heidelberg.

[12] Borgs, C. and Chayes, J. T. (2017). Graphons: A Nonparametric Method to Model, Estimate, and Design Algorithms for Massive Networks. arXiv:1706.01143 [cs].

[13] Borgs, C., Chayes, J. T., Lovasz, L., Sos, V. T., and Vesztergombi, K. (2007). Convergent Sequences of Dense Graphs I: Subgraph Frequencies, Metric Properties and Testing. arXiv:math/0702004.

[14] Braun, G., Tyagi, H., and Biernacki, C. (2022). An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees. arXiv:2112.10467 [stat].

[15] Cai, D., Ackerman, N., and Freer, C. (2015). An iterative step-function estimator for graphons. arXiv:1412.2129 [math].

[16] Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6(none):1847–1899. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[17] Chan, S. H. and Airoldi, E. M. (2014). A Consistent Histogram Estimator for Exchangeable Graph Models. arXiv:1402.1888 [stat].

[18] Chanda, K. C. (1954). A Note on the Consistency and Maxima of the Roots of Likelihood Equations. *Biometrika*, 41(1/2):56–61. Publisher: [Oxford University Press, Biometrika Trust].

[19] Chatterjee, S. (2015). Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1). arXiv:1212.1247 [math].

[20] Chen, J. (2023). *Statistical Inference Under Mixture Models.* ICSA Book Series in Statistics. Springer Nature, Singapore.

[21] Chen, Y. and Lei, J. (2024). Minimax Optimal Probability Matrix Estimation For Graphon With Spectral Decay. arXiv:2410.01073 [math].

[22] Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284. Publisher: [Oxford University Press, Biometrika Trust].

[23] Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

[24] De Nicola, G., Sischka, B., and Kauermann, G. (2022). Mixture models and networks: The stochastic blockmodel. *Statistical Modelling*, 22(1-2):67–94. Publisher: SAGE Publications India.

[25] Delmas, J.-F., Dronnier, D., and Zitt, P.-A. (2021). Optimal vaccination: Various (counter) intuitive examples. arXiv:2112.08756 [math].

[26] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 39(1):1–22.

[27] Doss, N., Wu, Y., Yang, P., and Zhou, H. H. (2021). Optimal estimation of high-dimensional location Gaussian mixtures. arXiv:2002.05818 [math].

[28] Dreveton, M., Fernandes, F., and Figueiredo, D. (2023). Exact recovery and Bregman hard clustering of node-attributed Stochastic Block Model. *Advances in Neural Information Processing Systems*, 36:37827–37848.

[29] Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate-Optimal Graphon Estimation. *The Annals of Statistics*, 43(6):2624–2652. Publisher: Institute of Mathematical Statistics.

[30] Gao, C. and Ma, Z. (2021). Minimax Rates in Network Analysis: Graphon Estimation, Community Detection and Hypothesis Testing. *Statistical Science*, 36(1):16–33. Publisher: Institute of Mathematical Statistics.

[31] Gaucher, S. and Klopp, O. (2021). Optimality of variational inference for stochastic block model with missing links. arXiv:2111.03305 [math].

[32] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870. Publisher: Institute of Mathematical Statistics.

[33] Hoff, P. D. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. arXiv:0711.1146 [stat].

[34] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098. Publisher: ASA Website _eprint: https://doi.org/10.1198/016214502388618906.

[35] Jaeger, M., Longa, A., Azzolin, S., Schulte, O., and Passerini, A. (2024). A Simple Latent Variable Model for Graph Learning and Inference. In *Proceedings of the Second Learning on Graphs Conference*, pages 26:1–26:18. PMLR. ISSN: 2640-3498.

[36] Janson, S. (2011). Graphons, cut norm and distance, couplings and rearrangements. arXiv:1009.2376 [math].

[37] Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles.* Probability and Its Applications. Springer-Verlag, New York.

[38] Klopp, O., Tsybakov, A. B., and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. arXiv:1507.04118.

[39] Klopp, O. and Verzelen, N. (2019). Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174(3-4):1033–1090. Publisher: Springer Verlag.

[40] Lafferty, J., Wasserman, L., and Liu, H. (2008). Lecture notes on Minimax theory, Carnegie Melon University, [https://www.stat.cmu.edu/˜larry/=sml/Minimax.pdf].

[41] Latouche, P., Birmele, E., and Ambroise, C. (2010). Variational Bayesian Inference and Complexity Control for Stochastic Block Models. arXiv:0912.2873 [stat].

[42] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2–es.

[43] Levie, R. (2023). A graphon-signal analysis of graph neural networks. arXiv:2305.15987.

[44] Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[45] Lovász, L. (2012). *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications.* American Mathematical Society, Providence, Rhode Island.

[46] Lovász, L. and Szegedy, B. (2007). Szemerédi's Lemma for the Analyst. *GAFA Geometric And Functional Analysis*, 17(1):252–270.

[47] Lu, Y. and Zhou, H. H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. arXiv:1612.02099.

[48] Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283. Publisher: Institute of Mathematical Statistics.

[49] Massoulié, L. (2014). Community detection thresholds and the weak Ramanujan property. In *STOC 2014: 46th Annual Symposium on the Theory of Computing*, pages 1–10, New York, United States.

[50] Mele, A., Hao, L., Cape, J., and Priebe, C. E. (2021). Spectral inference for large Stochastic Blockmodels with nodal covariates. arXiv:1908.06438 [stat].

[51] Mossel, E., Neeman, J., and Sly, A. (2016). Consistency Thresholds for the Planted Bisection Model. *Electronic Journal of Probability*, 21(none). arXiv:1407.1591 [math].

[52] Neykov, M. (2022). On the minimax rate of the Gaussian sequence model under bounded convex constraints. arXiv:2201.07329.

[53] Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727. arXiv:1312.5306 [stat].

[54] Orbanz, P. and Roy, D. M. (2015). Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. arXiv:1312.7857 [math].

[55] Parise, F. and Ozdaglar, A. (2019). Graphon Games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pages 457–458, New York, NY, USA. Association for Computing Machinery.

[56] Peters, B. C. and Walker, H. F. (1978). An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM Journal on Applied Mathematics*, 35(2):362–378. Publisher: Society for Industrial and Applied Mathematics.

[57] Peyré, G. and Cuturi, M. (2020). Computational Optimal Transport. arXiv:1803.00567 [stat].

[58] Ponti, A. (2024). Graph data augmentation with Gromow-Wasserstein Barycenters. arXiv:2404.08376 [cs].

[59] Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review*, 26(2):195–239. Publisher: Society for Industrial and Applied Mathematics.

[60] Rozemberczki, B. and Sarkar, R. (2021). Twitch Gamers: a Dataset for Evaluating Proximity Preserving and Structural Role-based Node Embeddings. arXiv:2101.03091 [cs].

[61] Ruiz, L., Chamon, L., and Ribeiro, A. (2020). Graphon Neural Networks and the Transferability of Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 1702–1712. Curran Associates, Inc.

[62] Ruiz, L., Chamon, L. F. O., and Ribeiro, A. (2021). Graphon Signal Processing. *IEEE Transactions on Signal Processing*, 69:4961–4976. arXiv:2003.05030 [eess].

[63] Saha, S. and Guntuboyina, A. (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *The Annals of Statistics*, 48(2):738–762. Publisher: Institute of Mathematical Statistics.

[64] Sischka, B. and Kauermann, G. (2022). EM-based smooth graphon estimation using MCMC and spline-based approaches. *Social Networks*, 68:279–295.

[65] Sischka, B. and Kauermann, G. (2024). Stochastic Block Smooth Graphon Model. *Journal of Computational and Graphical Statistics*, 0(0):1–15. Publisher: ASA Website _eprint: https://doi.org/10.1080/10618600.2024.2374571.

[66] Snijders, T. A. and Nowicki, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100.

[67] Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M., and Mucha, P. J. (2018). Stochastic Block Models with Multiple Continuous Attributes. arXiv:1803.02726 [cs].

[68] Su, Y., Wong, R. K. W., and Lee, T. C. M. (2020). Network Estimation via Graphon With Node Features. *IEEE Transactions on Network Science and Engineering*, 7(3):2078–2089. Conference Name: IEEE Transactions on Network Science and Engineering.

[69] Tarone, R. E. and Gruenhage, G. (1975). A Note on the Uniqueness of Roots of the Likelihood Equations for Vector-Valued Parameters. *Journal of the American Statistical Association*, 70(352):903–904. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1975.10480321.

[70] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York, NY.

[71] Villani, C. (2009). *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften.* Springer, Berlin, Heidelberg.

[72] Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601. Publisher: Institute of Mathematical Statistics.

[73] Wasserman, L., editor (2006). *Normal Means and Minimax Theory.* Springer, New York, NY.

[74] Weibel, J. (2024). *Graphons de probabilités, limites de graphes pondérés aléatoires et chaînes de Markov branchantes cachées.* phdthesis, Université d'Orléans.

[75] Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. arXiv:1309.5936 [math].

[76] Wu, W., Olhede, S., and Wolfe, P. (2025). Tractably modelling dependence in networks beyond exchangeability. *Bernoulli*, 31(1):584–608. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

[77] Wu, Y. and Yang, P. (2019). Optimal estimation of Gaussian mixtures via denoised method of moments. arXiv:1807.07237 [math].

[78] Xu, H., Luo, D., Carin, L., and Zha, H. (2020). Learning Graphons via Structured Gromov-Wasserstein Barycenters. arXiv:2012.05644 [cs].

[79] Zhang, Y., Levina, E., and Zhu, J. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2).

[80] Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighborhood smoothing. arXiv:1509.08588 [stat].