Markov Random Geometric Graph (MRGG): A Growth Model for Temporal Dynamic Networks

Yohann De Castro[1]

joint with: Quentin Duchemin[2] & Claire Lacour[3]

November 9th, 2022

[1] Institut Camille Jordan (ICJ) & Institut Universitaire de France (IUF)

[2] SDSC, EPFL

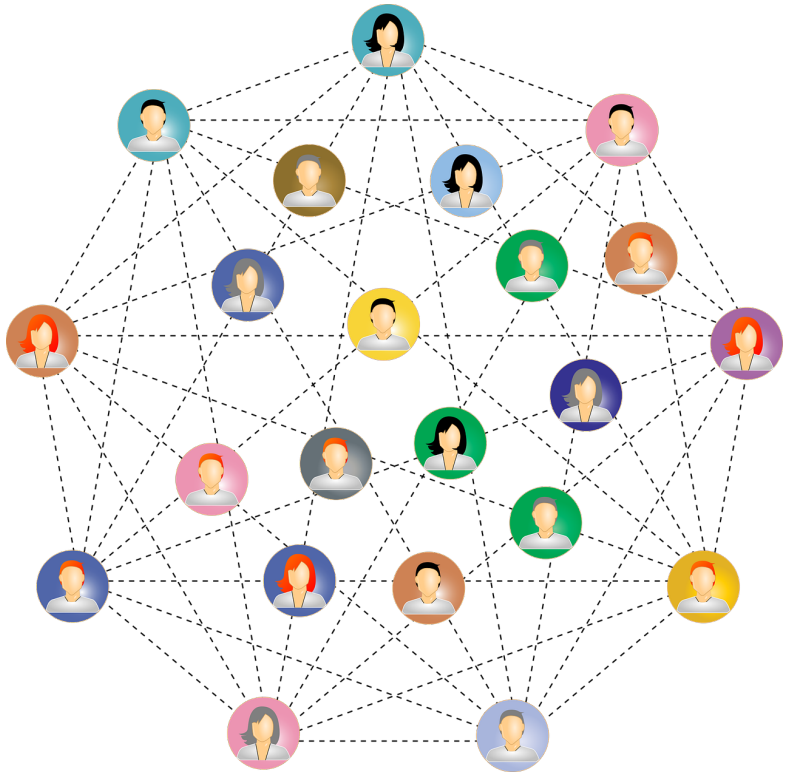
[3] LAMA, Univ. Gustave Eiffel

What is common to . . .

- Social networks
- Protein-protein interactions
- Freight transport between cities
- Spread of the virus in a population

?

All these datasets have a **graph structure**.



What can be tackled from graph data?

What can be tackled from graph data?

- **Security issues**

Is the user privacy preserved?

What can be tackled from graph data?

- **Security issues**

  Is the user privacy preserved?

- **Prediction tasks**

  What is the probability for two new users to be connected?

What can be tackled from graph data?

- **Security issues**

  Is the user privacy preserved?

- **Prediction tasks**

  What is the probability for two new users to be connected?

- **Estimation tasks**

  Can we learn the preference of each user?

What can be tackled from graph data?

- **Security issues** $\longrightarrow$ Graph alignment

  Is the user privacy preserved?

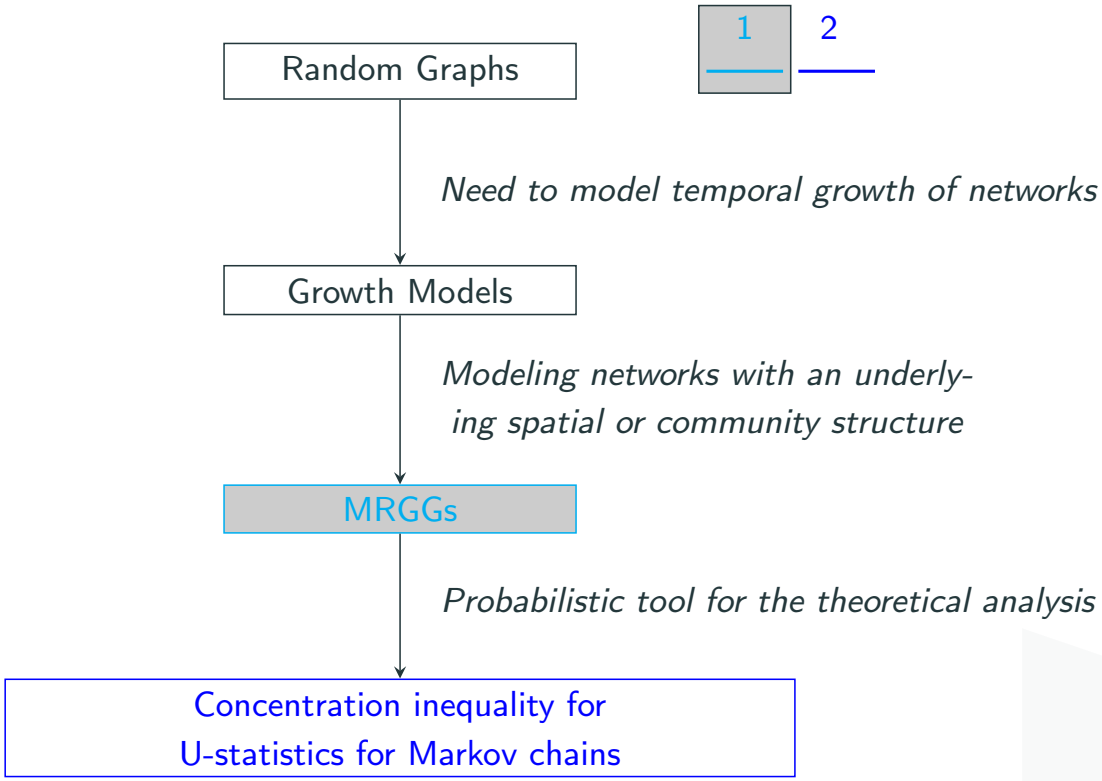
- **Prediction tasks** $\longrightarrow$ Link prediction

  What is the probability for two new users to be connected?

- **Estimation tasks** $\longrightarrow$ Community detection

  Can we learn the preference of each user?

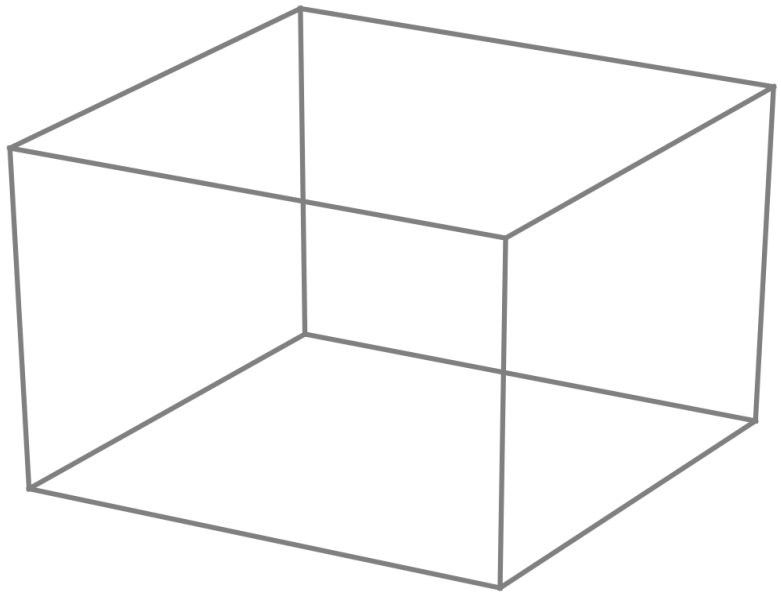
Outline of the presentation



I.1 Random Geometric Graphs: the problem of *geometry detection*

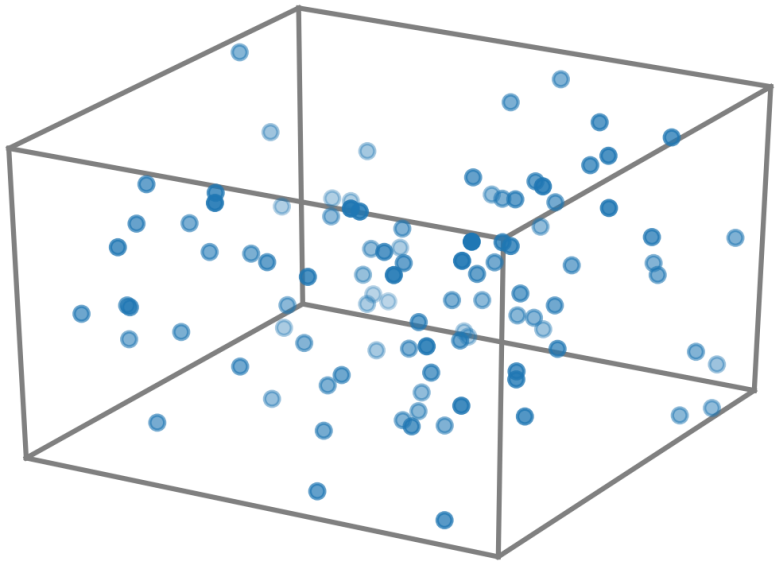
Presentation of the Random Geometric Graph (RGG)

1. Choose a latent space.



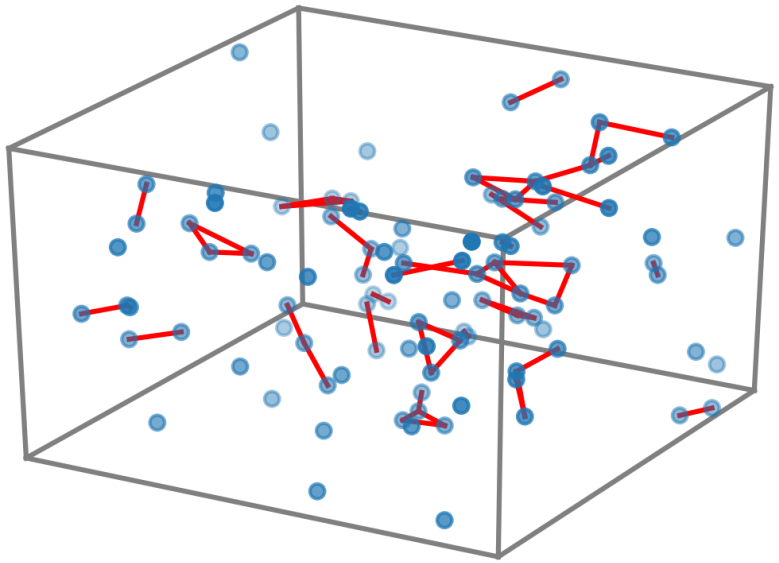
Presentation of the Random Geometric Graph (RGG)

(2) Sample points.



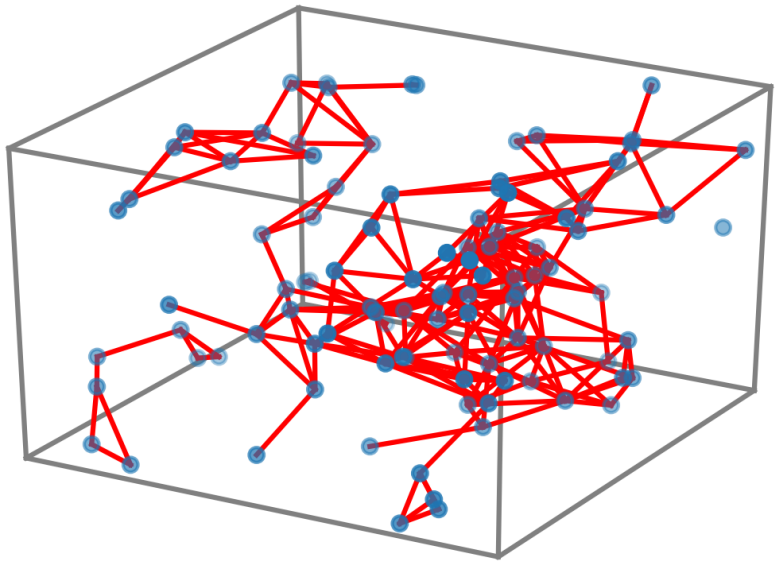
Presentation of the Random Geometric Graph (RGG)

3 Draw edges.



Presentation of the Random Geometric Graph (RGG)

3. Draw edges.



RGGs: The high-dimensional regime

- **Hard RGGs on $\mathbb{S}^{d-1}$:**

$$X_i \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$$

$$i \sim j \quad \Leftrightarrow \quad \langle X_i, X_j \rangle \geq t^*$$

Y.D.C. & Quentin Duchemin. Random Geometric Graph: Some recent developments and perspectives. *High Dimensional Probability IX, Progress in Probability series*, 2022.

RGGs: The high-dimensional regime

- **Hard RGGs on $\mathbb{S}^{d-1}$:**

$$X_i \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$$
$$i \sim j \quad \Leftrightarrow \quad \langle X_i, X_j \rangle \geq t^*$$

- Studying the behaviour of Hard RGGs in **the high-dimensional setting** (i.e. when $d = d(n)$ goes to $+\infty$ as $n \to +\infty$):

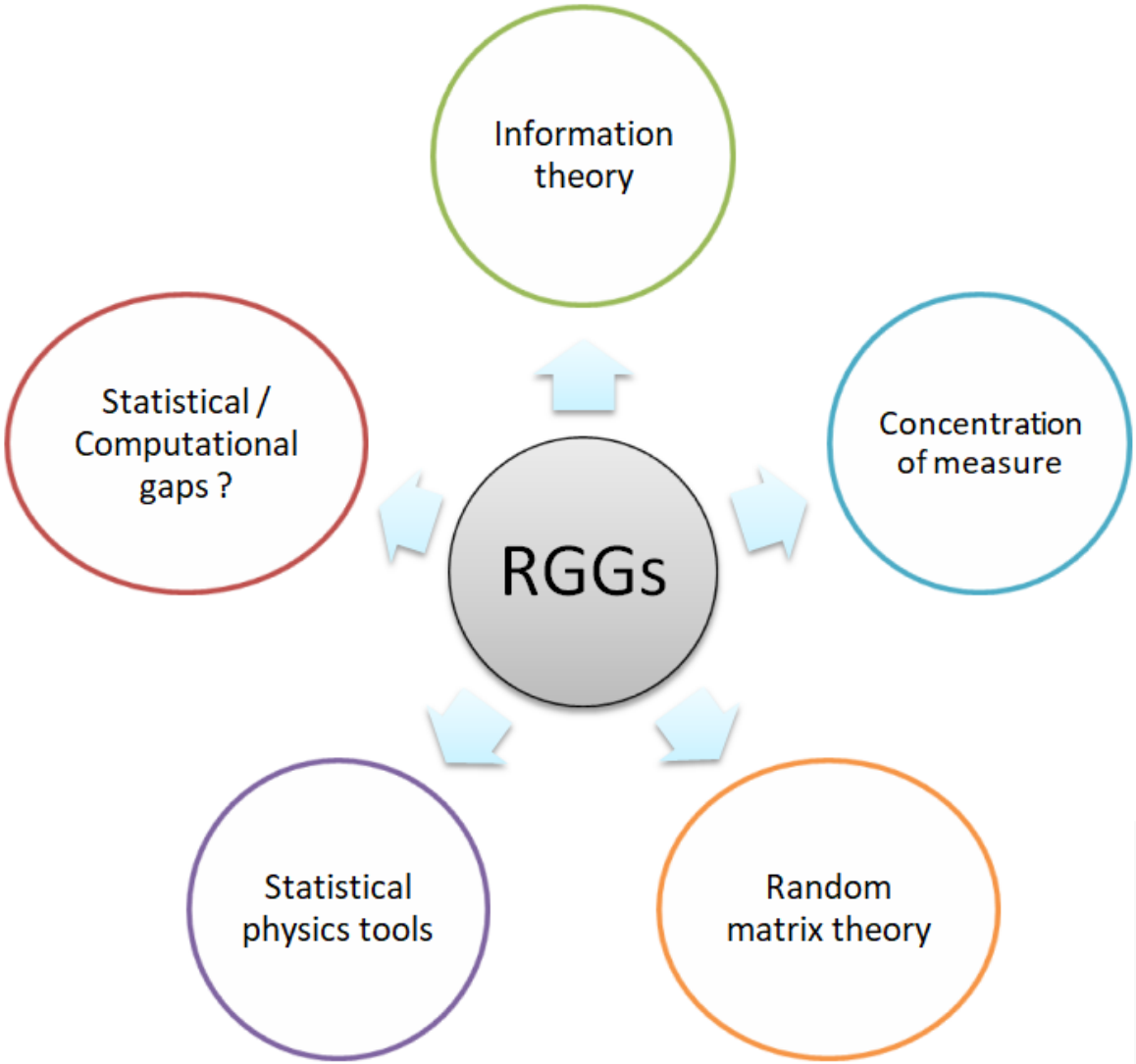
  Given some graph G with n points, the *geometry detection problem* reads as

$$\mathbb{H}_0 : "G \sim G(n,p)" \quad \text{VS} \quad \mathbb{H}_1 : "G \sim \mathrm{RGG}(n,p,d)",$$

  where $G(n,p)$ is the Erdös-Renyi distribution and $\mathrm{RGG}(n,p,d)$ is the distribution of Hard-RGGs on $\mathbb{S}^{d-1}$ with matching expected degree.

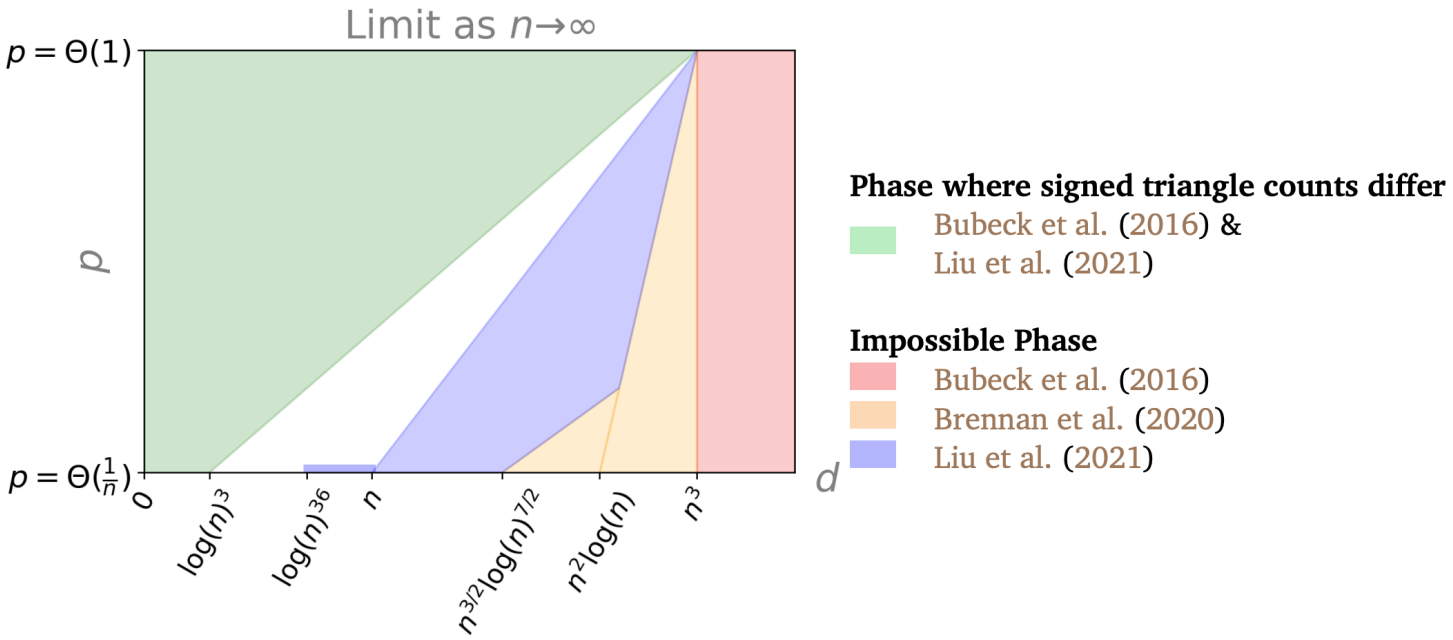
Y.D.C. & Quentin Duchemin. Random Geometric Graph: Some recent developments and perspectives. *High Dimensional Probability IX, Progress in Probability series*, 2022.

Mathematical Tools to study RGGs in high dimensions



Phase diagram for the problem of geometry detection

Task	Current state of knowledge	Ref.
Recognizing if a graph can be realized as a RGG	NP-hard	1998
Testing between $G(n,p,d)$ and $G(n,p)$ in high-dimension for $p \in (0,1)$ fixed	0 — Polynomial time test — n^3 — Undistinguishable — d	2016
Testing between $G(n,\frac{c}{n},d)$ and $G(n,\frac{c}{n})$ in high-dimension for $c>0$	0 — $\log^3 n$ — $\log^{36} n$ — d; Polynomial time test — ? — Undistinguishable	2016 & 2021



Y.D.C. & Quentin Duchemin. Random Geometric Graph: Some recent developments and perspectives. *High Dimensional Probability IX, Progress in Probability series*, 2022.

Literature in RGGs

- First motivated to model **wireless communication systems**.
- An active line of research is studying the behaviour of RGGs in **the high-dimensional setting** (i.e. when $d = d(n)$ goes to $+\infty$ as $n \to +\infty$).

Literature in RGGs

- First motivated to model **wireless communication systems**.
- An active line of research is studying the behaviour of RGGs in **the high-dimensional setting** (i.e. when $d = d(n)$ goes to $+\infty$ as $n \to +\infty$).

  In this work, we adopt another point of view and we focus on non-parametric estimation in a new *growth model* for RGGs.

I.2 Non-parametric estimation in RGGs

Non-parametric estimation in RGGs (1/2)

We work with the Euclidean d-dimensional sphere

$$\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\| = 1\}.$$

Non-parametric estimation in RGGs (1/2)

We work with the Euclidean d-dimensional sphere

$$\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\| = 1\}.$$

The graph is built by

- sampling n points $X_1, \ldots, X_n$ on $\mathbb{S}^{d-1}$.

Non-parametric estimation in RGGs (1/2)

We work with the Euclidean d-dimensional sphere

$$\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\| = 1\}.$$

The graph is built by

- sampling n points $X_1, \ldots, X_n$ on $\mathbb{S}^{d-1}$.
- setting an edge between nodes i and j with $i \neq j$ with probability $\mathsf{p}(\langle X_i, X_j \rangle)$ where $\mathsf{p} : [-1, 1] \to [0, 1]$ is called the *envelope function*.

Non-parametric estimation in RGGs (1/2)

We work with the Euclidean d-dimensional sphere

$$\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \ : \ \|x\| = 1\}.$$

The graph is built by

- sampling n points $X_1, \dots, X_n$ on $\mathbb{S}^{d-1}$.
- setting an edge between nodes i and j with $i \neq j$ with probability $\mathsf{p}(\langle X_i, X_j \rangle)$ where $\mathsf{p} : [-1, 1] \to [0, 1]$ is called the *envelope function*.

  The adjacency matrix A is such that

$$\forall i, j \in [n], \quad \begin{cases} A_{i,j} & = 0 \text{ if } i = j \\ A_{i,j} & \sim \mathrm{Ber}(\mathsf{p}(\langle X_i, X_j \rangle)) \text{ if } i \neq j \end{cases}.$$

This RGG corresponds to a graphon-type model where the graphon W is

$$\forall x, y \in \mathbb{S}^{d-1}, \quad W(x, y) := \mathsf{p}(\langle x, y \rangle).$$

Typical example: $\mathsf{p} : t \mapsto \mathbb{1}_{t \geq t^*}$, $t^* \in (-1, 1)$.

Non-parametric estimation in RGGs (2/2)

Can we recover p **from the observation of the graph (i.e.** A**)?**

Previous work In some work[1], the authors address the non-parametric estimation of the envelope function p in the independent setting.

[1] Y.D.C., C. Lacour and T.M Pham Ngoc, Adaptive Estimation of Nonparametric Geometric Graphs. *Mathematical Statistics and Learning*, 2020.

Non-parametric estimation in RGGs (2/2)

Can we recover p from the observation of the graph (i.e. A)?

Previous work In some work[1], the authors address the non-parametric estimation of the envelope function p in the independent setting.

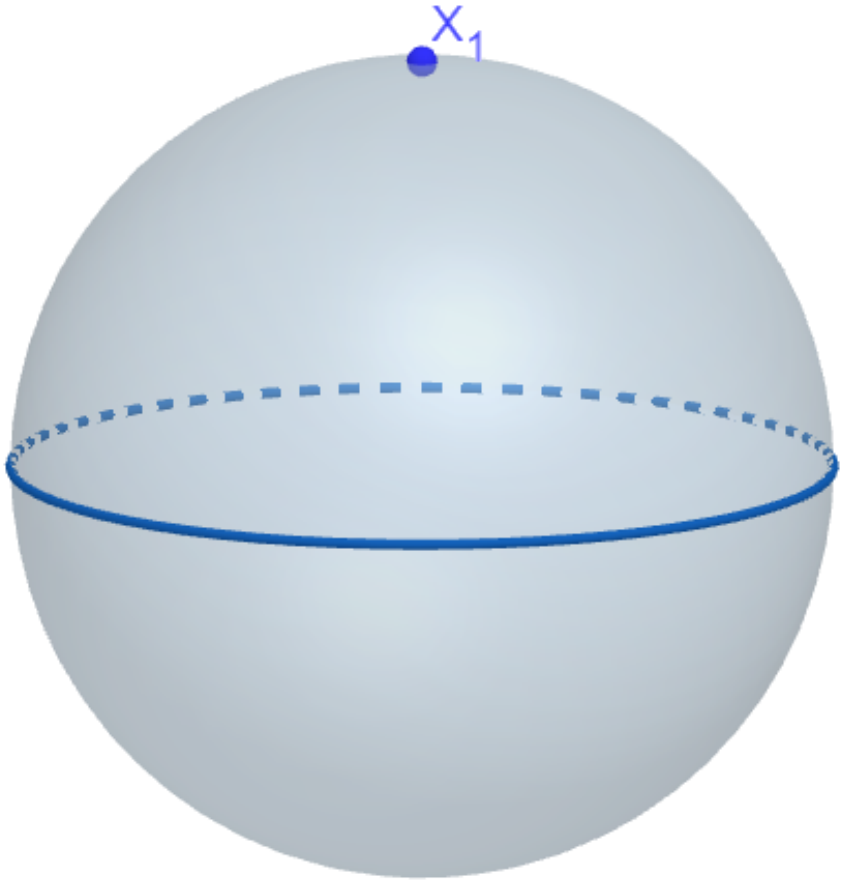
Goal

- Extend the previous work to a dependent framework.
  $\rightsquigarrow$ We sample the latent space using a Markovian dynamic.
- Estimate the Markov transition kernel.
  $\rightsquigarrow$ We propose a heuristic to solve link prediction problems.

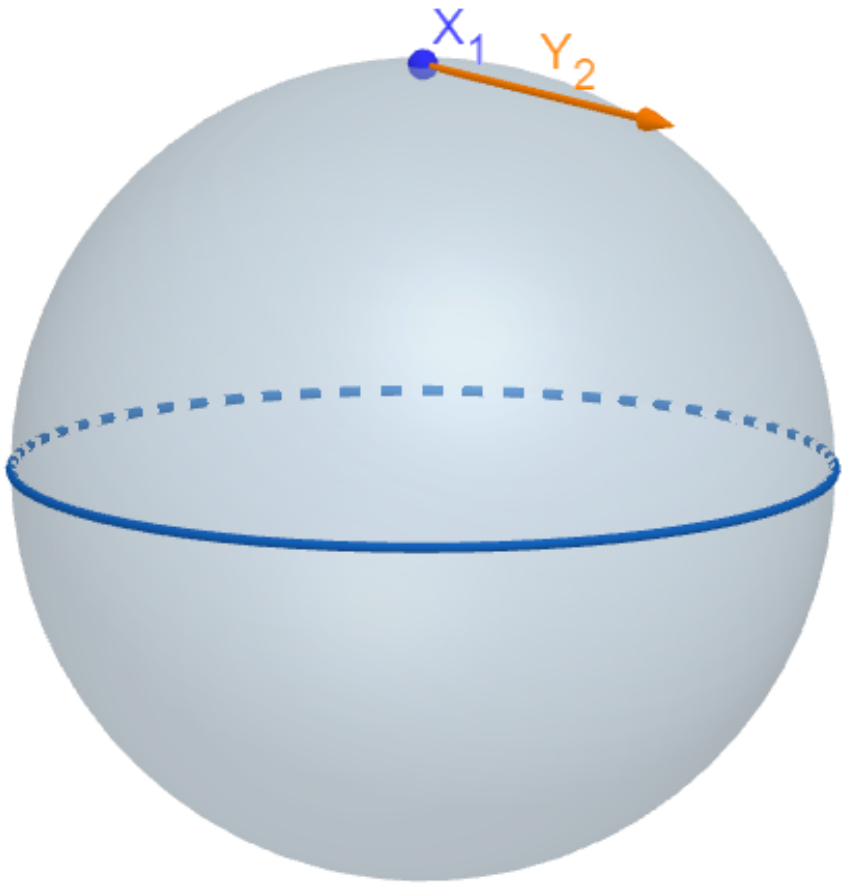
[1] Y.D.C., C. Lacour and T.M Pham Ngoc, Adaptive Estimation of Nonparametric Geometric Graphs. *Mathematical Statistics and Learning*, 2020.

I.3 The Markovian dynamic

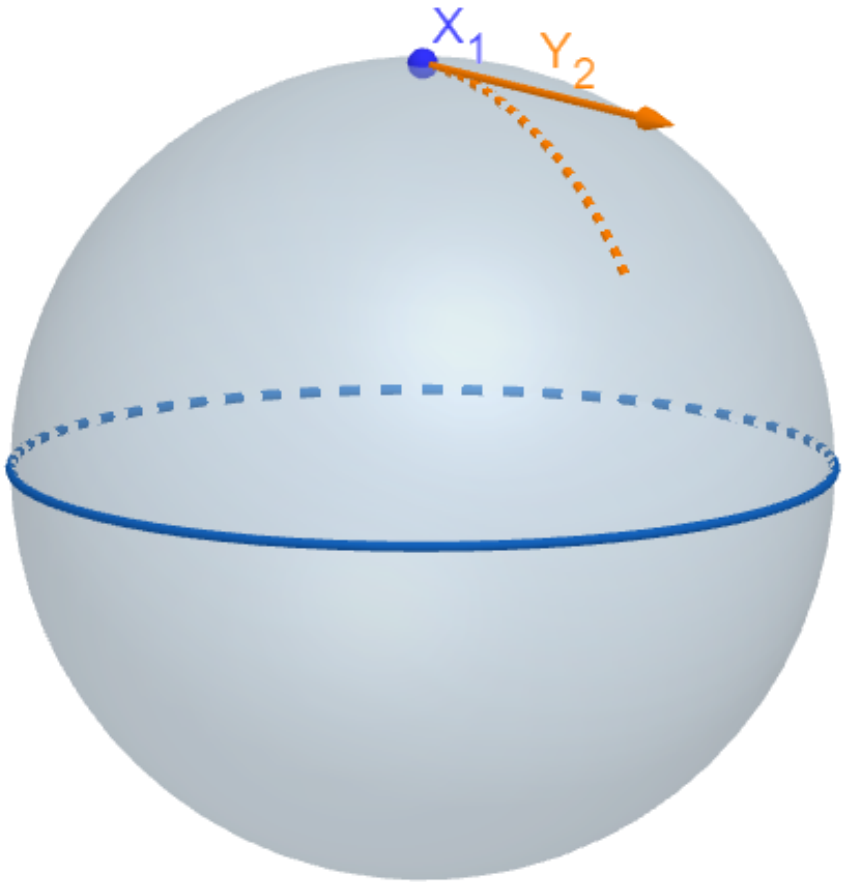
Latent Markovian dynamic: Isotropic sampling



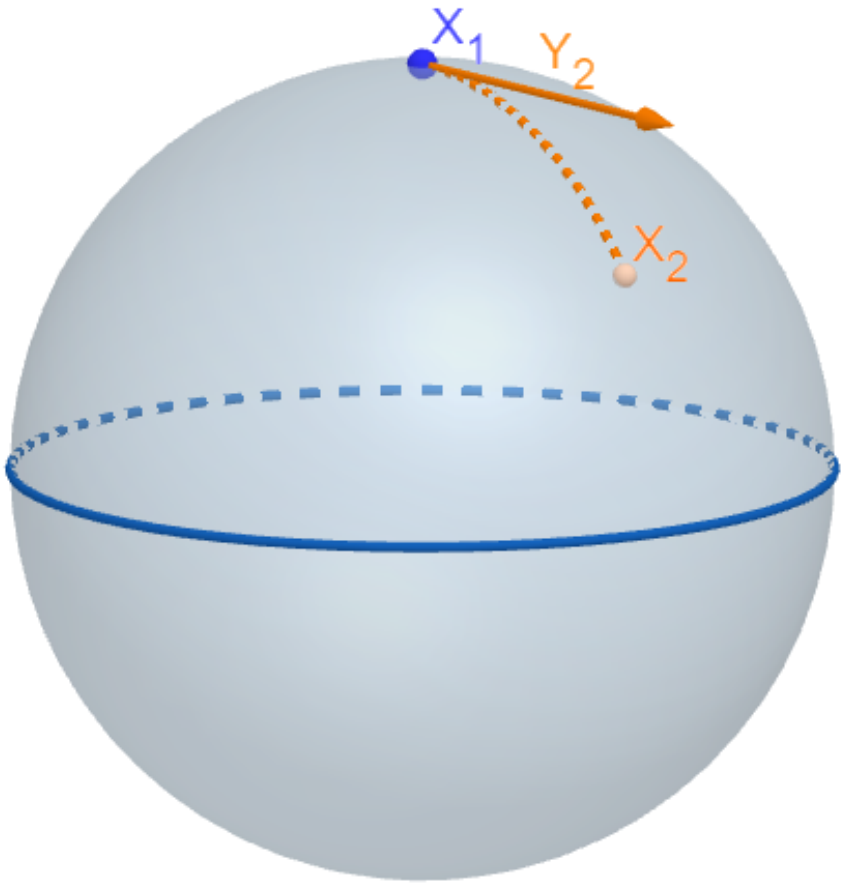
Latent Markovian dynamic: Isotropic sampling



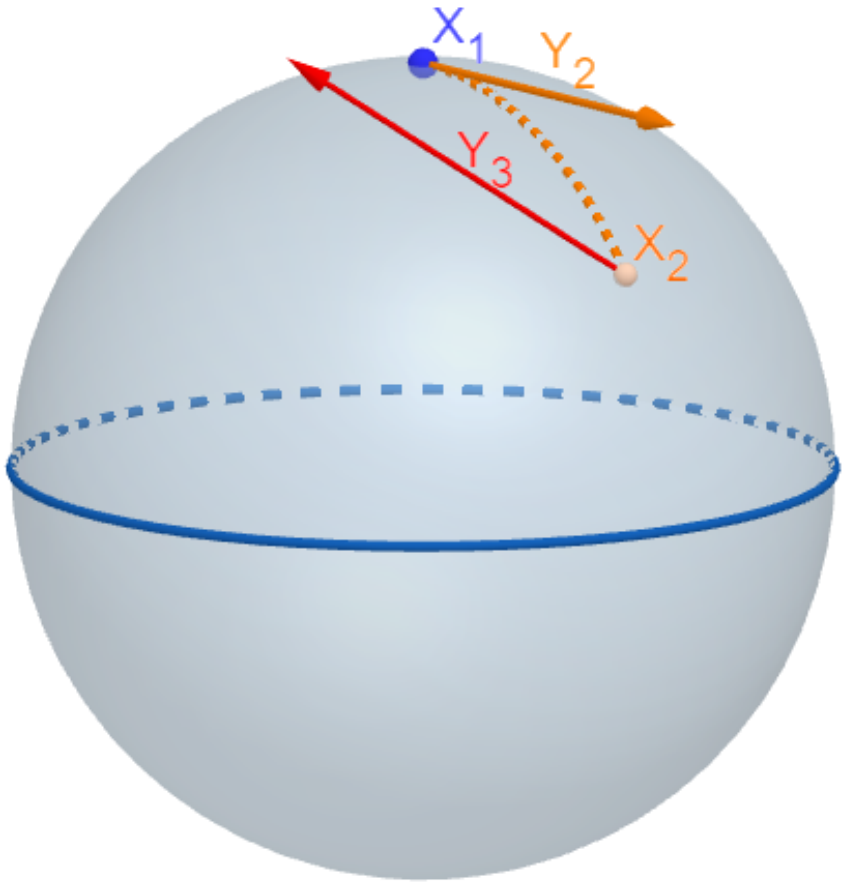
Latent Markovian dynamic: Isotropic sampling



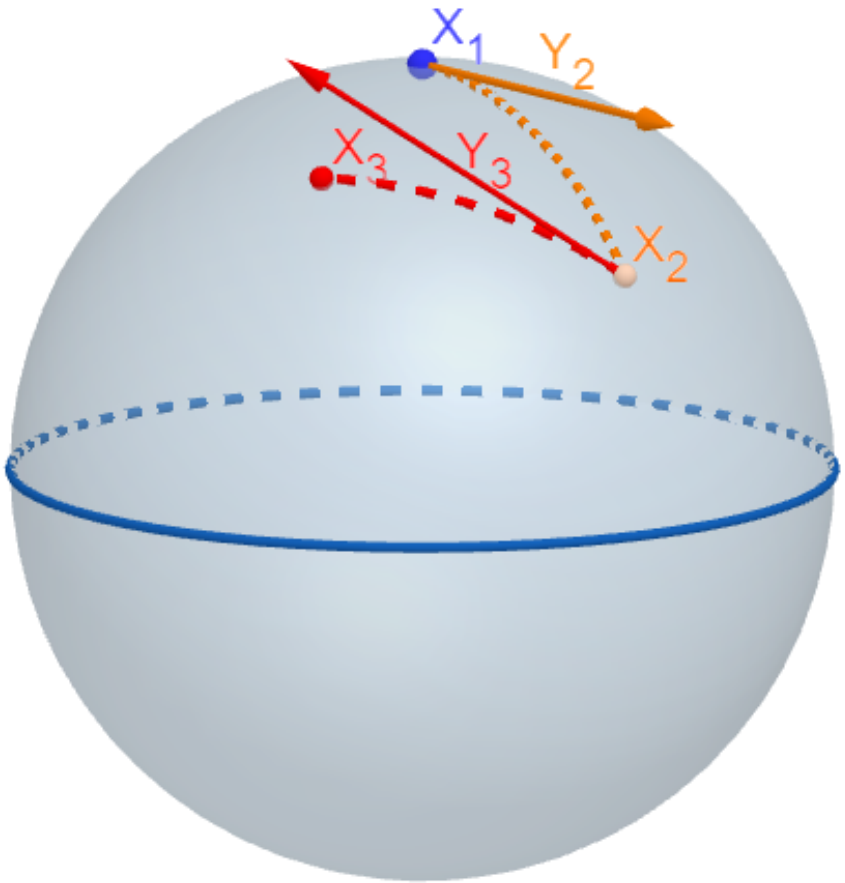
Latent Markovian dynamic: Isotropic sampling



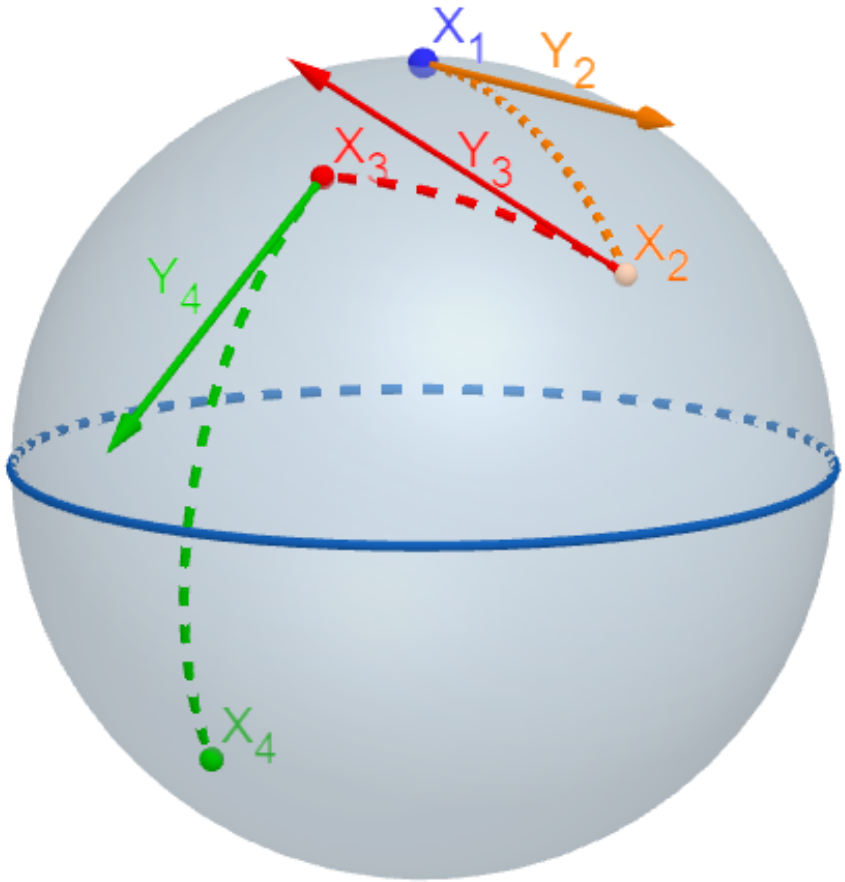
Latent Markovian dynamic: Isotropic sampling



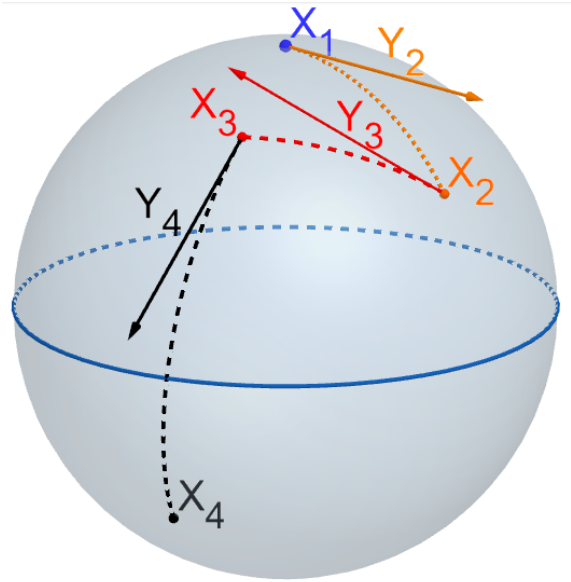
Latent Markovian dynamic: Isotropic sampling



Latent Markovian dynamic: Isotropic sampling



Sampling procedure: Isotropic Markovian dynamic

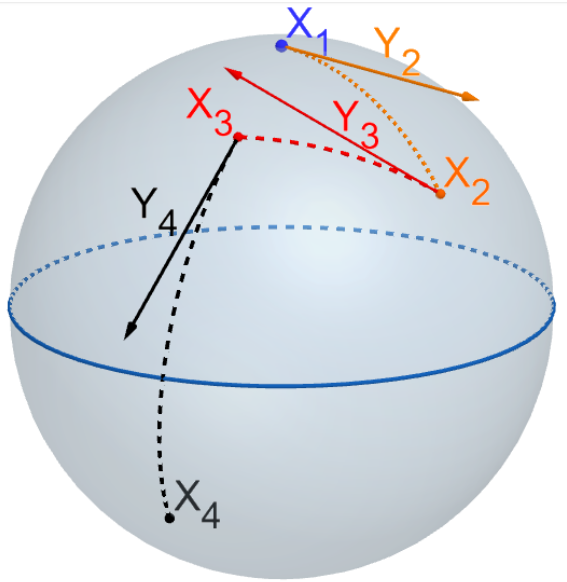


Sampling procedure: Isotropic Markovian dynamic

- $X_1 \sim \mathcal{U}(\mathbb{S}^{d-1})$
- $\forall i \in \{2, \dots, n\}$,
  - $Y_i \sim \mathcal{U}(\mathbb{S}^{d-1}) \mid Y_i \perp X_{i-1}$.
  - $r_i \sim f_{\mathcal{L}} : [-1, 1] \to [0, 1]$.
    $f_{\mathcal{L}}$ is called the *latitude function*.

  Then X_i is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1 - r_i^2} \times Y_i \, .$$

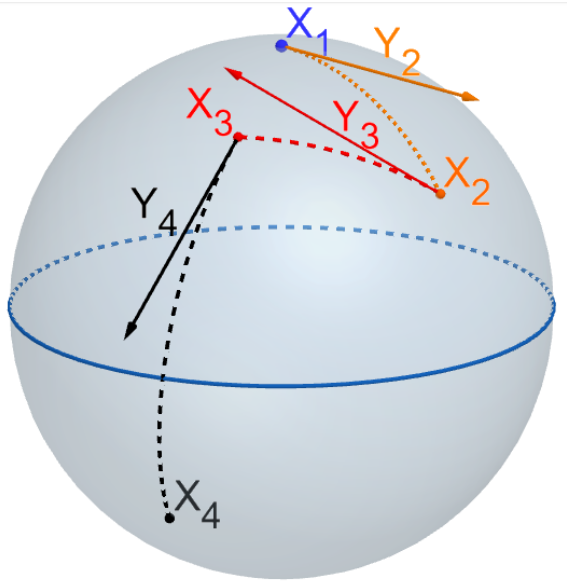


Sampling procedure: Isotropic Markovian dynamic

- $X_1 \sim \mathcal{U}(\mathbb{S}^{d-1})$
- $\forall i \in \{2, \dots, n\}$,
  - $Y_i \sim \mathcal{U}(\mathbb{S}^{d-1}) \mid Y_i \perp X_{i-1}$.
  - $r_i \sim f_{\mathcal{L}} : [-1, 1] \to [0, 1]$.
    $f_{\mathcal{L}}$ is called the *latitude function*.

Then X_i is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1 - r_i^2} \times Y_i \,.$$



Construction of the graph

$$i \leftrightarrow j \;\; \text{(i.e. } A_{i,j} = 1\text{)} \;\; \text{with proba.} \;\; W(X_i, X_j) := \mathrm{p}(\langle X_i, X_j \rangle).$$

I.3 Graphon estimation via Harmonic decomposition and Matrix concentration

Estimation of p: A plug-in approach

- W **can be viewed as a translation-invariant integral operator.**

$$\mathbb{T}_W : f \in L^2(\mathbb{S}^{d-1}) \mapsto \int_{\mathbb{S}^{d-1}} p(\langle x, \cdot \rangle) f(x) \sigma(dx) \in L^2(\mathbb{S}^{d-1}).$$

Estimation of p: A plug-in approach

- W **can be viewed as a translation-invariant integral operator.**

$$\mathbb{T}_W : f \in L^2(\mathbb{S}^{d-1}) \mapsto \int_{\mathbb{S}^{d-1}} p(\langle x, \cdot \rangle) f(x) \sigma(dx) \in L^2(\mathbb{S}^{d-1}).$$

- $\mathbb{T}_W$ **is Hilbert-Schmidt.**
  - Spectrum of $\mathbb{T}_W$: $\lambda^* = \{p_0^*, p_1^*, \dots, p_1^*, \dots, p_l^*, \dots, p_l^*, \dots\}$, with known multiplicities that depend only on the dimension of the sphere.
  - Decomposition of p: $\mathrm{p}(t) = \sum_{k \geq 0} p_k^* \phi_k(t)$ where $(\phi_k)_k$ do not depend on the graphon W, they are the Gegenbauer polynomials.
  - $k = 1$ corresponds to linear kernel: $\mathrm{p}(\langle x, y \rangle) = c_0 p_0^* + c_1 p_1^* \langle x, y \rangle + \dots$ (with c_k explicit constants) and p_1^* has multiplicity d in λ^*.

Estimation of p: A plug-in approach

- W **can be viewed as a translation-invariant integral operator.**

$$\mathbb{T}_W : f \in L^2(\mathbb{S}^{d-1}) \mapsto \int_{\mathbb{S}^{d-1}} p(\langle x, \cdot \rangle) f(x) \sigma(dx) \in L^2(\mathbb{S}^{d-1}).$$

- $\mathbb{T}_W$ **is Hilbert-Schmidt.**
  - Spectrum of $\mathbb{T}_W$: $\lambda^* = \{p_0^*, p_1^*, \ldots, p_1^*, \ldots, p_l^*, \ldots, p_l^*, \ldots\}$, with known multiplicities that depend only on the dimension of the sphere.
  - Decomposition of p: $\mathrm{p}(t) = \sum_{k \geq 0} p_k^* \phi_k(t)$ where $(\phi_k)_k$ do not depend on the graphon W, they are the Gegenbauer polynomials.
  - $k = 1$ corresponds to linear kernel: $\mathrm{p}(\langle x, y \rangle) = c_0 p_0^* + c_1 p_1^* \langle x, y \rangle + \ldots$ (with c_k explicit constants) and p_1^* has multiplicity d in λ^*.
- **Goal**:
  Show that $\lambda(T_n) \to \lambda^*$ where $T_n = \frac{1}{n} \left((1 - \delta_{i,j}) \mathrm{p}\left(\langle X_i, X_j \rangle \right) \right)_{1 \leq i,j \leq n}$.
  Show that $\lambda(\widehat{T}_n) \to \lambda(T_n)$ where $\widehat{T}_n = \frac{1}{n} A$.

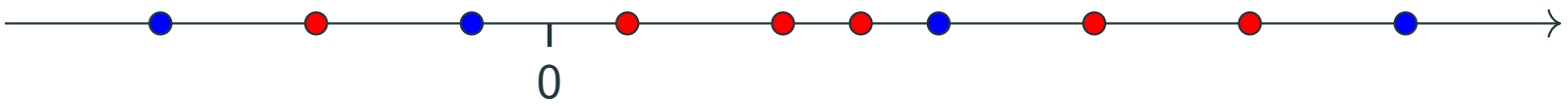
Estimation of p: Main result

Definition (Distance between spectra)

Given two sequences x, y of reals such that $\sum_i x_i^2 + y_i^2 < \infty$, we define

$$\delta_2^2(x, y) := \inf_{p \in \mathcal{S}} \sum_i (x_i - y_{p(i)})^2 ,$$

where $\mathcal{S}$ is the set of permutations of $\mathbb{N}$.



Estimation of p: Main result

Definition (Distance between spectra)

Given two sequences x, y of reals such that $\sum_i x_i^2 + y_i^2 < \infty$, we define

$$\delta_2^2(x, y) := \inf_{p \in \mathcal{S}} \sum_i (x_i - y_{p(i)})^2 ,$$

where $\mathcal{S}$ is the set of permutations of $\mathbb{N}$.

Estimation of p: Main result

Definition (Distance between spectra)

Given two sequences x, y of reals such that $\sum_i x_i^2 + y_i^2 < \infty$, we define

$$\delta_2^2(x, y) := \inf_{p \in \mathcal{S}} \sum_i (x_i - y_{p(i)})^2 ,$$

where $\mathcal{S}$ is the set of permutations of $\mathbb{N}$.

- $\delta_2^2 \left(\lambda(\widehat{T}_n), \lambda(T_n) \right) \to 0$, $\rightsquigarrow$ follows from one nice work[2].
- It remains to prove that $\delta_2^2 \left(\lambda(T_n), \lambda^* \right) \to 0$.

[2] Bandeira & Van Handel, Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *AoP*, 2016.

Estimation of p: Main result

Definition (Distance between spectra)

Given two sequences x, y of reals such that $\sum_i x_i^2 + y_i^2 < \infty$, we define

$$\delta_2^2(x, y) := \inf_{p \in \mathcal{S}} \sum_i (x_i - y_{p(i)})^2 ,$$

where $\mathcal{S}$ is the set of permutations of $\mathbb{N}$.

- $\delta_2^2\left(\lambda(\widehat{T}_n), \lambda(T_n)\right) \to 0$, $\rightsquigarrow$ follows from one nice work[2].
- It remains to prove that $\delta_2^2\left(\lambda(T_n), \lambda^*\right) \to 0$.

Theorem (EJS: 2022)

If
- p has regularity $s > 0$
- $\|f_{\mathcal{L}}\|_\infty < \infty$ and $\inf_{r \in [-1,1]} f_{\mathcal{L}}(r) > 0$

then for n large enough it holds

$$\mathbb{E}[\delta_2^2\left(\lambda(T_n), \lambda^*\right) \vee \delta_2^2(\lambda^{R_{opt}}(\widehat{T}_n), \lambda^*)] \lesssim \left[\frac{n}{\log^2 n}\right]^{-\frac{2s}{2s+(d-1)}},$$

where $\lambda^{R_{opt}}(\widehat{T}_n) = (\hat{\lambda}_1^{sort}, \dots, \hat{\lambda}_{\tilde{R}_{opt}}^{sort}, 0, \dots)$ and $R_{opt} = \lfloor \left(\frac{n}{\log^2(n)}\right)^{\frac{1}{2s+d-1}} \rfloor$.

[2] Bandeira & Van Handel, Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *AoP*, 2016.

I.4 Finding eigenvalues clusters

Estimation of p: In practice

The previous theorem proves that we can recover $(p_k^*)_{k\geq 1}$ up to a permutation! But ...

Estimation of p: In practice

The previous theorem proves that we can recover $(p_k^*)_{k\geq 1}$ up to a permutation! But ...

- **... finding such a permutation is NP-hard.**
  We develop an algorithm (the SCCHEi) based on a Hierarchical Agglomerative Clustering of the eigenvalues $\lambda(\widehat{T}_n)$ to estimate the true partition.

The previous theorem proves that we can recover $(p_k^*)_{k\geq 1}$ up to a permutation! But ...

- **... finding such a permutation is NP-hard.**
  We develop an algorithm (the SCCHEi) based on a Hierarchical Agglomerative Clustering of the eigenvalues $\lambda(\widehat{T}_n)$ to estimate the true partition.

- **... the optimal resolution level R^{opt} is unknown.**
  We use a data-driven choice of model size R based on the *slope heuristic*.

I.5 Link prediction

The characters of the Gold Linear Harmonic Rush

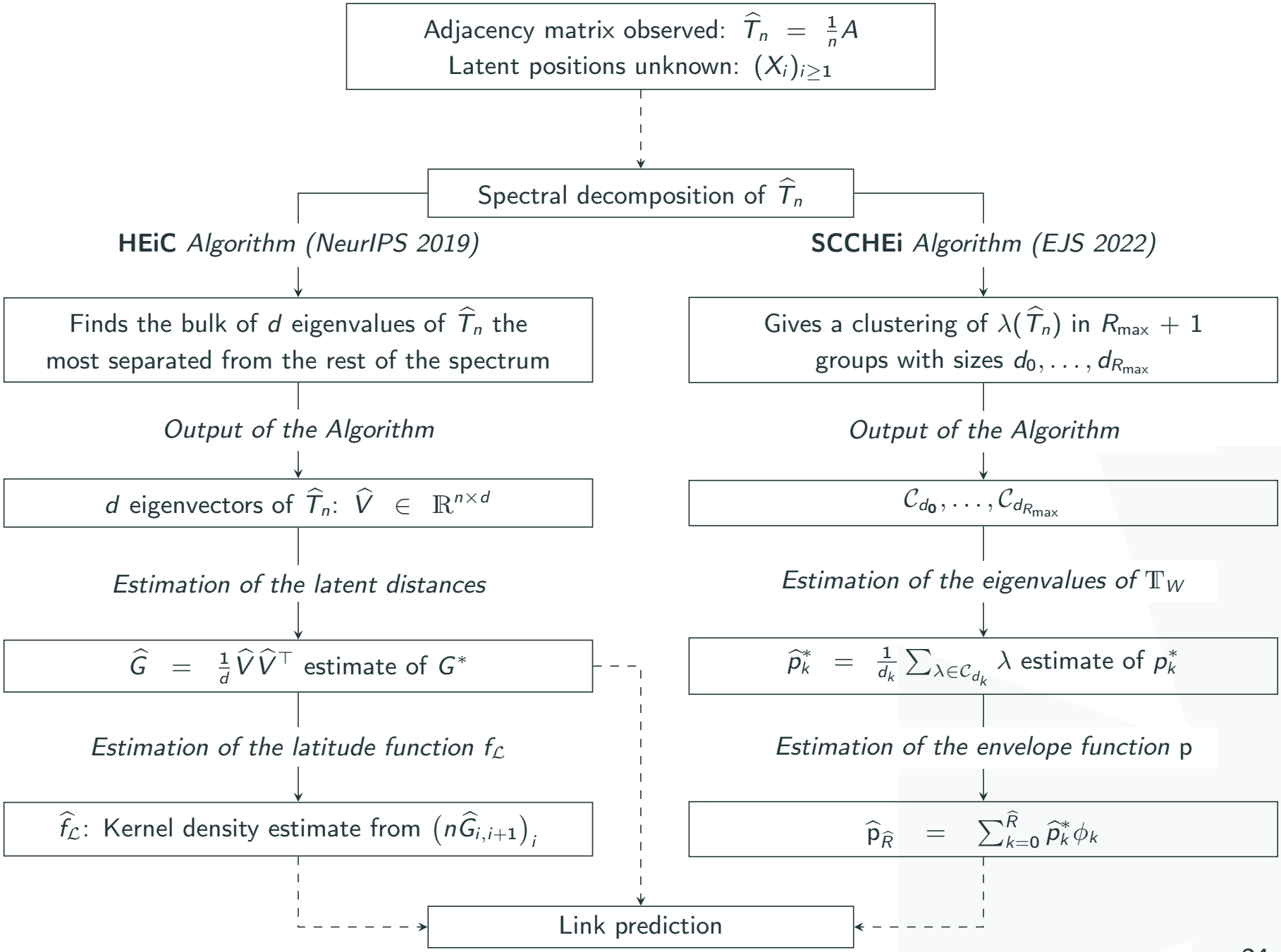
- Adjacency matrix observed: $\widehat{T}_n = \frac{1}{n} A$;
- Latent positions (unknown): $(X_i)_{i \geq 1}$;

The characters of the Gold Linear Harmonic Rush

- Adjacency matrix observed: $\widehat{T}_n = \frac{1}{n}A$;
- Latent positions (unknown): $(X_i)_{i\geq 1}$;
- Latent distances (unknown): $G^* := \frac{1}{n}\left(\langle X_i, X_j\rangle\right)_{1\leq i,j\leq n}$ (has rank d);
- Linear part of the kernel (unknown): $\mathrm{p}(\langle x, y\rangle) = c_0 p_0^* + c_1 p_1^* \langle x, y\rangle + \ldots$ (with c_k explicit constants) and p_1^* has multiplicity d in λ^*;
- The empirical eigenvectors corresponding to the linear Harmonic space (to be found): $\widehat{V} \in \mathbb{R}^{n\times d}$;

The characters of the Gold Linear Harmonic Rush

- Adjacency matrix observed: $\widehat{T}_n = \frac{1}{n}A$;
- Latent positions (unknown): $(X_i)_{i\geq 1}$;
- Latent distances (unknown): $G^* := \frac{1}{n}\left(\langle X_i, X_j\rangle\right)_{1\leq i,j\leq n}$ (has rank d);
- Linear part of the kernel (unknown): $\mathrm{p}(\langle x, y\rangle) = c_0 p_0^* + c_1 p_1^* \langle x, y\rangle + \dots$ (with c_k explicit constants) and p_1^* has multiplicity d in λ^*;
- The empirical eigenvectors corresponding to the linear Harmonic space (to be found): $\widehat{V} \in \mathbb{R}^{n\times d}$;
- The key remark (our hope): p_1^* is the ONLY eigenvalue with multiplicity d. Note that the multiplicities are pairwise disjoints as soon as $d \geq 3$ (otherwise we resort to classical Fourier analysis on the torus);


Adjacency matrix observed: $\widehat{T}_n = \frac{1}{n}A$
Latent positions unknown: $(X_i)_{i\geq 1}$
Spectral decomposition of $\widehat{T}_n$
HEiC *Algorithm (NeurIPS 2019)*
SCCHEi *Algorithm (EJS 2022)*
Finds the bulk of d eigenvalues of $\widehat{T}_n$ the most separated from the rest of the spectrum
Gives a clustering of $\lambda(\widehat{T}_n)$ in $R_{\max}+1$ groups with sizes $d_0, \ldots, d_{R_{\max}}$
Output of the Algorithm
Output of the Algorithm
d eigenvectors of $\widehat{T}_n$: $\widehat{V} \in \mathbb{R}^{n\times d}$
$\mathcal{C}_{d_0}, \ldots, \mathcal{C}_{d_{R_{\max}}}$
Estimation of the latent distances
Estimation of the eigenvalues of $\mathbb{T}_W$
$\widehat{G} = \frac{1}{d}\widehat{V}\widehat{V}^\top$ estimate of G^*
$\widehat{p}_k^* = \frac{1}{d_k}\sum_{\lambda\in\mathcal{C}_{d_k}} \lambda$ estimate of p_k^*
Estimation of the latitude function $f_{\mathcal{L}}$
Estimation of the envelope function p
$\widehat{f}_{\mathcal{L}}$: Kernel density estimate from $\left(n\widehat{G}_{i,i+1}\right)_i$
$\widehat{\mathrm{p}}_{\widehat{R}} = \sum_{k=0}^{\widehat{R}} \widehat{p}_k^*\phi_k$
Link prediction

From latent distances to link prediction

Denoting $\mathrm{proj}_{X_n^\perp}(\cdot)$ the orthogonal projection onto $\mathrm{Span}(X_n)^\perp$, it holds

$$\langle X_i, X_{n+1}\rangle = \langle X_i, X_n\rangle\langle X_n, X_{n+1}\rangle + \sqrt{1-\langle X_n, X_{n+1}\rangle^2}\sqrt{1-\langle X_i, X_n\rangle^2}\left\langle \frac{\mathrm{proj}_{X_n^\perp}(X_i)}{\|\mathrm{proj}_{X_n^\perp}(X_i)\|_2}, Y_{n+1}\right\rangle,$$

$\rightsquigarrow$ Latent distances $\mathsf{D}_{1:n} = (\langle X_i, X_j\rangle)_{1\le i,j\le n} \in [-1,1]^{n\times n}$ are enough for link prediction.

Indeed, $\forall i \in [n]$,

$$\eta_i(\mathsf{D}_{1:n}) = \mathbb{P}\left(A_{i,n+1} = 1 \mid \mathsf{D}_{1:n}\right)$$

$$= \int_{r,u\in(-1,1)} \mathsf{p}\left(\langle X_i, X_n\rangle r + \sqrt{1-r^2}\sqrt{1-\langle X_i, X_n\rangle^2}u\right) f_{\mathcal{L}}(r)dr(1-u^2)^{\frac{d-4}{2}}\frac{du}{b_d},$$

where $A_{i,n+1} \in \{0,1\}$ is one if and only if node $n+1$ is connected to node i.

Estimation of the latitude function and link prediction

Theorem

Assume that

$$\Delta^* := \min_{k \in \mathbb{N},\ k \neq 1} |p_1^* - p_k^*| > 0.$$

Then, we can compute $\widehat{G} \in \mathbb{R}^{n \times n}$ *s.t. with high probability*

$$\|G^* - \widehat{G}\|_F \lesssim \left(\frac{n}{\log^2(n)}\right)^{\frac{-s}{2s+d-1}},$$

with $G^* := \frac{1}{n}\left(\langle X_i, X_j \rangle\right)_{1 \leq i,j \leq n}$.

Heuristic from link prediction Denoting $\widehat{r} = n\widehat{G}$,

- estimate p with $\widehat{\mathrm{p}} \equiv \sum_{i=1}^{\widehat{R}_{opt}} \widehat{p}_i \phi_i$.
- let $\widehat{f_{\mathcal{L}}}$ be a kernel density estimator of $f_{\mathcal{L}}$ based on $(\widehat{r}_{i,i+1})_{1 \leq i \leq n-1}$.

Use the plug-in estimate

$$\hat{\eta}_i(\mathrm{D}_{1:n}) = \int_{r,u \in (-1,1)} \widehat{\mathrm{p}}\left(\widehat{r}_{i,n} r + \sqrt{1-r^2}\sqrt{1-\widehat{r}_{i,n}{}^2}\, u\right) \widehat{f_{\mathcal{L}}}(r) dr \frac{(1-u^2)^{\frac{d-4}{2}} du}{b_d},$$

Link prediction: experiments

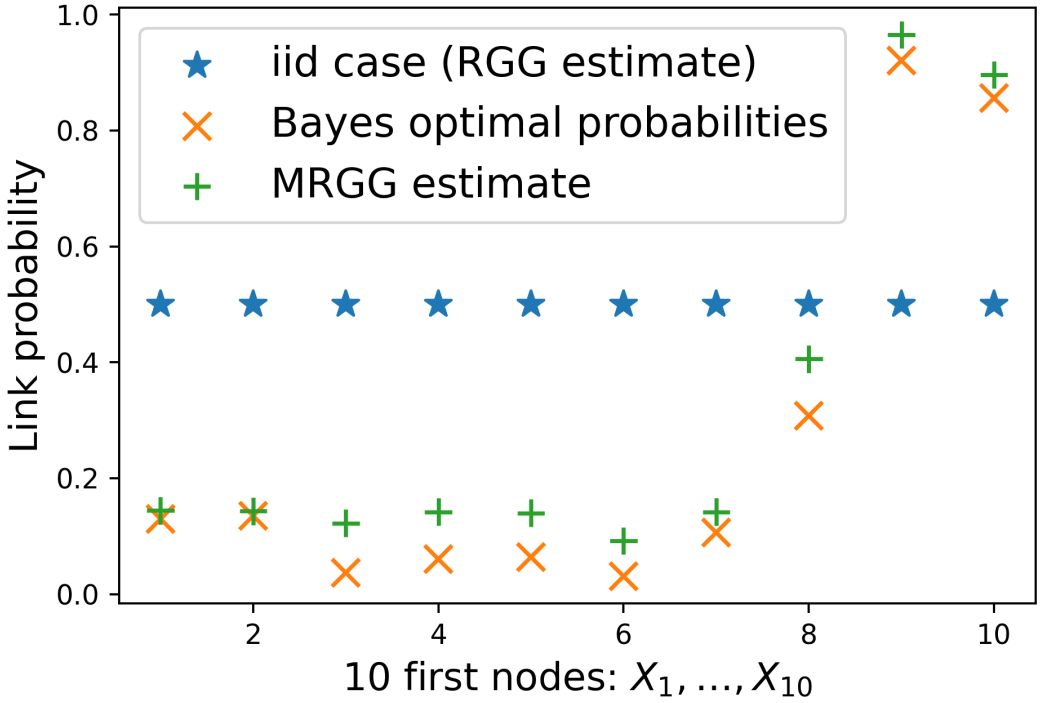
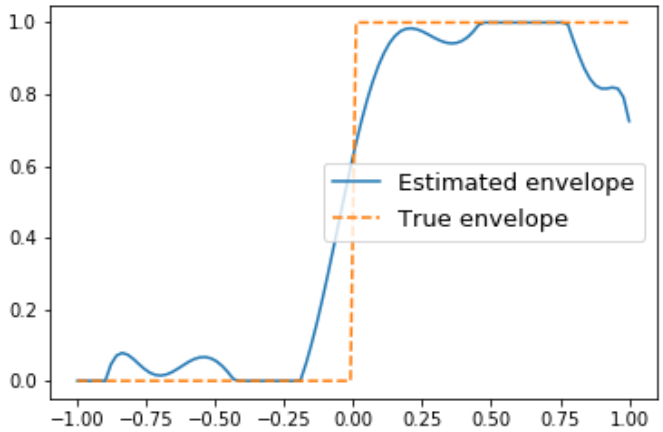
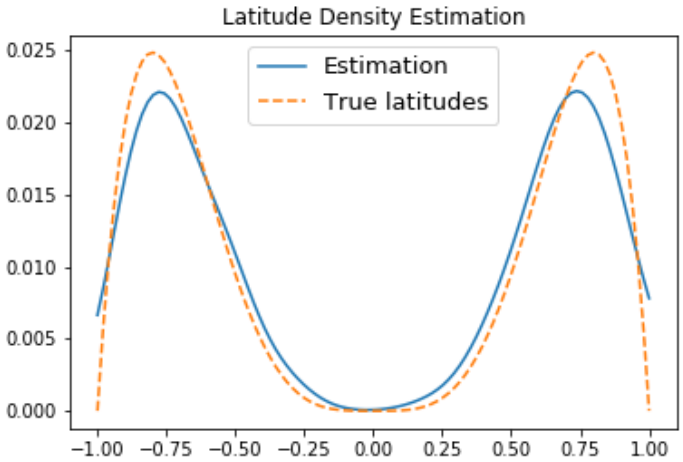


Figure 1: Link predictions between the future node X_{n+1} and the 10 first nodes $X_1, \ldots, X_{10}$. We work with $n = 2\,000$, $d = 3$, $\mathrm{p}(t) = \mathbb{1}_{t \geq \frac{1}{2}}$ and $f_{\mathcal{L}}(r) = \frac{1}{2} f_{(5,1)}(\frac{r+1}{2})$ where $f_{(5,1)}$ is the pdf of the Beta distribution with parameter $(5, 1)$.

Estimation results



(a) Envelope function



(b) Latitude function

Figure 2: Non-parametric estimation of envelope and latitude functions.

Transition towards the concentration toolbox

Transition: Proof and use of U-statistic concentration

Theorem

If
- p has regularity $s > 0$,
- $\|f_{\mathcal{L}}\|_\infty < \infty$ and $\inf_{r \in [-1,1]} f_{\mathcal{L}}(r) > 0$,

then for n large enough it holds

$$\mathbb{E}[\delta_2^2(\lambda(T_n), \lambda^*)] \lesssim \left[\frac{n}{\log^2 n}\right]^{-\frac{2s}{2s+(d-1)}}.$$

Transition: Proof and use of U-statistic concentration

We recall that $T_n = \frac{1}{n}\left((1-\delta_{i,j})p\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n}$ and we define

$$T_{R,n} = \frac{1}{n}\left((1-\delta_{i,j})p_R\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n},$$

where

$$p_R(t) := \sum_{k=0}^{R} p_k^* \phi_k(t), \quad R \ge 0.$$

Transition: Proof and use of U-statistic concentration

We recall that $T_n = \frac{1}{n}\left((1-\delta_{i,j})p\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n}$ and we define

$$T_{R,n} = \frac{1}{n}\left((1-\delta_{i,j})p_R\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n},$$

where

$$p_R(t) := \sum_{k=0}^{R} p_k^* \phi_k(t), \quad R \ge 0.$$

Then

$$\delta_2(\lambda(T_n), \lambda^*)^2 \lesssim \delta_2(\lambda(T_n), \lambda(T_{R,n}))^2 + \delta_2(\lambda(T_{R,n}), \lambda^*)^2.$$

Transition: Proof and use of U-statistic concentration

We recall that $T_n = \frac{1}{n}\left((1-\delta_{i,j})p\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n}$ and we define

$$T_{R,n} = \frac{1}{n}\left((1-\delta_{i,j})p_R\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n},$$

where

$$p_R(t) := \sum_{k=0}^{R} p_k^* \phi_k(t), \quad R \ge 0.$$

Then

$$\delta_2(\lambda(T_n), \lambda^*)^2 \lesssim \delta_2(\lambda(T_n), \lambda(T_{R,n}))^2 + \delta_2(\lambda(T_{R,n}), \lambda^*)^2.$$

The first term involves a U-statistic.

$$\delta_2\left(\lambda(T_{R,n}), \lambda(T_n)\right)^2 \underbrace{\le}_{\substack{\text{Hoffman-Wielandt}\\ \text{inequality}}} \|T_{R,n} - T_n\|_F^2 = \frac{1}{n^2}\sum_{i\neq j}(p-p_R)^2(\langle X_i, X_j\rangle).$$

Transition: Proof and use of U-statistic concentration

We recall that $T_n = \frac{1}{n}\left((1-\delta_{i,j})p\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n}$ and we define

$$T_{R,n} = \frac{1}{n}\left((1-\delta_{i,j})p_R\left(\langle X_i, X_j\rangle\right)\right)_{1\le i,j\le n},$$

where

$$p_R(t) := \sum_{k=0}^{R} p_k^* \phi_k(t), \quad R \ge 0.$$

Then

$$\delta_2(\lambda(T_n), \lambda^*)^2 \lesssim \delta_2(\lambda(T_n), \lambda(T_{R,n}))^2 + \delta_2(\lambda(T_{R,n}), \lambda^*)^2.$$

The first term involves a U-statistic.

$$\delta_2\left(\lambda(T_{R,n}), \lambda(T_n)\right)^2 \underbrace{\le}_{\substack{\text{Hoffman-Wielandt}\\ \text{inequality}}} \|T_{R,n} - T_n\|_F^2 = \frac{1}{n^2}\sum_{i\neq j}(p-p_R)^2(\langle X_i, X_j\rangle).$$

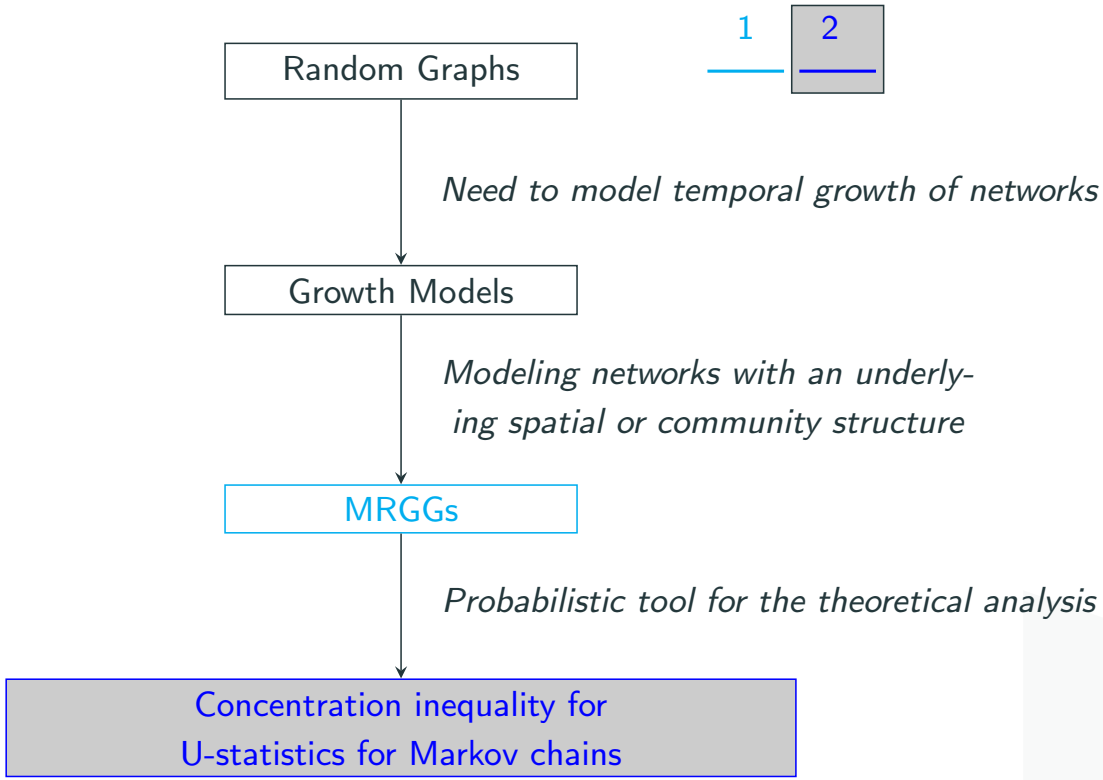
Definition (U-statistic of order 2)

A U-statistic of order 2 is a sum of the form

$$\sum_{1\le i<j\le n} h_{i,j}(X_i, X_j),$$

where $X_1, \dots, X_n$ are r.v. taking values in a measurable space (E, Σ) and where $h_{i,j} : E^2 \to \mathbb{R}$.

II. Concentration Inequality for U-Statistics of order 2 for Uniformly Ergodic Markov Chains



Review on concentration in the independent framework

- Arcones & Giné 1993

$$\left\|h_{i,j}\right\|_{\infty} < \infty \quad \text{and}$$
$$\forall i,j \in [n]\,,\ \forall x,\ \mathbb{E}_X\left[h_{i,j}(x,X)\right] = \mathbb{E}_X\left[h_{i,j}(X,x)\right] = 0\,.$$

Review on concentration in the independent framework

- Arcones & Giné 1993

$$\left\|h_{i,j}\right\|_{\infty} < \infty \quad \text{and}$$

$$\forall i,j \in [n]\,,\ \forall x,\ \mathbb{E}_X\left[h_{i,j}(x,X)\right] = \mathbb{E}_X\left[h_{i,j}(X,x)\right] = 0\,.$$

- Houdré & Reynaud-Bouret 2002

  Improved the constants in the exponential inequality of Arcones & Giné for U-statistics of order two.

Review on concentration in the independent framework

- Arcones & Giné 1993

$$\left\|h_{i,j}\right\|_{\infty} < \infty \quad \text{and}$$

$$\forall i,j \in [n]\,,\ \forall x,\ \mathbb{E}_X\left[h_{i,j}(x,X)\right] = \mathbb{E}_X\left[h_{i,j}(X,x)\right] = 0\,.$$

- Houdré & Reynaud-Bouret 2002

  Improved the constants in the exponential inequality of Arcones & Giné for U-statistics of order two.

- Giné, Latala & Zinn 2000

  Proved that exponential inequality can still be obtained for U-statistics with sufficiently light tails of arbitrary order.

Review on concentration in the independent framework

- Arcones & Giné 1993

$$\left\|h_{i,j}\right\|_{\infty} < \infty \quad \text{and}$$
$$\forall i,j \in [n]\,,\ \forall x,\ \mathbb{E}_X\left[h_{i,j}(x,X)\right] = \mathbb{E}_X\left[h_{i,j}(X,x)\right] = 0\,.$$

- Houdré & Reynaud-Bouret 2002

  Improved the constants in the exponential inequality of Arcones & Giné for U-statistics of order two.

- Giné, Latala & Zinn 2000

  Proved that exponential inequality can still be obtained for U-statistics with sufficiently light tails of arbitrary order.

- Joly & Lugosi in 2016

  Working with kernels that have finite p-th moment for some $p \in (1,2]$, they construct an estimator of the mean of the U-process using the median-of-means technique.

Assumptions

Assumption 1 (Uniform ergodicity)

The Markov chain $(X_i)_{i\geq 1}$ valued in the measurable space (E, Σ) is uniformly ergodic with transition kernel $P : E^2 \to \mathbb{R}_+$ and with invariant distribution π.

Assumptions

Assumption 1 (Uniform ergodicity)

The Markov chain $(X_i)_{i\geq 1}$ valued in the measurable space (E, Σ) is uniformly ergodic with transition kernel $P : E^2 \to \mathbb{R}_+$ and with invariant distribution π.

$\Leftrightarrow \exists \rho \in (0,1)$ and $\exists L > 0$ such that

$$\|P^n(x,\cdot) - \pi\|_{TV} \leq L\rho^n, \qquad \forall n \geq 0,\ \pi\text{−a.e } x \in E,$$

Assumptions

Assumption 1 (Uniform ergodicity)

The Markov chain $(X_i)_{i\geq 1}$ valued in the measurable space (E, Σ) is uniformly ergodic with transition kernel $P : E^2 \to \mathbb{R}_+$ and with invariant distribution π.

Assumption 2 ("Upper-bounded" transition kernel)

$$\exists \delta_M > 0,\ \exists\ \nu \text{ proba. meas. on } E \text{ s.t. } \forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$

Assumptions

Assumption 1 (Uniform ergodicity)

The Markov chain $(X_i)_{i\geq 1}$ valued in the measurable space (E, Σ) is uniformly ergodic with transition kernel $P : E^2 \to \mathbb{R}_+$ and with invariant distribution π.

Assumption 2 ("Upper-bounded" transition kernel)

$$\exists \delta_M > 0,\ \exists\, \nu \text{ proba. meas. on } E \text{ s.t. } \forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$

This holds for

- Aperiodic and irreducible Markov chains on finite state space.
- AR(1) process with mild conditions.
- ARCH process with mild conditions.

Assumptions

Assumption 1 (Uniform ergodicity)

The Markov chain $(X_i)_{i\geq 1}$ valued in the measurable space (E, Σ) is uniformly ergodic with transition kernel $P : E^2 \to \mathbb{R}_+$ and with invariant distribution π.

Assumption 2 ("Upper-bounded" transition kernel)

$$\exists \delta_M > 0,\ \exists\ \nu \text{ proba. meas. on } E \text{ s.t. } \forall x \in E,\ \forall A \in \Sigma, \quad P(x, A) \leq \delta_M \nu(A).$$

Assumption 3 (Bounded and π-canonical kernels)

$\forall i, j,\ h_{i,j} : (E^2, \Sigma \otimes \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *is measurable, bounded and π-canonical, i.e.*

$$\forall x \in E, \quad \mathbb{E}_\pi h_{i,j}(X, x) = \mathbb{E}_\pi h_{i,j}(x, X) = 0.$$

Main Result

Theorem (Bernoulli: 2022)

We consider a stationary Markov chain satisfying Assumptions 1, 2 and 3. Then there exist $\beta, \kappa > 0$ s.t. $\forall u > 0$, it holds with probability at least $1 - \beta e^{-u} \log n$,

$$U_{\mathrm{stat}}(n) := \sum_{1 \le i < j \le n} \left(h_{i,j}(X_i, X_j) - \mathbb{E}\left[h_{i,j}(X_i, X_j) \right] \right)$$

$$\le \kappa \log(n) \Big(\left[C_n + A \log(n) \sqrt{n} \right] \sqrt{u} + \left[A + B_n \sqrt{n} \right] u + \left[2A\sqrt{n} \right] u^{3/2} + A \left[u^2 + \log n \right] \Big).$$

Theorem (Bernoulli: 2022)

We consider a stationary Markov chain satisfying Assumptions 1, 2 and 3. Then there exist $\beta, \kappa > 0$ s.t. $\forall u > 0$, it holds with probability at least $1 - \beta e^{-u} \log n$,

$$U_{\mathrm{stat}}(n) := \sum_{1 \le i < j \le n} \left(h_{i,j}(X_i, X_j) - \mathbb{E}\left[h_{i,j}(X_i, X_j) \right] \right)$$

$$\le \kappa \log(n) \Big(\left[C_n + A \log(n) \sqrt{n} \right] \sqrt{u} + \left[A + B_n \sqrt{n} \right] u + \left[2A\sqrt{n} \right] u^{3/2} + A \left[u^2 + \log n \right] \Big).$$

with for some t_n scaling with $\log n$,

$$A := 2 \max_{i,j} \| h_{i,j} \|_\infty, \quad C_n^2 := \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbb{E}\left[\mathbb{E}_{X' \sim \nu}[h_{i,j}^2(X_i, X')] \right]$$

$$B_n^2 := \max \Bigg[\max_{0 \le k \le t_n} \max_i \sup_x \sum_{j=i+1}^{n} \mathbb{E}_{X' \sim \nu} \left(\mathbb{E}_{X \sim P^k(X', \cdot)} h_{i,j}(x, X) \right)^2,$$

$$\max_{0 \le k \le t_n} \max_j \sup_y \sum_{i=1}^{j-1} \mathbb{E}_{\tilde{X} \sim \pi} \left(\mathbb{E}_{X \sim P^k(y, \cdot)} h_{i,j}(\tilde{X}, X) \right)^2 \Bigg]$$

Theorem (Bernoulli: 2022)

We consider a stationary Markov chain satisfying Assumptions 1, 2 and 3. Then there exist $\beta, \kappa > 0$ s.t. $\forall u > 0$, it holds with probability at least $1 - \beta e^{-u} \log n$,

$$U_{\mathrm{stat}}(n) := \sum_{1 \le i < j \le n} \left(h_{i,j}(X_i, X_j) - \mathbb{E}\left[h_{i,j}(X_i, X_j) \right] \right)$$

$$\le \kappa \log(n) \Big(\left[C_n + A \log(n) \sqrt{n} \right] \sqrt{u} + \left[A + B_n \sqrt{n} \right] u + \left[2A\sqrt{n} \right] u^{3/2} + A \left[u^2 + \log n \right] \Big).$$

with for some t_n scaling with $\log n$,

$$A := 2 \max_{i,j} \| h_{i,j} \|_\infty, \quad C_n^2 := \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbb{E}\left[\mathbb{E}_{X' \sim \nu}[h_{i,j}^2(X_i, X')] \right] \le A^2 n^2,$$

$$B_n^2 := \max \Bigg[\max_{0 \le k \le t_n} \max_i \sup_x \sum_{j=i+1}^{n} \mathbb{E}_{X' \sim \nu} \left(\mathbb{E}_{X \sim P^k(X', \cdot)} \, h_{i,j}(x, X) \right)^2,$$

$$\max_{0 \le k \le t_n} \max_j \sup_y \sum_{i=1}^{j-1} \mathbb{E}_{\tilde{X} \sim \pi} \left(\mathbb{E}_{X \sim P^k(y, \cdot)} h_{i,j}(\tilde{X}, X) \right)^2 \Bigg] \le A^2 n.$$

Discussion

- "**Constants** *look pretty ugly*" ✗

Discussion

- *"***Constants** *look pretty ugly"* ✗

  In the specific case where $\nu = \pi$ (which includes the independent setting), we get that

$$C_n^2 = \sum_{i<j} \mathbb{E}\left\{ \mathrm{Var}_{\tilde{X}\sim\pi}\left[h_{i,j}(X_i, \tilde{X}) | X_i \right] \right\},$$

  and using Jensen inequality that

$$B_n^2 \leq \max\left[\sup_{x,i} \sum_{j=i+1}^{n} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(x, \tilde{X})] \,,\, \sup_{y,j} \sum_{i=1}^{j-1} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(\tilde{X}, y)] \right].$$

Discussion

- *"***Constants** *look pretty ugly"* ✗

  In the specific case where $\nu = \pi$ (which includes the independent setting), we get that

$$C_n^2 = \sum_{i<j} \mathbb{E}\left\{ \mathrm{Var}_{\tilde{X}\sim\pi}\left[h_{i,j}(X_i, \tilde{X}) | X_i \right] \right\},$$

  and using Jensen inequality that

$$B_n^2 \leq \max\left[\sup_{x,i} \sum_{j=i+1}^{n} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(x, \tilde{X})] \,,\; \sup_{y,j} \sum_{i=1}^{j-1} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(\tilde{X}, y)] \right].$$

- *"Not too bad but* **stationarity** *is restrictive"*

Discussion

- *"***Constants** *look pretty ugly"* ✗

  In the specific case where $\nu = \pi$ (which includes the independent setting), we get that

$$C_n^2 = \sum_{i<j} \mathbb{E}\left\{ \mathrm{Var}_{\tilde{X}\sim\pi}\left[h_{i,j}(X_i, \tilde{X}) | X_i \right] \right\},$$

  and using Jensen inequality that

$$B_n^2 \leq \max\left[\sup_{x,i} \sum_{j=i+1}^{n} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(x, \tilde{X})] \, , \, \sup_{y,j} \sum_{i=1}^{j-1} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(\tilde{X}, y)] \right].$$

- *"Not too bad but* **stationarity** *is restrictive"* ✓

Discussion

- *"***Constants** *look pretty ugly"* ✗

  In the specific case where $\nu = \pi$ (which includes the independent setting), we get that

$$C_n^2 = \sum_{i<j} \mathbb{E}\left\{ \mathrm{Var}_{\tilde{X}\sim\pi}\left[h_{i,j}(X_i, \tilde{X}) | X_i \right] \right\},$$

  and using Jensen inequality that

$$B_n^2 \leq \max\left[\sup_{x,i} \sum_{j=i+1}^{n} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(x, \tilde{X})] \, , \, \sup_{y,j} \sum_{i=1}^{j-1} \mathrm{Var}_{\tilde{X}\sim\pi}[h_{i,j}(\tilde{X}, y)] \right].$$

- *"Not too bad but* **stationarity** *is restrictive"* ✓

Theorem (Bernoulli: 2022)

Suppose Assumptions 1, 2 and 3. Assume further that

- either $h_{i,j} \equiv h_{1,j}$, $\forall i, j$.
- or a mild condition on the initial distribution of the chain.

Then there exist $\beta, \kappa > 0$ s.t. $\forall u > 0$, it holds w.p. $\geq 1 - \beta e^{-u} \log n$,

$$\frac{2}{n(n-1)} U_{\mathrm{stat}}(n) \leq \kappa \max_{i,j} \|h_{i,j}\|_\infty \log n \left\{ \frac{u}{n} + \left[\frac{u}{n}\right]^2 \right\}.$$

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\,U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} (\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)])}_{=\,U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} (\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)])$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

Lemma

Let $(U_m)_{m\in\mathbb{N}}$ be a martingale such that $U_0 = U_1 = 0$.

For each $m \geq 1$ and $\ell \geq 2$, let

$$A^\ell_m = \sum_{i=1}^{m} \mathbb{E}_{i-1}[(U_i - U_{i-1})^\ell] \quad \text{and } A^\ell_1 = 0.$$

If $A^\ell_m \leq w^\ell_m$ for $w^\ell_m \geq 0$ then

$$\mathbb{E}[e^{\alpha U_m}] \leq e^{\sum_{\ell\geq 2} \alpha^\ell w^\ell_m/\ell!}.$$

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

Lemma

Let $(U_m)_{m\in\mathbb{N}}$ be a martingale such that $U_0 = U_1 = 0$.

For each $m \geq 1$ and $\ell \geq 2$, let

$$A_m^\ell = \sum_{i=1}^{m} \mathbb{E}_{i-1}[(U_i - U_{i-1})^\ell] \quad \text{and } A_1^\ell = 0.$$

If $A_m^\ell \leq w_m^\ell$ for $w_m^\ell \geq 0$ then

$$\mathbb{E}[e^{\alpha U_m}] \leq e^{\sum_{\ell\geq 2} \alpha^\ell w_m^\ell/\ell!}.$$

$U^{(k)}_{\mathrm{stat}}(n)$ is a martingale !

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$p_{i,j}(x, z) := h_{i,j}(x, z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')]$$

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\, U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$\leq \delta_M \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell, \quad \text{with} \quad p_{i,j}(x,z) := h_{i,j}(x,z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')]$$

$$\left(X'_j\right)_j \overset{i.i.d.}{\sim} \nu$$

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\, U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$\leq \delta_M \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell, \quad \text{with} \quad p_{i,j}(x,z) := h_{i,j}(x,z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')]$$

$$\left(X'_j\right)_j \overset{i.i.d.}{\sim} \nu$$

duality lemma

$$\lesssim \left[\sup_{(\xi_j)_j \in \mathcal{L}_\ell} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{j-1}\left[p_{i,j}(X_i, X'_j)\xi_j(X'_j)\right]\right]^\ell$$

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\, U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$\leq \delta_M \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell, \quad \text{with} \quad p_{i,j}(x,z) := h_{i,j}(x,z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')]$$

$$\left(X'_j\right)_j \overset{i.i.d.}{\sim} \nu$$

duality lemma

$$\lesssim \left[\sup_{(\xi_j)_j \in \mathcal{L}_\ell} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{j-1}\left[p_{i,j}(X_i, X'_j)\xi_j(X'_j)\right]\right]^\ell \quad \rightarrow$$

Use of a Talagrand inequality for Markov chains

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\, U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$\leq \delta_M \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell, \quad \text{with} \quad p_{i,j}(x,z) := h_{i,j}(x,z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')], \quad \left(X'_j\right)_j \overset{i.i.d.}{\sim} \nu$$

duality lemma

$$\lesssim \left[\sup_{(\xi_j)_j \in \mathcal{L}_\ell} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{j-1}\left[p_{i,j}(X_i, X'_j)\xi_j(X'_j)\right]\right]^\ell \quad \rightarrow$$

converse duality lemma

Use of a Talagrand inequality for Markov chains

w.p. $\geq 1 - e^{-u}$

$$\lesssim \sum_{j=2}^{n} \mathbb{E}\Big|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\Big|^\ell + n\left(A\ell u\right)^\ell + B_n^2\left(A^2 n\ell u\right)^{\ell/2}.$$

Main proof arguments

Telescopic decomposition of $U_{\mathrm{stat}}(n)$.

$$U_{\mathrm{stat}}(n) = \sum_{k=1}^{t_n} \underbrace{\sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)]\right)}_{=\, U^{(k)}_{\mathrm{stat}}(n)} + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}[h_{i,j}(X_i, X_j)]\right)$$

Bounding $U^{(k)}_{\mathrm{stat}}(n)$.

$$A_n^\ell = \sum_{j=2}^{n} \mathbb{E}_{j-1}\left[\left(U^{(1)}_{\mathrm{stat}}(j) - U^{(1)}_{\mathrm{stat}}(j-1)\right)^\ell\right] \lesssim \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j)\right|^\ell$$

$$\leq \delta_M \sum_{j=2}^{n} \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell, \quad \text{with} \quad p_{i,j}(x,z) := h_{i,j}(x,z) - \mathbb{E}_{X' \sim \nu}[h_{i,j}(X_i, X')], \quad \left(X'_j\right)_j \overset{i.i.d.}{\sim} \nu$$

duality lemma

$$\lesssim \left[\sup_{(\xi_j)_j \in \mathcal{L}_\ell} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{j-1}\left[p_{i,j}(X_i, X'_j)\xi_j(X'_j)\right]\right]^\ell \rightarrow$$

converse duality lemma

Use of a Talagrand inequality for Markov chains

w.p. $\geq 1 - e^{-u}$

$$\lesssim \sum_{j=2}^{n} \boxed{\mathbb{E}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^\ell} + n\left(A\ell u\right)^\ell + B_n^2\left(A^2 n \ell u\right)^{\ell/2}.$$

Perspectives

RGGs

- Extension of the MRGG to more general Markovian sampling schemes.
- Study of the robustness of the estimation methods in MRGGs.
- A large set of open questions related to geometry detection in high dimensional RGGs.

Concentration for U-statistics in a dependent framework and applications

- Extension to U-statistics of higher order ($m > 2$).
- Replacing the Markovian assumption with a mixing condition.

Thanks for your attention!