LATENT DISTANCE ESTIMATION FOR RANDOM GEOMETRIC GRAPHS

Ernesto Araya Valdivia† and Yohann De Castro‡

†Laboratoire de Mathématiques d'Orsay, Université Paris-Sud

‡École Centrale de Lyon

Abstract

We observe a simple graph which has a hidden representation in a Euclidean space, that is the presence of edges depends on the latent distance between nodes. We give an spectral algorithm which recovers the latent distances. As byproduct we can recover the dimension of the underlying space as well.

Random geometric graph model

- **Latent points**: $\{X_i\}_{i=1}^n$ are i.i.d uniformly distributed in the unit sphere $\mathbb{S}^{d-1}$. The population Gram matrix is defined by $(\mathcal{G})_{ij} := \langle X_i, X_j \rangle$.
- **Connection function**: $f : [-1, 1] \to [0, 1]$ such that
$$\Theta_{ij}(n) = \rho_n f(\langle X_i, X_j \rangle)$$
where ρ_n is the sparsity parameter. We assume $\rho_n = \Omega(\frac{\log n}{n})$.
- **Random geometric graph**: The adjacency matrix $A(n)$ is given by
$$\mathbb{P}(A_{ij}(n) = 1 | \{X_1, \cdots, X_n\}) = \Theta_{ij}$$

Problem statement

- We observe a simple graph $G = (V, E)$, which we assume was generated by the angular RGG model. The connection function is unknown.
- **Objective**: Recover the Gram matrix $\mathcal{G}$ only using the adjacency matrix of G.
- **Error criteria**: We define the error $\mathbf{E}(A, B) = \frac{1}{n^2}\|A - B\|_F^2$ for A, B matrices of the same size.

Strategy

- Use the theory of graph limits [2]: $\frac{1}{\rho_n}A(n)$ converges towards the compact integral operator $T_f : L^2(\mathbb{S}^{d-1}, \mu) \to L^2(\mathbb{S}^{d-1}, \mu)$.
$$T_f g(x) = \int_{\mathbb{S}^{d-1}} f(\langle x, y \rangle) g(y) d\mu(y)$$
with μ = the uniform measure on $\mathbb{S}^{d-1}$. In particular, the spectrum of $\frac{1}{\rho_n}\Theta(n)$ converges to the spectrum of T_f in the ℓ_2 sense [4] and $\frac{1}{n}\|A(n) - \Theta(n)\|_{op} \to 0$ by [1].
- Use harmonic analysis for T_f on $\mathbb{S}^{d-1}$:
  - Non zero eigenvalues have fixed multiplicities that depend only on d. In particular, $\exists \lambda_1^* \in \lambda(T_f)$, such that $\dim E_1 = d$, where E_1 is the associated eigenspace.
  - **Reconstruction formula**: we have $\forall x, y \in \mathbb{S}^{d-1}$, $P_{E_j}(x, y) \propto G_j(\langle x, y \rangle)$, where $G_j(\langle x, y \rangle)$ is a Gegenbauer polynomial of degree j. Where P_{E_i} is the $L^2(\mu)$ orthogonal projector onto E_i.

Main assumptions

- **Eigengap condition**: $\exists \Lambda_1^* \subset \lambda(T_f)$ with $|\Lambda_1^*| = d$ and such that, for $\Delta^* > 0$
$$\min_{\lambda \in \lambda(T_f), \lambda \notin \Lambda_1^*} |\lambda - \lambda^*| > \Delta^*$$
- **Regularity**: f has regularity s if the eigenvalues of T_f satisfy $\sum_k |\lambda_k|^2 k^{2s} < \infty$.

Main result

Theorem 1 (Informal statement) *Let G be a graph generated with the RRG model with connection function f (unknown), which satisfies the eigengap condition. Then*

- *$\exists \Lambda_1 = (v_1, \cdots, v_d)$ subset of the eigenvectors of $A(n)$, such that $\mathbf{E}(\mathcal{G}, \hat{\mathcal{G}}) \xrightarrow[n \to \infty]{} 0$, where $\hat{\mathcal{G}} := \frac{1}{c}\sum_{i=1}^d v_i v_i^T$ and c is a known constant.*
- *Furthermore, if has regularity s, we have $\mathbf{E}(\mathcal{G}, \hat{\mathcal{G}}) = O(\Delta^{*-1} n^{-s/2s+d-1})$.*

Gram matrix recovery

Harmonic Eigen-Cluster (HEiC) algorithm:

- **Input**: (A, d) adjacency matrix and dimension of the sphere.
- $\lambda^{\text{sort}}(A) \leftarrow$ Compute $\lambda(A) = \{\lambda_0^{\text{sort}}, \cdots, \lambda_n^{\text{sort}}\}$ and sort decreasingly.
- $\text{Gap}_1 \leftarrow \max\left\{\max_{1 \le i \le n-d-1} \min\{\text{L}(i), \text{R}(i+d)\}, \text{L}(n-d+1)\right\}$, with
$$\text{L}(i) := |\hat{\lambda}_i^{\text{sort}} - \hat{\lambda}_{i-1}^{\text{sort}}| \quad \text{and} \quad \text{R}(i) := \text{L}(i+1)$$
- $$i^* \leftarrow \begin{cases} n - d \text{ if } \max_{1 \le i \le n-d-1} \min\{\text{L}(i), \text{R}(i+d)\} < \text{L}(n-d+1) \\ \arg\max_{1 \le i \le n-d-1} \min\{\text{L}(i), \text{R}(i+d)\}, \text{ otherwise} \end{cases}$$
- **Return**: (Gap_1, i^*)

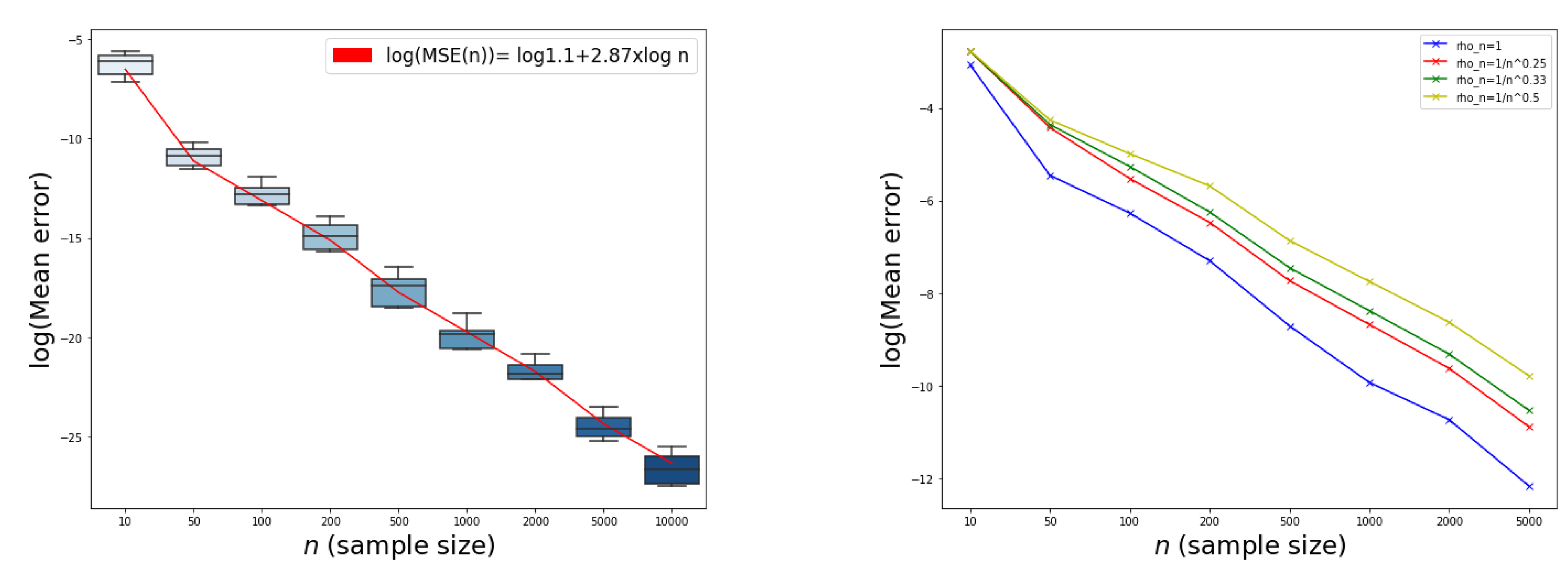

Figure 1: Simulations with the connection function $f(t) = 1_{t \le 0}$

Dimension recovery

Algorithm to recover the dimension if it belongs to a candidate set $\mathcal{D}$:

- **Input**: $(A, \mathcal{D})$ the adjacency matrix and $\mathcal{D} = \{d_1, \cdots, d_{\max}\}$
- $\text{Score}(d) \leftarrow \text{Gap}_1(d)$ obtained running $\text{HEiC}(A, d)$ as subroutine.
- **Return**: $d^* \leftarrow \arg\max_{d \in \mathcal{D}} \text{score}(d)$.

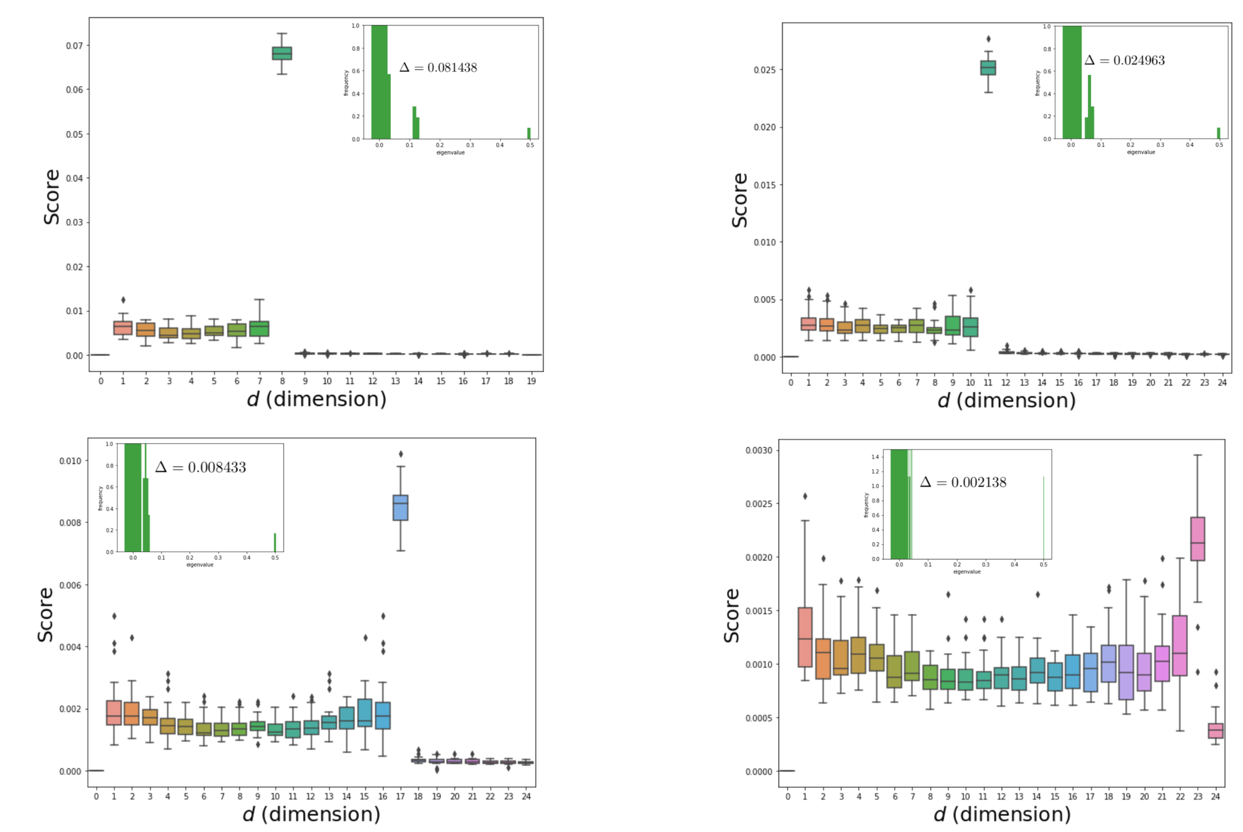

Figure 2: Simulations with the connection function $f(t) = \frac{1}{1+e^{-3t}}$ and $n = 1000$

Conclusions and extensions

- Dense random geometric graphs have an structured spectrum, which allow us to recover the latent distances with HEiC algorithm if the eigengap condition holds.
- The algorithm can be extended, almost without modification to compact lie groups and symmetric spaces [3].
- Consider the closed unit ball in $\mathbb{R}^d$ as underlying space and probability matrix $\Theta_{ij} = f(\langle x, y \rangle)$. The harmonic analysis on the ball allows for an extension of the previous methods, with slight modifications.

References

[1] A Bandeira and R. Van Handel. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries". In: *Annals of Probability* 44.4 (2016), pp. 2479–2506.

[2] C Borgs et al. "Convergent sequences of dense graphs II. Multiway cuts and statistical physics". In: *Annals of Mathematics* 176.1 (2012), pp. 151–219.

[3] Y. De Castro, C. Lacour, and T.M Pham Ngoc. "Adaptive estimation of nonparametric geometric graphs". In: (2017). URL: https://arxiv.org/pdf/1708.02107.pdf.

[4] V Koltchinskii and E Giné. "Random matrix approximation of spectra of integral operators". In: *Bernoulli* (2000), pp. 113–167.