Estimation non-paramétrique pour des graphes géométriques

Thanh Mai Pham Ngoc
en collaboration avec Yohann de Castro et Claire Lacour
Université Paris Sud - Ecole Ponts ParisTech - UPEM

4e Journée Statistique et Machine Learning de Paris Saclay
IHES 30/01/2019

Statistical model

We observe a random undirected graph G with n nodes and its adjacency matrix A. We assume that

$$A_{ij} = 1 \quad \text{with probability} \quad \theta_{ij} = W(X_i, X_j) = \mathbf{p}(\langle X_i, X_j \rangle)$$

with X_i latent variables (unobservable) lying in $\mathbb{S}^{d-1}$, drawn w.r.t σ uniform measure on $\mathbb{S}^{d-1}$.

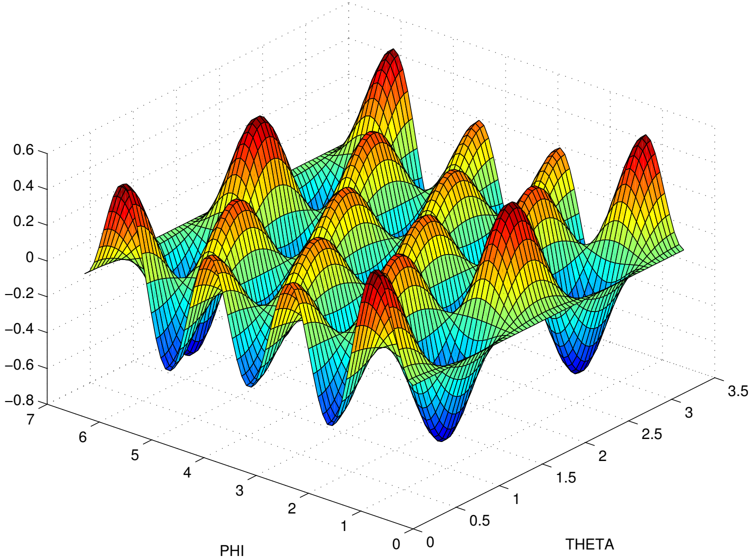
- Interest in latent metric space with distance invariant by isometry.
- The dimension d is supposed to be known.
- Aim : estimate the envelope function $\mathbf{p} : [-1, 1] \to [0, 1]$.

Harmonic Analysis on $\mathbb{S}^{d-1}$

- Real spherical harmonics $Y_{\ell j}$ are o.n.b of $\mathbb{L}^2(\mathbb{S}^{d-1})$
- $$\mathbb{L}^2(\mathbb{S}^{d-1}) = \bigoplus_{\ell=0}^{\infty} \mathbb{H}_\ell$$

  $\mathbb{H}_\ell$ space spanned by $\{ Y_{\ell j}, j = 1, \ldots, d_\ell \}$ and $d_\ell = \dim(\mathbb{H}_\ell)$
- For $\mathbb{S}^2$, $d_\ell = 2\ell + 1$.
- We have in $\mathbb{L}^2$- sense :

$$f(x) = \sum_{\ell \geq 0} \sum_{j=1}^{d_l} f^\star_{lj} Y_{lj}(x)$$


0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
7
6
5
4
3
2
1
0
PHI
0
0.5
1
1.5
2
2.5
3
3.5
THETA

Harmonic Analysis on $\mathbb{S}^{d-1}$

- The orthogonal projection on space $\mathbb{H}_\ell$ is

$$P_{\mathbb{H}_\ell}(f)(x) = \int_{\mathbb{S}^{d-1}} Z(x,y)f(y)dy$$

with

$$Z(x,y) := \sum_{j=1}^{d_\ell} Y_{\ell j}(x)Y_{\ell j}(y) = Z(\langle x,y\rangle)$$

and

$$Z(\langle x,y\rangle) = c_\ell G_\ell^\beta(\langle x,y\rangle) \quad \text{with } c_\ell = \frac{2\ell+d-2}{d-2}$$

and G_ℓ^β Gegenbauer polynomial on $[-1,1]$ of degree ℓ orthogonal for the weight function $w_\beta(t) = (1-t^2)^{\beta-1/2}$. Here $\beta = \frac{d-2}{2}$.

- For $\mathbb{S}^2$ $(d=3, \beta=\frac{1}{2})$: Legendre polynomials.

Finally

$$W(x,y) = \mathbf{p}(\langle x,y\rangle) = \sum_{\ell\geq 0} \mathbf{p}_\ell^\star c_\ell G_\ell^\beta(\langle x,y\rangle))$$

with

$$\mathbf{p}_\ell^\star = a_\ell \int_{-1}^{1} \mathbf{p}(t) G_\ell^\beta(t) w_\beta(t) dt$$

$$\|\mathbf{p}\|^2 = \sum_\ell d_\ell |\mathbf{p}_\ell^\star|^2$$

Regularity on p

- Weighted Sobolev Spaces of order s

$$\sum_{\ell=0}^{\infty} d_\ell |\mathbf{p}_\ell^\star|^2 (1 + \ell(\ell+d))^s) < +\infty.$$

- Approximation with R first coefficients of $\mathbf{p}$ is of order R^{-2s}.

Estimation of $\mathbf{p}$ - Overall view

Denote $\Theta = (\theta_{ij})$.

$$\lambda\left(\frac{A}{n}\right) \approx \lambda\left(\frac{\Theta}{n}\right) \approx \lambda(\mathbb{T}_W) = \{\underbrace{\mathbf{p}_0^\star}_{d_0}, \underbrace{\mathbf{p}_1^\star, \dots, \mathbf{p}_1^\star}_{d_1}, \dots, \underbrace{\mathbf{p}_\ell^\star, \dots, \mathbf{p}_\ell^\star}_{d_\ell}, \dots\}$$

with

$$(\mathbb{T}_W g)(x) = \int_{\mathbb{S}^{d-1}} W(x,y) g(y) d\sigma(y)$$

Estimation of **p** - step 1

We work conditionally to X_i and suppose that $\Theta = (\theta_{ij})$ is fixed.

Proposition (Bandeira and van Handel (2016))

With probability $1 - \alpha$ *we have*

$$\left\| \frac{A}{n} - \frac{\Theta}{n} \right\| \leq \frac{3}{\sqrt{n}} + C_0 \frac{\sqrt{\log(n/\alpha)}}{n}$$

$\|\cdot\|$ operator norm. Hence with probability $1 - \exp(-n)$

$$\forall k \in [n], \quad \left| \lambda_k \left(\frac{A}{n} \right) - \lambda_k \left(\frac{\Theta}{n} \right) \right| = \mathcal{O}\left(\frac{1}{\sqrt{n}} \right)$$

Estimation of $\mathbf{p}$ - step 2

We use that $\theta_{ij} = W(X_i, X_j)$.

Suppose that the kernel W is in $\mathbb{L}^2$ and is symmetric. Let us define

$$\forall x \in \mathbb{S}^{d-1}, \forall g \in \mathbb{L}^2(\mathbb{S}^{d-1}), \quad (\mathbb{T}_W g)(x) = \int_{\mathbb{S}^{d-1}} W(x,y) g(y) d\sigma(y)$$

The spectral theorem states that

$$W(x,y) = \sum_k \lambda_k^\star \phi_k(x) \phi_k(y),$$

with $\lambda^\star$ eigenvalues of $\mathbb{T}_W$ and ϕ eigenvectors of $\mathbb{T}_W$.

Estimation of **p** - step 2 - Large law of numbers

Consider $\lambda^\star$ eigenvalue of $\mathbb{T}_W$, and denote $v = (\phi(X_1), \dots, \phi(X_n))$ then

$$\begin{aligned}
\lambda^\star v_i = \lambda^\star \phi(X_i) = \mathbb{T}_W \phi(X_i) &= \int_{\mathbb{S}^{d-1}} W(X_i, y)\phi(y)dy & (1)\\
&\approx \frac{1}{n} W(X_i, X_j)\phi(X_j) & (2)\\
&= \left(\frac{\Theta}{n} v\right)_i & (3)
\end{aligned}$$

- $\lambda^\star$ is almost an eigenvalue of $\frac{\Theta}{n}$
- Spectrum of $\mathbb{T}_W$ is close to spectrum of $\frac{\Theta}{n}$.

Estimation of **p** - step 2 - Large law of numbers

To compare spectra, we use the distance

$$\delta_2(x,y) := \inf_{\pi \in \mathcal{P}} \left[\sum (x_i - y_{\pi(i)} \right]^{\frac{1}{2}} = \lim_{N \to \infty} \left[\sum_{k=-N}^{N} (x_k - y_k) \right]^{\frac{1}{2}},$$

where $\mathcal{P}$ is the set of permutations with finite support and $x_{-1} \leq x_{-2} \leq \cdots \leq 0 \leq \cdots \leq x_2 \leq x_1$, completing with zeros if necessary.

Proposition (Koltchinski-Giné (2000))

Si $\mathbb{E}|W(X,Y)|^2 \leq \infty$

$$\delta_2 \left(\lambda \left(\frac{\Theta \mathbb{1}_{i \neq j}}{n} \right), \lambda(\mathbb{T}_W) \right) \to 0 \quad p.s$$

Also proved of a rate of convergence in $\sqrt{n}$.

Estimation of **p** - step 3

Proposition

If W only depends on the scalar product $\langle . , . \rangle$ then

- *The eigenvectors of $\mathbb{T}_W$ are the spherical harmonics.*
- *The eigenvalues of $\mathbb{T}_W$ are the Fourier coefficients of **p** in the Gegenbauer basis with multiplicity d_ℓ*

$$\lambda(\mathbb{T}_W) = \{ \underbrace{\mathbf{p}_0^\star}_{d_0}, \underbrace{\mathbf{p}_1^\star, \dots, \mathbf{p}_1^\star}_{d_1}, \dots, \underbrace{\mathbf{p}_\ell^\star, \dots, \mathbf{p}_\ell^\star}_{d_\ell}, \dots \}$$

Estimator of $\mathbf{p}$

Let R be a level of approximation and $\widetilde{R} = \sum_{\ell=0}^{R} d_\ell$

We set $\mathcal{M}_R = \left\{ (\underbrace{u_0^\star}_{d_0}, \underbrace{u_1^\star, \dots, u_1^\star}_{d_1}, \dots, \underbrace{u_R^\star, \dots, u_R^\star}_{d_R}) \in \mathbb{R}^{\widetilde{R}} \right\}$

We choose the sequence of $\mathcal{M}_R$ closest to A/n :

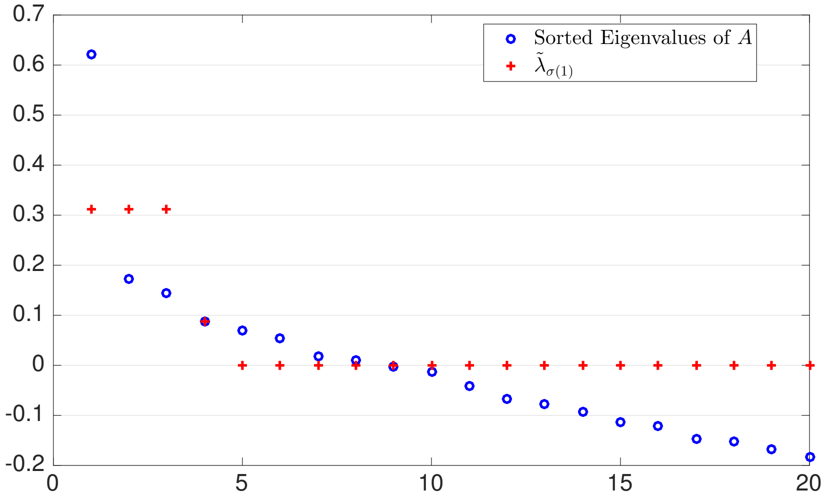
$$\widehat{\lambda}^R = \underset{u \in \mathcal{M}_R}{\operatorname{argmin}}\ \delta_2 \left(u, \lambda \left(\frac{A}{n} \right) \right)$$

where $\delta_2^2(u, v) = \min_{\sigma \in \mathfrak{S}_n} (u_k - \lambda_{\sigma(k)}(v))^2$

Proposition

Time complexity is in $n^3 + (R+2)!$

Example with $\mathbb{S}^2$, $R = 1$ and $n = 20$



Notations

- $\mathbb{P}(A_{ij} = 1) = \theta_{ij} = W(X_i, X_j) = \mathbf{p}(\langle X_i, X_j \rangle)$
- n : number of vertices of the observed graph
- $\lambda^\star$: spectrum of kernel W : $W(x, y) = \sum_k \lambda_k^\star \phi_k(x) \phi_k(y)$
- R : level of approximation, corresponds to $\widetilde{R} = d_0 + d_1 + \ldots d_R$ eigenvalues
- $\lambda^{\star R}$: set of $\widetilde{R}$ first eigenvalues of W
- $\widehat{\lambda}^R$: estimator of $\lambda^{\star R}$ and consequently of $\lambda^\star$
- $\delta_2(u, v)$ distance ℓ^2 up to permutations between two sequences u et v

Results

Theorem

$\exists\ \kappa_0 > 0$ *such that for all* $\alpha \in (0,1)$, *if* $n^3 \geq (2\widetilde{R})^3 \vee \widetilde{R}\log(2\widetilde{R}/\alpha)$ *with probability greater than* $1 - 3\alpha$,

$$\delta_2(\widehat{\lambda}^R, \lambda^{\star R}) \leq 4\delta_2(\lambda^{\star R}, \lambda^\star) + \kappa_0 \sqrt{\widetilde{R}\left(1 + \log\left(\widetilde{R}/\alpha\right)\right)/n}.$$

Moreover $\exists\ \kappa_1 > 0$ *such that if* $n \geq 2\widetilde{R}$ *then*

$$\mathbb{E}[\delta_2^2(\widehat{\lambda}^R, \lambda^{\star R})] \leq \kappa_1 \left\{ \delta_2^2(\lambda^{\star R}, \lambda^\star) + \frac{\widetilde{R}\log n}{n} \right\}.$$

Convergence rate

Corollary

Assume that $\mathbf{p}$ *belongs to the weighted Sobolev space of order* $s > 0$ *and set* $R_{\rm o} = \lfloor (n/\log n)^{\frac{1}{2s+d-1}} \rfloor$. *Then*

$$\mathbb{E}\left[\delta_2^2(\widehat{\lambda}^{R_{\rm o}}, \lambda^\star)\right] \leq C \left[\frac{n}{\log n}\right]^{-\frac{2s}{2s+(d-1)}} .$$

Usual nonparametric rate of convergence in dimension $(d-1)$.

Adaptive of choice of R ?

Adaptation to smoothness s

Let $\mathcal{R} = \{1, 2, \ldots, R_{\max}\}$ an admissible set of values of R with $2\widetilde{R}_{\max} \leq n$.

Goldenshluger-Lepski procedure :

$$B(R) := \max_{R' \in \mathbb{R}} \left\{ \delta_2(\widehat{\lambda}^{R'}, \widehat{\lambda}^{R' \wedge R}) - \kappa \sqrt{\frac{\widetilde{R'} \log n}{n}} \right\},$$

$$\widehat{R} \in \underset{R \in \mathbb{R}}{\operatorname{argmin}} \left\{ B(R) + \kappa \sqrt{\frac{\widetilde{R} \log n}{n}} \right\}.$$

Final estimator : $\widehat{\lambda}^{\widehat{R}}$

Adaptation to smoothness s

Theorem

If $\kappa \geq \kappa_0\sqrt{5}$, $\exists\ C' > 0$ *such that*

$$\mathbb{E}[\delta_2^2(\widehat{\lambda}^{\widehat{R}}, \lambda^\star)] \leq C' \min_{R \in \mathbb{R}} \left\{ \delta_2^2(\lambda^{\star R}, \lambda^\star) + \kappa^2 \frac{\widetilde{R} \log n}{n} \right\}.$$

And if $\mathbf{p}$ *belongs to the weighted Sobolev space of order* $s > 0$ *then*

$$\mathbb{E}\left[\delta_2^2(\widehat{\lambda}^{\widehat{R}}, \lambda^\star)\right] \leq C \left[\frac{n}{\log n}\right]^{-\frac{2s}{2s+(d-1)}}.$$

A problem of identifiability

Example for $\mathbb{S}^2$: $d = 3,\ d_\ell = 2\ell + 1,\ G_\ell$ Legendre polynomials

$$\mathbf{p}_a = \frac{1}{2} c_0 G_0 + \mu c_1 G_1 + 0 \times c_2 G_2 + 0 \times c_3 G_3 + \mu c_4 G_4 \,,$$
$$\mathbf{p}_b = \frac{1}{2} c_0 G_0 + 0 \times c_1 G_1 + \mu c_2 G_2 + \mu c_3 G_3 + 0 \times c_4 G_4$$

with $0 < \mu \leq 1/24$ (polynomials of degree 4 taking values in $[0,1]$)

$$\lambda_a^\star = (1/2,\ \underbrace{\mu,\mu,\mu}_{3},\ \underbrace{0,0,0,0,0}_{5},\ \underbrace{0,0,0,0,0,0,0}_{7},\ \underbrace{\mu,\mu,\mu,\mu,\mu,\mu,\mu,\mu,\mu}_{9})$$
$$\lambda_b^\star = (1/2,\ \underbrace{0,0,0}_{3},\ \underbrace{\mu,\mu,\mu,\mu,\mu}_{5},\ \underbrace{\mu,\mu,\mu,\mu,\mu,\mu,\mu}_{7},\ \underbrace{0,0,0,0,0,0,0,0,0}_{9})$$

$$3 + 9 = 5 + 7$$

Polynomial case

Proposition

Assume that $\mathbf{p}$ *is a polynomial of degree* D *and the* $\mathbf{p}_\ell^\star$*'s are non null and distincts. If* $R \geq D$ *and* n *are large enough then*

$$\mathbb{E}[\|\widehat{\mathbf{p}}^R - \mathbf{p}\|_2^2] \leq (18 + 4\kappa_0^2)\frac{\widetilde{R}\log n}{n},$$

Extension

These results can be extended to S (latent space) compact Lie groups (or compact symmetric space) equipped with a distance γ

$$W(x, y) = \mathbf{p}(\cos \gamma(x, y))$$

- spheres : $\mathbb{S}^{\mathbf{d}-1} = \mathrm{SO}(\mathbf{d})/\mathrm{SO}(\mathbf{d}-1)$
- real projective spaces : $\mathbb{RP}^{\mathbf{d}-1} = \mathrm{SO}(\mathbf{d})/\mathrm{O}(\mathbf{d}-1)$
- complex projective spaces : $\mathbb{CP}^{\mathbf{d}-1} = \mathrm{SU}(\mathbf{d})/\mathrm{U}(\mathbf{d}-1)$

Numerical results

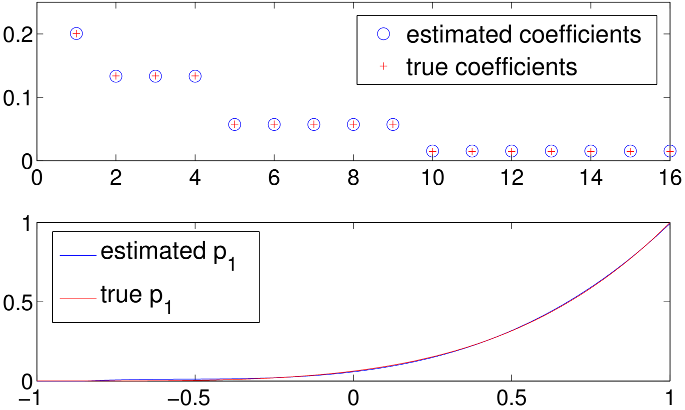
$\mathbb{S}^2$

$d = 3$

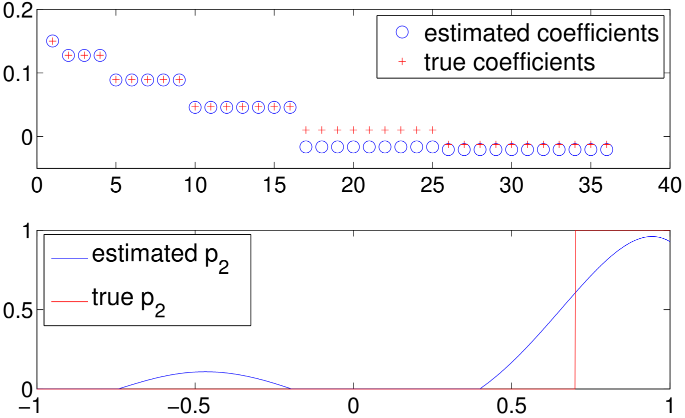
$n = 5000$

$R_{\max} = 4$

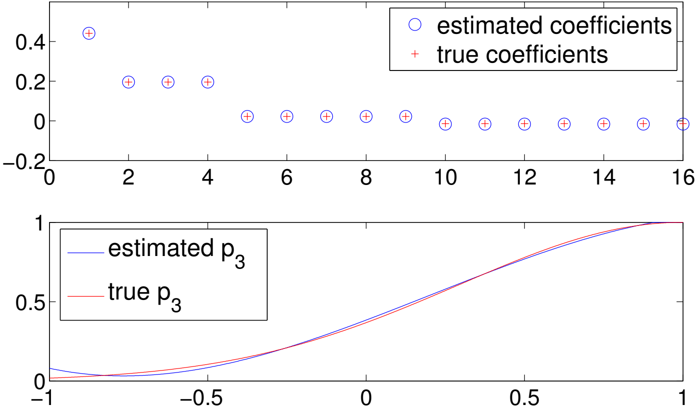
Estimation of $\mathbf{p}_1(t) = \left(\frac{1+t}{2}\right)^4$



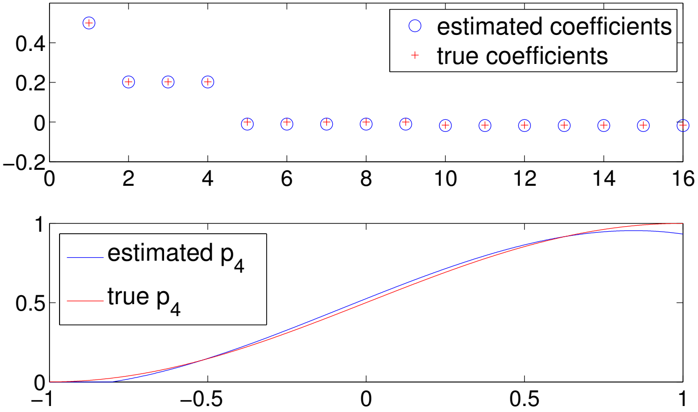
Estimation of $\mathbf{p}_2(t) = \mathbb{1}_{t>0.7}$



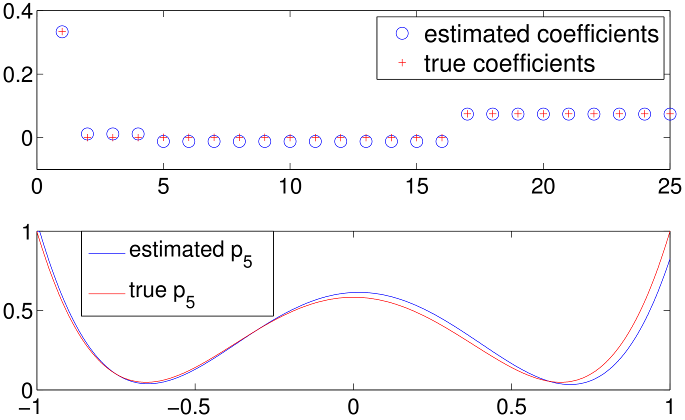
Estimation of $\mathbf{p}_3(t) = e^{-(t-1)^2}$



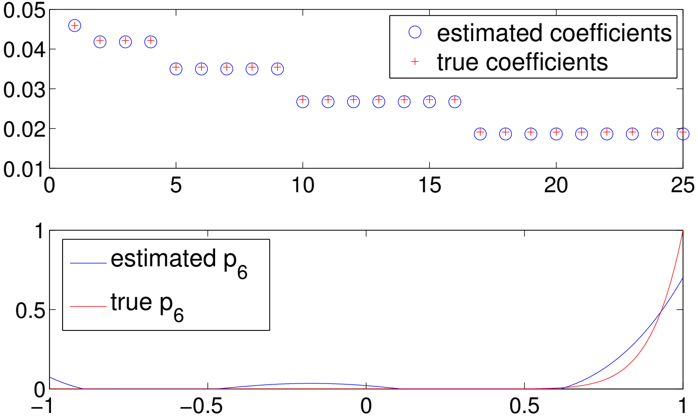
Estimation of $\mathbf{p}_4(t) = 0.5 + 0.5\sin\left(\pi t/2\right)$



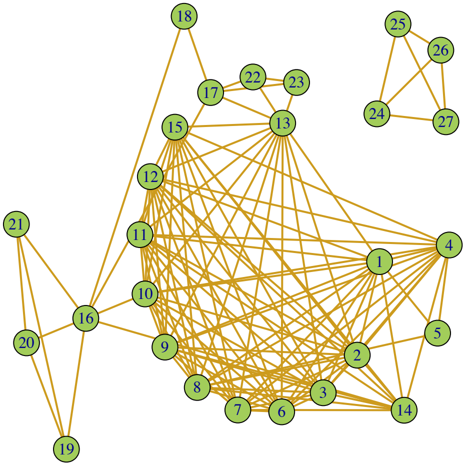
Estimation of $\mathbf{p}_5(t) = \frac{1}{3} + \frac{1}{12}\left(35t^4 - 30t^2 + 3\right)$



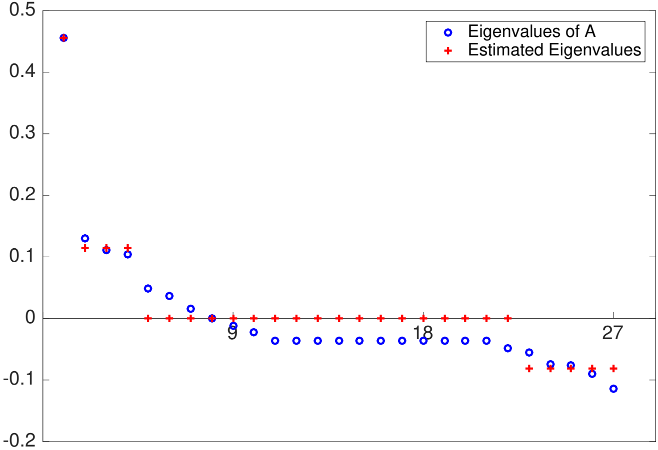
Estimation of $\mathbf{p}_6(t) = t^{10}\mathbb{1}_{t>0}$



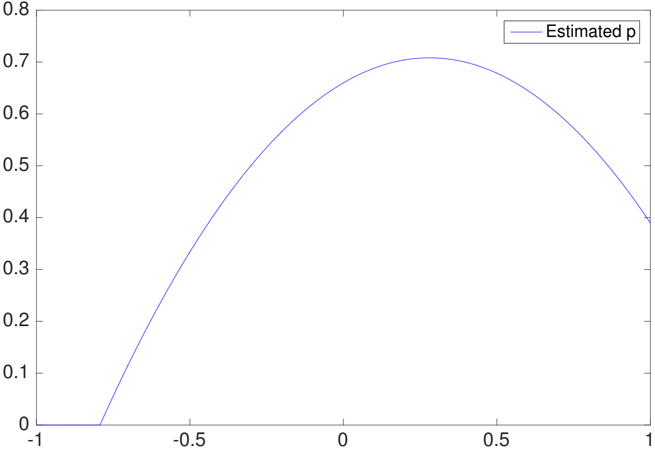
Ecological interactions : Zebra



Ecological interactions : Zebra



Ecological interactions : Zebra



Link to the article " Adaptive Estimation of Nonparametric Geometric Graphs"

https ://arxiv.org/abs/1708.02107