Model selection for simplicial approximation

C. Caillerie et B. Michel

INRIA Geometrica Team

ATMCS, the 25th June 2010

Problem

Topological and Geometric Inference: shape estimation from an approximation.

Shape: $\mathcal{O}$ open **bounded** set of $\mathbb{R}^n$ *or* K its boundary.

Approximation: K' finit set of approximating points of K.

2 views:

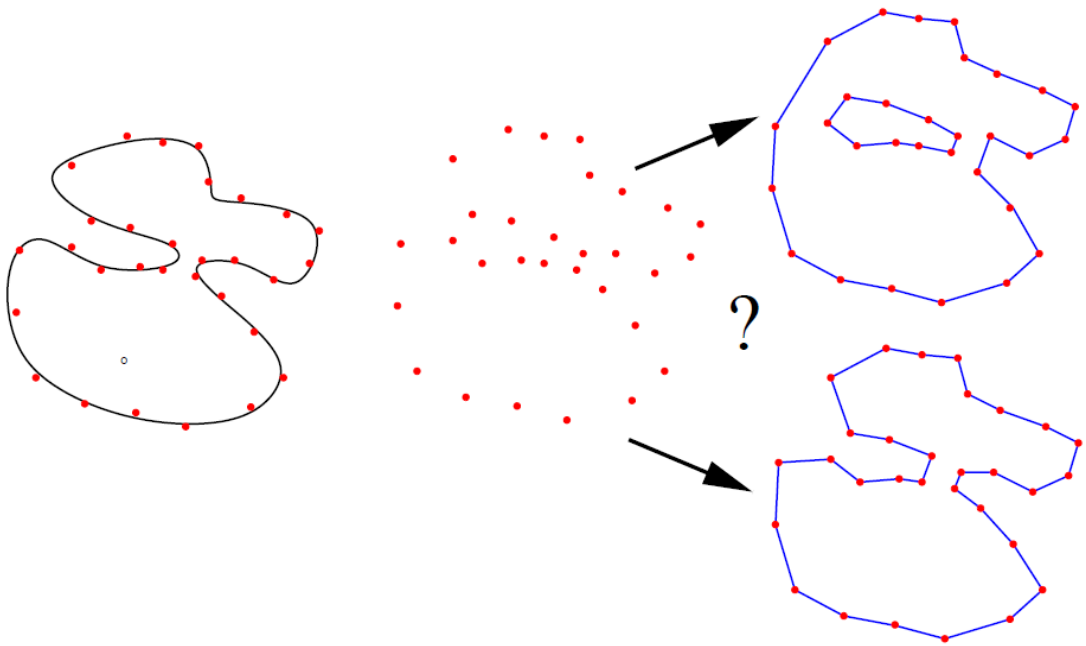
- Reconstruction.
- Estimation of geometric or topological properties.

2 motivations:

- Computer science and geometric computation (e.g. surface reconstructon in $\mathbb{R}^3$).
- Geometric Data Analysis(submanifold of $\mathbb{R}^D$, with a large D).

Model selection, what for?

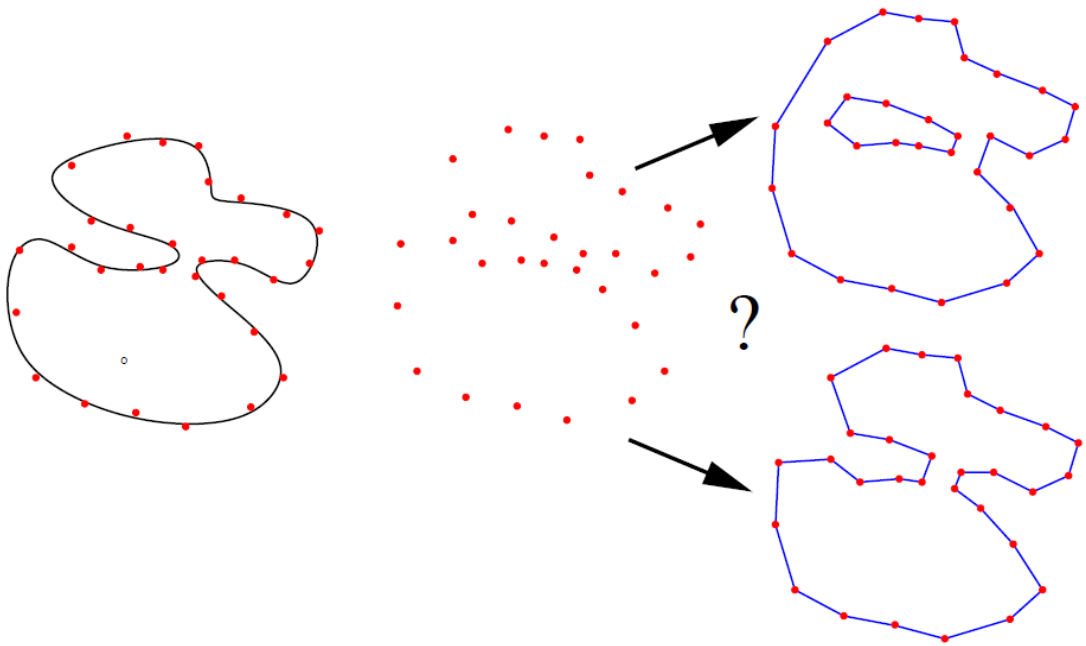
An open question: several possible choices.



→ choice of the "best" !
→ statistical tool : model selection!

Model selection, what for?

An open question: several possible choices.



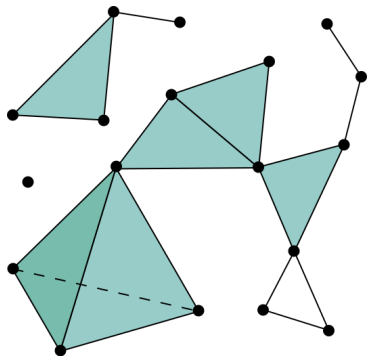
$\longrightarrow$ choice of the "best" !

$\longrightarrow$ statistical tool : model selection!

Simplicial complexes

- Simplicial complexes (s.c.) $\mathcal{C}$:
  - Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
  - The intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{C}$ is a face of both σ_1 and σ_2.

  Ex : Čech complexes, Rips complexes, α-shape, witness complex ...
  In general : one has a filtration $(\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$ of s.c. where α is a regularization parameter.

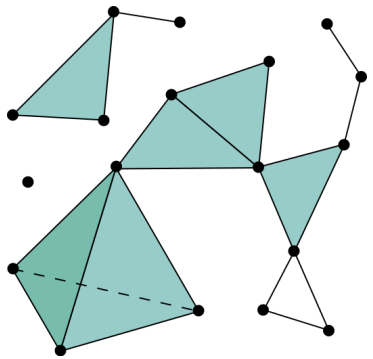


- Under proper assumptions, the persistent Betti numbers of $\mathcal{C}_\alpha$, for a well chosen α, coincide with the Betti numbers of the approximated shape (Chazal & Oudot 2008) .

Simplicial complexes

- Simplicial complexes (s.c.) $\mathcal{C}$:
  - Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
  - The intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{C}$ is a face of both σ_1 and σ_2.

  Ex : Čech complexes, Rips complexes, α-shape, witness complex ...
  In general : one has a filtration $(\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$ of s.c. where α is a regularization parameter.



- Under proper assumptions, the persistent Betti numbers of $\mathcal{C}_\alpha$, for a well chosen α, coincide with the Betti numbers of the approximated shape (Chazal & Oudot 2008) .

Our problem

- Input $X_1, \ldots, X_n \in \mathbb{R}^D$.
- Choose landmarks : $\mathcal{L}$
- One define a s.c. family over $\mathcal{L}$, indexed with a parameter α
- How to choose α ?

Two goals :

- Define a statistical framework for simplicial approximation.
- Use selection model tools to select a s.c. in the collection.

Our problem

- Input $X_1, \ldots, X_n \in \mathbb{R}^D$.
- Choose landmarks : $\mathcal{L}$
- One define a s.c. family over $\mathcal{L}$, indexed with a parameter α
- How to choose α ?

Two goals :

- Define a statistical framework for simplicial approximation.
- Use selection model tools to select a s.c. in the collection.

Data analysis

- "approximation" version of PCA :

  Model: $x_i = z_i + \varepsilon_i$ where $z_i \in E_d$ is an affine subspace of $\mathbb{R}^D$.

  PCA: least-squares minimization to find E_d.

- limitation : linearity of the appoximating spaces E_d.

- Idea: replace affine spaces by simplicial complexes.

Statistical model

- Let $\mathcal{G}$ an unknown geometric object in $\mathbb{R}^D$,

$$\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma\xi_i \quad \text{avec} \quad \bar{x}_i \in \mathcal{G}$$

$\bar{x}_i$:original points (unknown).
ξ_i iid $\mathcal{N}(0, I_D)$.

- For a s.c. $\mathcal{C}$, the best approximating point $\bar{\mathbf{x}}$ belonging to $\mathbf{C} = \mathcal{C}^n$ in the sense of least squares (LSE):

$$\hat{\mathbf{x}}_{\mathcal{C}} := \text{argmin}_{\mathbf{t} \in \mathcal{C}^n} \|\mathbf{X} - \mathbf{t}\|^2.$$

Notation : $\forall u \in \mathbb{R}^{nD}$, $\|u\|^2 := \frac{1}{nD} \sum_{i=1}^{nD} u_i^2$.

- A collection of s.c. $(\mathcal{C}_{\alpha \in \mathcal{A}}) \rightarrow$ A collection of LSE: $(\hat{\mathbf{x}}_\alpha)_{\alpha \in \mathcal{A}}$

Statistical model

- Let $\mathcal{G}$ an unknown geometric object in $\mathbb{R}^D$,

$$\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma\xi_i \quad \text{avec} \quad \bar{x}_i \in \mathcal{G}$$

  $\bar{x}_i$:original points (unknown).
  ξ_i iid $\mathcal{N}(0, I_D)$.

- For a s.c. $\mathcal{C}$, the best approximating point $\bar{\boldsymbol{x}}$ belonging to $\mathbf{C} = \mathcal{C}^n$ in the sense of least squares (LSE):

$$\hat{\mathbf{x}}_{\mathcal{C}} := \text{argmin}_{\mathbf{t} \in \mathcal{C}^n} \|\mathbf{X} - \mathbf{t}\|^2.$$

  Notation : $\forall u \in \mathbb{R}^{nD}$, $\|u\|^2 := \frac{1}{nD} \sum_{i=1}^{nD} u_i^2$.

- A collection of s.c. $(\mathcal{C}_{\alpha \in \mathcal{A}}) \rightarrow$ A collection of LSE: $(\hat{\boldsymbol{x}}_\alpha)_{\alpha \in \mathcal{A}}$

Statistical model

- Let $\mathcal{G}$ an unknown geometric object in $\mathbb{R}^D$,

$$\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma\xi_i \quad \text{avec} \quad \bar{x}_i \in \mathcal{G}$$

  $\bar{x}_i$:original points (unknown).
  ξ_i iid $\mathcal{N}(0, I_D)$.

- For a s.c. $\mathcal{C}$, the best approximating point $\bar{\boldsymbol{x}}$ belonging to $\mathbf{C} = \mathcal{C}^n$ in the sense of least squares (LSE):

$$\hat{\mathbf{x}}_{\mathcal{C}} := \text{argmin}_{\mathbf{t} \in \mathcal{C}^n} \|\mathbf{X} - \mathbf{t}\|^2.$$

  Notation : $\forall u \in \mathbb{R}^{nD}$, $\|u\|^2 := \frac{1}{nD} \sum_{i=1}^{nD} u_i^2$.

- A collection of s.c. $(\mathcal{C}_{\alpha \in \mathcal{A}}) \rightarrow$ A collection of LSE: $(\hat{\boldsymbol{x}}_\alpha)_{\alpha \in \mathcal{A}}$

Risk minimization

- Risk $\hat{\boldsymbol{x}}_\alpha$: $\mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Oracle (unknown): $\alpha_{or} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Selected model :

$$\hat{\alpha} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \left\{\|\mathbf{X} - \hat{\boldsymbol{x}}_\alpha\|^2 + \operatorname{pen}(\alpha)\right\},$$

- Find a penalty function for which the risk of $\hat{\boldsymbol{x}}_{\hat{\alpha}}$ is close to the target $\min_{\alpha\in\mathcal{A}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.

Risk minimization

- Risk $\hat{\boldsymbol{x}}_\alpha$: $\mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Oracle (unknown): $\alpha_{or} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Selected model :

$$\hat{\alpha} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \left\{\|\mathbf{X} - \hat{\boldsymbol{x}}_\alpha\|^2 + \operatorname{pen}(\alpha)\right\},$$

- Find a penalty function for which the risk of $\hat{\boldsymbol{x}}_{\hat{\alpha}}$ is close to the target $\min_{\alpha\in\mathcal{A}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.

Risk minimization

- Risk $\hat{\boldsymbol{x}}_\alpha$: $\mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Oracle (unknown): $\alpha_{or} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Selected model :

$$\hat{\alpha} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \left\{\|\mathbf{X} - \hat{\boldsymbol{x}}_\alpha\|^2 + \operatorname{pen}(\alpha)\right\},$$

- Find a penalty function for which the risk of $\hat{\boldsymbol{x}}_{\hat{\alpha}}$ is close to the target $\min_{\alpha\in\mathcal{A}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.

Risk minimization

- Risk $\hat{\boldsymbol{x}}_\alpha$: $\mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Oracle (unknown): $\alpha_{or} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.
- Selected model :

$$\hat{\alpha} := \underset{\alpha\in\mathcal{A}}{\operatorname{argmin}} \left\{\|\mathbf{X} - \hat{\boldsymbol{x}}_\alpha\|^2 + \operatorname{pen}(\alpha)\right\},$$

- Find a penalty function for which the risk of $\hat{\boldsymbol{x}}_{\hat{\alpha}}$ is close to the target $\min_{\alpha\in\mathcal{A}} \mathbb{E}_{\bar{\boldsymbol{x}}}\left(\|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2\right)$.

Model selection on simplicial complexes

- $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$: a collection of k-homogeneous s.c. in $\mathbb{R}^D$
- Notation :
  $\mathcal{C}_\alpha^+$ set of the simplices of maximal dimension k
  Δ_s : diameter of the smallest enclosing ball of s.
  $|\mathcal{C}|_k := (\sum_{s\in\mathcal{C}^+} \Delta_s^k)^{1/k}$
  $\delta_{\mathcal{C}} := \inf_{s\in\mathcal{C}_\alpha^+} \Delta_s$
- Noise level / complexity of the s.c. : for all $\alpha \in \mathcal{A}$,

$$\sigma \leq \delta_{\mathcal{C}_\alpha} \sqrt{\frac{D}{k}} \left[4\kappa \left(\sqrt{\ln \frac{4|\mathcal{C}_\alpha|_k}{\delta_{\mathcal{C}_\alpha}}} + \sqrt{\pi} \right) \right]^{-1}.$$

Model selection on simplicial complexes

Theorem - C. & Michel, 2009

There exists some constants c1 and c2 such that for all $\eta > 1$, if

$$\text{pen}(\alpha) \geq \eta \frac{\sigma^2}{nD} \left(c_1 nk \left[\ln \frac{|\mathcal{C}_\alpha|_k \sqrt{D}}{\sigma \sqrt{k}} + c_2 \right] \right),$$

then, almost surely, there exists a minimizer $\hat{\alpha}$ of the penalized criterion

$$\text{crit}(\alpha) = \|\bar{\boldsymbol{x}} - \hat{\boldsymbol{x}}_\alpha\|^2 + \text{pen}(\alpha)$$

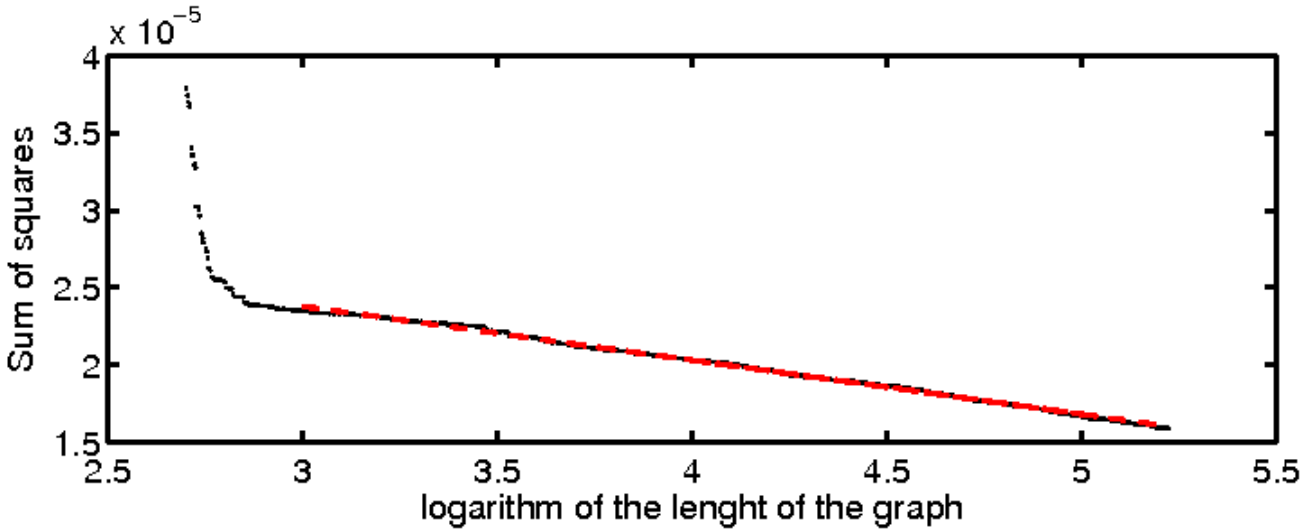
and the penalized estimator $\hat{\boldsymbol{x}}_{\hat{\alpha}}$ satisfies the following risk bound:

$$\mathbb{E}_{\bar{\boldsymbol{x}}} \|\hat{\boldsymbol{x}}_{\hat{\alpha}} - \bar{\boldsymbol{x}}\|^2 \leq c_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{ d(\bar{\boldsymbol{x}}, \mathcal{C}_\alpha^n)^2 + \text{pen}(\alpha) \right\} + \frac{\sigma^2}{nD} (\Sigma + 1) \right].$$

where $\sum_{\alpha \in \mathcal{A}} \frac{1}{|\mathcal{C}_\alpha|_k^L} = \Sigma < \infty$.

Slope heuristic : (penalty calibration)

- The penalty shape is known : $\mathrm{pen}(\alpha) = c \ln |\mathcal{C}_\alpha|_k$, but c is unknown.
- Slope heuristic : "optimal penalty = 2 times minimal penalty"

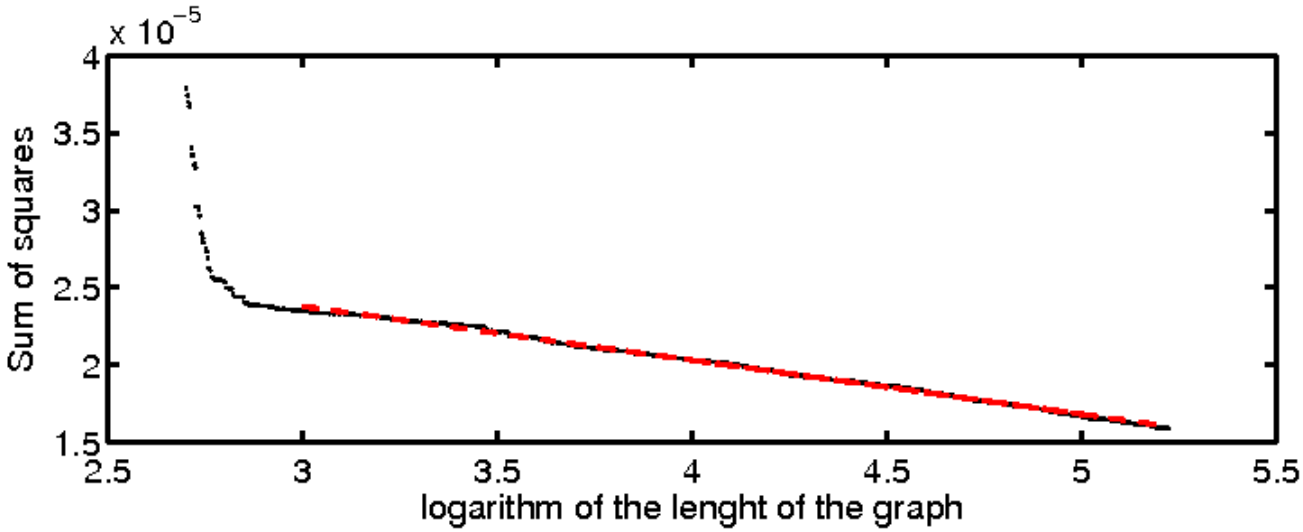


- Theoretical justifications of the slope method : Arlot & Massart, 2009 / Birgé & Massart 2007 .

- Validates the penalty shape suggested by the theorem.

Slope heuristic : (penalty calibration)

- The penalty shape is known : $\text{pen}(\alpha) = c \ln |\mathcal{C}_\alpha|_k$, but c is unknown.
- Slope heuristic : "optimal penalty = 2 times minimal penalty"

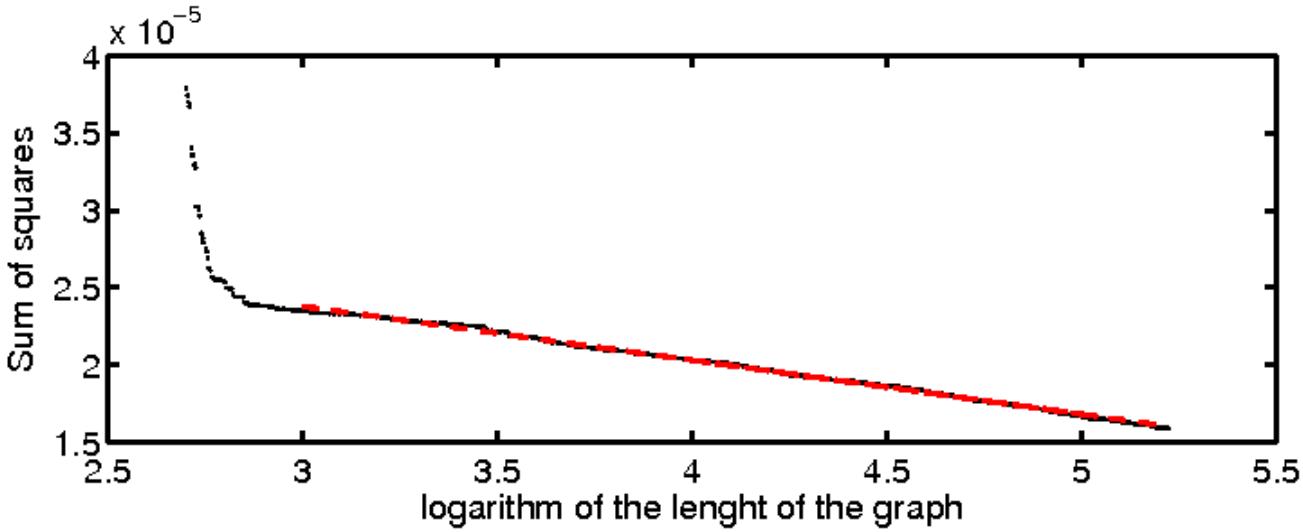


- Theoretical justifications of the slope method : Arlot & Massart, 2009 / Birgé & Massart 2007 .

- Validates the penalty shape suggested by the theorem.

Slope heuristic : (penalty calibration)

- The penalty shape is known : $\text{pen}(\alpha) = c \ln |\mathcal{C}_\alpha|_k$, but c is unknown.
- Slope heuristic : "optimal penalty = 2 times minimal penalty"

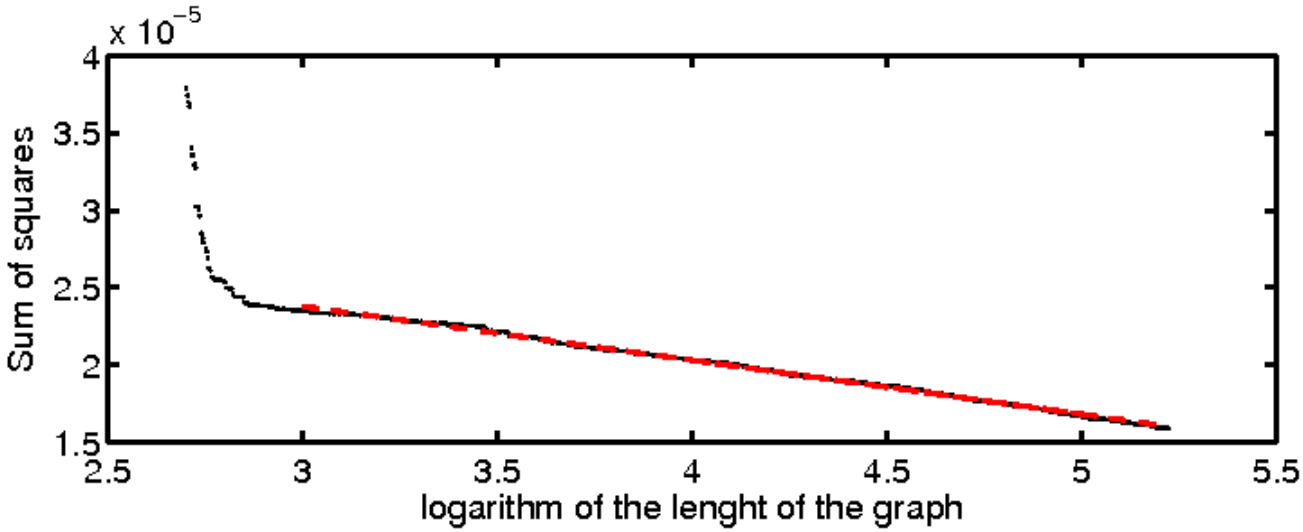


- Theoretical justifications of the slope method : Arlot & Massart, 2009 / Birgé & Massart 2007 .

- Validates the penalty shape suggested by the theorem.

Slope heuristic : (penalty calibration)

- The penalty shape is known : $\text{pen}(\alpha) = c \ln |\mathcal{C}_\alpha|_k$, but c is unknown.
- Slope heuristic : "optimal penalty = 2 times minimal penalty"



- Theoretical justifications of the slope method : Arlot & Massart, 2009 / Birgé & Massart 2007 .

- Validates the penalty shape suggested by the theorem.

Lissajous curve

- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, $(\sigma = 0.5\%)$ where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

Lissajous curve

- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, ($\sigma = 0.5\%$) where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

Lissajous curve

- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, ($\sigma = 0.5\%$) where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

Lissajous curve

- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, ($\sigma = 0.5\%$) where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

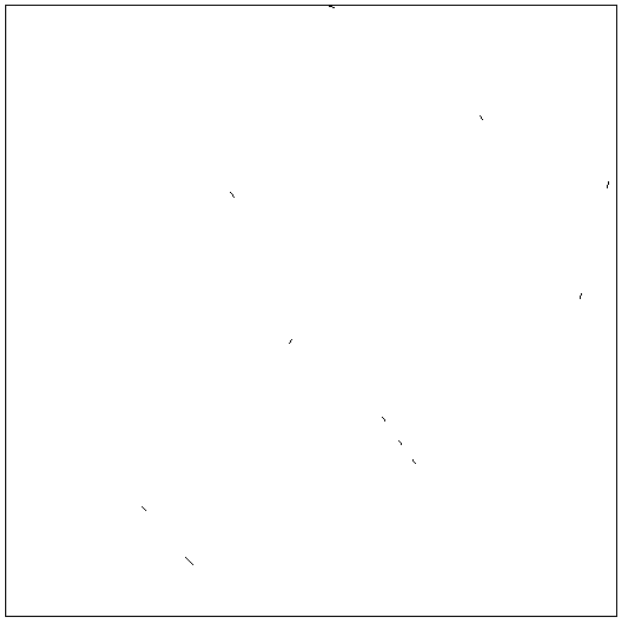
Lissajous curve

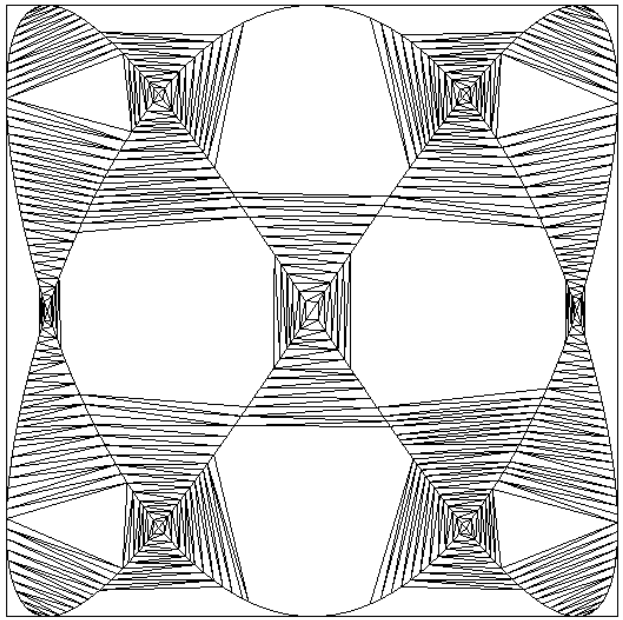
- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, ($\sigma = 0.5\%$) where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

Lissajous curve

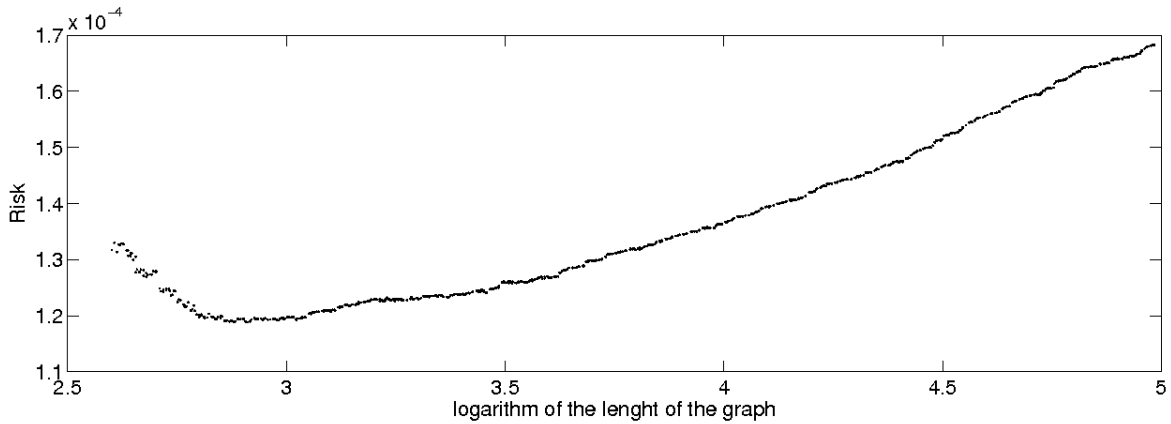
- Initial points $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, $(\sigma = 0.5\%)$ where $\bar{x}_i$ sampled on the Lissajous curve.
- $\mathcal{P}$ randomly separated between $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observations : $\mathcal{P}_o$
- Landmarks : 500 landmarks defined from $\mathcal{P}_l$ (neural-gas algorithm).
- Generation of α-shape graphs over the landmark points.
- 500 simulations of $\mathcal{P}_o$ in order to estimate the oracle graph (fixed landmarks).

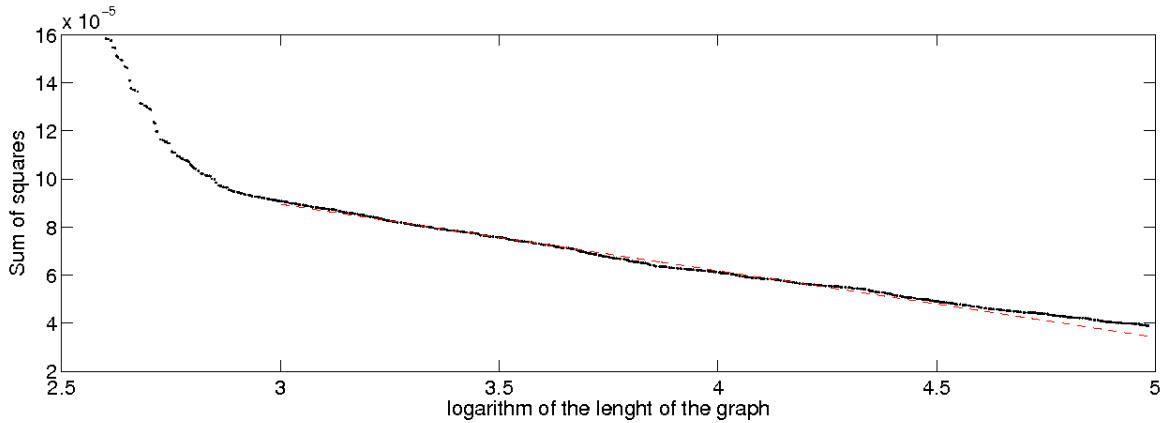
Lissajous curve- Extremal complexes





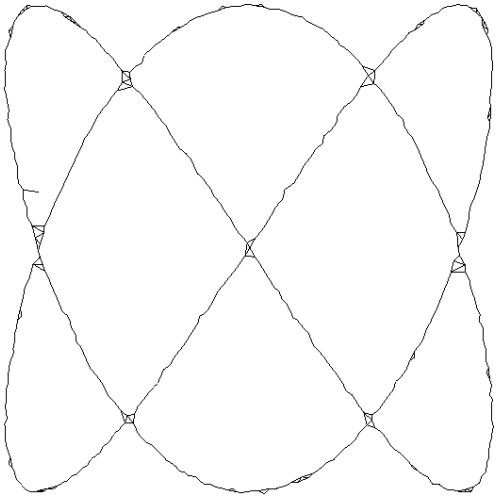
Lissajous curve - Risk and $SS(\alpha)$

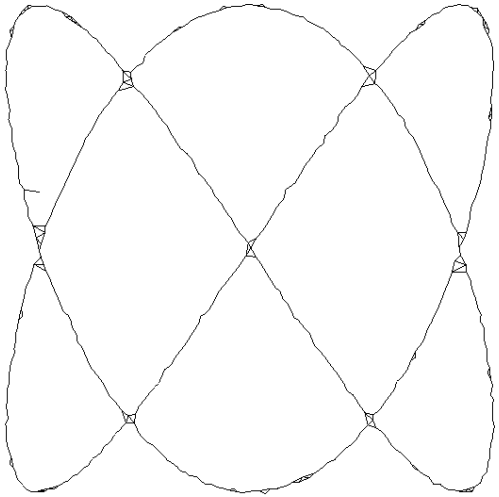




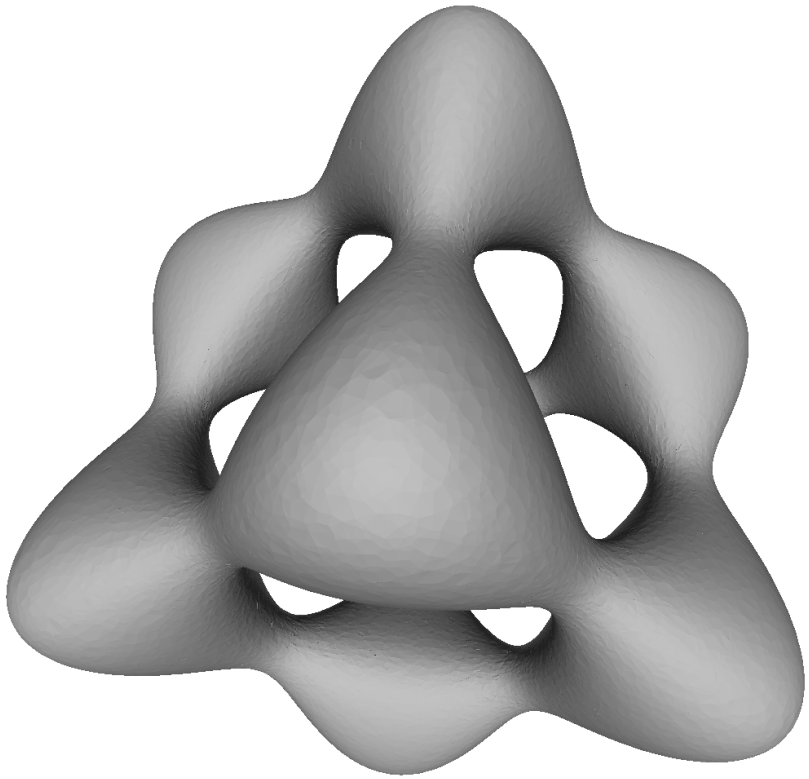
$\Rightarrow$ slope heuristic can be used.

Lissajous curve - oracle and selected graph



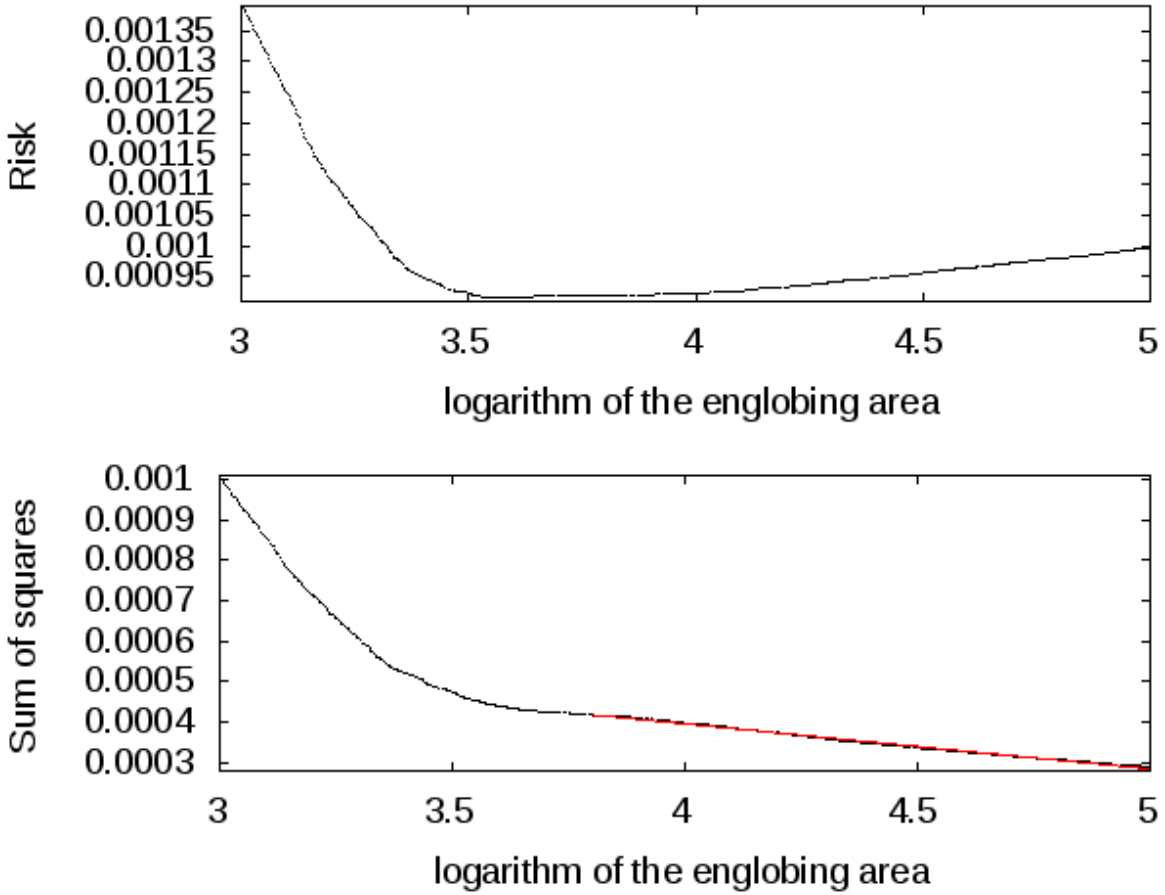


Tangle-Cube



$$x^4 + y^4 + z^4 - 5(x^2 + y^2 + z^2) + 11.8 = 0$$

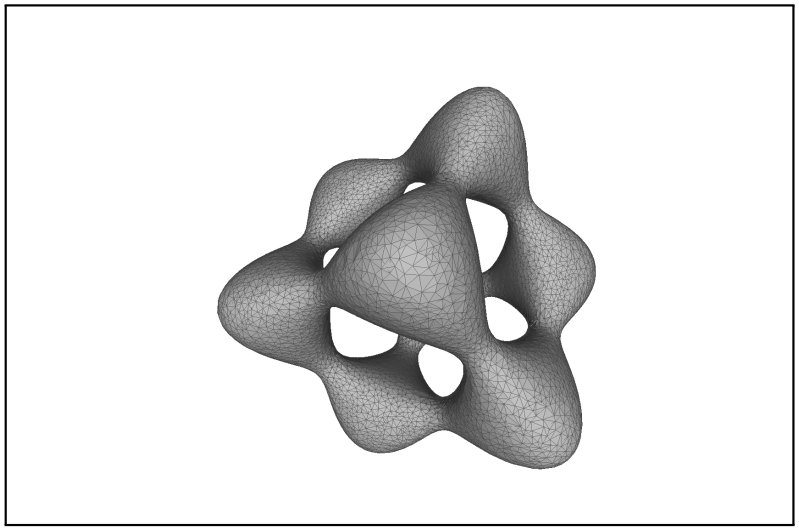
Tangle Cube : $SS(\alpha)$



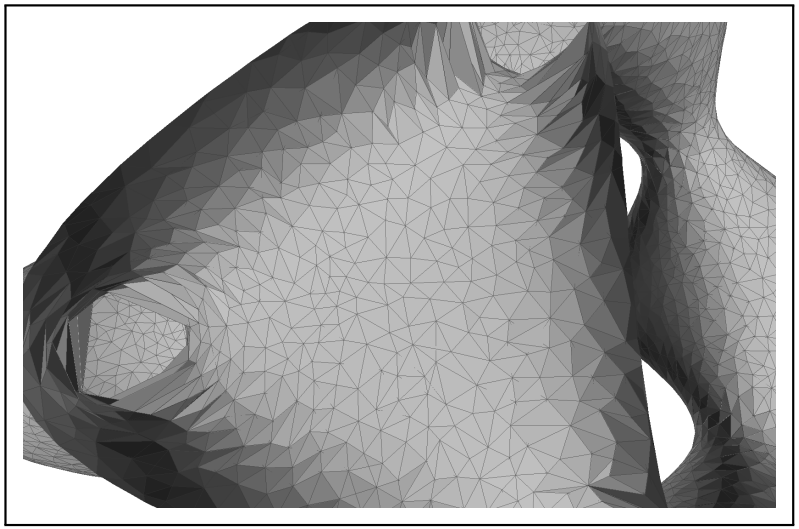
$\alpha_{\text{or}} = 0.051517$, $\ln |\mathcal{C}_{\alpha_{\text{or}}}|_2 = 3.5481$

$\hat{\alpha} = 0.053659$ and $\ln |\mathcal{C}_{\hat{\alpha}}|_2 = 3.5656$.

Tangle-Cube: selected surface

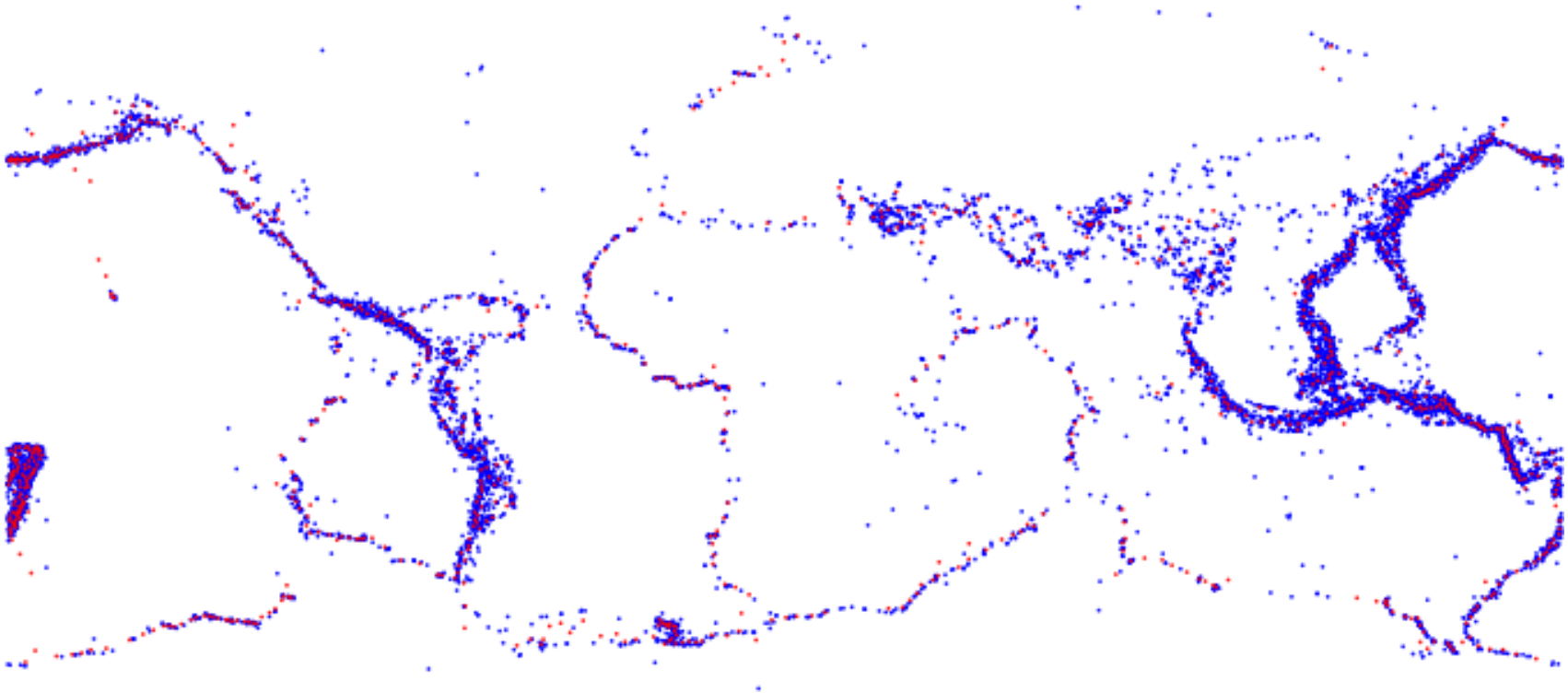


(a) Selected complex.

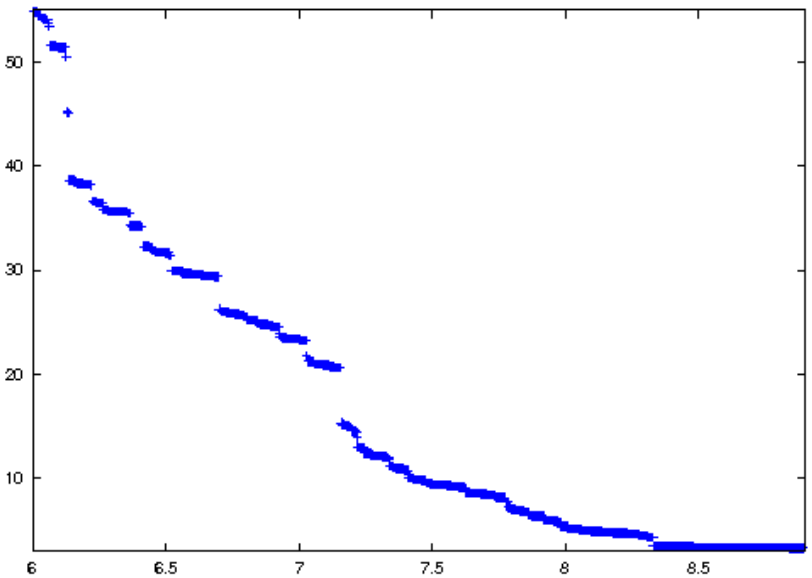


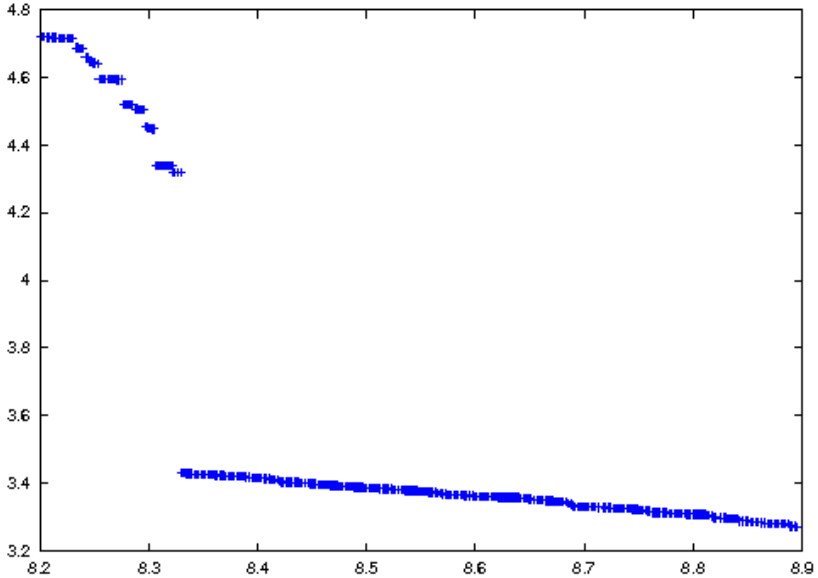
(b) Selected complex - zoom.

Earthquakes

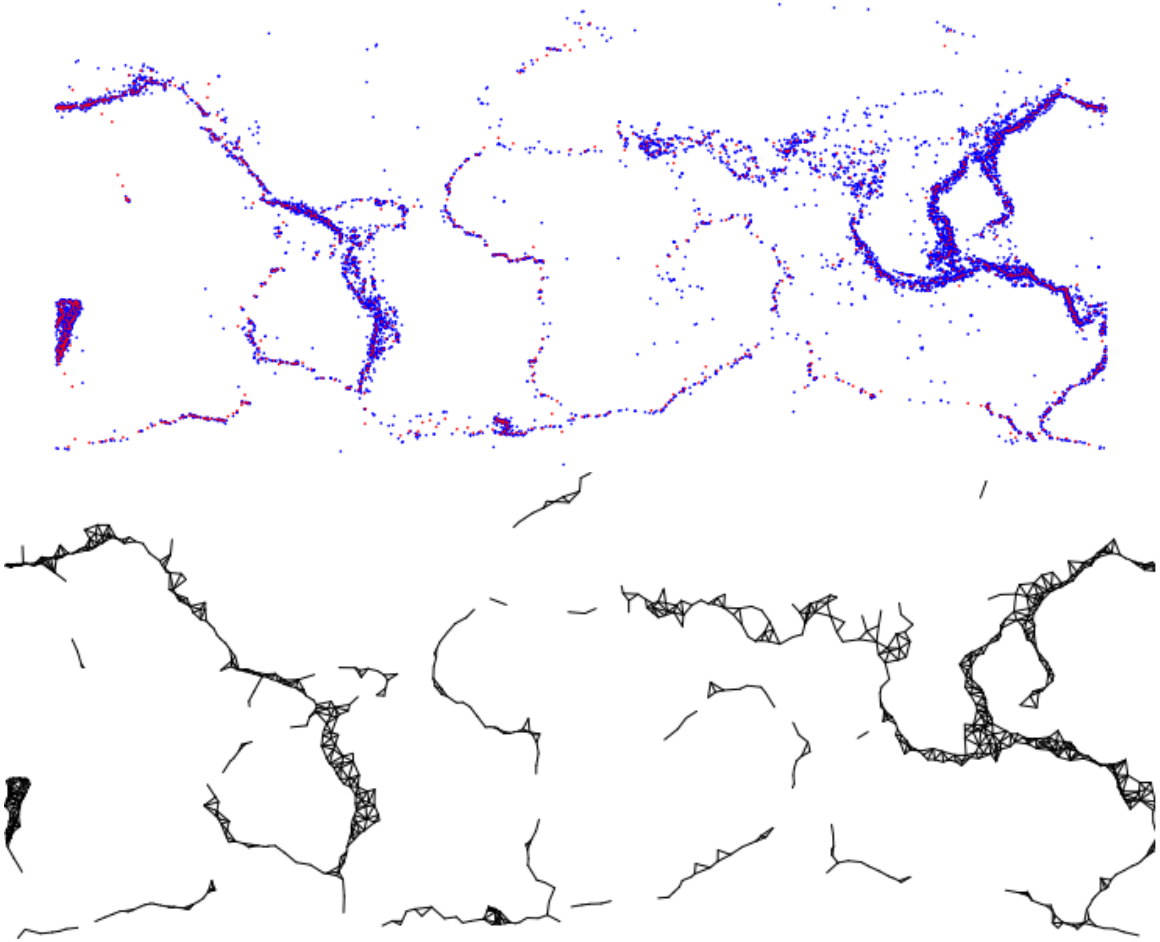


Earthquakes : $SS(\alpha)$

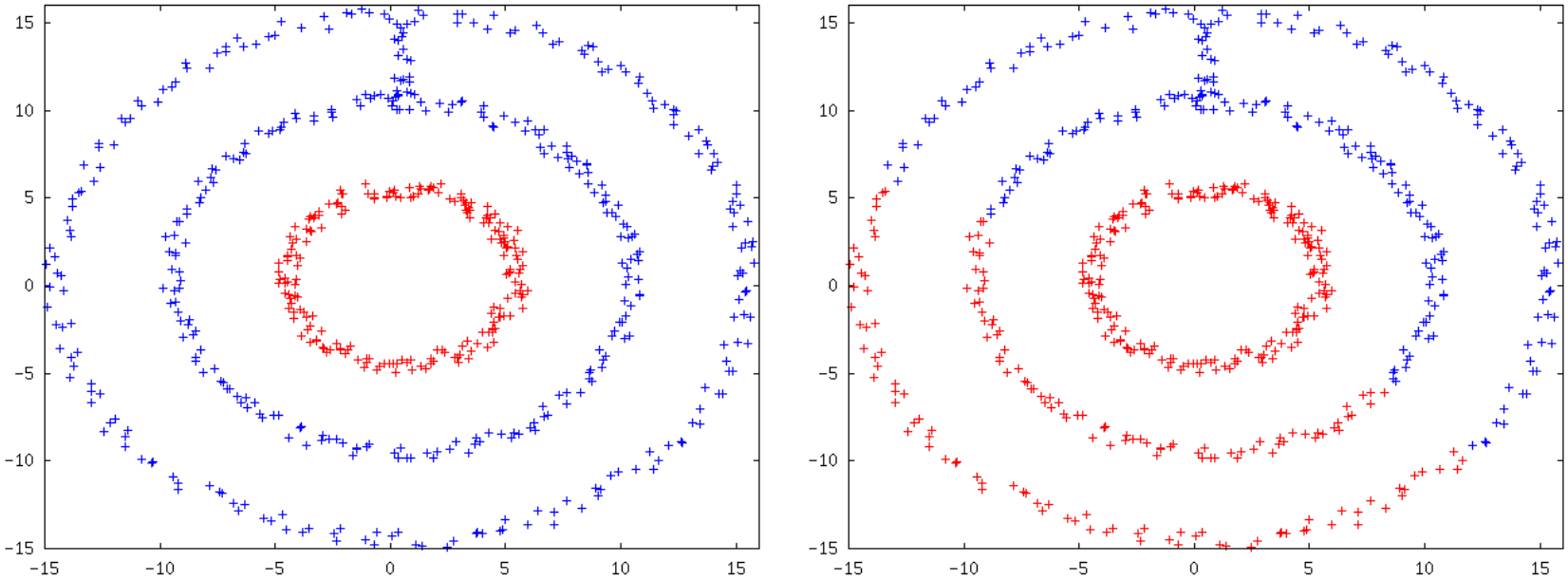




Earthquakes : selected graph

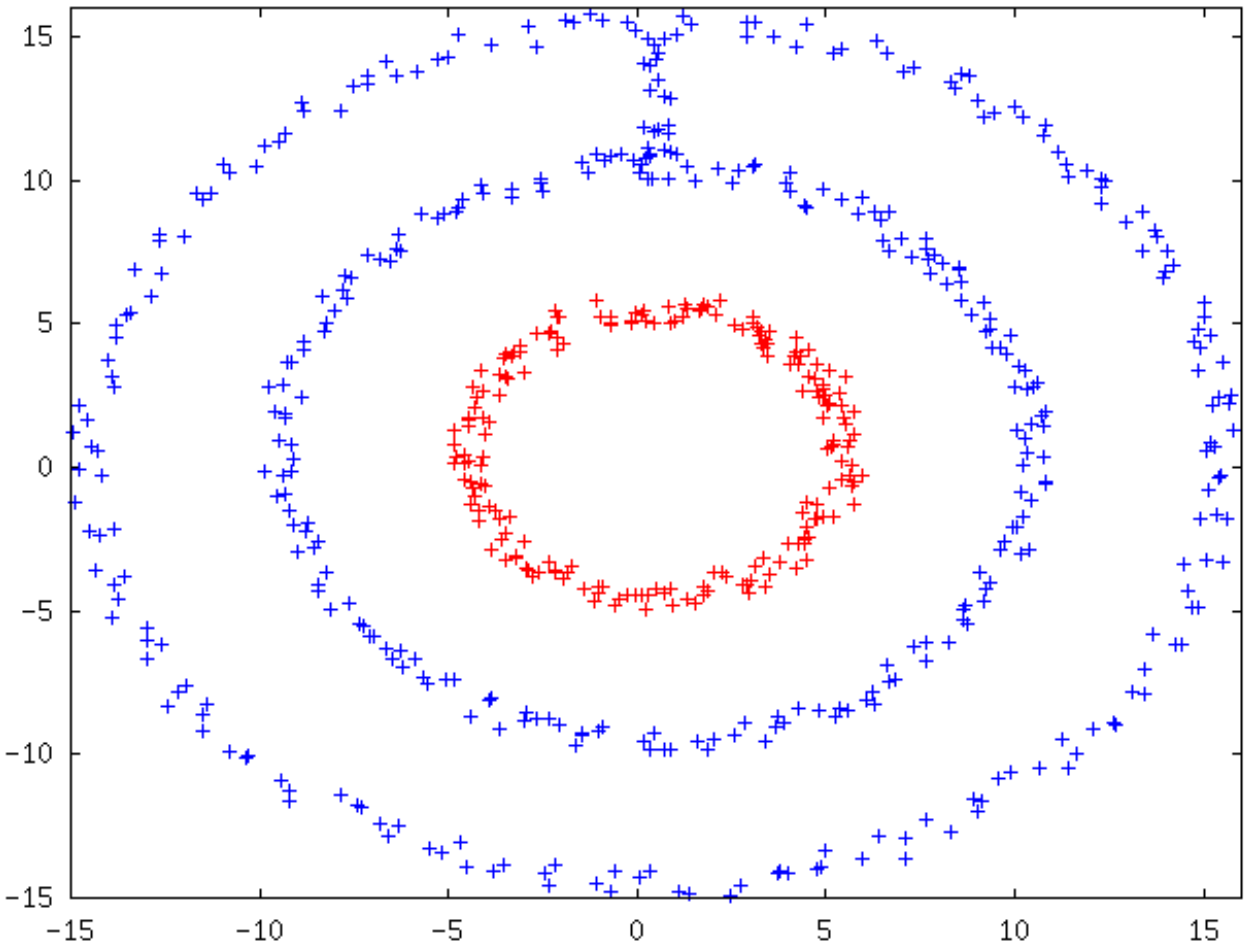


Spectral clustering



Classical spectral clustering based on a k-nearest neighbor with $k = 25(\textit{left})$ and $k = 30$ (right) : the clustering depends on k.

Spectral clustering: results



Spectral clustering based on the graph selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection

Conclusion and future works

- Minimization of a penalized least squares criterion $\longrightarrow$ selection of a simplicial complex, with THEORETICAL GUARANTEES.
- Minimization of the risk of the estimators associated to the complex collection $\longrightarrow$ penalty proportional to $\ln |\mathcal{C}_\alpha|_k$.
- Calibration of the penalty from the data : slope heuristics method.
- Experimental issues:
  1. Landmarks choice.
  2. Calibration : all the s.c. do not allow us to use the slope heuristics.
- Next Step: statistical/ topological selection


S. Arlot and P Massart.
Data-driven calibration of penalties for least-squares regression.
J.Mach.Learn.Res., 10:245–279, 2009.


Lucien Birgé and Pascal Massart.
Minimal penalties for Gaussian model selection.
Probab. Theory Related Fields, 138:33–73, 2007.


C. Caillerie and B. Michel.
Model selection for simplicial approximation.
Technical Report 6981, INRIA, 2009.


F. Chazal and S. Oudot.
Towards persistence-based reconstruction in euclidean spaces.
In *Proc. 24th ACM Sympos. on Comput. Geom.*, pages 232–241, 2008.


Pascal Massart.
Concentration Inequalities and Model Selection, volume Lecture Notes in Mathematics.
Springer-Verlag, 2007.