Model selection for simplicial approximation

C. Caillerie and B. Michel

INRIA Geometrica team

TGDA, Paris, july 2009

Summary

Outline

1. Motivations
2. Model selection and simplicial complexes
3. Experimental results
4. Discussion

Principal Component Analysis

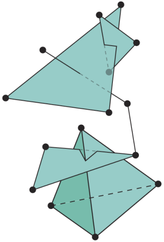
- Observations $X_1, \ldots, X_n \in \mathbb{R}^D$.
- probabilist version of PCA :

  Model : $x_i = z_i + \varepsilon_i$ where $z_i \in E_d$ affine subspace of R^Q

  PCA : least square minimization to find E_d.
- Main limitation : linearity of E_d.
- Extension : principal curves.

Data analysis with simplicial complexes

- Simplicial complex (s.c.) $\mathcal{C}$:
  - Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
  - The intersection of any two simplices s_1, $s_2 \in \mathcal{C}$ is either a face of both s_1 and s_2, or empty.

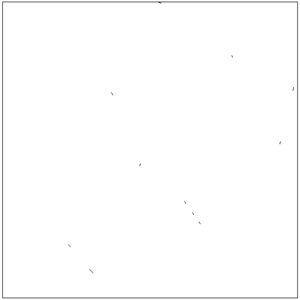


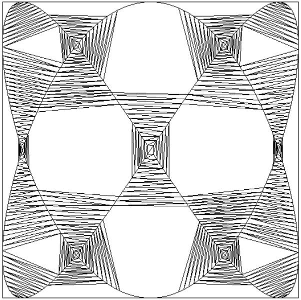
- Ex : Delauney, Rips complex, α-shape, witness complex ...
- s.c. are used for:
  - dimension estimation,
  - topological inference,
  - reconstruction.
- Initial idea of this work : fit a s.c. on the data.

General framework of the talk

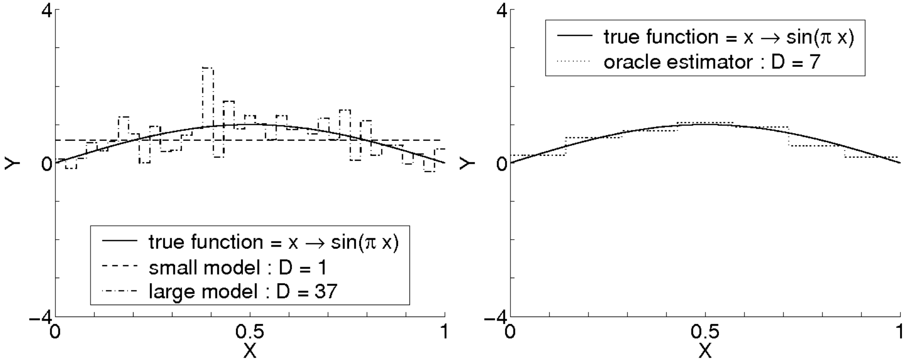
- Observations $X_1, \dots, X_n$.
- Choose some landmarks points.
- Several possible s.c. can be defined on the landmarks :
  $\longrightarrow$ a collection of s.c. $(\mathcal{C}_{\alpha \in \mathcal{A}})$ indexed by a scale parameter α.
- Which s.c should be chosen ?

Framework of the talk

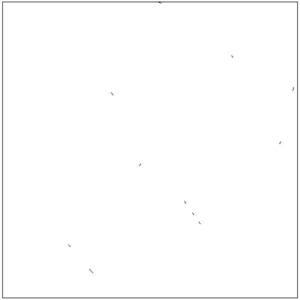


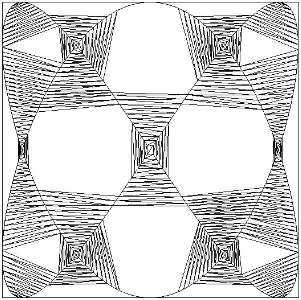


bias-variance tradeoff



Framework of the talk





Aims of this work :

- Define a statistical framework for the simplicial approximation.
- Use some model selection tools to find a “convenient” s.c. in the collection.

Outline

Geometric model

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

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

  where the original points $\bar{x}_i$ are unknown. The r.v. ξ_i are independent standard Gaussian vectors.

- equivalent statement :

$$\mathbf{X} = \bar{\mathbf{x}} + \sigma \boldsymbol{\xi} \quad \text{with} \quad \bar{\mathbf{x}} \in \mathcal{G}^n,$$

- Best approximating point of $\bar{\mathbf{x}}$ belonging to $\mathcal{C}^n$ is the least square estimator (LSE) of $\bar{\mathbf{x}}$ associated to $\mathcal{C}^n$:

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

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

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

Asymptotic model selection criterion

- Model selection via penalization :

$$\mathrm{crit}(m) = \gamma_n(\hat{\mathbf{x}}_m) + \mathrm{pen}(m)$$

  γ_n : empirical contrast: least squares or log likelihood.
  pen : $\mathcal{A} \to \mathbb{R}^+$: penalty function.

- C_p Mallows : penalized least square regression, pen $= 2D\sigma^2/n$.
- AIC : density estimation, pen $= D/n$
- BIC : density estimation pen $= D \log n/n$

- All these criterion are based on asymptomatic results.
- In our context : can be hardly applied since
  - no theoretical justifications,
  - what is D ?

Non asymptotic Gaussian model selection

- Birgé and Massart : non asymptotic model selection theory.
- Gaussian model selection (in our context)

$$\mathbf{X} = \bar{\mathbf{x}} + \sigma \boldsymbol{\xi} \quad \text{with } \bar{\mathbf{x}} \in \mathbb{R}^{Q}$$

- Collection of models $(C_\alpha)_{\alpha \in \mathcal{A}}$, where $C_\alpha \subset \mathbb{R}^Q$
  $\rightarrow$ LSE estimators $(\hat{\mathbf{x}}_\alpha)_{\alpha \in \mathcal{A}}$.
- Risk of $\hat{\mathbf{x}}_\alpha$: $\mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2\right)$.
- Oracle (unknown): $\alpha_{or} := \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} \mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2\right)$.
- Aim : find a penalty function pen such that the risk of $\hat{\mathbf{x}}_{\hat{\alpha}}$ where

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

is close to the benchmark $\min_{\alpha \in \mathcal{A}} \mathbb{E}_{\bar{\mathbf{x}}}\left(\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2\right)$.

Non asymptotic Gaussian model selection

- The penalty function depends on (see the theorem hereafter)
  1. the "size" of the model collection,
  2. the complexity of the models.
- Hypothesis on the model collection "size" : some weights w_α fulfills

$$\sum_{\alpha \in \mathcal{A}} e^{-w_\alpha} = \Sigma < \infty.$$

- Complexity of each model : entropy measure. For all $\alpha \in \mathcal{A}$, the auxiliary entropic function Φ_α is defined by

$$\Phi_\alpha(u) = \kappa \int_0^u \sqrt{\mathrm{H}(C_\alpha, \|\cdot\|, r)}\, dr.$$

For all $\alpha \in \mathcal{A}$ let d_α defined by the equation (if it exits)

$$\Phi_\alpha\left(\frac{2\sigma\sqrt{Q}}{\sqrt{d_\alpha}}\right) = \frac{\sigma d_\alpha}{\sqrt{Q}}.$$

Non asymptotic Gaussian model selection

Theorem 1 - Birgé Massart 01 [2]

Let $\eta > 1$. For a penalty such that

$$\mathrm{pen}(\alpha) \geq \eta\, \sigma^2 \left(\sqrt{d_\alpha} + \sqrt{2 w_\alpha} \right)^2$$

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

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\mathbf{x}}_\alpha\|^2 + \mathrm{pen}(\alpha).$$

Furthermore, the following risk bound holds for all $\bar{\mathbf{x}} \in \mathbb{R}^Q$

$$\mathbb{E}_{\bar{\mathbf{x}}} \|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \leq c_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{ d(\bar{\mathbf{x}}, C_\alpha)^2 + \mathrm{pen}(\alpha) \right\} + \sigma^2 (\Sigma + 1) \right]$$

where c_η depends only on η and $d(\bar{\mathbf{x}}, C_\alpha) := \inf_{\mathbf{y} \in C_\alpha} \|\bar{\mathbf{x}} - \mathbf{y}\|$.

Non asymptotic linear Gaussian model selection

- The models C_α are linear subspaces of $\mathbb{R}^{nD}$.
- d_α is equal to the dimension of C_α.
- Risk bound : true oracle inequality :

$$\mathbb{E}_{\bar{\mathbf{x}}} \|\hat{\mathbf{x}}_{\hat{\alpha}} - \bar{\mathbf{x}}\|^2 \leq c'_\eta \left[\inf_{\alpha \in \mathcal{A}} \left\{ \mathbb{E}_{\bar{\mathbf{x}}} \left(\|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2 \right) \right\} + \sigma^2 (\Sigma + 1) \right]$$

- But if $\mathcal{C}_\alpha$ is a simplicial complex, of course $C_\alpha = \mathcal{C}_\alpha^n$ is not a linear subspace.

Model selection on simplicial complexes

- $(\mathcal{C}_\alpha)_{\alpha\in\mathcal{A}}$ a collection of k-homogeneous s.c. in $\mathbb{R}^D$
- Hypothesis on the collection "size". Weights : $w_\alpha = L \ln |\mathcal{C}_\alpha|_k$ with

$$\sum_{\alpha\in\mathcal{A}} \frac{1}{x_\alpha^L} = \Sigma < \infty$$

- Notation :
  Δ_s : diameter of the smallest including ball of the simplex s.
  $|\mathcal{C}|_k := (\sum_{s\in\mathcal{C}^+} \Delta_s^k)^{1/k}$ and $\delta_\mathcal{C} := \inf_{s\in\mathcal{C}_\alpha^+} \Delta_s$ where $\mathcal{C}_\alpha^+$ is the subset of simplices of $\mathcal{C}_\alpha$ of maximal dimension k.
- Hypothesis on the s.c. complexity. 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 2 - Caillerie and M. (2009) [4]

There exists some absolute constants c_1 and c_2 such that for all $\eta > 1$, if

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

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

$$\mathrm{crit}(\alpha) = \|\mathbf{X} - \hat{\boldsymbol{x}}_\alpha\|^2 + \mathrm{pen}(\alpha)$$

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

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

Remarks

- A quite general result.
- A qualitative result.
- Roughly speaking : pen is proportional to $\ln |\mathcal{C}_\alpha|_k$.
  For a collection of graph : $|\mathcal{C}_\alpha|_1$ is the graph length.
- Not exactly an oracle inequality ... additional work necessary to control the shape of the risk.
- If the true positions $\bar{\mathbf{x}}$ are sampled on $\mathcal{G}$ according to μ, for the integrated risk :

$$\int_{\bar{x}\in\mathcal{G}} \mathbb{E}_{\bar{x}} \|\hat{x}_{\hat{\alpha}} - \bar{x}\|^2 \, d\mu(\bar{x}) \leq c_\eta \left[\inf_{\alpha\in\mathcal{A}} \left\{ \int_{\bar{x}\in\mathcal{G}} d(\bar{x}, \mathcal{C}_\alpha)^2 \, d\mu(\bar{x}) + \mathrm{pen}(\alpha) \right\} + \sigma^2(\Sigma + 1) \right]$$

Sketch of the proof

- $C_\alpha = \mathcal{C}_\alpha^n$
- $Q = nD$
- Based on the following entropic result :

Proposition

For all k-homogeneous simplicial complex $\mathcal{C}$ of $\mathbb{R}^D$ and all $r \leq \delta_{\mathcal{C}}$

$$\mathrm{N}\left(\mathcal{C}^n, \|\cdot\|, r\right) \leq \left(\frac{4|\mathcal{C}|_k}{r}\right)^{nk}.$$

Outline

Slope heuristics

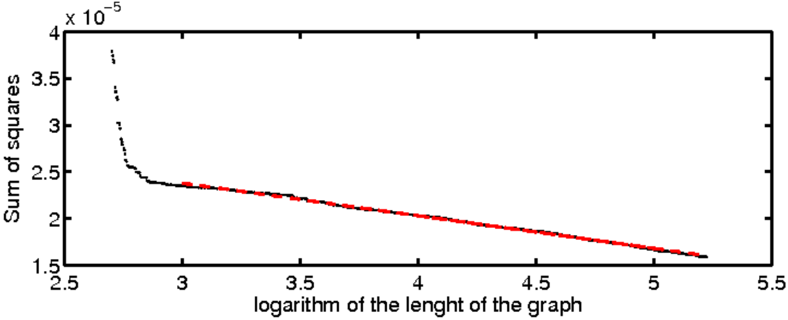
- The penalty type is known : $\text{pen}(\alpha) = c \log |\mathcal{C}_\alpha|_k$, but c is unknown.
- Slope heuristics method :
  1. For each simplicial complex, compute the sum of squares $SS(\alpha) := \|\hat{\mathbf{x}}_\alpha - \mathbf{X}\|^2$.
  2. Plot the point cloud $\{\ln |\mathcal{C}_\alpha|_k, SS(\alpha)\}_{\alpha \in \mathcal{A}}$ and check that a linear trend is observed for large α.
  3. Compute the slope $\hat{\beta}$ of the linear regression of $SS(\alpha)$ on $\ln |\mathcal{C}_\alpha|_k$ for large α.
  4. Select the simplicial complex in the collection minimizing

$$\text{crit}(\alpha) = \|\bar{\mathbf{x}} - \hat{\mathbf{x}}_\alpha\|^2 - 2\hat{\beta} \ln |\mathcal{C}_\alpha|_k \ .$$

- Theoretical results on the slope heuristics [3, 1].

Slope heuristics for graphs

"the optimal penalty is twice the minimal penalty"

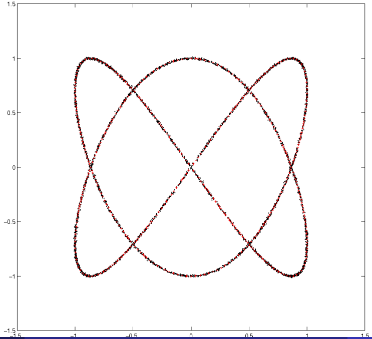


Selection of α-graphs : framework

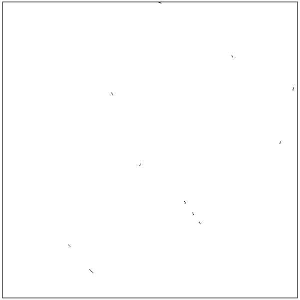
- An observed sample $X_1, \ldots, X_n$.
- Define a set of landmarks from the X_i.
- Define a collection of α-graphs (α-shape of dim 1).
- For each graph, compute the length $l(\alpha)$ and $SS(\alpha) := \|\hat{\mathbf{x}}_\alpha - \mathbf{X}\|^2$.
- Proceed the slope heuristics method to select a graph.

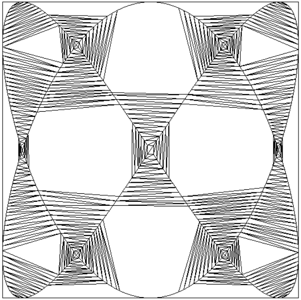
Lissajous curve (1)

- True points : $\bar{x}_1, \ldots, \bar{x}_n$ ($n = 5000$) sampled on the Lissajous curve.
- Observed points : $\forall i = 1, \ldots, n, \quad X_i = \bar{x}_i + \sigma\xi_i$, $\sigma = 0.005$.
- Landmarks points : Furthest point strategy on a set of true points (located on the Lissajous curve) $\rightarrow$ 500 landmark points.
- Compute the α-graphs on the landmark points.
- Compute the same experience 500 times to estimate the oracle graph (with fixed landmarks).

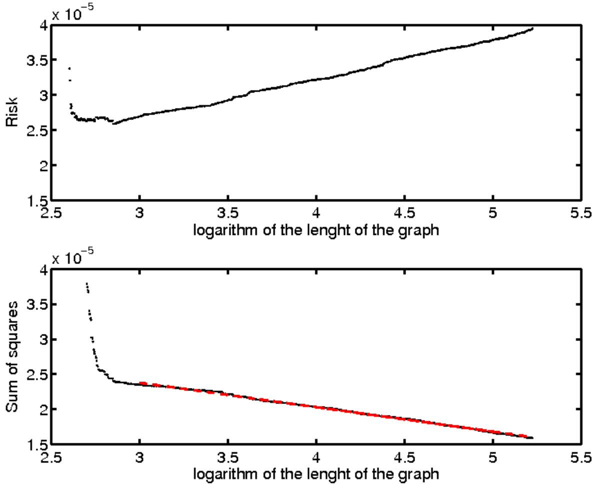


Lissajous curve (1) - extremal graphs



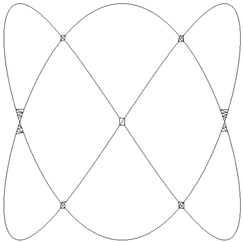


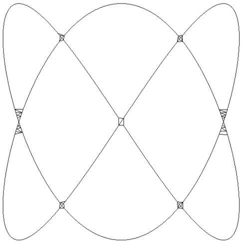
Lissajous curve (1) - risk and $SS(\alpha)$



$\Rightarrow$ the slope heuristics can be applied.

Lissajous curve (1) oracle and selected graphs





Lissajous curve (1) - 500 experiences

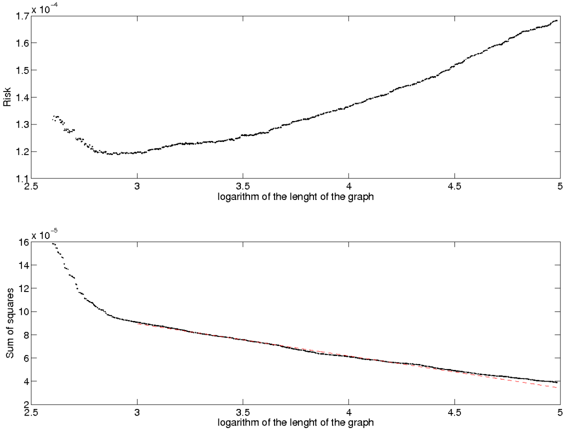
$\alpha \times 10^{-3}$	$\alpha_{\min}\ldots$	1.129	1.255	1.256	1.283	1.286	1.298	1.344
$N(\alpha)$	0	1	369	6	19	77	3	10
Selection perc.	0	0.2	73.8	1.2	3.8	15.4	0.6	2
Length	0.0394	16.86	17.30	17.37	17.45	17.50	17.57	17.64
Risk $\times 10^{-5}$	29841	2.627	2.589	2.588	2.591	2.594	2.594	2.596

$\alpha \times 10^{-3}$	1.493	1.603	1.643	1.669	1.672	1.748	$\ldots\alpha_{\max}$
$N(\alpha)$	6	4	1	2	1	1	0
Selection perc.	1.2	0.8	0.2	0.4	0.2	0.2	0
Length	17.71	17.97	18.10	18.31	18.46	18.61	185.8
Risk $\times 10^{-5}$	2.606	2.618	2.623	2.639	2.641	2.642	3.946

Lissajous curve (2)

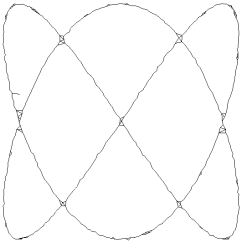
- Initial point set $\mathcal{P}$: $X_i = \bar{x}_i + \sigma\xi_i$, ($\sigma = 0.005$) where the $\bar{x}_i$ are sampled on the Lissajous curve.
- $\mathcal{P}$ is randomly separated into $\mathcal{P}_o$ (5000 points) and $\mathcal{P}_l$ (5000 points)
- Observed points : $\mathcal{P}_o$
- Landmark points : 500 landmarks points defined from $\mathcal{P}_l$ thanks to the neural-gas algorithm.
- Compute the α-graphs on the landmark points.
- Simulate $\mathcal{P}_o$ 500 times to estimate the oracle graph (with fixed landmarks).

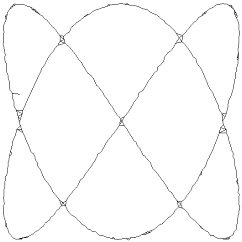
Lissajous curve (2) - risk and $SS(\alpha)$



$\Rightarrow$ the slope heuristics can be applied.

Lissajous curve (2) oracle and selected graphs



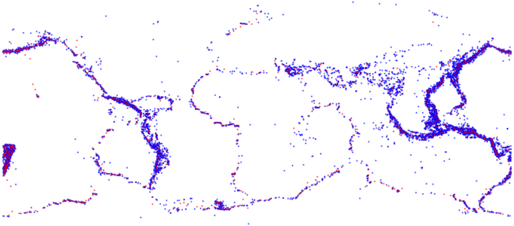


Lissajous curve (2) - 500 experiences

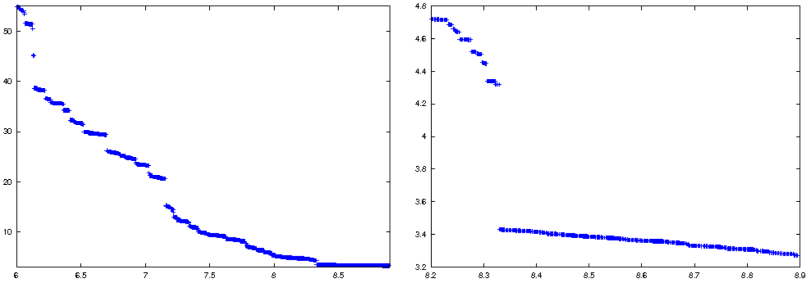
$\alpha \times 10^{-3}$	$\alpha_{min}\ldots$	0.9537	0.9891	1.051	1.076	1.078	1.084
$N(\alpha)$	0	38	3	107	36	281	2
Selection perc.	0	7.6	0.6	21.4	7.2	56.2	0.4
Length	0.03083	17.45	17.64	17.87	17.97	18.02	18.09
Risk $\times 10^{-4}$	308	1.1910	1.1899	1.1897	1.1942	1.1939	1.1937

$\alpha \times 10^{-3}$	1.126	1.183	1.187	1.200	1.205	1.271	$\ldots\alpha_{max}$
$N(\alpha)$	13	12	0	4	1	3	0
Selection perc.	2.6	2.4	0	0.8	0.2	0.6	0
Length	18.29	18.34	18.38	18.49	18.55	18.82	146.1
Risk $\times 10^{-4}$	1.1898	1.1886	1.1885	1.1899	1.1932	1.1944	1.6823

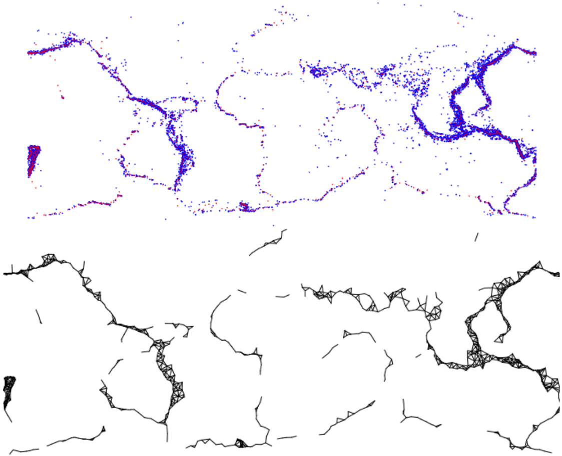
Real data : locations of earthquakes



Earthquakes : $SS(\alpha)$



Real data : selected graph



Outline

Discussion

- A first attempt to use modern model selection tools for geometric inference.
- Model selection via penalization : a general result gives the penalty form.
- For application : the slope heuristics does not work all the times (α -Rips)
- Future works :
  - theoretical aspects : a theory on s.c. approximation to control the bias.
  - heterogeneous s.c. ?
  - application : the same procedure in higher dimensions, other s.c families...

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.
Gaussian model selection.
J. Eur. Math. Soc. (JEMS), 3:203–268, 2001.

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.

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