Dimension Estimation using Random Connection Models

Paulo Serra

KdV Institute

18th March 2016

UNIVERSITEIT VAN AMSTERDAM

Summary

1. What is meant by dimension
2. Why estimate the dimension?
3. Previous work
4. Modelling the data
5. Intuition behind- and definition of the estimators
6. Consistency of the estimators for the intrinsic dimension
7. Numerical results
8. Conclusions

What is meant by dimension

Our setup is the following:

- There is data $X_1, \ldots, X_n$, where $X_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}^D$, for some $D \in \mathbb{N}$ which we call the *ambient dimension*.
- Actually the dimension might be much smaller; eg.,

$$X_i = \varphi(\tilde{X}_i) + \sigma\, \epsilon_i, \quad \sigma \geq 0,$$

  where $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^D$, is some smooth embedding.
- The number $d \leq D$ is the *intrinsic dimension* of the dataset.
- I will talk about the estimation of the intrinsic dimension d.

What is meant by dimension

Our setup is the following:

- There is data $X_1, \dots, X_n$, where $X_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}^D$, for some $D \in \mathbb{N}$ which we call the *ambient dimension*.
- Actually the dimension might be much smaller; eg.,

$$X_i = \varphi(\tilde{X}_i) + \sigma\,\epsilon_i, \quad \sigma \geq 0,$$

  where $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^D$, is some smooth embedding.
- The number $d \leq D$ is the *intrinsic dimension* of the dataset.
- I will talk about the estimation of the intrinsic dimension d.

What is meant by dimension

Our setup is the following:

- There is data $X_1, \ldots, X_n$, where $X_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}^D$, for some $D \in \mathbb{N}$ which we call the *ambient dimension*.

- Actually the dimension might be much smaller; eg.,

$$X_i = \varphi(\tilde{X}_i) + \sigma\, \epsilon_i, \quad \sigma \geq 0,$$

  where $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^D$, is some smooth embedding.

- The number $d \leq D$ is the *intrinsic dimension* of the dataset.

- I will talk about the estimation of the intrinsic dimension d.

What is meant by dimension

Our setup is the following:

- There is data $X_1, \ldots, X_n$, where $X_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}^D$, for some $D \in \mathbb{N}$ which we call the *ambient dimension*.
- Actually the dimension might be much smaller; eg.,

$$X_i = \varphi(\tilde{X}_i) + \sigma\, \epsilon_i, \quad \sigma \geq 0,$$

  where $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^D$, is some smooth embedding.
- The number $d \leq D$ is the *intrinsic dimension* of the dataset.
- I will talk about the estimation of the intrinsic dimension d.

What is meant by dimension

Our setup is the following:

- There is data $X_1, \ldots, X_n$, where $X_i \overset{i.i.d.}{\sim} F$ on $\mathbb{R}^D$, for some $D \in \mathbb{N}$ which we call the *ambient dimension*.

- Actually the dimension might be much smaller; eg.,

$$X_i = \varphi(\tilde{X}_i) + \sigma\, \epsilon_i, \quad \sigma \geq 0,$$

  where $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^D$, is some smooth embedding.

- The number $d \leq D$ is the *intrinsic dimension* of the dataset.

- I will talk about the estimation of the intrinsic dimension d.

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK$^+$06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK+06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK+06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK^{+}06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK^{+}06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK$^+$06]

Why estimate the intrinsic dimension?

There are plenty of reasons to do this:

- Dimensionality reduction[1] (eg., PCA, SOM, MDS, ISOMAP, LLE, Hessian and Laplacian eigenmaps, LLP);
- Independent component analysis ([HKO01]);
- Adaptation;
- Avoid *curse of dimensionality* (if possible);
- Compressibility;
- Speed of algorithms;

[1] [Koh90, CC00, TDSL00, RS00, DG03, HC02, GK^{+}06]

Previous work

Main approaches

- Multidimensional scaling; [She62a, She62b, Kru64a, Kru64b, Ben69]
- Testing approach; [Tru68]
- Karhunen–Loève expansions; [FO71, Fuk82]
- AIC, BIC; [Aka74, Sch78]
- Correlation integral based; [CV02, Kég02, GP04, HA05, SRHI10]
- Clustering approaches; [EC12]
- Based on graphs; [CH04, FSA07, LPS+08]
- KNN; [LB04, KvL15]

Previous work

Limitations

- They require extensive knowledge about distances or similarities between observations, sometimes perturbations thereof, and about F;
- Sometimes only limited information is available;
- Computationally heavy, typically at least $\mathcal{O}(Dn^2)$;
- No results on consistency or rates;
- The scale at which we look at the data affects the dimension (not always noted in the literature);

Previous work

Limitations

- They require extensive knowledge about distances or similarities between observations, sometimes perturbations thereof, and about F;
- Sometimes only limited information is available;
- Computationally heavy, typically at least $\mathcal{O}(Dn^2)$;
- No results on consistency or rates;
- The scale at which we look at the data affects the dimension (not always noted in the literature);

Previous work

Limitations

- They require extensive knowledge about distances or similarities between observations, sometimes perturbations thereof, and about F;
- Sometimes only limited information is available;
- Computationally heavy, typically at least $\mathcal{O}(Dn^2)$;
- No results on consistency or rates;
- The scale at which we look at the data affects the dimension (not always noted in the literature);

Previous work

Limitations

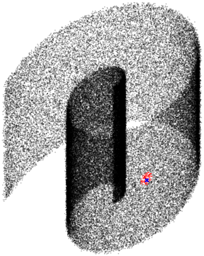
- They require extensive knowledge about distances or similarities between observations, sometimes perturbations thereof, and about F;
- Sometimes only limited information is available;
- Computationally heavy, typically at least $\mathcal{O}(Dn^2)$;
- No results on consistency or rates;
- The scale at which we look at the data affects the dimension (not always noted in the literature);

Previous work

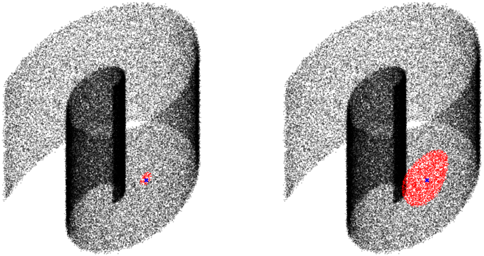
Limitations

- They require extensive knowledge about distances or similarities between observations, sometimes perturbations thereof, and about F;
- Sometimes only limited information is available;
- Computationally heavy, typically at least $\mathcal{O}(Dn^2)$;
- No results on consistency or rates;
- The scale at which we look at the data affects the dimension (not always noted in the literature);

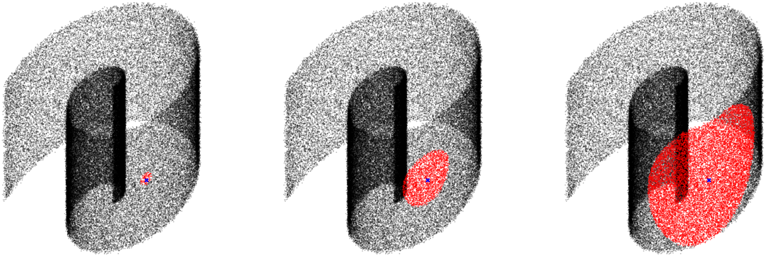
Example of scale dependent dimension



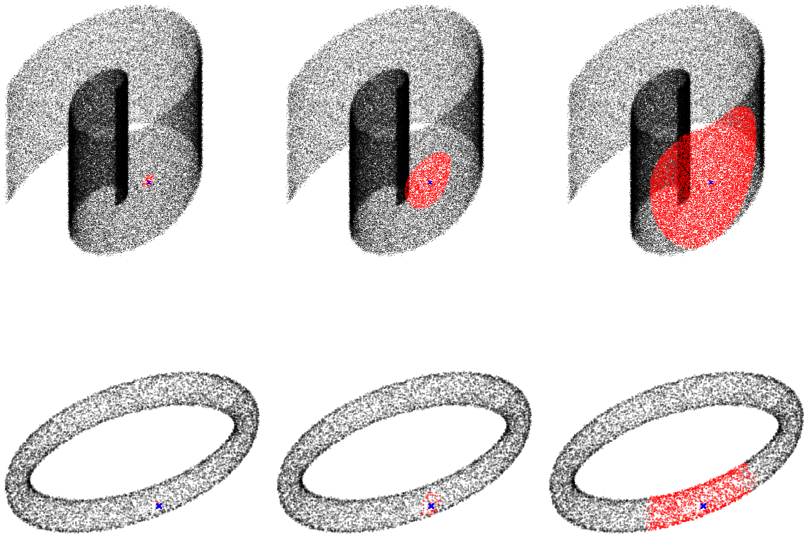
Example of scale dependent dimension



Example of scale dependent dimension



Example of scale dependent dimension



Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = \boldsymbol{A}_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = A_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = A_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = \boldsymbol{A}_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = \boldsymbol{A}_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = \boldsymbol{A}_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Sampling

- We only assume that we can observe adjacency matrices $\boldsymbol{\mathcal{A}}$.
- Each $\mathcal{A}_{i,j} = 1$ iif X_i and X_j are "close".
- We model $\boldsymbol{\mathcal{A}}$ (or the corresponding graph) as a random connection model:
- For some metric r and some number ϵ we assume that $\boldsymbol{\mathcal{A}} = \boldsymbol{A}_\epsilon$, where $A_{i,j} = 1_{\{r(X_i,X_j)\leq\epsilon\}}$, $i < j$, completed by symmetry, no self-loops.
- This is a model from continuum percolation.
- r and ϵ may be unknown.
- The parameter ϵ represents the scale at which we look at the data.

Modelling the data

Some definitions

- We actually work with $\boldsymbol{B} = \boldsymbol{B}_\epsilon = \boldsymbol{A}_\epsilon^2$:

$$B_i \triangleq B_{i,i} = \sum_{j=1}^{n} A_{i,j}, \quad \text{and} \quad B_{i,j} = \sum_{k=1}^{n} A_{i,k} A_{k,j}, \quad i,j = 1,\ldots,n,\, i \neq j,$$

- Define the functions $p(x)$ and $p(x, y)$,

$$p(x) = \mathbb{P}\{r(X,x) \le \epsilon\}, \quad \text{and} \quad p(x,y) = \mathbb{P}\{r(X,x) \le \epsilon, r(X,y) \le \epsilon\},$$

- The B_i are equally distributed, not independent; same holds for the $B_{i,j}$:

$$B_i | X_i \sim \mathrm{Bin}\{n-1,\, p(X_i)\}, \quad \text{and} \quad B_{i,j} | (X_i, X_j) \sim \mathrm{Bin}\{n-2,\, p(X_i, X_j)\}.$$

Modelling the data

Some definitions

- We actually work with $\boldsymbol{B} = \boldsymbol{B}_\epsilon = \boldsymbol{A}_\epsilon^2$:

$$B_i \triangleq B_{i,i} = \sum_{j=1}^{n} A_{i,j}, \quad \text{and} \quad B_{i,j} = \sum_{k=1}^{n} A_{i,k} A_{k,j}, \quad i, j = 1, \ldots, n,\ i \neq j,$$

- Define the functions $p(x)$ and $p(x, y)$,

$$p(x) = \mathbb{P}\{r(X, x) \leq \epsilon\}, \quad \text{and} \quad p(x, y) = \mathbb{P}\{r(X, x) \leq \epsilon, r(X, y) \leq \epsilon\},$$

- The B_i are equally distributed, not independent; same holds for the $B_{i,j}$:

$$B_i | X_i \sim \text{Bin}\{n-1,\, p(X_i)\}, \quad \text{and} \quad B_{i,j} | (X_i, X_j) \sim \text{Bin}\{n-2,\, p(X_i, X_j)\}.$$

Modelling the data

Some definitions

- We actually work with $\boldsymbol{B} = \boldsymbol{B}_\epsilon = \boldsymbol{A}_\epsilon^2$:

$$B_i \triangleq B_{i,i} = \sum_{j=1}^{n} A_{i,j}, \quad \text{and} \quad B_{i,j} = \sum_{k=1}^{n} A_{i,k}A_{k,j}, \quad i,j = 1,\dots,n,\ i \neq j,$$

- Define the functions $p(x)$ and $p(x,y)$,

$$p(x) = \mathbb{P}\{r(X,x) \le \epsilon\}, \quad \text{and} \quad p(x,y) = \mathbb{P}\{r(X,x) \le \epsilon, r(X,y) \le \epsilon\},$$

- The B_i are equally distributed, not independent; same holds for the $B_{i,j}$:

$$B_i | X_i \sim \mathsf{Bin}\{n-1,\ p(X_i)\}, \quad \text{and} \quad B_{i,j}|(X_i,X_j) \sim \mathsf{Bin}\{n-2,\ p(X_i,X_j)\}.$$

Modelling the data

Some definitions

- We define for i, j, k mutually different,

$$p_1 = p_1(\epsilon) = \mathbb{E}A_{i,j}, \qquad \text{and} \qquad p_2 = p_2(\epsilon) = \mathbb{E}A_{i,k}A_{k,j}.$$

We see that

- $p_1 = \mathbb{P}\{r(X,Y) \leq \epsilon\} = \mathbb{E}\mathbb{P}\{r(X,Y) \leq \epsilon | X\} = \mathbb{E}p(X)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | X, Y\} = \mathbb{E}p(X,Y)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | Z\} = \mathbb{E}\{p(Z)^2\} \geq \mathbb{E}\{p(Z)\}^2 = p_1^2$.

Modelling the data

Some definitions

- We define for i, j, k mutually different,

$$p_1 = p_1(\epsilon) = \mathbb{E}A_{i,j}, \qquad \text{and} \qquad p_2 = p_2(\epsilon) = \mathbb{E}A_{i,k}A_{k,j}.$$

We see that

- $p_1 = \mathbb{P}\{r(X,Y) \leq \epsilon\} = \mathbb{E}\mathbb{P}\{r(X,Y) \leq \epsilon | X\} = \mathbb{E}p(X)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | X, Y\} = \mathbb{E}p(X,Y)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | Z\} = \mathbb{E}\{p(Z)^2\} \geq \mathbb{E}\{p(Z)\}^2 = p_1^2$.

Modelling the data

Some definitions

- We define for i, j, k mutually different,

$$p_1 = p_1(\epsilon) = \mathbb{E}A_{i,j}, \qquad \text{and} \qquad p_2 = p_2(\epsilon) = \mathbb{E}A_{i,k}A_{k,j}.$$

We see that

- $p_1 = \mathbb{P}\{r(X,Y) \le \epsilon\} = \mathbb{E}\mathbb{P}\{r(X,Y) \le \epsilon | X\} = \mathbb{E}p(X)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \le \epsilon, r(Z,Y) \le \epsilon | X, Y\} = \mathbb{E}p(X,Y)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \le \epsilon, r(Z,Y) \le \epsilon | Z\} = \mathbb{E}\{p(Z)^2\} \ge \mathbb{E}\{p(Z)\}^2 = p_1^2$.

Modelling the data

Some definitions

- We define for i, j, k mutually different,

$$p_1 = p_1(\epsilon) = \mathbb{E}A_{i,j}, \qquad \text{and} \qquad p_2 = p_2(\epsilon) = \mathbb{E}A_{i,k}A_{k,j}.$$

We see that

- $p_1 = \mathbb{P}\{r(X,Y) \leq \epsilon\} = \mathbb{E}\mathbb{P}\{r(X,Y) \leq \epsilon | X\} = \mathbb{E}p(X)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | X,Y\} = \mathbb{E}p(X,Y)$,
- $p_2 = \mathbb{E}\mathbb{P}\{r(X,Z) \leq \epsilon, r(Z,Y) \leq \epsilon | Z\} = \mathbb{E}\{p(Z)^2\} \geq \mathbb{E}\{p(Z)\}^2 = p_1^2$.

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- Consider, for $x \in \mathcal{X} \subseteq \mathbb{R}^D$, the ball $V(x, \epsilon, D) = \{y \in \mathbb{R}^D : r(x, y) \leq \epsilon\}$, and denote $V(\epsilon, D) = V_\epsilon(0, \epsilon, D)$.

- If ϵ is small (or if $\epsilon \to 0$) and if F admits a continuous density f with respect to the Lebesgue measure μ

$$p(x) = \int_{\mathcal{X}} 1_{V(x,\epsilon,D)}(y)\, f(y)\, d\mu(y) \approx f(x) \int_{\mathcal{X}} 1_{V(\epsilon,D)}(y)\, d\mu(y) = f(x)\, v_\epsilon,$$

where

$$v_\epsilon = \int_{V(\epsilon,D)\cap\mathcal{X}} d\mu \approx \int_{V(\epsilon,d)} d\mu = \mu\{V(\epsilon, d)\}.$$

Defining the estimators

Intuition behind the estimators

- So $p(x)$ should depend on d (and x, and ϵ) but not D.

- Since $p_1 = \mathbb{E}\{p(X,\epsilon)\}$ and $p_2 = \mathbb{E}\{p(X,\epsilon)^2\}$, we can approximate

$$p_1 \approx \mathbb{E}f(X)\,\mu\{V(\epsilon,d)\} \quad \text{and} \quad p_2 \approx \mathbb{E}\{f(X)^2\}\,\mu\{V(\epsilon,d)\}^2.$$

- Using estimators for p_1 or p_2 we could invert this to get estimates for d.

- Instead we can get rid of the constants by considering

$$\frac{p_1(2\epsilon)}{p_1(\epsilon)} \approx \frac{\mu\{V(2\epsilon,d)\}}{\mu\{V(\epsilon,d)\}}, \text{ and } \frac{p_2(2\epsilon)}{p_2(\epsilon)} \approx \frac{\mu\{V(2\epsilon,d)\}^2}{\mu\{V(\epsilon,d)\}^2}.$$

Defining the estimators

Intuition behind the estimators

- So $p(x)$ should depend on d (and x, and ϵ) but not D.

- Since $p_1 = \mathbb{E}\{p(X,\epsilon)\}$ and $p_2 = \mathbb{E}\{p(X,\epsilon)^2\}$, we can approximate

$$p_1 \approx \mathbb{E} f(X)\, \mu\{V(\epsilon, d)\} \quad \text{and} \quad p_2 \approx \mathbb{E}\{f(X)^2\}\, \mu\{V(\epsilon, d)\}^2.$$

- Using estimators for p_1 or p_2 we could invert this to get estimates for d.

- Instead we can get rid of the constants by considering

$$\frac{p_1(2\epsilon)}{p_1(\epsilon)} \approx \frac{\mu\{V(2\epsilon, d)\}}{\mu\{V(\epsilon, d)\}}, \text{ and } \frac{p_2(2\epsilon)}{p_2(\epsilon)} \approx \frac{\mu\{V(2\epsilon, d)\}^2}{\mu\{V(\epsilon, d)\}^2}.$$

Defining the estimators

Intuition behind the estimators

- So $p(x)$ should depend on d (and x, and ϵ) but not D.

- Since $p_1 = \mathbb{E}\{p(X,\epsilon)\}$ and $p_2 = \mathbb{E}\{p(X,\epsilon)^2\}$, we can approximate

$$p_1 \approx \mathbb{E} f(X)\, \mu\{V(\epsilon, d)\} \quad \text{and} \quad p_2 \approx \mathbb{E}\{f(X)^2\}\, \mu\{V(\epsilon, d)\}^2.$$

- Using estimators for p_1 or p_2 we could invert this to get estimates for d.

- Instead we can get rid of the constants by considering

$$\frac{p_1(2\epsilon)}{p_1(\epsilon)} \approx \frac{\mu\{V(2\epsilon, d)\}}{\mu\{V(\epsilon, d)\}}, \text{ and } \frac{p_2(2\epsilon)}{p_2(\epsilon)} \approx \frac{\mu\{V(2\epsilon, d)\}^2}{\mu\{V(\epsilon, d)\}^2}.$$

Defining the estimators

Intuition behind the estimators

- So $p(x)$ should depend on d (and x, and ϵ) but not D.

- Since $p_1 = \mathbb{E}\{p(X,\epsilon)\}$ and $p_2 = \mathbb{E}\{p(X,\epsilon)^2\}$, we can approximate

$$p_1 \approx \mathbb{E} f(X)\, \mu\{V(\epsilon,d)\} \quad \text{and} \quad p_2 \approx \mathbb{E}\{f(X)^2\}\, \mu\{V(\epsilon,d)\}^2.$$

- Using estimators for p_1 or p_2 we could invert this to get estimates for d.

- Instead we can get rid of the constants by considering

$$\frac{p_1(2\epsilon)}{p_1(\epsilon)} \approx \frac{\mu\{V(2\epsilon,d)\}}{\mu\{V(\epsilon,d)\}}, \text{ and } \frac{p_2(2\epsilon)}{p_2(\epsilon)} \approx \frac{\mu\{V(2\epsilon,d)\}^2}{\mu\{V(\epsilon,d)\}^2}.$$

Defining the estimators

Intuition behind the estimators

- If $\hat{p}_1(\epsilon)$ and $\hat{p}_2(\epsilon)$ are estimators for $p_1(\epsilon)$ and $p_2(\epsilon)$, respectively, then we implicitly define $\hat{d}_1$, $\hat{d}_2$ as any solutions to

$$\frac{\hat{p}_1(2\epsilon)}{\hat{p}_1(\epsilon)} = g(\epsilon, \hat{d}_1) \qquad \text{and} \qquad \frac{\hat{p}_2(2\epsilon)}{\hat{p}_2(\epsilon)} = g(\epsilon, \hat{d}_2)^2,$$

- We should expect in general that $g(\epsilon, d) \approx g(d) = 2^d$, and so d:

$$\hat{d}_1 = \frac{\log \hat{p}_1(2\epsilon) - \log \hat{p}_1(\epsilon)}{\log 2}, \quad \text{and} \quad \hat{d}_2 = \frac{\log \hat{p}_2(2\epsilon) - \log \hat{p}_2(\epsilon)}{\log 4}.$$

- If d is an integer, then define also $\tilde{d}_1 = [\hat{d}_1]$ and $\tilde{d}_2 = [\hat{d}_2]$.

Defining the estimators

Intuition behind the estimators

- If $\hat{p}_1(\epsilon)$ and $\hat{p}_2(\epsilon)$ are estimators for $p_1(\epsilon)$ and $p_2(\epsilon)$, respectively, then we implicitly define $\hat{d}_1$, $\hat{d}_2$ as any solutions to

$$\frac{\hat{p}_1(2\epsilon)}{\hat{p}_1(\epsilon)} = g(\epsilon, \hat{d}_1) \qquad \text{and} \qquad \frac{\hat{p}_2(2\epsilon)}{\hat{p}_2(\epsilon)} = g(\epsilon, \hat{d}_2)^2,$$

- We should expect in general that $g(\epsilon, d) \approx g(d) = 2^d$, and so d:

$$\hat{d}_1 = \frac{\log \hat{p}_1(2\epsilon) - \log \hat{p}_1(\epsilon)}{\log 2}, \quad \text{and} \quad \hat{d}_2 = \frac{\log \hat{p}_2(2\epsilon) - \log \hat{p}_2(\epsilon)}{\log 4}.$$

- If d is an integer, then define also $\tilde{d}_1 = [\hat{d}_1]$ and $\tilde{d}_2 = [\hat{d}_2]$.

Defining the estimators

Intuition behind the estimators

- If $\hat{p}_1(\epsilon)$ and $\hat{p}_2(\epsilon)$ are estimators for $p_1(\epsilon)$ and $p_2(\epsilon)$, respectively, then we implicitly define $\hat{d}_1$, $\hat{d}_2$ as any solutions to

$$\frac{\hat{p}_1(2\epsilon)}{\hat{p}_1(\epsilon)} = g(\epsilon, \hat{d}_1) \qquad \text{and} \qquad \frac{\hat{p}_2(2\epsilon)}{\hat{p}_2(\epsilon)} = g(\epsilon, \hat{d}_2)^2,$$

- We should expect in general that $g(\epsilon, d) \approx g(d) = 2^d$, and so d:

$$\hat{d}_1 = \frac{\log \hat{p}_1(2\epsilon) - \log \hat{p}_1(\epsilon)}{\log 2}, \quad \text{and} \quad \hat{d}_2 = \frac{\log \hat{p}_2(2\epsilon) - \log \hat{p}_2(\epsilon)}{\log 4}.$$

- If d is an integer, then define also $\tilde{d}_1 = [\hat{d}_1]$ and $\tilde{d}_2 = [\hat{d}_2]$.

Estimates of p_1 and p_2

Definition and relation to correlation integral

- The obvious estimators for p_1 and p_2 are

$$\hat{p}_1 = \frac{1}{m_n}\sum_{i=1}^{m_n}\frac{B_i}{n-1}, \quad \text{and} \quad \hat{p}_2 = \frac{2}{m_n(m_n-1)}\sum_{i=1}^{m_n-1}\sum_{j=i+1}^{m_n}\frac{B_{i,j}}{n-2}.$$

- Since $\mathbb{E}B_i/(n-1) = p_1$, and $\mathbb{E}B_{i,j}/(n-2) = p_2$, $\hat{p}_1$ and $\hat{p}_2$ are unbiased.

- As a function of ϵ, if $r(x,y) = \|x-y\|_2$, $\hat{p}_1$ is called the correlation integral[2]

$$C(\epsilon) = \lim_{n\to\infty}\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} 1_{\{\|x_i-x_j\|_2\le\epsilon\}}.$$

- The limit as $\epsilon \to 0$ of $-\log\{C(\epsilon)\}/\log(\epsilon)$ is called correlation dimension.

[2] [CV02]

Estimates of p_1 and p_2

Definition and relation to correlation integral

- The obvious estimators for p_1 and p_2 are

$$\hat{p}_1 = \frac{1}{m_n} \sum_{i=1}^{m_n} \frac{B_i}{n-1}, \quad \text{and} \quad \hat{p}_2 = \frac{2}{m_n(m_n-1)} \sum_{i=1}^{m_n-1} \sum_{j=i+1}^{m_n} \frac{B_{i,j}}{n-2}.$$

- Since $\mathbb{E}B_i/(n-1) = p_1$, and $\mathbb{E}B_{i,j}/(n-2) = p_2$, $\hat{p}_1$ and $\hat{p}_2$ are unbiased.

- As a function of ϵ, if $r(x,y) = \|x-y\|_2$, $\hat{p}_1$ is called the correlation integral[2]

$$C(\epsilon) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1_{\{\|x_i - x_j\|_2 \le \epsilon\}}.$$

- The limit as $\epsilon \to 0$ of $-\log\{C(\epsilon)\}/\log(\epsilon)$ is called correlation dimension.

[2] [CV02]

Estimates of p_1 and p_2

Definition and relation to correlation integral

- The obvious estimators for p_1 and p_2 are

$$\hat{p}_1 = \frac{1}{m_n}\sum_{i=1}^{m_n}\frac{B_i}{n-1}, \quad \text{and} \quad \hat{p}_2 = \frac{2}{m_n(m_n-1)}\sum_{i=1}^{m_n-1}\sum_{j=i+1}^{m_n}\frac{B_{i,j}}{n-2}.$$

- Since $\mathbb{E}B_i/(n-1) = p_1$, and $\mathbb{E}B_{i,j}/(n-2) = p_2$, $\hat{p}_1$ and $\hat{p}_2$ are unbiased.

- As a function of ϵ, if $r(x,y) = \|x-y\|_2$, $\hat{p}_1$ is called the correlation integral[2]

$$C(\epsilon) = \lim_{n\to\infty}\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}1_{\{\|x_i-x_j\|_2\le\epsilon\}}.$$

- The limit as $\epsilon \to 0$ of $-\log\{C(\epsilon)\}/\log(\epsilon)$ is called correlation dimension.

[2] [CV02]

Estimates of p_1 and p_2

Definition and relation to correlation integral

- The obvious estimators for p_1 and p_2 are

$$\hat{p}_1 = \frac{1}{m_n}\sum_{i=1}^{m_n}\frac{B_i}{n-1}, \quad \text{and} \quad \hat{p}_2 = \frac{2}{m_n(m_n-1)}\sum_{i=1}^{m_n-1}\sum_{j=i+1}^{m_n}\frac{B_{i,j}}{n-2}.$$

- Since $\mathbb{E}B_i/(n-1) = p_1$, and $\mathbb{E}B_{i,j}/(n-2) = p_2$, $\hat{p}_1$ and $\hat{p}_2$ are unbiased.

- As a function of ϵ, if $r(x,y) = \|x-y\|_2$, $\hat{p}_1$ is called the correlation integral[2]

$$C(\epsilon) = \lim_{n\to\infty}\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}1_{\{\|x_i-x_j\|_2\le\epsilon\}}.$$

- The limit as $\epsilon \to 0$ of $-\log\{C(\epsilon)\}/\log(\epsilon)$ is called correlation dimension.

[2] [CV02]

Estimates of p_1 and p_2

Asymptotics

Theorem

Let $m_n \leq n$ such that $m_n \to \infty$ as $n \to \infty$. If $m_n = o(n)$, and $p_2 > p_1^2$, then

$$S_1^{-1/2}\left(\frac{\hat{p}_1}{p_1} - 1\right) \xrightarrow{d} N(0,1), \qquad \textit{where} \qquad S_1 = \frac{p_2 - p_1^2}{m_n\, p_1^2}.$$

If $m_n = n$ then the previous display also holds if we assume that $n^2 p_1$ is bounded away from 0, $p_2 \lesssim n p_1^2$, $n^2(p_2 - p_1^2) \to \infty$, $p_{s,3} - p_1 p_2 \lesssim n(p_2 - p_1^2)^2$, and $p_{s,4} - p_1^4 \lesssim (p_2 - p_1^2)^2$.

Theorem

Assume that p_2 is such that as $n \to \infty$, $n^3 p_2$ is bounded away from zero, and that $p_{s,3} + p_{l,3} \lesssim n^3 p_2^2$. Then,

$$S_2^{-1/2}\left(\frac{\hat{p}_2}{p_2} - 1\right) = O_p(1), \qquad \textit{where} \qquad S_2 = \frac{p_{s,4} + 4p_{l,4} + 4p_{0,2} - p_2^2}{n p_2^2}.$$

Estimates of p_1 and p_2

Asymptotics

Theorem

Let $m_n \le n$ such that $m_n \to \infty$ as $n \to \infty$. If $m_n = o(n)$, and $p_2 > p_1^2$, then

$$S_1^{-1/2}\left(\frac{\hat{p}_1}{p_1} - 1\right) \xrightarrow{d} N(0,1), \qquad \text{where} \qquad S_1 = \frac{p_2 - p_1^2}{m_n\, p_1^2}.$$

If $m_n = n$ then the previous display also holds if we assume that $n^2 p_1$ is bounded away from 0, $p_2 \lesssim n p_1^2$, $n^2(p_2 - p_1^2) \to \infty$, $p_{s,3} - p_1 p_2 \lesssim n(p_2 - p_1^2)^2$, and $p_{s,4} - p_1^4 \lesssim (p_2 - p_1^2)^2$.

Theorem

Assume that p_2 is such that as $n \to \infty$, $n^3 p_2$ is bounded away from zero, and that $p_{s,3} + p_{l,3} \lesssim n^3 p_2^2$. Then,

$$S_2^{-1/2}\left(\frac{\hat{p}_2}{p_2} - 1\right) = O_p(1), \qquad \text{where} \qquad S_2 = \frac{p_{s,4} + 4p_{l,4} + 4p_{0,2} - p_2^2}{n p_2^2}.$$

Consistency of the estimators

Asymptotics of $\hat{d}_1$: implicit estimator

Theorem

Assume that the conditions required for the convergence of $\hat{p}_1(\epsilon)$ and $\hat{p}_1(2\epsilon)$ with rate $m_n^{1/2}$ hold. For that ϵ, d, and m_n, assume that, as $n \to \infty$,

$$p_1(2\epsilon) = p_1(\epsilon)\, g\Big(\epsilon,\, d + o(m_n^{-1/2})\Big). \tag{B}$$

Assume that the derivative of $d \mapsto g(\epsilon, d)$ exists, is continuous and non-zero at d. Then, as $n \to \infty$,

$$m_n^{1/2}\left\{\hat{d}_1 - d\right\} \xrightarrow{d} N\left(0,\, \left\{\frac{\partial \log g(\epsilon, d)}{\partial d}\right\}^{-2} V\right).$$

where $V = \frac{p_2(\epsilon) - p_1(\epsilon)^2}{p_1(\epsilon)^2} + \frac{p_2(2\epsilon) - p_1(2\epsilon)^2}{p_1(2\epsilon)^2} - 2\frac{Cov\left\{\hat{p}_1(\epsilon), \hat{p}_1(2\epsilon)\right\}}{p_1(\epsilon)\, p_1(2\epsilon)}$.

Consistency of the estimators

Asymptotics of $\hat{d}_1$: explicit estimator

Theorem

Assume that the conditions required for the convergence of $\hat{p}_1(\epsilon)$ and $\hat{p}_1(2\epsilon)$ with rate $m_n^{1/2}$ hold. For that ϵ, d, and m_n, assume that, as $n \to \infty$,

$$p_1(2\epsilon) = p_1(\epsilon)\, \overbrace{g\Big(\epsilon,\, d + o(m_n^{-1/2})\Big)}^{2^{d+o(m_n^{-1/2})}}. \tag{B}$$

Assume that the derivative of $d \mapsto g(\epsilon, d)$ exists, is continuous and non-zero at d. Then, as $n \to \infty$,

$$m_n^{1/2}\left\{\hat{d}_1 - d\right\} \xrightarrow{d} N\left(0,\, \overbrace{\left\{\frac{\partial \log g(\epsilon, d)}{\partial d}\right\}^{-2}}^{\log(2)^{-2}} V\right).$$

where $V = \frac{p_2(\epsilon) - p_1(\epsilon)^2}{p_1(\epsilon)^2} + \frac{p_2(2\epsilon) - p_1(2\epsilon)^2}{p_1(2\epsilon)^2} - 2\frac{Cov\left\{\hat{p}_1(\epsilon), \hat{p}_1(2\epsilon)\right\}}{p_1(\epsilon)\, p_1(2\epsilon)}$.

Consistency of the estimators

Asymptotics of $\hat{d}_2$

Theorem

Suppose that for some $\delta > 0$ and some $\kappa > 1$ (eventually depending on ϵ),

$$\kappa^2 \, g(\epsilon, d - \delta/2)^2 \leq \frac{p_2(2\epsilon)}{p_2(\epsilon)} \leq \frac{1}{\kappa^2} \, g(\epsilon, d + \delta/2)^2. \tag{I}$$

uniformly in ϵ (or if ϵ is know, for that ϵ). Then

$$\mathbb{P}\Big\{\big|\hat{d}_2 - d\big| < \delta/2\Big\} \geq 1 - \kappa^2 \, \frac{S_2(\epsilon) + S_2(2\epsilon)}{(\kappa - 1)^2}.$$

If d is an integer and we take $\delta = 1$, then we get a lower bound for $\mathbb{P}\big(\tilde{d}_2 = d\big)$.

Consistency of the estimators

Bound for specific design: price of high intrinsic dimension

- For Gaussian design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{2/\sqrt{3-2\epsilon}\right\}^d e^{-\epsilon(1-2\epsilon)} - 1}{m_n}.$$

- For uniform design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{1/(1-2\epsilon)^2\right\}^d - 1}{m_n}.$$

- So in general we need rather large sample size if d is large.

Consistency of the estimators

Bound for specific design: price of high intrinsic dimension

- For Gaussian design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{2/\sqrt{3-2\epsilon}\right\}^d e^{-\epsilon(1-2\epsilon)} - 1}{m_n}.$$

- For uniform design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{1/(1-2\epsilon)^2\right\}^d - 1}{m_n}.$$

- So in general we need rather large sample size if d is large.

Consistency of the estimators

Bound for specific design: price of high intrinsic dimension

- For Gaussian design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{2/\sqrt{3-2\epsilon}\right\}^d e^{-\epsilon(1-2\epsilon)} - 1}{m_n}.$$

- For uniform design we can bound, for appropriately small ϵ,

$$S_1(\epsilon) \leq \frac{\left\{1/(1-2\epsilon)^2\right\}^d - 1}{m_n}.$$

- So in general we need rather large sample size if d is large.

Numerical results

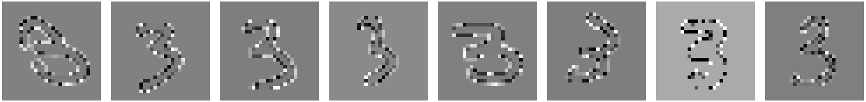
Comparison with other estimators: the real data

We compared our estimators with some competing estimators with some simulated- and real data. The real data:

- 'Isomap faces' dataset

- 'Hands' dataset

- 'MNIST' dataset



Numerical results

Comparison with other estimators: the results

	n	d	D	Dataset	$\hat{d}$	E_{CAP}	MLE	CorrDim	RegDim
1	1000	1	3	Unif. on Helix	0.99	1.00	1.00	1.00	0.99
2	1000	2	3	Swiss roll	1.94	2.14	1.94	1.99	1.87
3	1000	5	5	Gaussian	5.06	5.33	5.00	4.91	4.86
4	1000	7	8	Unif. on $\mathbb{S}^7$	6.81	5.88	6.53	6.85	6.23
5	5000	7	8	Unif. on $\mathbb{S}^7$	6.88	6.85	6.72	6.95	6.46
6	1000	12	12	U$\{[0,1]^{12}\}$	9.45	7.74	9.32	10.66	8.78
7	5000	12	12	U$\{[0,1]^{12}\}$	10.08	9.24	9.76	10.83	9.26
8	698	–	64×64	Isomap faces	3.99	3.04	3.99	3.53	4.22
9	481	–	512×480	Hands	2.75	1.27	2.88	3.92	2.56
10	7141	–	28×28	MNIST "3"	14.98	8.92	15.95	14.17	14.75
11	6824	–	28×28	MNIST "4"	13.68	8.13	14.44	9.54	13.16
12	6313	–	28×28	MNIST "5"	15.94	8.40	15.55	18.00	14.28

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Recap / Conclusions

- Our approach combines the notion of correlation integral with the doubling property of the Lebesgue measure.
- This gives us (essentially) parameter free estimators of intrinsic dimension.
- We can estimate scale dependent intrinsic dimensions.
- We give assumptions under which we derive a bound on the probability of recuperating the true dimension.
- The simulations show that the estimators compare well with competing estimators for different types of real- and simulated data.
- In particular, the estimators do well without using distance data.
- For large (intrinsic) dimension, we need large sample sizes to get accuracy.
- The dimension is often underestimated (based on the simulations).

Thanks for listening.

References I

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[Ben69] Robert S Bennett. The intrinsic dimensionality of signal collections. *Information Theory, IEEE Transactions on*, 15(5):517–525, 1969.

[CC00] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC press, 2000.

[CH04] Jose A Costa and Alfred O Hero. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In *Signal Processing Conference, 2004 12th European*, pages 369–372. IEEE, 2004.

[CV02] Francesco Camastra and Alessandro Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(10):1404–1407, 2002.

[DG03] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[EC12] Brian Eriksson and Mark Crovella. Estimating intrinsic dimension via clustering. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 760–763. IEEE, 2012.

[FO71] Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. *Computers, IEEE Transactions on*, 100(2):176–183, 1971.

[FSA07] Amir M Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th International Conference on Machine Learning (ICML-07)*, pages 265–272, 2007.

[Fuk82] Keinosuke Fukunaga. Intrinsic dimensionality extraction. *Handbook of Statistics*, 2:347–360, 1982.

[GK+06] Evarist Giné, Vladimir Koltchinskii, et al. Empirical graph laplacian approximation of laplace–beltrami operators: Large sample results. In *High dimensional probability*, pages 238–259. Institute of Mathematical Statistics, 2006.

[GP04] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, 2004.

[HA05] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in r d. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296. ACM, 2005.

[HC02] Xiaoming Huo and Jihong Chen. Local linear projection (llp). In *Proc. of First Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, 2002.

[HKO01] Aapo Hyviirinen, Juha Karhunen, and Erki Oja. Independent component analysis. *Wileyand Sons*, 2001.

[Kég02] Balázs Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems*, pages 681–688, 2002.

[Koh90] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[Kru64a] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

References II

[Kru64b] Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[KvL15] Matthäus Kleindessner and Ulrike von Luxburg. Dimensionality estimation without distances. In *AISTATS*, 2015.

[LB04] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2004.

[LPS+08] Nikolai Leonenko, Luc Pronzato, Vippal Savani, et al. A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, 2008.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[She62a] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

[She62b] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika*, 27(3):219–246, 1962.

[SRHI10] Kumar Sricharan, Raviv Raich, and Alfred O Hero III. Optimized intrinsic dimension estimator using nearest neighbor graphs. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5418–5421. IEEE, 2010.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[Tru68] Gerard V Trunk. Statistical estimation of the intrinsic dimensionality of data collections. *Information and Control*, 12(5):508–525, 1968.

Moments of $\hat{p}_1$ and $\hat{p}_2$

- The variance of $\hat{p}_1$ can be expressed in terms of polynomials in n and $\mathbb{E}B_i$, $\mathbb{E}B_i^2$, and $\mathbb{E}B_iB_j$.
- The variance of $\hat{p}_2$ can be expressed in terms of polynomials in n and $\mathbb{E}B_{i,j}$, $\mathbb{E}B_{i,j}^2$, $\mathbb{E}B_{i,j}B_{i,k}$, and $\mathbb{E}B_{i,j}B_{k,l}$.
- Some of these have general formulas

$$\mathbb{E}B_i^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{i,j_1}\cdots A_{i,j_k}}^{p_{s,k}},$$

$$\mathbb{E}B_{i,j}^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{1,k+1}A_{1,k+2}\cdots A_{k,k+1}A_{k,k+2}}^{p_{q,k}}.$$

- In general it is a lot of work to count the graphs.

Moments of $\hat{p}_1$ and $\hat{p}_2$

- The variance of $\hat{p}_1$ can be expressed in terms of polynomials in n and $\mathbb{E}B_i$, $\mathbb{E}B_i^2$, and $\mathbb{E}B_iB_j$.
- The variance of $\hat{p}_2$ can be expressed in terms of polynomials in n and $\mathbb{E}B_{i,j}$, $\mathbb{E}B_{i,j}^2$, $\mathbb{E}B_{i,j}B_{i,k}$, and $\mathbb{E}B_{i,j}B_{k,l}$.
- Some of these have general formulas

$$\mathbb{E}B_i^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{i,j_1}\cdots A_{i,j_k}}^{p_{s,k}},$$

$$\mathbb{E}B_{i,j}^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{1,k+1}A_{1,k+2}\cdots A_{k,k+1}A_{k,k+2}}^{p_{q,k}}.$$

- In general it is a lot of work to count the graphs.

Moments of $\hat{p}_1$ and $\hat{p}_2$

- The variance of $\hat{p}_1$ can be expressed in terms of polynomials in n and $\mathbb{E}B_i$, $\mathbb{E}B_i^2$, and $\mathbb{E}B_iB_j$.
- The variance of $\hat{p}_2$ can be expressed in terms of polynomials in n and $\mathbb{E}B_{i,j}$, $\mathbb{E}B_{i,j}^2$, $\mathbb{E}B_{i,j}B_{i,k}$, and $\mathbb{E}B_{i,j}B_{k,l}$.
- Some of these have general formulas

$$\mathbb{E}B_i^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{i,j_1}\cdots A_{i,j_k}}^{p_{s,k}},$$

$$\mathbb{E}B_{i,j}^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k) \overbrace{\mathbb{E}A_{1,k+1}A_{1,k+2}\cdots A_{k,k+1}A_{k,k+2}}^{p_{q,k}}.$$

- In general it is a lot of work to count the graphs.

Moments of $\hat{p}_1$ and $\hat{p}_2$

- The variance of $\hat{p}_1$ can be expressed in terms of polynomials in n and $\mathbb{E}B_i$, $\mathbb{E}B_i^2$, and $\mathbb{E}B_iB_j$.
- The variance of $\hat{p}_2$ can be expressed in terms of polynomials in n and $\mathbb{E}B_{i,j}$, $\mathbb{E}B_{i,j}^2$, $\mathbb{E}B_{i,j}B_{i,k}$, and $\mathbb{E}B_{i,j}B_{k,l}$.
- Some of these have general formulas

$$\mathbb{E}B_i^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k)\, \overbrace{\mathbb{E}A_{i,j_1}\cdots A_{i,j_k}}^{p_{s,k}},$$

$$\mathbb{E}B_{i,j}^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k)\, \overbrace{\mathbb{E}A_{1,k+1}A_{1,k+2}\cdots A_{k,k+1}A_{k,k+2}}^{p_{q,k}}.$$

- In general it is a lot of work to count the graphs.

Moments of $\hat{p}_1$ and $\hat{p}_2$

- The variance of $\hat{p}_1$ can be expressed in terms of polynomials in n and $\mathbb{E}B_i$, $\mathbb{E}B_i^2$, and $\mathbb{E}B_iB_j$.
- The variance of $\hat{p}_2$ can be expressed in terms of polynomials in n and $\mathbb{E}B_{i,j}$, $\mathbb{E}B_{i,j}^2$, $\mathbb{E}B_{i,j}B_{i,k}$, and $\mathbb{E}B_{i,j}B_{k,l}$.
- Some of these have general formulas

$$\mathbb{E}B_i^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k)\, \overbrace{\mathbb{E}A_{i,j_1}\cdots A_{i,j_k}}^{p_{s,k}},$$

$$\mathbb{E}B_{i,j}^r = \sum_{k=1}^{r} \left\{ {r \atop k} \right\} (n-1)\cdots(n-k)\, \overbrace{\mathbb{E}A_{1,k+1}A_{1,k+2}\cdots A_{k,k+1}A_{k,k+2}}^{p_{q,k}}.$$

- In general it is a lot of work to count the graphs.

Moments of $\hat{p}_1$ and $\hat{p}_2$

General moments involving entires of $\boldsymbol{B}$

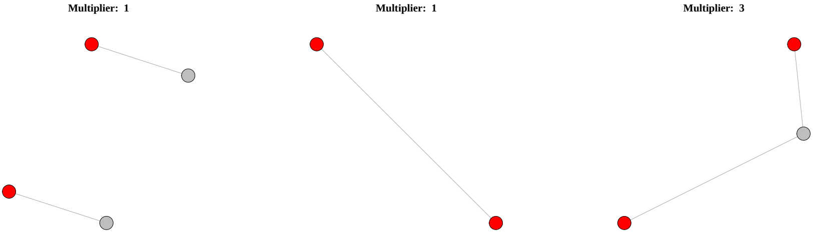
$$\mathbb{E}B_iB_j = \sum_{l_1=1}^{n}\sum_{l_2=1}^{n}\mathbb{E}A_{i,l_1}A_{j,l_2} \qquad \longrightarrow \qquad I = \{(1,3),(2,4)\};\ C = \begin{bmatrix} \cdot & 1 & 1 & 0 \\ \cdot & \cdot & 0 & 1 \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

$$\mathbb{E}B_iB_j = p_1 + 3(n-2)p_2 + (n-2)(n-3)p_1^2.$$

Moments of $\hat{p}_1$ and $\hat{p}_2$

General moments involving entires of $\boldsymbol{B}$

$$\mathbb{E}B_iB_j = \sum_{l_1=1}^{n}\sum_{l_2=1}^{n}\mathbb{E}A_{i,l_1}A_{j,l_2} \qquad \longrightarrow \qquad I = \{(1,3),(2,4)\};\ C = \begin{bmatrix} \cdot & 1 & 1 & 0 \\ \cdot & \cdot & 0 & 1 \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

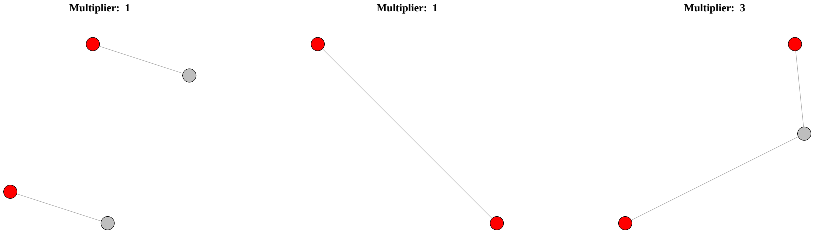


$$\mathbb{E}B_iB_j = p_1 + 3(n-2)p_2 + (n-2)(n-3)p_1^2.$$

Moments of $\hat{p}_1$ and $\hat{p}_2$

General moments involving entires of $\boldsymbol{B}$

$$\mathbb{E}B_iB_j = \sum_{l_1=1}^{n}\sum_{l_2=1}^{n}\mathbb{E}A_{i,l_1}A_{j,l_2} \longrightarrow I = \{(1,3),(2,4)\};\ C = \begin{bmatrix} \cdot & 1 & 1 & 0 \\ \cdot & \cdot & 0 & 1 \\ \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$



$$\mathbb{E}B_iB_j = p_1 + 3(n-2)p_2 + (n-2)(n-3)p_1^2.$$