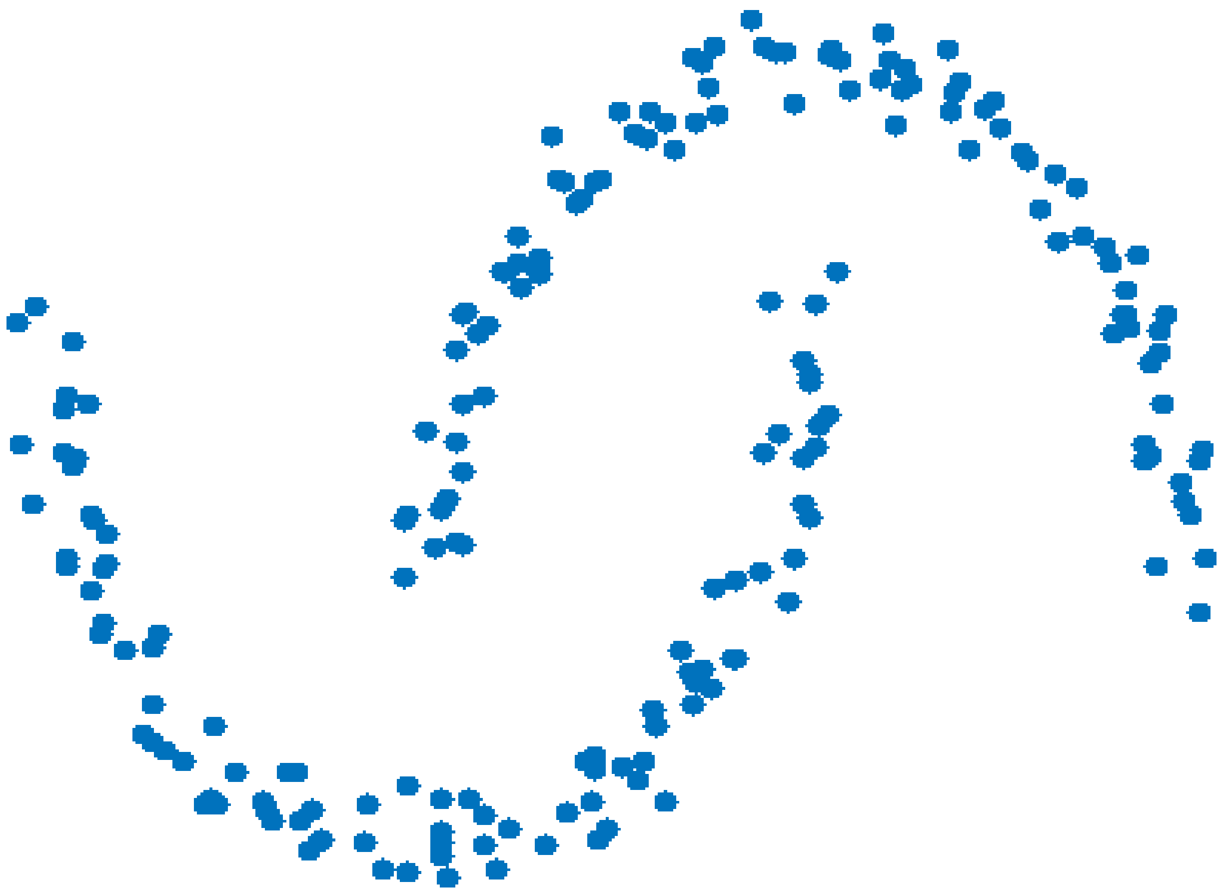
Density estimation from k-nn graphs.

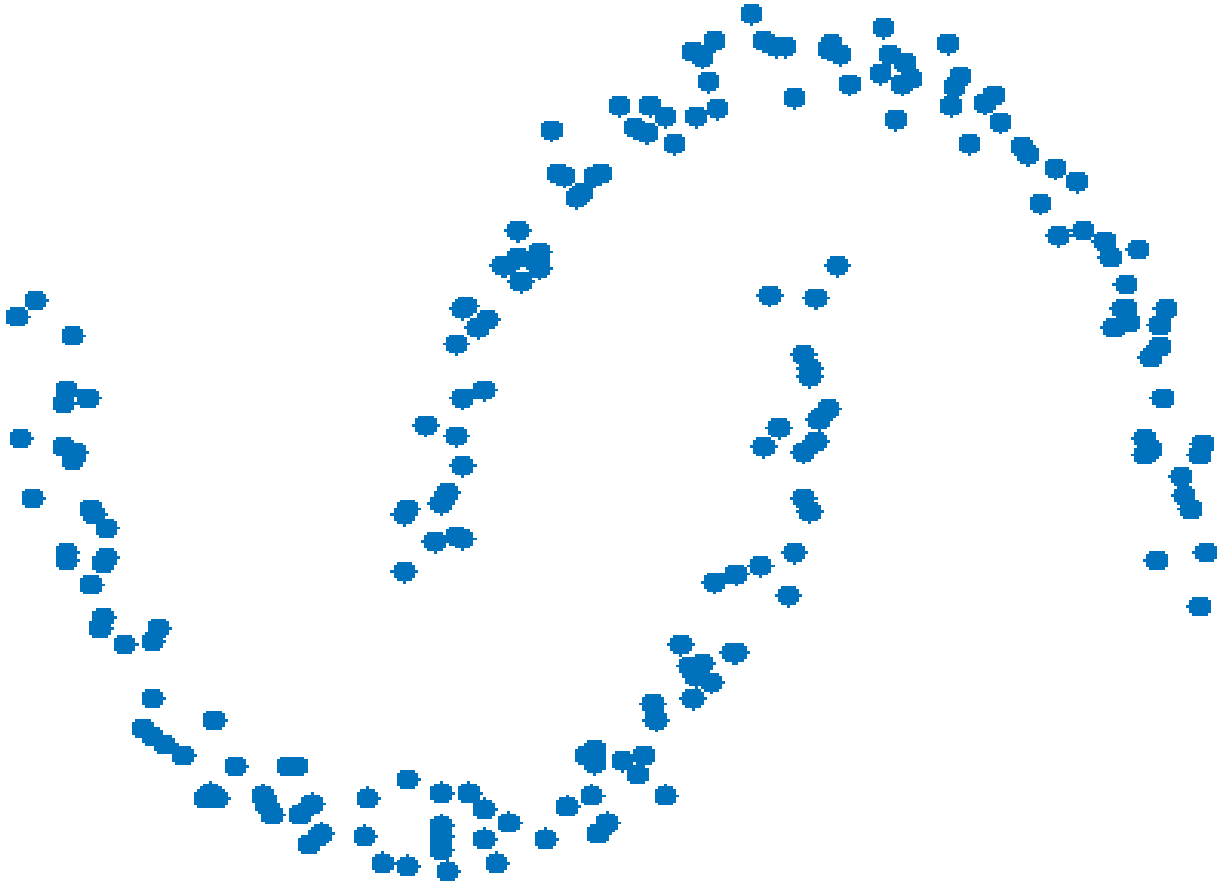
Thomas Bonis

Graphs for non-linear data analysis

Graphs for non-linear data analysis

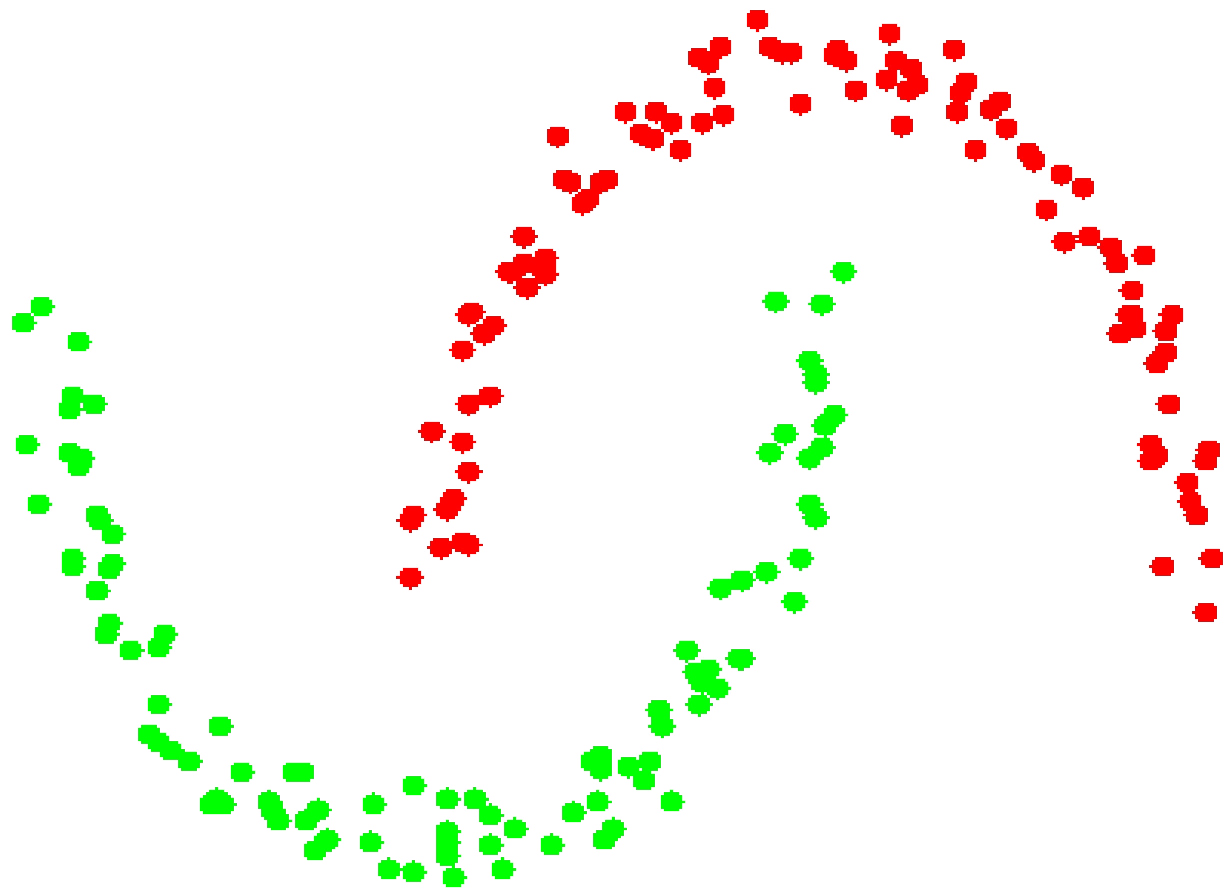


Graphs for non-linear data analysis



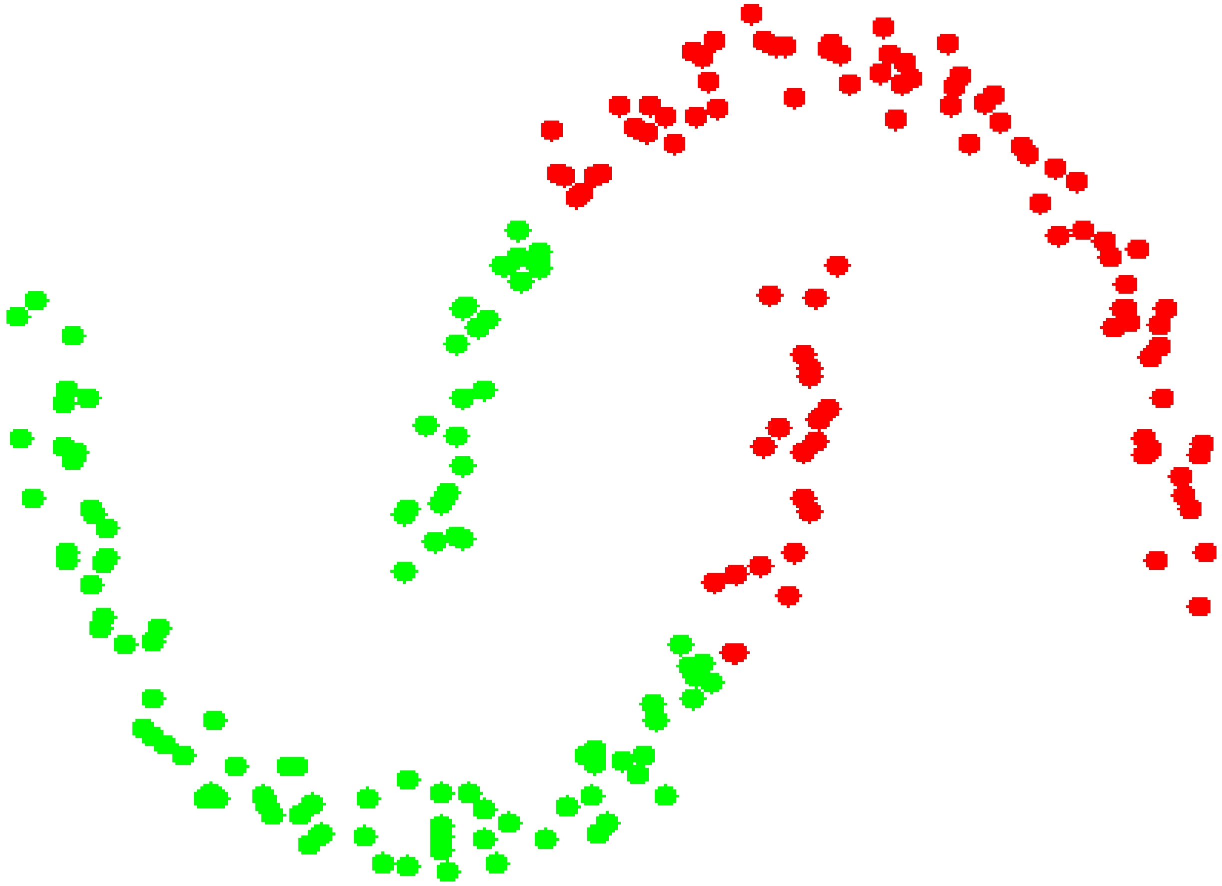
Objective: clustering

Graphs for non-linear data analysis



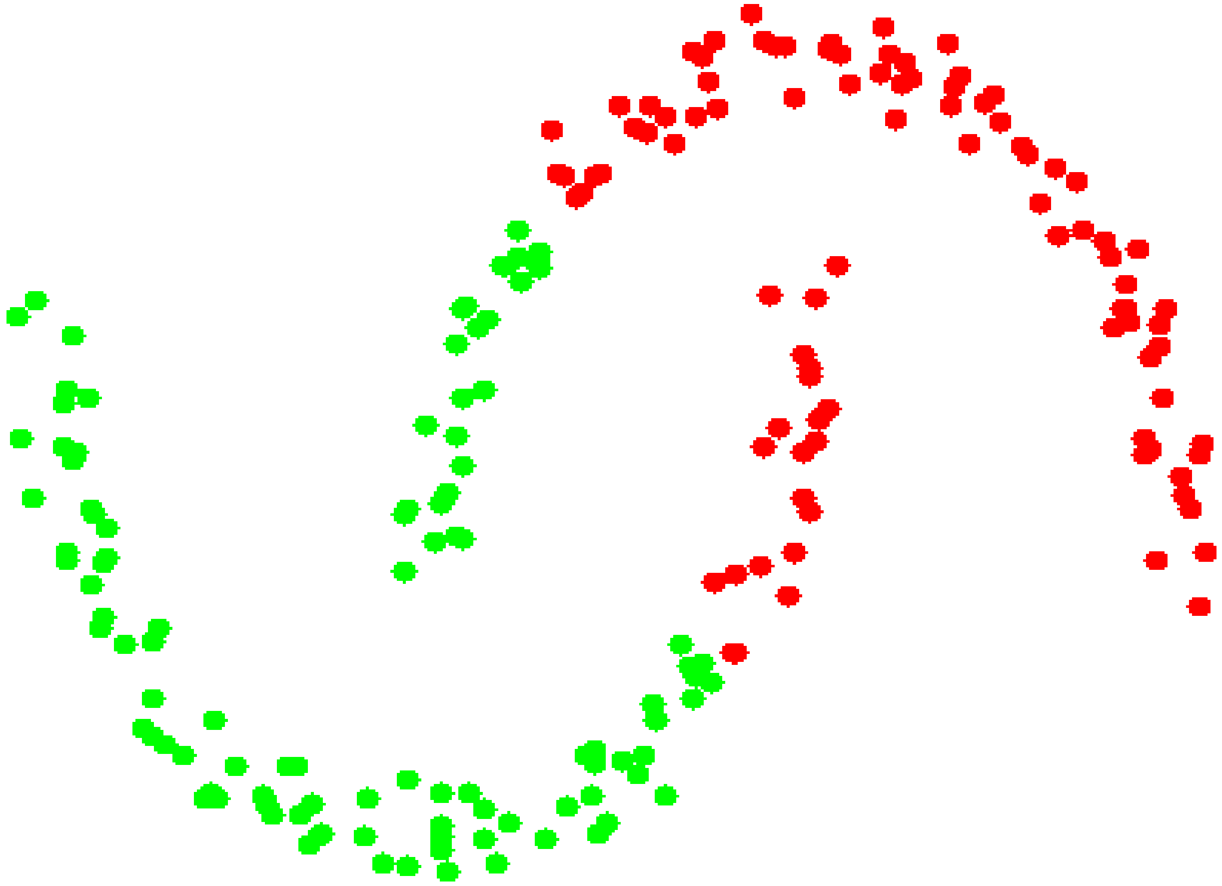
Objective: clustering

Graphs for non-linear data analysis



K-means fails as it produces convex clusters

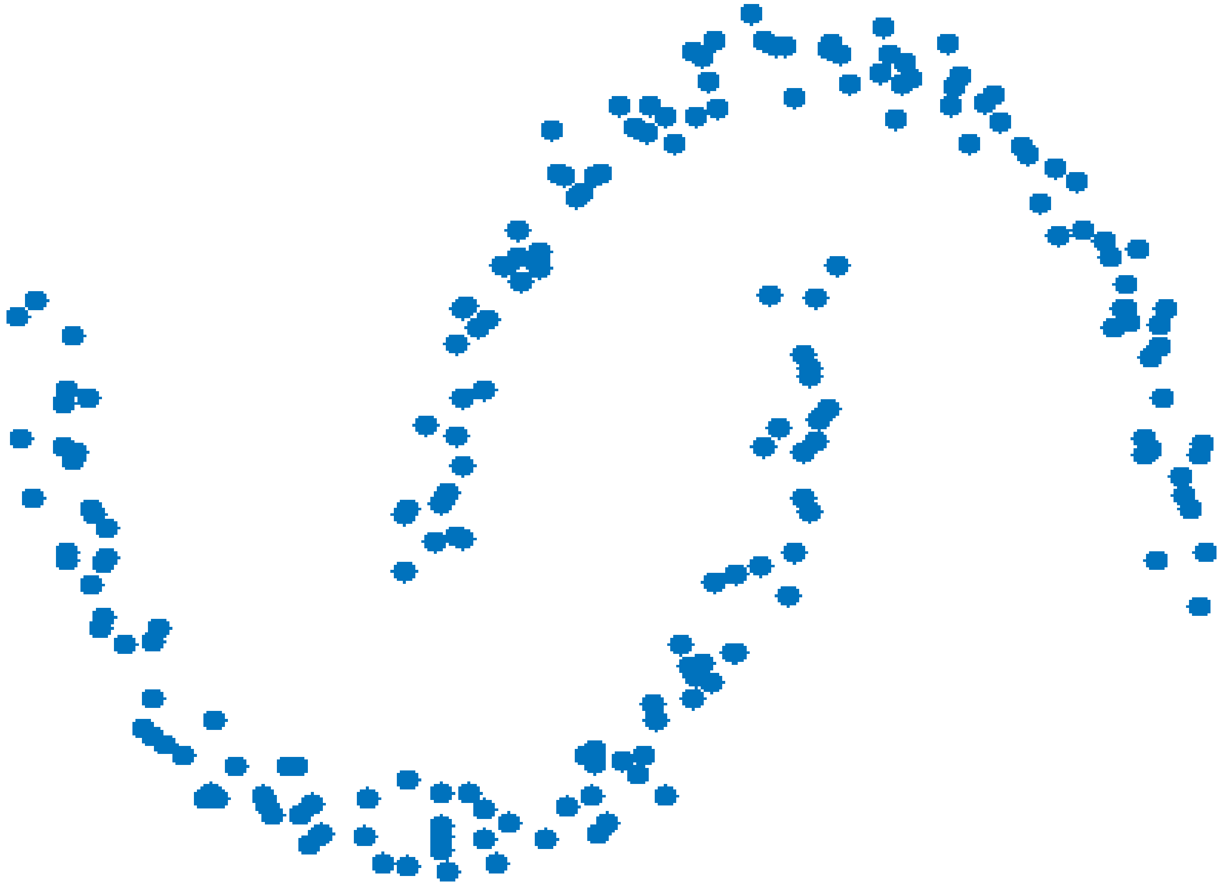
Graphs for non-linear data analysis



K-means fails as it produces convex clusters

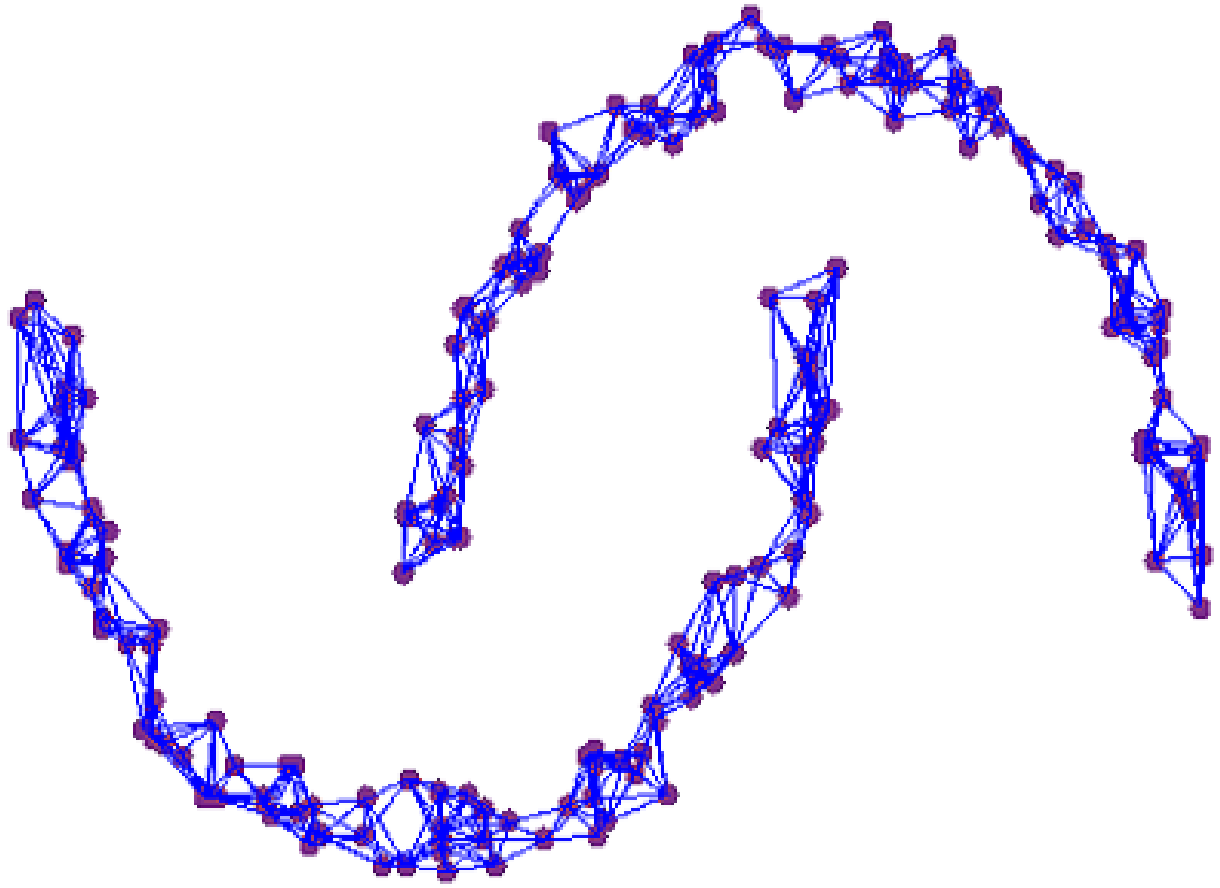
$\Rightarrow$ need a non-linear algorithm.

Graphs for non-linear data analysis



Non-linear data analysis in two steps:

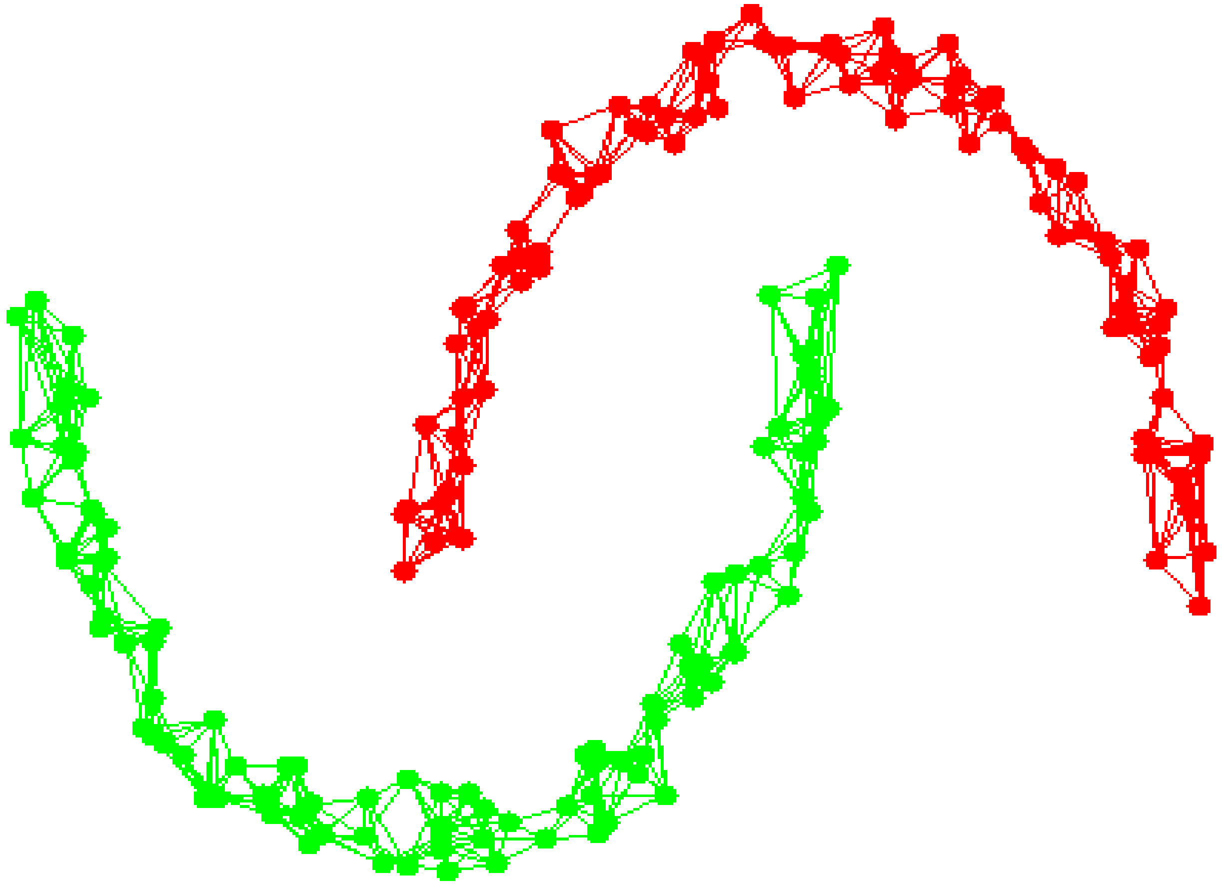
Graphs for non-linear data analysis



Non-linear data analysis in two steps:

$1)$ Build a neighborhood graph on the data.

Graphs for non-linear data analysis



Non-linear data analysis in two steps:

$1)$ Build a neighborhood graph on the data.

$2)$ Use a graph analysis algorithm.

Graphs for non-linear data analysis

Graph-based approach is at the center of many non-linear algorithms:

- (Spectral) Clustering
- Dimensionality reduction (Isomap, diffusion maps, etc.)
- Semi supervised learning
- Manifold learning

Graphs for non-linear data analysis

Graph-based approach is at the center of many non-linear algorithms:

- (Spectral) Clustering
- Dimensionality reduction (Isomap, diffusion maps, etc.)
- Semi supervised learning
- Manifold learning

Von Luxburg and Alamgir (2013): are we sure the graph contains all the relevant information?

The Problem

The Problem

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density $f > 0$.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Vertices: $X_1, \ldots, X_n$

Edges: (X_i, X_j) where X_j is one of the k-nearest neighbor of X_i.

The Problem

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density $f > 0$.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Vertices: $X_1, \ldots, X_n$

Edges: (X_i, X_j) where X_j is one of the k-nearest neighbor of X_i.

Question: can we estimate f from $G_{k,n}$?

The Problem

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density $f > 0$.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Vertices: $X_1, \ldots, X_n$

Edges: (X_i, X_j) where X_j is one of the k-nearest neighbor of X_i.

Question: can we estimate f from $G_{k,n}$?

Difficult problem as, locally, the graph is the same everywhere.

The Problem

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density $f > 0$.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Vertices: $X_1, \ldots, X_n$

Edges: (X_i, X_j) where X_j is one of the k-nearest neighbor of X_i.

Question: can we estimate f from $G_{k,n}$?

Difficult problem as, locally, the graph is the same everywhere.

Von Luxburg and Alamgir (2013): yes, with quantative guarantees, but only in dimension 1.

The Problem

$X_1, \dots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density $f > 0$.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \dots, X_n$.

Vertices: $X_1, \dots, X_n$

Edges: (X_i, X_j) where X_j is one of the k-nearest neighbor of X_i.

Question: can we estimate f from $G_{k,n}$?

Difficult problem as, locally, the graph is the same everywhere.

Von Luxburg and Alamgir (2013): yes, with quantative guarantees, but only in dimension 1.

Hashimoto et al. (2016): yes, but no quantitative guarantee.

Random walks on k-nn graphs

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density f.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Random walks on k-nn graphs

$X_1, \dots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density f.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \dots, X_n$.

Ting et al. (2010): a random walk on $G_{k,n}$ is an approximation of a diffusion process with generator:

$$\mathcal{L}_{\tilde{\mu}} = f^{-2/d} \left(\nabla \log f . \nabla + \tfrac{1}{2} \Delta \right)$$

and reversible measure $\tilde{\mu}$ with density proportional to $f^{2+2/d}$.

Random walks on k-nn graphs

$X_1, \ldots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density f.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \ldots, X_n$.

Ting et al. (2010): a random walk on $G_{k,n}$ is an approximation of a diffusion process with generator:

$$\mathcal{L}_{\tilde{\mu}} = f^{-2/d} \left(\nabla \log f . \nabla + \tfrac{1}{2} \Delta \right)$$

and reversible measure $\tilde{\mu}$ with density proportional to $f^{2+2/d}$.

$\pi_{k,n} :=$ invariant measure of a random walk on $G_{k,n}$ (measure on $(\mathbb{R}/\mathbb{Z})^d$).

Idea: if the invariant measure of a random walk on $G_{k,n}$ $\pi_{k,n} \approx \tilde{\mu}$ then it can be used to estimate f.

Random walks on k-nn graphs

$X_1, \dots, X_n \in (\mathbb{R}/\mathbb{Z})^d$ i.i.d $\sim \mu$ with density f.

$G_{k,n}$ is a k-nearest neighbors graph on $X_1, \dots, X_n$.

Ting et al. (2010): a random walk on $G_{k,n}$ is an approximation of a diffusion process with generator:

$$\mathcal{L}_{\tilde{\mu}} = f^{-2/d} \left(\nabla \log f . \nabla + \tfrac{1}{2} \Delta \right)$$

and reversible measure $\tilde{\mu}$ with density proportional to $f^{2+2/d}$.

$\pi_{k,n} :=$ invariant measure of a random walk on $G_{k,n}$ (measure on $(\mathbb{R}/\mathbb{Z})^d$).

Idea: if the invariant measure of a random walk on $G_{k,n}$ $\pi_{k,n} \approx \tilde{\mu}$ then it can be used to estimate f.

Problem: invariant measures of random walks on **directed** graphs are complex objects.

Convergence result

Convergence result

Hashmioto et al. (2016): $\pi_{k,n}$ converges weakly to $\tilde{\mu}$ when $n \to \infty, k/n \to 0, k >> n^{2/d+2} \log(n)^{d/d+2}$

Convergence result

Hashmioto et al. (2016): $\pi_{k,n}$ converges weakly to $\tilde{\mu}$ when $n \to \infty, k/n \to 0, k >> n^{2/d+2} \log(n)^{d/d+2}$

Can we obtain a quantitative version of this result?

Convergence result

Hashmioto et al. (2016): $\pi_{k,n}$ converges weakly to $\tilde{\mu}$ when $n \to \infty, k/n \to 0, k >> n^{2/d+2} \log(n)^{d/d+2}$

Can we obtain a quantitative version of this result?

Proposition
There exists $C > 0$ such that with probability $1 - \frac{C}{n}$,

$$W_2(\nu, \tilde{\mu}) \leq C \left(\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}} + \left(\frac{k}{n}\right)^{1/d} \right).$$

.

Convergence result

Hashmioto et al. (2016): $\pi_{k,n}$ converges weakly to $\tilde{\mu}$ when $n \to \infty, k/n \to 0, k >> n^{2/d+2} \log(n)^{d/d+2}$

Can we obtain a quantitative version of this result?

Proposition
There exists $C > 0$ such that with probability $1 - \frac{C}{n}$,

$$W_2(\nu, \tilde{\mu}) \leq C \left(\underbrace{\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}}}_{\text{Variance}} + \underbrace{\left(\frac{k}{n}\right)^{1/d}}_{\text{Bias}} \right).$$

.

Convergence result

Hashmioto et al. (2016): $\pi_{k,n}$ converges weakly to $\tilde{\mu}$ when $n \to \infty, k/n \to 0, k >> n^{2/d+2} \log(n)^{d/d+2}$

Can we obtain a quantitative version of this result?

Proposition
There exists $C > 0$ such that with probability $1 - \frac{C}{n}$,

$$W_2(\nu, \tilde{\mu}) \leq C \left(\underbrace{\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}}}_{\text{Variance}} + \underbrace{\left(\frac{k}{n}\right)^{1/d}}_{\text{Bias}} \right).$$

.

Idea behind the proof: show the measures of $(Y_t)_{t \geq 0}$, diffusion process with generator $\mathcal{L}_\mu$ and $Y_0 \sim \pi_{k,n}$ does not change much as t goes to infinity.

Open questions

Open questions

- Pointwise convergence of $\pi_{k,n}$ (we only have measure convergence)

Open questions

- Pointwise convergence of $\pi_{k,n}$ (we only have measure convergence)

  Invariant measure of random walks are used by graph algorithms such as PageRank.

  Obtain an actual density estimator which can be used in algorithms (e.g. graph embedding)

Open questions

- Pointwise convergence of $\pi_{k,n}$ (we only have measure convergence)

  Invariant measure of random walks are used by graph algorithms such as PageRank.

  Obtain an actual density estimator which can be used in algorithms (e.g. graph embedding)

- Variance term $\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}}$ is suboptimal, we expect $\sqrt{\frac{\log n}{k}}$.

Open questions

- Pointwise convergence of $\pi_{k,n}$ (we only have measure convergence)

  Invariant measure of random walks are used by graph algorithms such as PageRank.

  Obtain an actual density estimator which can be used in algorithms (e.g. graph embedding)

- Variance term $\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}}$ is suboptimal, we expect $\sqrt{\frac{\log n}{k}}$.

  $k >> n^{\alpha}$ $\qquad\qquad$ $k >> \log n$

  The smaller k, the sparse the graph.

Open questions

- Pointwise convergence of $\pi_{k,n}$ (we only have measure convergence)

  Invariant measure of random walks are used by graph algorithms such as PageRank.

  Obtain an actual density estimator which can be used in algorithms (e.g. graph embedding)

- Variance term $\frac{n^{1/d}\sqrt{\log n}}{k^{1/2+1/d}}$ is suboptimal, we expect $\sqrt{\frac{\log n}{k}}$.

  $k >> n^{\alpha}$ $\qquad$ $k >> \log n$

  The smaller k, the sparse the graph.

  Rate appearing in the convergence of other important quantities (spectra of graph Laplacians).

References

- Hashimoto, Sun and Jaakola, Metric recovery from directed unweighted graph, AISTATS 2016.

- Ting, Huand and Jordan, An analysis of the Convergence of Graph Laplacians, ICML 2010.

- Von Luxburg and Alamgir, Density estimation from unweighted kNN graphs: a roadmap, NIPS 2013.

Central Limit Theorem

Central Limit Theorem

The Gaussian measure γ is the invariant measure of the diffusion process

$$\mathcal{L}_\gamma = -x.\nabla + \Delta$$

Central Limit Theorem

The Gaussian measure γ is the invariant measure of the diffusion process

$$\mathcal{L}_\gamma = -x.\nabla + \Delta$$

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \sim \nu_n$

Central Limit Theorem

The Gaussian measure γ is the invariant measure of the diffusion process

$$\mathcal{L}_\gamma = -x.\nabla + \Delta$$

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim \nu_n$

Replacing one X_i at random with an independent copy X_i' we obtain a discrete process.

Central Limit Theorem

The Gaussian measure γ is the invariant measure of the diffusion process

$$\mathcal{L}_\gamma = -x.\nabla + \Delta$$

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim \nu_n$

Replacing one X_i at random with an independent copy X_i' we obtain a discrete process.

- ν_n is an invariant measure of this discrete process.
- this process approximates the diffusion process with generator $\mathcal{L}_\gamma$.

Central Limit Theorem

The Gaussian measure γ is the invariant measure of the diffusion process

$$\mathcal{L}_\gamma = -x.\nabla + \Delta$$

$X_1, \dots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim \nu_n$

Replacing one X_i at random with an independent copy X_i' we obtain a discrete process.

- ν_n is an invariant measure of this discrete process.
- this process approximates the diffusion process with generator $\mathcal{L}_\gamma$.

$\Rightarrow \nu_n \approx \gamma$.

Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

Theorem
If $\mathbb{E}[\|X_1\|^4] < \infty$, there exists $C > 0$ s.t.

$$W_2(\nu_n, \gamma) \leq n^{-1/2} d^{1/4} \mathbb{E}[X_1 X_1^T \|X_1\|^2]^{1/2}.$$

Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

Theorem
Let $p \geq 2$. If $\mathbb{E}[\|X_1\|^{p+2}] < \infty$, there exists $C_p > 0$ s.t.

$$W_2(\nu_n, \gamma) \leq C_p n^{-1/2} \left(\mathbb{E}[\|X\|^{p+2}] + d^{1/4}\mathbb{E}[X_1 X_1^T \|X_1\|^2]^{1/2}\right).$$

Central Limit Theorem

$X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[XX^T] = I_d$.

Theorem
Let $p \geq 2$. If $\mathbb{E}[\|X_1\|^{p+q}] < \infty, q \in [0, q]$, there exists $C_p > 0$ s.t., taking $m = \min(2, q)$,

$$W_2(\nu_n, \gamma) \leq C_p \Bigg(n^{-1/2+(2-q)/2p} \mathbb{E}[\|X_1\|^{p+q}]^{1/p} + \begin{cases} n^{-m/4} \mathbb{E}[\|X_1\|^{2+m}]^{1/2} + o(n^{-m/4}) \text{ if } m < 2 \\ n^{-1/2} d^{1/4} \|\mathbb{E}[X_1 X_1^T \|X_1\|^2]\|^{1/2} \text{ if } m = 2 \end{cases} \Bigg)$$

Probably (close to) optimal as it generalizes rates obtained in dimension 1.