A Nonlinear Filter for Markov Chains and its Effect on Diffusion Maps

Stefan Steinerberger
Yale University, stefan.steinerberger@yale.edu

A Nonlinear Filter for Markov Chains and its Effect on Diffusion Maps

Stefan Steinerberger

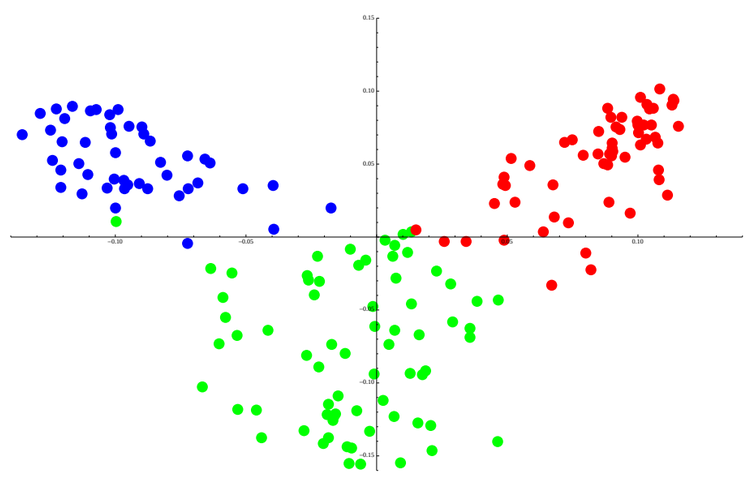
Department of Mathematics

Classical Diffusion Maps

We are interested in the interplay between Markov chains on a high-dimensional data set $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and the inner workings of spectral methods. There are many different methods, see e.g. the work of Belkin & Niyogi, Coifman & Lafon, Coifman & Maggioni, Donoho & Grimes. Usually, these techniques proceed by imposing a Markov chain on the data set and analyzing diffusion on the arising graph. A popular and natural choice for the Markov chain is to declare that the probability p_{ij} to move from point x_j to x_i is

$$p_{ij} = \frac{\exp\left(-\frac{1}{\varepsilon}\|x_i - x_j\|^2_{\ell^2(\mathbb{R}^d)}\right)}{\sum_{k=1}^n \exp\left(-\frac{1}{\varepsilon}\|x_k - x_j\|^2_{\ell^2(\mathbb{R}^d)}\right)},$$

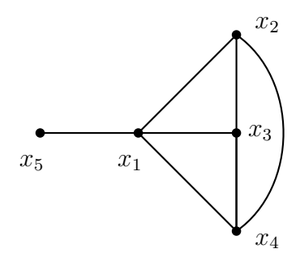
where the value of ε needs to be chosen depending on the given data as it induces a natural length scale $\sim \sqrt{\varepsilon}$ which should match the distance between neighbouring points.



This can work very well: see, for example, the wine data set mapped into two dimensions using a diffusion map (colors were added afterwards).

Rewarding Self-Consistency

We were motivated by the following example: suppose we are given



The standard diffusion paradigm proceeds by defining a random walk. However, given this local structure one would certainly believe that x_1, x_2, x_3, x_4 are well connected while x_5 seems to be an outlier. Consider a random walk starting in x_1. A simple computation yields

probability of being in	x_1	x_2	x_3	x_4	x_5
after 1 step	0	1/4	1/4	1/4	1/4
after 2 steps	1/2	1/6	1/6	1/6	0
minimum	0	1/6	1/6	1/6	0

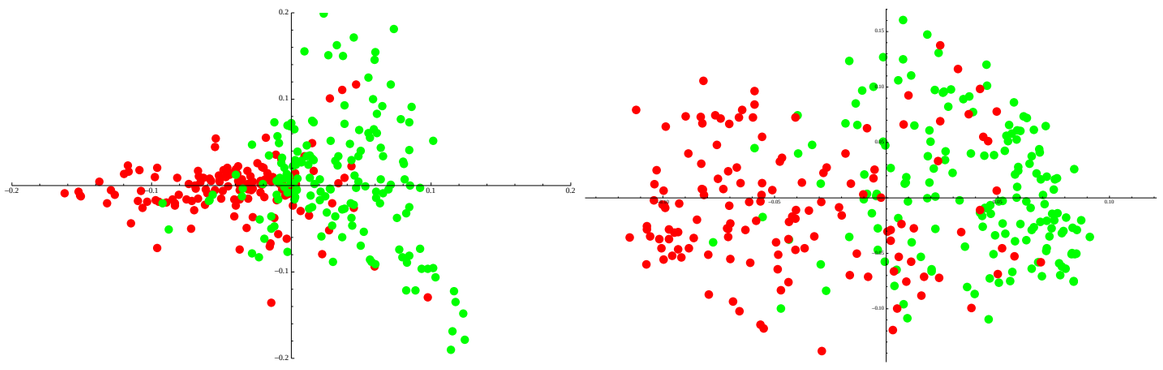
The minimum automatically detects unlikely outliers. Formally, assume we are given $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and an associated Markov chain described by the matrix $P = (p_{ij})_{i,j=1}^n$. We propose using another matrix Q instead: we obtain the matrix P^* by setting $p_{ii} = 0$ and rescaling every column of P so that to we are once again given a transition matrix. Q is then given by

$$\boxed{q_{ij} = \frac{\min((P^*)_{ij}, ((P^*)^2)_{ij}, \ldots, ((P^*)^k)_{ij}),}{\sum_{m=1}^n \min((P^*)_{mj}, ((P^*)^2)_{mj}, \ldots, ((P^*)^k)_{mj})}.}$$

We may then proceed with an analysis of the data set using Q instead of P. It seems that $k = 2$ is most effective in practice but there are certainly cases where a larger k may prove advantageous.

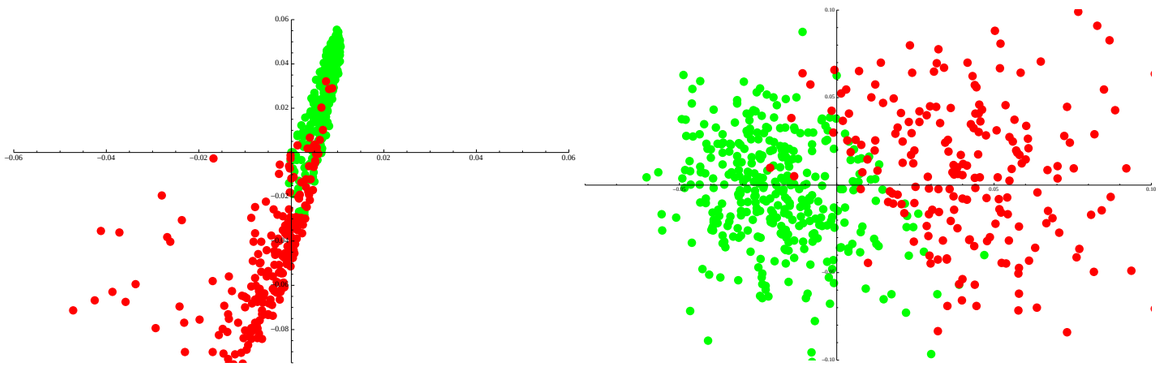
Examples

1 Cleveland Heart Disease Data Set.



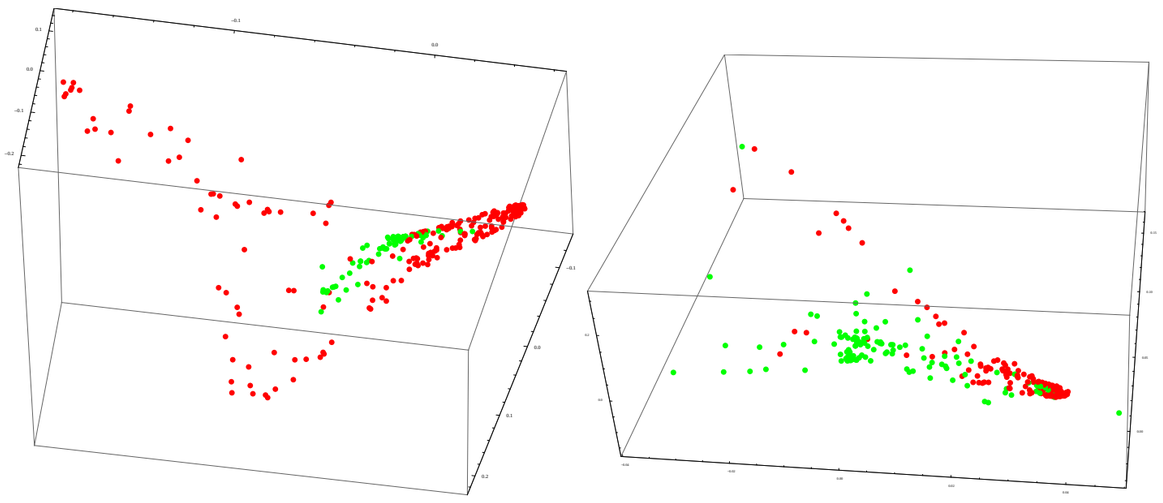
Classical method — New method

2 Wisconsin Breast Cancer Data Set.



Classical method — New method

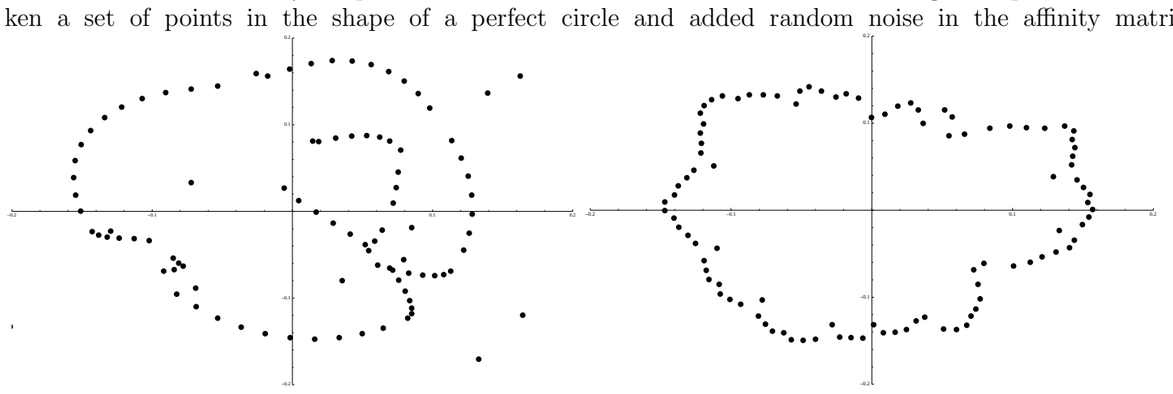
3 Ionosphere data set.



Classical method — New method

Error correction

The method can be easily adapt to detect and correct for error. In the following example, we have taken a set of points in the shape of a perfect circle and added random noise in the affinity matrix.



Classical method — New method

Theoretical results. Let $G = (V, E)$ be a finite graph with the property that every vertex has at most $c \in \mathbb{N}$ at distance at most 2 with transition matrix P. Construct G_1 by adding every possible edge with probability $0 < p \ll 1$ and let Q be the affinity assigned by the filter applied to the random walk on G_1.

Theorem. *The number N of vertices $(v_i, v_j) \in V \times V$ that are incorrectly thought of as present by the filter*

$$(P)_{i,j} = 0 \qquad \text{and} \qquad Q_{ij} > 0$$

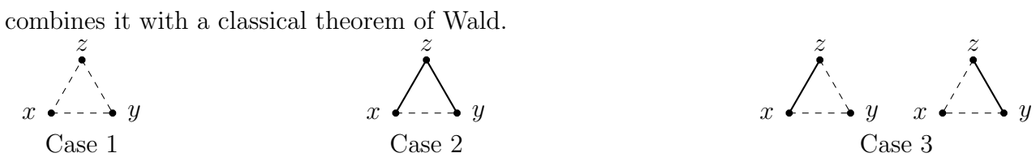
satisfies

$$\mathbb{E}N \leq cnp + cn^2p^2 + \frac{1}{2}n^3p^3.$$

This implies that the filter can successfully detect $|V|$ fake edges while only making $\mathcal{O}(1)$ mistakes on average. The trickiest part of the (not very complicated) proof uses the reproducing property of the binomial distribution

$$\mathcal{B}(\mathcal{B}(n,p),q) \sim \mathcal{B}(n,pq)$$

and combines it with a classical theorem of Wald.



References

- M. Belkin and P. Niyogi, Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation 15 (2003): 1373-1396.
- R. Coifman and S. Lafon, Diffusion maps. Appl. Comput. Harmon. Anal. 21 (2006), no. 1, 5–30.
- R. Coifman and M. Maggioni, Diffusion wavelets. (English summary) Appl. Comput. Harmon. Anal. 21 (2006), no. 1, 53-94.
- G. David and A. Averbuch, Hierarchical data organization, clustering and denoising via localized diffusion folders. Appl. Comput. Harmon. Anal. 33 (2012), no. 1, 1-23.
- D. Donoho and C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA 100 (2003), no. 10, 5591–5596.
- A. Wald, On cumulative sums of random variables, The Annals of Mathematical Statistics 15 (3): 283-296.
- A. Wald, Some generalizations of the theory of cumulative sums of random variables. The Annals of Mathematical Statistics 16 (3): 287-293.