ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

# Improved the locally aligned ant technique (LAAT) strategy to recover manifolds embedded in strong noise

Felipe Contreras[1,2*], Reynier Peletier[1] and Kerstin Bunte[2]

1- University of Groningen - Kapteyn Astronomical Institute
Broerstraat 5, 9712 CP Groningen - Netherlands

2- University of Groningen - Bernoulli Institute
for Mathematics, Computer Science and Artificial Intelligence
Broerstraat 5, 9712 CP Groningen - Netherlands

**Abstract**. The automatic detection, extraction, and modeling of manifold structures from large data-sets are of great interest, especially in Astronomy. Existing manifold learning techniques for feature extraction in Computer Vision, Bioinformatics and signal denoising typically fail in astronomical scenarios, since they mostly assume low levels of noise and one manifold of fixed dimension. Therefore, the Locally Aligned Ant Technique (LAAT) was recently proposed to discover multiple faint and noisy structures of varying dimensionality embedded in large amounts of background noise. Although it demonstrates excellent results in multiple scenarios, its performance depends on global thresholding and user tuning. Here, we improve LAAT and replace the global threshold by a flexible local strategy.

## 1 Introduction

In many fields such as Astronomy [1], medical science [2] and sensory activity recognition [3], nonlinear dimensionality reduction [4] and manifold learning techniques [5] are employed to find low dimensional structures inside potentially high-dimensional and big data point clouds. Astronomers study evolutionary processes and the history of cosmological interactions by analyzing the structures left behind, often employing N-body simulations [1]. These structures are typically nonlinear, plentiful, diffuse, they intersect and are surrounded by noise and outliers, which depreciate the result of conventional techniques [4, 5]. Detection methods designed specifically for one-dimensional astronomical structures find the medial axis by topology and Delaunay tessellation, but they are computationally expensive, and their result highly depends on the subset presented [6, 7]. More general approaches include Graph-based methods combined with Markov chain [8] and approaches to find elongated noisy clusters, such as Longest Leg Path Distance (LLPD) and Hierarchical Clustering algorithm Based on Noise Removal (HCBNR) [9, 10]. However, the graph-based methods require substantial amounts of memory and time, and the clustering techniques often require to set the number of clusters and struggle with ubiquitous noise.

Recently, the Locally Aligned Ant Technique (LAAT) [11] was proposed to detect an arbitrary number of diffuse manifolds of different dimensionality and varying density, embedded in large amounts of noise and outliers. It is inspired by the heuristic ant colony algorithm [12], which incorporates local alignment information and pheromone dynamics that act as reinforcement for faint structures. It has been shown to outperform alternative techniques in synthetic and

*Felipe Contreras thanks the National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO NACIONAL/2020-21200114 for their support.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

real-world astronomical N-body simulations while being efficient with resources [11]. LAAT is the first step in a pipeline of subsequent techniques to separate and model each detected manifold structure demonstrated for simulations of the cosmic web, jellyfish galaxies, and streams [13, 14]. LAAT's output is a pheromone value for each of the original data, with large values indicating points likely being part of a manifold and low values marking noise. Originally a global user-defined threshold is applied to determine the structures and remove all noise. This is often problematic, since a generous threshold that preserves very faint structures, also keeps a lot of noise around very dense structures, making subsequent techniques more costly, while a strict threshold reduces the noise, but also faint manifolds. Since LAAT represents the first step of the pipeline, its performance is crucial for subsequent methods. Therefore, we improve the implementation of the algorithm and propose a novel strategy to separate structures from noise.

## 2 Methodology

The *Locally Aligned Ant Technique* (LAAT) discovers diverse manifolds from large, potentially high-dimensional point clouds full of noise $\{\boldsymbol{x}_i \in \mathbb{R}^{\mathrm{D}}\}_{\mathrm{i=1}}^{\mathrm{n}}$. Due to space limitations, we refer the interested reader to [11] for details.

### 2.1 Modifications to LAAT

To enhance the LAAT for manifold detection, we identify two areas of improvement: 1) in the practical implementation and 2) post-processing for the separation of structures and noise. The latter concerns replacing the global thresholding by a local strategy based on gradients in the pheromone distribution to remove noise while preserving faint structures. And the former pertains practical implementation choices, such as the starting position of the ants. The current LAAT version places the initial $N_{\mathrm{ant}}$ ants randomly on particles, which have a neighborhood size bigger than the median of all particles. While this makes sure that ants start in regions of higher density, it makes the discovery of structures fainter than the median difficult. To solve this issue, we propose to put the initial ants in the boundary of the main and fainter structures using a local criterion that measures the variations of densities locally. Therefore, we compute for all points $\boldsymbol{x}_i$ the density of neighbors $\rho_i$ and its normalized standard deviation $s_i$:

$$\rho_i = \frac{|\mathcal{N}_r^i|}{\mathcal{V}_{\mathcal{N}_r^i}}; \quad s_i = \frac{1}{\bar{\rho}_i}\sqrt{\frac{1}{|\mathcal{N}_r^i|}\sum_{j\in\mathcal{N}_r^i}(\rho_j - \bar{\rho}_i)^2}, \tag{1}$$

where $|\mathcal{N}_r^i|$ and $\mathcal{V}_{\mathcal{N}_r^i}$ denote the number of particles and volume of the radius $r$ neighborhood centered at $\boldsymbol{x}_i$. Finally, the starting probability at point $\boldsymbol{x}_i$ is:

$$P(i) = \frac{e^{s_i\beta_{\mathrm{start}}}}{\sum_{j=1}^{n} e^{s_j\beta_{\mathrm{start}}}},$$

with inverse temperature parameter $\beta_{\mathrm{start}}$. We choose this expression because is simple, and every particle has a non-zero probability to receive an ant at the start, which increases with growing variation in density compared to its neighbors.

**Local separation of structures and noise:** The output of LAAT is a pheromone value for each data point, with high values marking particles close to the center of structures that are dense and/or elongated. However, a global

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

threshold for the final removal of noise, as originally proposed, is difficult. A conservative choice preserves faint structures, but also a lot of noise, which makes subsequent processing unnecessarily expensive and may keep some spurious random noise accumulations. Higher thresholds, on the other hand, remove more noise, but fainter structures disappear as well. Therefore, we propose a local criterion based on the pheromone distribution in the neighborhood of particles.

In every particle $\boldsymbol{x}_i$ we sort its neighbors $\mathcal{N}_r^i$ by their pheromone level in ascending order to obtain a local pheromone distribution curve. The discrete pheromone distribution is smoothed by fitting a twice differentiable piece-wise cubic spline, which preserves the global properties of the distribution, thus avoiding strong changes in the derivative generated by the interpolation method, with the smoothness determined by the number of pieces. The new local strategy then finds the first local minimum of the first derivative of this distribution, which corresponds to the first inflection point defined by its negative gradient. The pheromone value $T(p_0)$ in the inflection point $p_0$ detects the first steep change in the pheromone distribution and is used as the local threshold. When a particle is close to a structure, its pheromone level typically approximates the above inflection point. We define an aggressiveness parameter $\eta$ that allows to increase or decrease the local threshold, making the selection more aggressive or more conservative, respectively. Assume the domain of the smoothed distribution is $\mathbb{D} = [p_{\min}, p_{\max}]$, then the final point $p_f$ with corresponding local threshold $T(p_f)$ is given by:

$$p_f = \begin{cases} p_0 + (p_{\max} - p_0)\eta & \text{if } \eta \geq 0, \\ p_0 + (p_0 - p_{\min})\eta & \text{if } \eta < 0 \ . \end{cases} \tag{2}$$

The exact determination of when a particle can be considered noise or part of a very faint noisy structure is data set and application dependent. Hence, it cannot be decided in general and requires the ability of the user to make that choice. This is implemented by a filter $\psi$ denoting the minimum pheromone level that discards every point $\boldsymbol{x}_k$ when none of its neighbors reach that level.

## 3 Experiments and Discussion

Since LAAT was extensively shown to outperform alternative techniques in previous publications, we demonstrate and compare the improved LAAT to its old implementation in two scenarios: 1) A synthetic data set of a faint filament of varying density connecting two dense clusters embedded in uniform noise that exceeds the number of manifold particles by four. And 2) a real-world Dark Matter-only N-body cosmological simulation of the cosmic web as used in [14].

*Synthetic data:* This three-dimensional data set consists of a single manifold of $8 \cdot 10^4$ particles buried in $3.2 \cdot 10^5$ uniform noise points, see Fig. 1a). The structure is created using 11 multivariate Gaussians centered on a polynomial curve and evenly spaced. The covariances are aligned with the polynomial, and the number of points generated by them is minimal midway between the clusters and increases towards them. The middle Gaussian contains only 1% of the total manifold points, while the clusters gather 40%. We run the old LAAT using a neighborhood radius $r = 3$, a $\beta_{\text{start}} = 5.0$, with $7^3$ ants and 5000 iterations for 50 epochs, and with standard values for the remaining user parameters. Fig. 1b) shows the surviving selection with global pheromone thresholds $0.5\bar{F}$ and $1.0\bar{F}$, where $\bar{F}$ is the average pheromone over all particles. Panel c) shows the

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

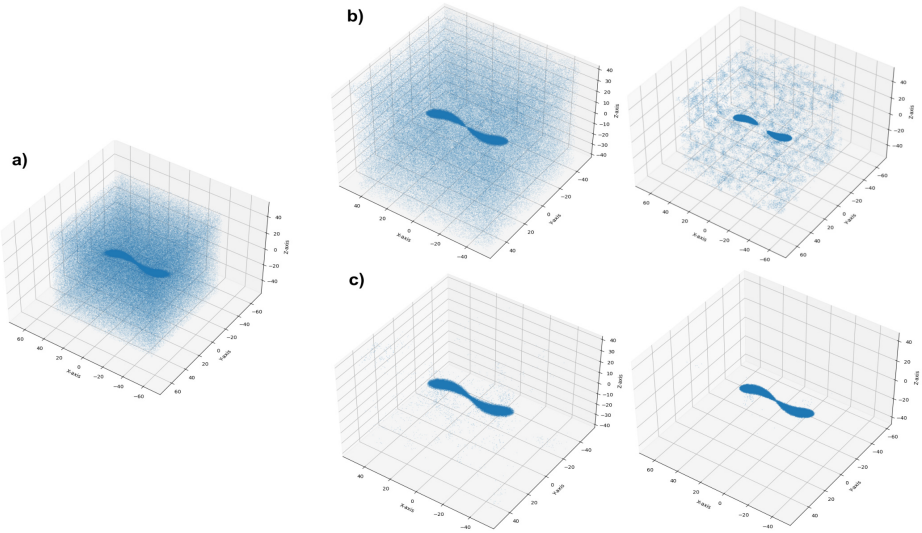


Fig. 1: **a)**: Synthetic data with a 80k points manifold embedded in 320k uniform noise particles. **b)**: Old LAAT result with global thresholds $0.5\bar{F}$ (left) and $1.0\bar{F}$ (right). **c)**: Same as **b)** but using the improved LAAT starting positions.

surviving data selection using the improved LAAT with identical parameters and fixed global thresholds instead of local post-processing, demonstrating that already the new starting positions of the ants help clear out most of the noise. A sample of results after the application of the local pheromone with filter $\psi = \bar{F}$ and varying number of splines and aggressiveness $\eta$ are shown in Fig. 2, with corresponding percentages of recovered manifold and noise points in Table 1.

The improved LAAT initializes the ants in zones with higher gradients in density, which reduces more background noise and recovers the faint filament center where the old LAAT with identical parameters disconnects it. The proposed local thresholding strategy, even in the most conservative case considered, recovers less than 1% percent of the original noise, while keeping the struc-

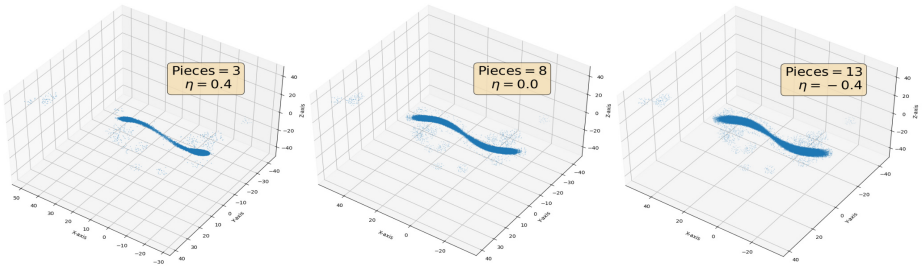


Fig. 2: Sample of the sensitivity response of resulting the selected data when changing the fitting number of pieces and the aggressiveness $\eta$ parameters.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Table 1: Percentages of manifold (left) versus noise (right) recovered by LAAT.

| Pieces\Aggressiveness | -0.4 | 0.0 | 0.4 |
|---|---|---|---|
| 3 | 62%/0.4% | 43%/0.3% | 29%/0.2% |
| 8 | 85%/0.6% | 72%/0.4% | 40%/0.2% |
| 13 | 92%/0.7% | 85%/0.6% | 44%/0.3% |

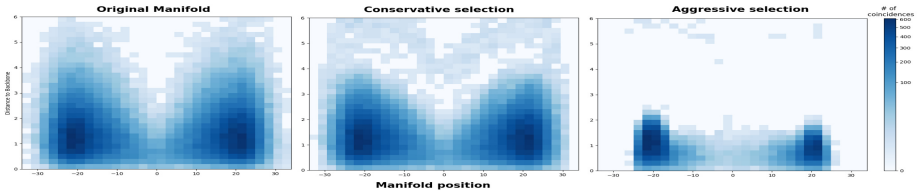


Fig. 3: 2D-Histograms of the minimum Euclidean distance (capped at 6) between selected particles and the backbone (x-axis) of the synthetic data. From left to right: the original manifold, the most conservative and most aggressive selection.

ture intact in all cases. Furthermore, the local post-processing strategy also reduces the noise of the structures themselves, which increases the efficiency of subsequent techniques in the pipeline for manifold modeling. Fig. 3 shows the 2D-histograms of the minimum Euclidean distance of the particles to the backbone of the original manifold (left) as well as the most conservative (middle) and the most aggressive cases for selection (right), respectively. Even requesting the most aggressive reduction preserves the entire structure. The number of pieces for the spline interpolation had no substantial effect on these results.

*Cosmic web:* A cube of $40^3$ Mpc$^3$/h is selected from a Dark Matter-only N-body cosmological simulation with $\approx 2.7 \cdot 10^5$ particles, shown in Fig. 4a). The full details of the N-body simulation can be founded in [14]. The improved LAAT is run with radius $r = 2$, $\beta_{\text{start}} = 5.0$, $6^3$ ants, for 12k iterations and 100 epochs, and with standard values for the remaining user parameters. For the post-processing two scenarios, namely a conservative and a liberal setting that both reduce the raw data by $\approx 50\%$ are used, to demonstrate the behavior. The conservative scheme uses 20 pieces with $\eta = -0.8$ and a high filter $\phi = \bar{F}$ and the liberal one uses 8 pieces with $\eta = 0.15$ and a low filter $\phi = 0.2\bar{F}$. Both results are shown in Fig. 4 **b)** and **c)**, respectively. This demonstrates how the local post-processing can be used to concentrate on the detection of the most dense structures or the recovery of faint structures that tend to be confused with noise.

## 4 Conclusions

This paper presents improvements to the Locally Aligned Ant Technique (LAAT) for the detection and noise removal of multiple diffuse manifolds in the presence of large amounts of noise and outliers. The new implementations reduce noise close to the structures without loss of fainter ones as often occurred in the original LAAT. The improved noise reduction allows for more efficient subsequent modeling of the structures with 1-DREAM [13]. We demonstrate the performance on synthetic data and a real-world cube of a cosmic web simulation based on

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

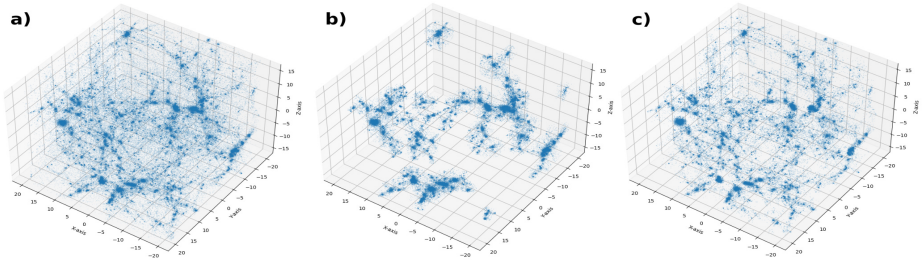


Fig. 4: **a)**: Cosmic web data cube of $40^3$ Mpc$^3$/h and $\approx 2.7 \cdot 10^5$ particles. **b)**: Improved LAAT result with aggressive filter $\phi = \bar{F}$ and **c)** with conservative filter $\phi = 0.2\bar{F}$, both reducing the original data by $\approx 50\%$.

two scenarios: a very conservative and aggressive setting for noise removal. In future work, we will investigate the use of Evolutionary Game Theory to replace the fixed radius to achieve further automation.

## References

[1] A. Knebe, F. R. Pearce, H. Lux, Y. Ascasibar, P. Behroozi, J. Casado, and et. al. Structure finding in cosmological simulations: the state of affairs. *MNRAS*, 435(2):1618–1658, 2013.

[2] J. L. Nielson, S. R. Cooper, J. K. Yue, M. D. Sorani, T. Inoue, and et. al. Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis. *PLOS ONE*, 12(3):e0169490, Mar. 2017.

[3] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling Methods for Persistent Homology. In F. Bach and D. Blei, editors, *Proc Int Conf Mach Learn (ICML)*, volume 37, pages 2143–2151, Lille, France, July 2015. PMLR.

[4] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, Dec. 2000.

[5] M. Hein and M. Maier. Manifold Denoising. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in NeurRIPS 19*, pages 561–568. MIT Press, 2007.

[6] N. Shivashankar, P. Pranav, V. Natarajan, R. v. d. Weygaert, E. P. Bos, and S. Rieder. Felix: A Topology Based Framework for Visual Exploration of Cosmic Filaments. *IEEE Trans Vis Comput Graph*, 22(6):1745–1759, June 2016.

[7] T. Sousbie. The persistent cosmic web and its filamentary structure I: Theory and implementation. *MNRAS*, 414(1):350–383, June 2011. arXiv: 1009.4015.

[8] P. D. Dixit. Introducing user-prescribed constraints in markov chains for nonlinear dimensionality reduction. *Neural Computation*, 31(5):980–997, May 2019.

[9] A. Little, M. Maggioni, and J. M. Murphy. Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms. *Journal of Machine Learning Research*, 21(6):1–66, 2020.

[10] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang. A hierarchical clustering algorithm based on noise removal. *International Journal of Machine Learning and Cybernetics*, 10(7):1591–1602, July 2019.

[11] A. Taghribi, K. Bunte, R. Smith, J. Shin, M. Mastropietro, R. F. Peletier, and P. Tino. Laat: Locally aligned ant technique for discovering multiple faint low dimensional structures of varying density. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[12] M. Dorigo, M. Birattari, and T. Stutzle. Ant colony optimization. *IEEE computational intelligence magazine*, 1(4):28–39, 2006.

[13] M. Canducci, P. Awad, A. Taghribi, M. Mohammadi, M. Mastropietro, S. De Rijcke, R. F. Peletier, R. Smith, K. Bunte, and P. Tino. 1-dream: 1d recovery, extraction and analysis of manifolds in noisy environments. *Astronomy & Computing*, page 100658, 2022.

[14] P. Awad, R. Peletier, M. Canducci, R. Smith, A. Taghribi, M. Mohammadi, J. Shin, P. Tino, and K. Bunte. Swarm intelligence-based extraction and manifold crawling along the large-scale structure. *MNRAS*, 520(3):4517–4539, 02 2023.