

Multi-Manifold learning in comparison of astronomical observations and numerical Simulations

By

Marco Canducci

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Artificial Intelligence and Natural Computation Group
School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
January 2021

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

© Copyright by MARCO CANDUCCI, 2021

All Rights Reserved

Abstract

The intrinsic nature of noisy and complex data sets is often concealed in low-dimensional structures embedded in a higher dimensional space. Number of methodologies have been developed to extract and represent such structures in the form of manifolds (i.e. geometric structures that locally resemble continuously deformable intervals of $\mathbb{R}^j$ [1]). Usually a-priori knowledge of the manifold's intrinsic dimensionality is required. Additionally, their performance can often be hampered by the presence of a significant high-dimensional noise aligned along the low-dimensional core manifold. In real-world applications, the data can contain several low-dimensional structures of different dimensionalities. This work describes a framework for dimensionality estimation and reconstruction of multiple noisy manifolds embedded in a noisy environment. The workings of the framework are demonstrated on different synthetic data sets, presenting challenging features for state-of-the-art techniques in Multi-Manifold learning. Through these worked examples it is shown how the proposed methodology is able to model abstract and topologically challenging manifolds such as Möbius strip and toroids (also higher dimensional). The comparison with existing techniques is organized along the two separate aspects of the methodology, namely manifold approximation and probabilistic modelling. The framework is then applied to astronomical complex data set containing simulated gas volume particles from a particle simulation of a dwarf galaxy interacting with its host galaxy cluster and a Dark Matter simulation of the Large Scale Structure (LSS) of the Universe (Cosmic Web). Detailed analysis of the recovered 1D and 2D manifolds can help in understanding the nature of Star Formation in such complex systems and to link the kinematic properties of Dark Matter filaments with their galaxies. An additional case study is presented for the

[1] j is the manifold dimensionality. Mathematically, a continuous deformation corresponds to a homeomorphism: one-to-one continuous mapping with continuous inverse.

modelling of cavities in a Jellyfish galaxy. Tracking their evolution through time may help in understanding their origin.

ACKNOWLEDGMENTS

When I first came to Birmingham, five years ago, I felt out of place. The only thing I knew about the city was that it was the second most populated town in UK and coming from a small town in Italy that was enough to feel overwhelmed. It took time to feel comfortable living in Birmingham but in retrospect, coming here was probably the best choice I could make. Not only have I acquired access to extensive knowledge during my time here, I also met many wonderful people with whom I shared this adventure, both in terms of work and everyday life. For all this, I have to thank my supervisor, Peter Tiňo. Indeed, he gave me the chance to prove myself in a field of work I knew almost nothing about. He has been patient and kind during these four years and extremely helpful in any situation that arose. His imagination and tenacity are qualities I often struggle to achieve, thus, having him as a guiding light to illuminate the path to take has always been of great comfort to me. I hope the next three years will be even more fruitful than the past four. I know I will make my best so that this happens. Thanks also to the examiners who gave me a chance to talk about my work extensively. In a world made of ten minutes presentations, your attention to detail and the interest you demonstrated in my work have truly made me feel appreciated.

While the pandemic heavily affected the fun parts of doing a PhD, I am extremely grateful that before March 2020 I had numerous chances to travel and meet passionate researchers and lovely people. These opportunities came by being part of a fantastic group of dedicated people: the SUNDIAL ITN. Funded by the European Research Council (ERC), the SUNDIAL ITN has been a one of a kind network, dedicated to improving the standard techniques in Astronomy for linking observations and simulations and pushing further the

boundary of knowledge in this field. I would like to thank Reynier Peletier, the PI of the project, who pushed us all to be precise and insightful in our investigations of the Universe; Sven De Rijcke, for the enormous help in understanding numerical simulations and the physics involved and Kerstin Bunte, for her availability and the precious discussions on how to improve my skills in many different research-related fields. Of course every single one of the supervisors within SUNDIAL have been essential to this journey and I would like to thank you all for sharing your time with me. I want to thank all the ESRs in the group, starting with Michele and Abolfazl, with whom I was lucky enough to collaborate the most, Mohammad, Angela, Bahar, Alan, Shivangee, Thanh, Teymoor, Alex, Aleke, Caroline and Nushkia. I wish you all the best and I hope we will meet again soon, at least to celebrate the end of the project we worked so hard to achieve. I would also like to thank all my coworker fellows at the University of Birmingham. I haven't been too much around I must say, but I enjoyed the time we spent together.

Thanks to the Italian gang I met in Birmingham: Andrea, Simone, Alice, Jorge, Gianluca, Marco, Giulio, Isotta and Giulio and Nora. Having someone to take my mind off work and simply being together with you all, having fun and discovering what Birmingham had to offer has helped me immensely. The same goes for the "Esta Noche" group, Céline, Emanuel, Kiki, Dimitra, Reem and Wen. I wouldn't have probably finished my PhD if it wasn't for your support and care. I am very happy that we met and although our lives may change in the future and different ways may be laid in front of us, having shared this journey with you has been a pleasure. The Dynamo Keynes football team also deserves to be mentioned here, having been a constant in my life throughout these four years. I also do not want to forget the people who cheered for me from a distance: Valentina, Davide, Filippo, Chiara, Denise, Simone, Bebe, Ester, Annalisa, Gioele, Raffaele, Giovanni and all my friends from Santarcangelo, Rimini and Bologna.

Thanks to my family, my parents Dante and Letizia and siblings, Martina, Michele

and Michael. Being afar for such a long time is not easy and it also doesn't happen frequently in a family. We seem to be used to it somehow. I've always felt your support and affection. I can't wait to be with you again, especially in these difficult times. Finally, I cannot thank enough the only person who kept me sane over the prolonged lockdown and the person I chose as a life companion. Federica, thank you for always having been on my side during these challenging times, for caring about me and carrying my insecurities. We make a wonderful team.

Birmingham, 08 April 2022

Marco Canducci

Contents

Page

1 Introduction **1**

1.1 Overview (motivation of the work) 1

1.1.1 A definition of manifolds 4

1.1.2 Proposed methodology 5

1.2 Organization of the work 8

2 Background on Manifold Learning **12**

2.1 Recovering low-dimensional representations of manifolds 13

2.1.1 Artificial Neural Networks 14

2.1.2 Additional Approaches 16

2.2 Density estimation and Model Selection 20

2.2.1 Finite Gaussian Mixture Models 20

2.2.2 Model Selection 21

2.2.3 Automatic GMM model selection 23

2.2.4 Generative Topographic Mapping 26

2.3 Multiple manifolds learning 29

I Methodology **32**

3 Adopted data sets **33**

3.1 Astronomical numerical simulations 33
3.1.1 Smoothed Particle Hydrodynamics (SPH) 33
3.2 Synthetic data sets 39
3.2.1 Manifolds equations and data sets generation 41
3.2.2 Behaviour of simulated physical properties 48

4 Single noisily sampled manifold **50**
4.1 Probabilistic modelling of a single manifold 51
4.2 Manifold crawling 56
4.2.1 Initialisation 57
4.2.2 Expansion Phase 59
4.2.3 Contraction Phase 59
4.2.4 Initialization of the latent graph embedding $\mathbf{f}(v; \mathbf{W})$ 60
4.3 Local on-edge curvature 62
4.3.1 Algebraic derivation 63
4.3.2 Geometric derivation 64
4.3.3 Demonstrative applications 65

5 Multiple noisily sampled manifolds **69**
5.1 Diffusion and background noise removal 69
5.2 Intrinsic dimensionality of manifolds 71
5.3 Multi-Manifold Crawling 76
5.3.1 An efficient alternative to Multi-Manifold Crawling 78
5.4 Comparison to existing methods 80
5.4.1 Multi-manifold learning 80
5.4.2 Probabilistic Modelling 86
5.5 Application to higher dimensional manifold 89

II Applications **92**

6 Spatial distribution along manifolds **93**

6.1 Experiments on a jellyfish galaxy 93

6.1.1 A Multi Manifold analysis of a dwarf jellyfish galaxy 97

7 Moving through manifolds **103**

7.1 Bi-dimensional profiles 104

7.1.1 Application to Jellyfish data set 109

7.2 Co-Moving orthonormal coordinate frames 110

7.2.1 Application to Jellyfish 116

7.2.2 Application to Cosmic Web 118

8 Modelling time-evolving manifolds **125**

8.1 One-dimensional GTM 126

8.2 Geodesic GTM for cavities 128

8.3 Temporal evolution 129

8.4 Experiments on a synthetic data set 130

8.5 Tracking the JellyFish galaxy's tail through time 132

8.6 Tracking supernovae explosions through time 133

9 Conclusions **139**

References **146**

List of Figures

1.1 Schematic representation of typical properties of the initial data set (fig. 1.1a) and the steps adopted in the framework proposed in this work for manifold extraction and modelling, in order: Diffusion and Filtering (fig. 1.1b), dimensionality index (fig. 1.1c), Multi Manifold Crawling (fig. 1.1d) and Abstract GTM (fig. 1.1e). For visualization purposes, the manifolds presented in the figures are all one-dimensional. However the methodology is widely applicable to higher dimensional manifolds as presented in the following. 9

2.1 Adapted from Bishop et al., 1998b: GTM construction model for a two dimensional non-linear manifold embedded in a three dimensional Euclidean space. The regular grid in the latent space (on the left) is mapped through the mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ onto the data space, where a spherical gaussian noise model is assumed. 28

3.1 Four different orbital stages of a single simulated dwarf galaxy in interaction with a Fornax-like galaxy cluster's halo. Note the typical filamentary structures departing from the "head" and forming the "tail" of the galaxy. . 35

3.2 Three samples from the simulated volume, projected on two of the three axis, showing clear intertwined one-dimensional structures. The snapshots are taken at the final stage of the simulation (current time). 38

3.3 Subfigures 5.1b - 5.1f: Ground-truth for each manifold used in dataset $\mathcal{D}$; 3.3h - 3.3l Corresponding noisy manifolds. Subfigure 3.3m: One- and three-dimensional noisy manifolds (left panel), two-dimensional noisy manifolds (central panel) and complete data set $\mathcal{Q}^1$, background noise included (grey points). 40

3.4 Data set $\mathcal{D}^2$ with the three interlocked noisy toroids (right panel) and its diffused and filtered counterpart $\tilde{\mathcal{D}}^2$ (left panel). 42

3.5 Distribution of variables V_1 and V_2 for the whole data set (sub-figures 3.5a and 3.5b) mimicking a stage of a simulated Jellyfish galaxy. The variables are designed to have peculiar profiles over the length of the manifolds or over their thickness (see description in the text). 46

4.1 Subfig. 4.1a shows the classic GTM setup; the leftmost panel is a representation of the latent space discretized in a grid of points x_i. The j-dimensional interval $[-1, +1]^j$ is mapped through the parametrized function $\mathbf{y}(\mathcal{X}; W)$ onto the data space with a spherical Gaussian centered on each point assumed as noise. The noise model aims at describing the true distribution of points in the data space (rightmost panel). Subfig. 4.1b is a sketch of *AGTM*. The latent space is substituted with an abstract graph, shown here for the Möbius strip, together with its topological representation (the two arrows pointing in opposite directions on the shortest edges of the graph). Function $f(\mathcal{V}; \mathbf{W})$ maps the graph onto the data space and manifold-aligned noise models are estimated on each graph's node $\overline{v}_i$. The true noisy distribution is displayed in the rightmost panel. 51

4.2 Left panel: Embedded graph of the Möbius strip used for initialisation of AGTM. Right panel: Probabilistic model obtained after 2 iterations of AGTM when initialised over graphs $\mathcal{G}$ and $\overline{\mathcal{G}}$ with a manifold aligned noise. 52

4.3 Visual representation of the Expansion and Contraction phases in the *Crawling* algorithm. Starting from the initial 1-edge nodes (blue points), the first set of node candidates (green points) are found (top panel) along the directions represented by black dashed lines. For every candidate $\tilde{\mathbf{t}}_c$, construct $\mathcal{B}(\tilde{\mathbf{t}}_c, \beta \cdot r) \cap \overline{\mathcal{V}}$ (red circles in bottom left panel). If the ball is non empty, its only element takes the place of $\tilde{\mathbf{t}}_c$ and is added to the graph (bottom right panel, red point). 58

4.4 Projection of tangent space onto new node's estimate. 59

4.5 Curvature computed on the recovered graphs of three concentric circles data sets. the radii of the circles are $1, 2$ and 4, respectively. The curvature values for the corresponding circles are shown in the colormap (algebraic and geometric on its left and right side respectively). Note the relation to the inverse of the radius. 66

4.6 Data set (left panel), extracted graph (central panel) and planar representation (right panel) for manifold $\mathcal{M}(x, y, z)$, and its on-edge curvature distribution in the two formulations: algebraic and geometric (left and right respectively). 67

5.1 Diffused manifolds using SAF technique. 70

5.2 Eigenvalue distribution on simplex and dataset labelled using the dimensionality index δ^G (upper row) and δ^S (lower row). 73

5.3 Confusion matrices for dimensionality indices δ_i^O (left), δ_i^G (center) and δ_i^S (right). The superior performance of δ_i^S is confirmed in fig. 5.2. 75

5.4 Both panels show the probabilistic model of each manifold in data set $\mathcal{D}$, recovered through AGTM, highlighted as an isosurface of the respective Probability Density Function. The black skeletons in each panel are the embedded graphs, as obtained after MultiManifold Crawling and training of AGTM. . . 78

5.5 Comparison of embeddings obtained for data set $\mathcal{Q}^2$ via LRNE (second column) Multi-Manifold Crawling (rightmost column) and its $\epsilon-$mesh alternative (third column). The clusters shown in the leftmost column are equally recovered by the three methods (figure 5.5a). Figure 5.5b shows the number of components (top row) and computational time (bottom row) for Multi-Manifold Crawling (black boxplots) and its $\epsilon-$mesh alternative (blue boxplots, identified by label "NG") evaluated over 30 samplings of synthetic data set $\mathcal{Q}^2$. 82

5.6 Same as figure 5.5a, but for data set $\mathcal{Q}^1$. 83

5.7 Same as figure 5.5b, but for data set $\mathcal{Q}^1$. 85

5.8 Top row shows the results obtained with SMCE on the data set described in Elhamifar et al., 2011, while bottom row shows the results obtained with Multi-Manifold Crawling. From left to right: two obtained clusters, embedding (bottom: graph) of the first cluster and embedding (bottom: graph) of the second cluster. 86

5.9 Subfigures 5.9a and 5.9b present the average log-likelihood per point of the probabilistic models generated by the four methodologies with different parameter settings (top row) and the computational time required to train the models (bottom row) for data sets $\mathcal{Q}^2$ and $\mathcal{Q}^1$ respectively. 87

5.10 Results of crawling on sampled $3-$torus (subfig. 5.10a) and comparison between ground truth graph (subfig. 5.10b, left) and crawled regularly sampled $3-$torus (subfig. 5.10b, right). 90

6.1 Gas particles of the simulated dwarf galaxy falling into the halo of the Fornax Galaxy Cluster. 94

6.2 1-D (left) and 2-D (right) distributions of the diffused particles in the tail of the jellyfish structure. Highlighted in black are the points belonging to two distinct 1-D and 2-D structures. 95

6.3 Average log-likelihood (top row) and log computational time of the probabilistic modelling techniques described in section 5.4.2 applied to the jellyfish data set. 99

6.4 Embedded graph (left panel) and planar representation (right panel) of a 2-dimensional manifold extracted from the jellyfish data set, and its on-edge "geometric" curvature distribution. 99

6.5 1D (6.5a) and 2D (6.5b) manifolds extracted from the data set in fig. 6.1. The graphs' nodes are colored based on the value of [C II] of the surrounding particles lying not further than 1 kpc from the local tangent space of the manifold, weighted by the nodes' responsibilities. 101

7.1 Partition of data sets $\tilde{\mathcal{Q}}^3$ (left) and $\mathcal{Q}^3$ (right) into their corresponding low dimensional subsets: $\tilde{\mathcal{Q}}^3_1$ ($\mathcal{Q}^3_1$, red), $\tilde{\mathcal{Q}}^3_2$ ($\mathcal{Q}^3_2$, blue), $\tilde{\mathcal{Q}}^3_3$ ($\mathcal{Q}^3_3$, green). 104

7.2 Sketch depicting the formation of the uniformly sampled concentric cylindrical volumes, aligned with the local tangent space of manifold $\mathcal{M}_k$ on the AGTM center $\overline{\mathbf{v}}_\ell$. 105

7.3 Bi-dimensional profiles for manifolds $\mathcal{M}_{13}$ (left column) and $\mathcal{M}_{14}$ (right column) with (top row) and without (bottom row) the normalization term in equation (6.1), for variables $\overline{\rho}_i, \overline{V}_{1,i}, \overline{V}_{2,i}$, and $\overline{m}_i$ (left and right panel respectively). 108

7.4 Bi-dimensional profiles of four elongated manifolds recovered via the methodology discussed in this work. The manifolds are located in different regions of the simulated Jellyfish galaxy, but show very similar behaviours w.r.t. the analyzed properties. 110

7.5 Sketch depicting the formation of co-moving orthonormal planes along a portion of manifold $\mathcal{M}_k$, enclosed within two centers of the corresponding AGTM. 111

7.6 Selected snapshots of the video clip generated by considering co-moving coordinate frames on the elongation of manifold $\mathcal{M}_{13}$. The top row panels show the distribution of variables ρ, V_1 and V_2 over the coordinate frame $\mathcal{T}_{\perp}^{\ell}$ defined on $\overline{\mathbf{v}}_{\ell}$, shown in bottom central and right panels as a red sphere. Bottom left panel presents the Probability Density Function obtained through AGTM on the same plane. 112

7.7 As in figure 7.6 but representing properties of manifold $\mathcal{M}_{14}$. 115

7.8 Equally spaced snapshots of manifold 1 recovered from Jellyfish data set. . . 116

7.9 Snapshots of manifold 2 recovered from Jellyfish data set. 117

7.10 Snapshots of manifold 1 recovered from Cosmic Web data set. 122

7.11 Snapshots of manifold 2 recovered from Cosmic Web data set. 123

8.1 Top row: 1D time-evolving data set with corresponding "Ground Truth". Bottom: centres obtained via independent (magenta line) or time-dependent modelling. 131

8.2 Mean and standard deviation of distances between updated centres and corresponding Ground Truth at time τ. 131

8.3 Bi-dimensional profiles of the JellyFish's tidal tail tracked thtough 8 consecutive temporal stages. 134

8.4 Continuation of figure 8.3. 135

8.5 Top row: Temporal stages of simulated dwarf galaxy. Grey dots are galaxy's point distribution, in cyan the recovered bubble propagated through time. Bottom row: Zoom in on the bubble with centres (black) representing opposite lobes (red and blue). 136

8.6 As in figure 8.2 but w.r.t. the centres obtained via $l1-$regularization "on centres", for the remaining optimization formulations. 137

8.7 Weighted mean of simulated quantities ($\log_{10}(\rho)$, T and Neutral Fraction (NF)), on centres $N8$ and $N42$ of the models propagated through time using costs (8.13)-(8.16). 138

List of Tables

3.1 Number of particles and attributes of scientific data sets 39

3.2 Manifolds' intrinsic dimensionality, underlying distribution and number of points. 41

5.1 Parameters used for all Manifold Learning methods in this analysis for both data sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$. 81

Chapter One

Introduction

1.1 Overview (motivation of the work)

The field of astrophysics is constantly presented with numerous challenges. Notable examples and proposed solutions are the following. **(1)** The discrepancy between the observed rotational velocity profile (with respect to the distance from the center of a galaxy) of stars and gas within a disc galaxy with respect to the profile predicted by classical Newtonian dynamics. Dark matter has been proposed as a possible cause for this discrepancy: larger than expected rotational velocity curve on the outskirts of the galaxy. **(2)** The observational evidence that all galaxies are constantly receeding from each other (recessional velocity), and that their recessional velocity is proportional to their distance. This observation led to the conclusion that the Universe is expanding. **(3)** The missing baryon problem: the observed discrepancy between the total mass of baryonic matter in the Universe between a short time after the Big Bang and more recent times.

Furthermore, as an extremely inter-disciplinary field, it has always been influenced by the developments in its founding disciplines (e.g. physics, chemistry, mathematics, quantum mechanics,...) and consequently driven their own advances. An outstanding example is

given by the introduction of General Relativity within the scientific landscape. As a de facto generalization of Newtonian Gravitational Theory, General Relativity[1] (A. Einstein, 1916) describes the structure of the Universe as a four-dimensional (spatio-temporal) space. Its derivation was fuelled by differential geometry and tensor calculus, frameworks that can be traced back to the 1600s as developed by Gottfried Leibniz, Isaac Newton and notably Carl Friedrich Gauss (Gauss, 1981), but later extended at the beginning of the XX century by David Hilbert, among others. The motivation behind the search of a new gravitational theory can be identified in the discovery of the finiteness of the speed of light, derived from the work on electromagnetism proposed in Maxwell (1873). This upper limit on the speed of light caused the crisis with Newtonian gravitational theory, since it was in disagreement with the assumption of instantaneous action at a distance of a force. However, this inconsistency was solved by introducing, with General Relativity, the idea of a curved time-space that is bent in the surroundings of a mass. The influence of General Relativity in the modern scientific world, and especially in astronomy, is still growing. As an example, the theoretical prediction of the existence of Gravitational Waves, due to the gravitational interaction of bodies in a binary system, was only confirmed by their first detection in Abbott et al., 2016, by the LIGO and VIRGO collaborations on February 11, 2016. Such a time delay between theoretical prediction and observational evidence is explained by the technical difficulty of these observations, and the necessity for instrumentation and technology that was simply not available at the time.

The technological advancement seen over the last decade has indeed opened new possibilities in many research fields. In particular, the increase in available computational power has allowed for a much larger amount of data being collected per experiment (or in the case of astronomy, observation), what is often referred to as Big Data. However, if on the one side the advent of Big Data brought excitement throughout the scientific community,

[1]The provided citation is only the end-point one of a series that aimed at fully describing the principles of General Relativity. The initial germs of the theory can be already found in Albert Einstein, 1908.

on the other it demanded more efficient (and less time consuming) ways to extract relevant information from it. In these high-dimensional (larger number of variables) and densely sampled (large number of observations) data sets however, information has often been found to hide in lower-dimensional structures. Driven by this, numerous techniques have been developed for Dimensionality Reduction (DR) and Visualization over the years (a review of these techniques is presented in chapter 2). These approaches can be particularly relevant when analyzing numerical simulations of astrophysical systems. A large variety of objects has been and is currently being simulated with different techniques. However, the underlying modelled physical processes are gravitational interactions and hydrodynamics. While inherently different, these two types of interactions have been frequently modelled as systems of N-body particles, discretizing the contributions to the gravitational field on one side and adopting a Lagrangian interpretation of fluids on the other. The large data sets produced by these simulations are sets of particles, having positional and dynamical properties in the simulated volume (especially for N-body gravitational simulations) and other quantities representing local properties of the fluid (for hydrodynamical simulations). In this sense, throughout this work, this kind of data set is referred to as a point-cloud. Given the extremely non-isotropic structure of the initial stage of the simulated object (due to assumptions on their formation) and the fact that its evolution through time is regulated by the equations of motions in a gravitational field and the laws of hydrodynamics, the deformations of such objects are non trivial at any given evolutionary stage. In particular, when the relation between variables is non linear, a depiction of the structures hidden within the high-dimensional data can be achieved by Manifold Learning, where the term "manifold" denotes specific connotations of the structures, defined within the fields of differential geometry and topology.

1.1.1 A definition of manifolds

From a topological perspective, manifolds are low-dimensional structures that can be smoothly covered with local, linear patches of the same dimension as the manifold. When the ambient space is $\mathbb{R}^3$, surfaces and curves are manifolds, while in higher dimensional spaces, appropriate generalizations are conveniently defined. A linear patch is called a "chart", while the union of all charts covering the manifold is called an "atlas". More formally, given a point $\mathbf{p}$ on a space $\mathcal{M} \subset \mathbb{R}^n$, a chart on $\mathcal{M}$ at point $\mathbf{p}$ is defined as the pair $(\mathcal{U}, \phi)$, where $\mathcal{U}$ is the neighbourhood of $\mathbf{p}$ and ϕ an diffeomorphism[2] from $\mathcal{U}$ to an open subset of $\mathbb{R}^j$ (j being the dimension of the manifold). If every point of $\mathcal{M}$ has a chart, then $\mathcal{M}$ is called a *locally Euclidean topological space of dimension j*. Two charts are compatible when their respective homeomorphisms are well defined in the intersection of the two neighbourhoods, meaning that given two charts $(\mathcal{U}, \phi;\ \mathcal{U} \to \mathbb{R}^j)$ and $(\mathcal{V}, \psi;\ \mathcal{V} \to \mathbb{R}^j)$, the transition functions $\phi \circ \psi^{-1}$ and $\psi \circ \phi^{-1}$ are C^∞. A collection of charts of $\mathcal{M}$ forms an atlas if their union covers $\mathcal{M}$; a maximal atlas is one that is not contained in any other atlas. A locally Euclidean topological space of dimension j is a topological manifold if two more conditions apply: Hausdorff and second-countable. A topological space is Hausdorff when for any pair of distinct points in the space, exist corresponding open neighborhoods whose intersection is empty. It is called second-countable if there exist a countable base of open sets in the space s.t. any open subset of the space can be represented as union of elements of some subfamily of the base. If a maximal atlas exists for the topological manifold of dimension j, then the object identified by the maximal atlas and the topological manifold is called a *smooth* manifold of dimension j. The maximal atlas is called a differentiable structure on $\mathcal{M}$. Furthermore, the atlas is oriented if the Jacobian determinant of the transition functions is positive. The functions are then said to be "orientation preserving". If all the transition functions of the atlas are

[2] A diffeomorphism is a bijective function continuously differentiable with continuously differentiable inverse.

orientation preserving, then the atlas is "oriented". A smooth manifold having an oriented atlas is said to be "orientable", "not orientable" in the opposite case (e.g. a toroid is orientable, while a Möbius band is non-orientable). An extensive review of this topic can be found in Tu (2010). Throughout this work only smooth, differentiable manifolds are considered, without however distinguishing by orientation. Furthermore in order to account for observational noise in the data sets, the considered manifolds are always intended as sampled differentiable manifolds, corrupted by noise. When generalizing the proposed work to the multi-manifold case, the underlying assumption is that the data set is sampled from a collection of differentiable manifolds of different dimensions, each one corrupted by transverse (observational) noise. The collection of noisily sampled manifolds is generally considered to be embedded in a higher-dimensional Euclidean space, filled by systematic (uniform) noise.

The purpose of this work is to develop a consistent methodology capable of analyzing data sets that inherently consist of multiple noisily sampled manifolds of different dimensionalities, embedded in a common higher-dimensional space filled with uniform noise. Astronomical simulations, for the reasons previously mentioned, are ideal test-beds for this study. Recovering the noisily sampled manifolds from the point-cloud, would allow for a detailed study of individual structures and aid scientific inference on phenomenon such as Star Formation (in the context of simulated individual galaxies) and galactic dynamics within the Cosmic Web structure (simulations of the Large Scale Structure of the Universe).

1.1.2 Proposed methodology

In the following, a framework is proposed for automated dimensionality estimation and reconstruction of multiple noisy manifolds embedded in a noisy environment. The Generative Topographic Mapping (GTM) model is generalized so that densities aligned along arbitrary manifolds (even not orientable ones - such as Möbius strip) can be captured. This is achieved

by replacing the Euclidean latent space (generally parameterized as a discretized interval of $\mathbb{R}^j$ and thus linear) with a latent graph reflecting the topology of the data manifold. This representation of the latent space is at the core of the generalization of GTM, since now the neurons are organized along an abstract structure (a graph is a pair of sets: the set of nodes and the set of edges and it does not have to be embedded in a Euclidean space) that closely captures the topology of the high-dimensional data set via connections through edges between nodes. When embedded in the data space, the graph provides a manifold skeleton around which the noise models can be organized. This work is inspired by Xiaoxia Wang et al., 2008, but extends and generalizes it by threefold: **(1)** it proposes a new robust dimensionality index estimation for data points, **(2)** through a dedicated manifold crawling mechanism it allows for completely abstract manifold representations (latent graph) in the GTM latent space (instead of a regular grid) and **(3)** it has Gaussian noise components naturally aligned along the manifold, unlike the spherical noise models in the original GTM and Xiaoxia Wang et al., 2008. Manifold aligned noise models in GTM were also considered in Bishop et al., 1998a, but under the assumption of simple latent space structure in the form of j-dimensional interval. The key idea was to impose a larger variance on the locally parallel directions to the manifold than on the perpendicular ones. In the following exposition instead, the latent space is a discrete structure (abstract graph representing a skeleton of a given manifold), local noise models are formulated through kernel based estimates of the local covariance matrix of the data, with trainable scale parameter to allow for optimized overlapping of the neighbouring Gaussian components.

This work presents a radical reformulation of GTM to capture and model low-dimensional general noisily sampled manifolds through a dedicated abstract graph-structured latent space reflecting the core manifold structure, specific to each manifold. Bacciu et al., 2013 present another radical generalization of GTM in the reverse direction - this time keeping the original simple latent space structure, but allowing for abstract structure in the data space - the

space of trees.

To summarize, the main contributions introduced in this Thesis are:

- Formulation of a new dimensionality index assigned to individual points based on which point-cloud can be partitioned into background points and sets representing cloud of points organised along noisy low-dimensional manifold structures;
- Development of a recursive crawling algorithm for the extraction of multiple low-dimensional, noisy manifolds embedded in a higher dimensional space: Multi-Manifold crawling;
- Extension of GTM's applicability to a broad class of manifolds (e.g. non-orientable or closed manifolds) by reformulating the latent space as an abstract graph and introducing a manifold-aligned noise attached to the embedded nodes of the latent graph: Abstract GTM. The abstract latent space and its embedding allows us to *understand the important global structural features of the underlying manifolds* - an important aspect of the proposed methodology bringing manifold learning under the umbrella of Artificial Intelligence.

The methodology[3] has been published in Canducci et al., 2022. A slight variation of the latent structure was presented in Taghribi et al., 2022 where the low-dimensional space is inherently spherical. It was furthermore extended to the case of time-evolving structures in Canducci et al., 2021 in the context of hydrodynamical simulations of a dwarf galaxy. In this work it is applied in section 6.1 to an astrophysical data set resulting from a mixed *N-body/Smoothed Particle Hydrodynamics* numerical simulation of a dwarf galaxy falling into the gaseous halo of a galaxy cluster. The point-cloud generated by the simulation presents nonlinear, noisy, low dimensional structures, providing for an ideal test bed for the

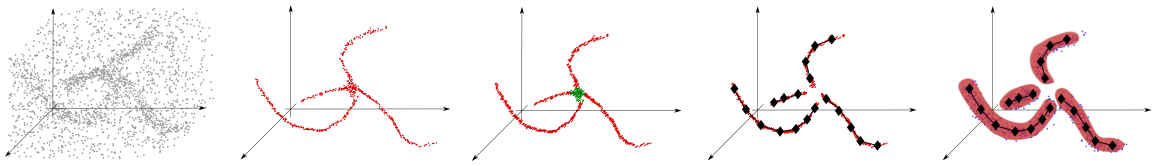
[3]A MATLAB implementation for the whole methodology and the data sets generation can be found at https://github.com/MarcoCanducci/general-noisy-multi-manifold-learning.git.

methodology. Two of the most significant manifolds are extracted and modelled, suggesting a possible scenario for formation of new stars in such a disrupted dwarf galaxy. The same analysis is performed on a different astronomical simulated data set, modelling the Large Scale Structure (LSS) of the Universe as seen via the dark matter distribution. This is theorized to be highly non-uniform, forming complicated filaments (here loosely intended as one-dimensional noisily sampled manifolds embedded in the three-dimensional simulated volume) that intersect in nodes and walls. This peculiar shape of the LSS is generally referred to as Cosmic Web. Two filaments are extracted and studied in terms of their kinematic properties in order to understand their influence on their galactic content.

1.2 Organization of the work

This work is divided in two parts. Part I presents the full proposed methodology, while Part II its application to scientifically relevant studies. Chapter 2 first presents the existing work on manifold learning, Principal Component Analysis (effectively the first, simplest manifold learning approach). A summary of the evolution of density estimation and probabilistic modelling of low-dimensional structures is then presented, referring in particular to Gaussian Mixture Models (GMM). By noting the importance of retaining a representation of the low-dimensional structures in data, Generative Topographic Mapping (GTM) was proposed as a natural alternative to the previously presented methodologies. Here, centers of the GMM are constrained to lie onto a low-dimensional manifold by means of a mapping function from a latent space. GTM serves as the starting point for the proposed methodology. As such, a detailed description is presented, before discussing the generalization of existing methodologies to the multi-manifold case.

Chapter 3 presents at first the synthetic data sets used as controlled experiments to



(a) (b) (c) (d) (e)

Figure 1.1: Schematic representation of typical properties of the initial data set (fig. 1.1a) and the steps adopted in the framework proposed in this work for manifold extraction and modelling, in order: Diffusion and Filtering (fig. 1.1b), dimensionality index (fig. 1.1c), Multi Manifold Crawling (fig. 1.1d) and Abstract GTM (fig. 1.1e). For visualization purposes, the manifolds presented in the figures are all one-dimensional. However the methodology is widely applicable to higher dimensional manifolds as presented in the following.

demonstrate and test the methodology. A controlled experiment is considered an experiment performed on a data set for which the ground-truth is known, and the performance easily evaluated by direct comparison. These data sets are purposefully built as challenging cases for both the existing and proposed methodology (a schematic representation is given in fig. 1.1a). The first one is a composition of different noisily sampled manifolds of different dimensionalities, comprising interesting topological structures such as the Möbius band and a hollow toroid. The second synthetic data set is composed of three interlocked toroids. The third data set is a simple approximation of a Jellyfish galaxy as simulated via a generic Smoothed Particle Hydrodynamics (SPH) simulation code. Each particle is identified by its coordinates in the 3−dimensional space, but it also retains information about two more properties. These properties have different behaviours along and across the two streams. The objective of this data set is to verify that the methodology offers a convenient tool for visualizing these properties in a compact and intuitive way. The following section of this chapter briefly describes the SPH formulation and the two astronomical simulated objects studied later in this work: Jellyfish galaxy and Cosmic Web.

Chapter 4 introduces the core model of the methodology, Abstract GTM (AGTM), representing density aligned along a single manifold (red regions in fig. 1.1e represent the probabilistic mass defined by the model). It is there explained how the latent space graph

representing topology of the data manifold is extracted through manifold crawling (black diamonds in fig. 1.1d represent embedded nodes of the latent graphs and lines their edges, as obtained via manifold crawling) and how this graph is then embedded in the data space along with the suitable set of noise models, aligned along the tangent space of the manifold. A discussion on how to calculate local curvature at the edges of the embedded graph is proposed, taking advantage of the smooth manifold description provided by AGTM.

Chapter 5 defines the adopted notion of dimensionality index for individual points and extends the framework of chapter 4 to the multi-manifold case (points of different dimensionality are shown in fig. 1.1c with different colors). A procedure for manifold denoising and filtering is also there described as a pre-processing step to the analysis performed at later stages (fig. 1.1b, red dots). A computationally efficient alternative to Multi-Manifold Crawling is also proposed. The price to pay for gains in efficiency is weaker detection stability of low dimensional manifold entities buried in the data. An experimental comparison on the first two synthetic data sets of the methodology with alternative multi-manifold learning and probabilistic modelling methods is also presented. Even though the methodology aims to provide density models of multiple low-dimensional manifolds buried in the data, the comparative experiments are organized separately for the multi-manifold capture and probabilistic modelling aspects of this work. This is because there is short of methods that can simultaneously recover low dimensional representations of an unknown number of manifolds while building their probabilistic models. An interesting case study of a $3-$dimensional torus embedded in $\mathbb{R}^4$ is at last shown in this chapter, proving that the proposed methodology is suitable to be applied to higher dimensional manifolds, of complicated topology in higher dimensional spaces than $\mathbb{R}^3$.

Part II starts with the analysis of extracted one- and two-dimensional manifolds from the Jellyfish galaxy data set (chapter 6). Since other manifold learning methodologies are not suited for this example, only a comparison of the probabilistic modelling aspect of the

methodology with the density estimation techniques identified in the previous chapters is a viable approach. Taking advantage of the low-dimensional representation of the structures recovered with the methodology and their corresponding high-dimensional embedding under the form of constrained GMMs, a visualization technique is proposed that compactly shows given simulated properties of the particles with respect to the manifold's position. Particularly interesting is the simulated intensity of a specific emission line associated to Star Formation. Both the one- and two-dimensional recovered manifolds show trends for this property, hinting at specific formation scenarios.

Chapter 7 describes two additional ways of visualizing properties of one-dimensional manifolds, by either performing cylindrical cuts along its spine (section 7.1) or by moving through it (section 7.2). In the first case a global depiction of the manifold's properties is obtained, while in the second case a more detailed analysis can be performed.

As a slightly different application, the case where certain topological properties of the low-dimensional manifolds are known is considered. In particular, cavities (chapter 8), as structures inherently isomorphic to spherical surfaces are treated here. Furthermore, a methodology for the time propagation of an initial model through the nearby evolutionary stages is proposed, by means of different cost functions. The framework for the temporal propagation of probabilistic models is a first-order Markovian assumption (Gagniuc, 2017). The performance of the model propagation is evaluated first on a synthetic one-dimensional time-evolving manifold and later applied on the temporal evolution of cavities in the Jellyfish data set. Having spatio-temporal models of time-evolving manifolds proves helpful in keeping track of the manifolds' properties and their evolution through time.

Finally, chapter 9 presents the conclusions on the proposed methodologies and its application to the adopted data sets. The whole work is briefly summarized and perspectives on future possibilities are presented.

Chapter Two

Background on Manifold Learning

Dimensionality reduction and Density Estimation of raw data, are commonly used tools to extract information from complex and noisy data sets. Due to dependencies among measured attributes of real world data, the data is often distributed along low-dimensional structures in a higher dimensional measurement space. The research field that focuses on extracting and describing these structures is usually referred to as *Manifold Learning*. In the following, we review the existing methodologies falling within this general framework, discussing their main assumptions, advantages and disadvantages. In particular, we are interested in describing the state-of-the-art techniques that efficiently deal with intrinsic noise within the manifolds. In fact, in real-world applications intrinsic noise is often a result of measurement errors. Preserving this information when performing Dimensionality reduction is essential in capturing the true nature of the underlying structure. In this scenario, density estimation techniques play an important role, since they describe the distribution of points via probabilistic models in a concise and mathematically well-defined manner. These two aspects of Manifold learning are treated separately in the next two sections, to better highlight their properties.

2.1 Recovering low-dimensional representations of manifolds

The task of recovering a low-dimensional representation of a structure lying in a high-dimensional space, has been historically divided in two main tasks: dimensionality reduction and manifold learning. While the former aims at reducing the number of variables and still preserve most of the information, the latter is devoted to obtaining a low-dimensional representation of the structure. However, the distinction between the two is not always clear as under specific assumptions (e.g. manifold as a linear subspace) these two terms are interchangeable. In fact, some dimensionality reduction techniques may even be considered primitive manifold learning ones. In line with this argument, all methodologies presented in the following discussion are revisited in the context of manifold learning. This includes possible visualization techniques like Self-Organizing Maps (SOM), where the "manifold" is intrinsically represented as isomorphic to a two-dimensional subspace.

As a first order attempt to manifold learning, Principal Component Analysis (PCA Pearson, 1901) is a well understood and widely used linear dimensionality reduction scheme. It assumes that the data lies on a linear low-dimensional space, described via a multivariate Gaussian. The recovered principal components capture the variance of the data along orthogonal directions. However, by design, PCA cannot appropriately capture non-linear low-dimensional structures. This lack of flexibility has been addressed by non-linear dimensionality reduction methodologies such as Classical Multidimensional Scaling (CMDS, Torgerson, 1952). Here, pairwise distances between points in the high-dimensional space are converted into coordinates over the low-dimensional space (of given dimensionality j), via the minimization of a specific loss function, called *strain*. Multidimensional Scaling is a visualization technique, where the low-dimensional space has dimension $j = 2$, typically. The extension of Classical/Metric MDS to non-Metric MDS (n-mMDS, J. Kruskal, 1964;

J. B. Kruskal, 1964), generalizes the methodology to deal with dissimilarities between points instead of pairwise distances. The *strain* function is here replaced by a *stress* function that can be further generalized using different Minkowski $r-$metrics (also called l_p-norms). The *stress* function is ultimately a "residual sum of squares" between corresponding pairwise distances in both spaces.

2.1.1 Artificial Neural Networks

In the context of visualization techniques and their interpretation as manifold learning methods, Self-Organizing Map (SOM, Kohonen, 1982) plays an essential role as a precursor of its probabilistic reformulation: Generative Topographic Mapping (GTM, Bishop et al., 1998b, see section 2.2.4). SOM is an Artificial Neural Network (ANN) formulation where a low-dimensional (tipically $j = 2$) space is represented as a regular grid of neurons (prototypes), arranged in either a rectangular or a hexagonal shape. The low-dimensional space is embedded in the ambient space of the data, covering it. The topological ordering of the low-dimensional space is then propagated into the ambient space. Generally, the map is initialized by aligning its axis with the first two principal components of the data (by means of PCA). The embedded neurons are then left to be trained. The training process presents an input data point at a time to the map. The distance (usually Euclidean) between the input point and all the neurons in the map is computed and the Best Matching Unit (BMU) selected as the closest neuron to the input data point. The positions of the BMU and (partially) its neighbors are shifted towards the target data point, by a small amount. Both the size of the neihgbourhood in the map and the shift of the BMU are decreased gradually over the training phase, in order for the methodology to converge to a stable solution. The lack of a generative model for the data and a cost function for optimization, the long computational time for training and the uncertainty over the initial configuration of the map are often re-

garded as disadvantages of SOM over other unsupervised learning algorithms (e.g. $k-$NN). However, some of these problems have been addressed by methodologies such as Generative Topographic Mapping or Growing SOM (Alahakoon et al., 2000). A detailed discussion of the statistical and convergence properties of SOM has been presented in Yin et al. (1995). Neural Gas (Martinetz et al., 1993), in particular, by relaxing the topological constraint on the low-dimensional space, is able to achieve a lower reconstruction error. Here, the embedded neurons (prototypes) are all influenced by every input data point (presented sequentially to the system), according to a specific adaptation equation which has two important dependencies. First, for each data point, "neighborhood-ranking" is performed, by sorting the prototypes according to their distance to the point. The adaptation function then moves each prototype towards the selected point by an amount proportional to their ranking. It has been shown how this process is related to the expansion of gas in a volume (hence the name Neural-Gas). The dynamics of the gas (prototypes) is guided by stochastic gradient descent on a specific cost function, representing the energy of the gaseous system. This formulation is closely related to "fuzzy clustering" (Dunn, 1973), where the hard assignment, typical in classical clustering, is replaced by a probability measure for each prototype given the data point. A generalization to a "batch" and "median" versions have been proposed in Cottrell et al., 2006, aimed at improving its performance.

Having mentioned ANN, particularly interesting is the case of Autoencoders (Rumelhart et al., 1986). Autoencoders (AE), are ANNs whose architecture creates a bottleneck towards the hidden layer, that comprises of fewer neurons than the input and output layers. In particular, the input and output layers have the same number of neurons. By minimizing the reconstruction error (originally defined as the ℓ_2-norm of the difference between reconstructed and input data), these ANNs are able to obtain a "significant" representation of the high-dimensional data in the low-dimensional hidden layer. The encoder (from input to hidden layer) and decoder (from hidden to output layer) are usually Neural Networks

themselves, but when linear operators are dropped (linear autoencoder), AEs are equivalent to PCA. Thus, AEs are a generalization of PCA, where the low-dimensional linear subspace is replaced by a manifold. More complicated architectures (e.g. Convolutional AEs, Masci et al., 2011) aim at recovering more complex low-dimensional representations on the hidden layers. Their probabilistic formulation converts the methodology to a generative one. Variational Autoencoders, (Kingma et al., 2014; Larsen et al., 2015) assume a generative model on the observed data points $\mathbf{x}_i$ conditioned on an unobserved latent variable $\mathbf{z}_i$ (probabilistic decoder). A posterior over the latent variable $\mathbf{z}_i$ is also assumed, conditioned on the input $\mathbf{x}_i$ (probabilistic encoder) and a prior on $\mathbf{z}_i$. With this structure, VAEs trained on specific data sets are able to generate new observations closely related to the training data set, by sampling from the prior on the latent variables. Lately, deep versions of VAEs have been proposed (Vahdat et al., 2020; Vancura et al., 2021). Despite their large success in both supervised and unsupervised learning tasks, when the data is aligned along a low-dimensional manifold, these methods are not suited for the case treated in this work: a collection of noisily sampled manifolds of different dimensionality, embedded in a noisy environment.

2.1.2 Additional Approaches

Kernel PCA (KPCA, Schölkopf et al., 1998) and Principal Polynomial Analysis (PPA, Laparra et al., 2016) have been proposed as generalizations of PCA to non-linear dimensionality reduction. As a way of mapping non-linear features over a linear subspace, KPCA transforms the input data, projecting it onto a Hilbert space (complete vector space with suitable inner product) via a specific Kernel function that can be chosen by the user. Eigen-decomposition is then performed on the Kernel matrix (matrix containing mutual dot products between points mapped in the Hilbert space) in replacement of the covariance matrix of the original data. PPA's main difference w.r.t. its predecessor (PCA) is the replacement of a

$j-$dimensional orthonormal basis for the low-dimensional space with a set of curves parameterized as polynomials of different degree.

Appeared in the same issue of *Science*, Isomap (Tenenbaum et al., 2000) and Locally Linear Embedding (LLE, Roweis et al., 2000) are two widely used non-linear dimensionality reduction algorithms. Isomap is considered an extension of metric MDS to points lying on a low-dimensional manifold embedded in a higher dimensional space. Given a point-cloud, Isomap constructs a neighbourhood graph, where each point is connected to either its first K neighbors or all points within a radius ϵ. For any pair of points, their geodesic distance is then approximated by the shortest path (e.g. through the Dijkstra algorithm, Dijkstra, 1959) connecting them, computed on the graph. The edge connecting any pair of nodes within the graph is weighted by the their Euclidean distance, so that the total geodesic distance between any pair of points is obtained as a weighted sum of short distances. Metric MDS is then applied to the matrix containing all pairwise geodesic distances and a low-dimensional representation of the manifold obtained as in the original method. In order to increase the computational efficiency of Isomap, Landmark-MDS (LMDS) has been proposed. The whole point-cloud representing the manifold is replaced by a suitable sample and MDS applied on the corresponding distance matrix. A further extension of Isomap to non-uniformly sampled manifolds is Conformal-Isomap (C-Isomap), where the edges are additionally weighted by a scalar function of a point's local density Silva et al., 2002.

While Isomap and its extensions can be considered global methods, since they represent the whole manifold in the form of a graph, Locally Linear Embedding (LLE), has a different approach to the recovery of a manifold's low-dimensional representation. It selects a small neighbourhood centered on each point of the data set. The points belonging to each neighbourhood are assumed to lie on a linear subspace of $\mathbb{R}^j$ (j being the dimensionality of the manifold). It then finds, for each neighbourhood, the linear coefficients that can reconstruct its center from its members. This is achieved by minimization of a cost function, contain-

ing the reconstruction errors due to the formulation of the local hyperplane ("patch"). The optimal weights are recovered by solving a least-squares problem. The low-dimensional representation ("embedding") of the manifold is found by minimization of the embedding cost function, where the reconstruction weights are now fixed and optimization is performed on the low-dimensional coordinates. LLE is considered a local method, as opposed to Isomap being considered global, since it models the whole manifold by covering it with small linear patches.

Many other Manifold Learning algorithms aiming to provide suitable approximations to a low-dimensional data manifold have been proposed. Laplacian eigenmaps Belkin et al., 2001 builds a neighbourhood graph of the manifold similarly to Isomap. However, taking advantage of spectral graph theory (e.g. Chung, 1997), it first computes the Laplacian of the graph and then performs eigen-decomposition to recover the optimal embedding of the manifold. Hessian eigenmaps (Donoho et al., 2003) computes for each neighbourhood the Hessian matrix. The Hessian matrix represents the second-derivatives of a mapping function from low- to high-dimensional space, w.r.t. coordinates defined on the local tangent space of the neighbourhood. Eigendecomposition of the Hessian matrix yields the low-dimensional space coordinates.

A different approach in manifold learning is represented by Stochastic Neighbour Embedding (SNE, Hinton et al., 2003). In this methodology, an asymmetric probability is introduced for two points to be neighbors in the high-dimensional space. At the same time, the probability of the two projected points to still be neighbors in the low-dimensional space is evaluated. The optimal embedding is obtained by minimizing the sum of the Kullback-Leibler divergences Kullback et al., 1951 between the two distributions (minimize the statistical distance between the two). By replacing the low-dimensional Gaussian distributions with a Student's t-distribution Student, 1908, and symmetrizing the cost function, t-distributed Stochastic Neighbour Embedding (t-SNE) aims at solving the "crowding problem": nearby

points in the high-dimensional space are condensed in small regions when mapped into low-dimensional space due to the necessity of representing farther away points on the same low-dimensional space.

Local Tangent Space Alignment (LTSA, Zhang et al., 2004) represents the local geometry of the manifold by estimating its tangent space over small neighborhoods around each point of the data set. The recovered tangent spaces (considered affine subspaces in the respective neighbourhoods) are aligned via an optimization procedure over the overall reconstruction error. Once again, the optimal alignment is found through eigen-decomposition of a specific matrix, containing information about the *local geometry* of the manifold.

In Boissonnat et al., 2018 (and bibliographic references therein), a different angle, based on computational geometry, has been proposed in order to extract low-dimensional manifolds from data samples. Here, simplicial complexes (such as Delaunay triangulations, α-complexes and filtrations) are used for manifold reconstruction. With these techniques it is possible to infer geometrical and topological properties of data points that uniformly sample a single manifold embedded in a higher dimensional space (Delingette, 1994). Simplicial complexes have been widely used in Computer Graphics's task, such as object reconstruction from 3D-scanners (H. Delingette, 1997) and RGB videos (C.-H. Lin et al., 2019), or face reconstruction from RGB images (Zollhöfer et al., 2018). While such techniques are potentially powerful and theoretically well grounded, their capability to naturally handle high-dimensional noise aligned along low-dimensional manifolds is limited. Besides the sensitivity to the noise issues, it is often required that the intrinsic data dimensionality is known a priori.

Riemannian Manifold Learning (RML, Tong Lin et al. (2008)), presented as an update to the algorithm introduced in Tony Lin et al., 2006), first builds a simplicial complex of the manifold using the available points. For each local neighbourhood it then constructs a

normal coordinate chart, computing the local tangent space via PCA. While the previous version of the algorithm used Dijkstra's shortest path algorithm to approximate geodesic distances on the simplicial complex, the new version uses a second-order polynomial curve, with weights estimation via least-squares.

2.2 Density estimation and Model Selection

Density estimation is the task of recovering (or more realistically approximating) the unknown true distribution that generated the observed data. As mentioned earlier, PCA is based on the assumption that the distribution of data can be in first analysis described by a multivariate Gaussian. The respective Principal components and their variances are nonother than the eigen-vectors and eigen-values (respectively) of the Gaussian's Covariance matrix. In this sense, the underlying, unknown, true distribution of the data is approximated by a multivariate Gaussian distribution. However, when data is distributed in a more complicated way, e.g. lies on a low-dimensional non-linear manifold, this assumption is too restrictive and a more general framework is needed.

2.2.1 Finite Gaussian Mixture Models

Finite Gaussian Mixture models (McLachlan et al., 1988; Bishop, 2006; McLachlan et al., 2019), hereon referred to as GMM, attempt to approximate a general distribution with a linear combination of Gaussians. The positioning (mean) and size (covariance matrix) of each "component", as well as the mixing weights of the linear combination can be adjusted to provide complicated distributions at will. In particular, consider data set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} = \mathbf{X} \in \mathbb{R}^n$, modelled as a GMM with K components. Each component is identified by its mean ($\boldsymbol{\mu}_k$) and covariance matrix ($\boldsymbol{\Sigma}_k$), plus its dedicated mixture coefficient π_k. The global

density model is then obtained by:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k), \tag{2.1}$$

where $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0,\ k = 1, \ldots, K$, forming a convex combination of the components, and

$$p(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) = \left[(2\pi)^n |\mathbf{\Sigma}_k|\right]^{-\frac{1}{2}} \exp\left[(\mathbf{x} - \mu_k)^\top \mathbf{\Sigma}_k^{-1} (\mathbf{x} - \mu_k)\right] \tag{2.2}$$

is the *probability density function* (PDF) of gaussian component k. However, when interested in recovering an unknown distribution from observed data, it is possible to train GMM via optimization algorithms, such as the Expectation-Maximization algorithm Dempster et al., 1977, applied on the log-likelihood as the cost function to be minimized:

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) = \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k) \tag{2.3}$$

Despite its versatility, GMM has one fundamental flaw when applied to density modelling: the estimation of the number of the hyperparamters (i.e. number of components).

2.2.2 Model Selection

In order to establish which set of hyper-parameters best models the data distribution, a number of criteria have been developed to aid the choice. In particular, all these criteria may be interpreted as the realization of Ockham's razor, *"Numquam ponenda est pluralitas sine necessitate"*: Plurality must never be posited without necessity. This principle, despite it not being rigorously derived within a scientific framework, constitutes an approach in model formulation generally used in science and extended to physics by Ernst Mach Becher, 1905. In the context of probabilistic modelling, we can interpret Ockham's razor principle as: *the simplest model that minimizes a cost function is to be preferred to any other more complex*

model. Where by simplicity and complexity we refer to the number of hyper-parameters. This principle is incorporated in mainly two criteria for probabilistic modelling, both depending on the log-likelihood of the data w.r.t. the model (eq. (2.3)): the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). The main difference between the two is the importance of the model's hyper-parameters: models with a higher number of hyper-parameters are more penalized by BIC than by AIC.

These two goodness-of-fit measures are generally used in the comparison between probabilistic models. However, in order to avoid over-fitting, a preemptive optimization procedure is required. This methodology consists in splitting randomly the data set into three or more (in the case of cross-validation) sub sets. Each subset is considered as an independent realization of the unknown distribution that generated the original data set. When the data set is large, it is generally sufficient to divide it randomly into three subsets, called *Training* ($\sim$ 50%), *Validation* ($\sim$ 25 %) and *Test* (or *holdout*, $\sim$ 25%) sets. Given M GMMs, each one having a fixed number of components, the *Training* set is used to obtain the Maximum-Likelihood-Estimates (MLE) of each model's parameters $(\boldsymbol{\pi}^m_{MLE}, \boldsymbol{\mu}^m_{MLE}, \boldsymbol{\Sigma}^m_{MLE}$, for $m = 1, \ldots, M)$. The MLE of the parameters are obtained via any optimization algorithm, i.e. Expectation-Maximization. The performance of each model is then evaluated on the *Validation* set. It is in this step that AIC/BIC come to fruition, since models with different number of components (thus with a different "architecture") have to be compared. After the model with the best performance on the *Validation* set has been found, it can be evaluated on the *Test* set, obtaining an unbiased measure of the model's performance on the data set.

This approach generally assumes that the data set is populated enough to sample the unknown distribution equally in all three sub sets, once it is split. In the case where this condition does not apply, $k-$fold cross-validation is advised Hastie et al., 2001. The available data is split into k (equally populated) parts, generally $k = 5$ or $k = 10$. The model is trained and evaluated k times, where at each iteration a different sub set is used

for validation and the remaining $k-1$ sets for training. Finally, the performance is obtained by averaging on the k trials.

2.2.3 Automatic GMM model selection

In some cases, model selection is not required for approximating the unknown distribution sampled by a set of observations. Parzen–Rosenblatt Window Parzen, 1962; Rosenblatt, 1956, or Kernel-Density Estimation (KDE) assumes that it is possible to estimate the underlying distribution of available data by placing a smoothing kernel $\Phi(x)$ on each observation. The kernel may vary, but a spherical Gaussian density function is often used. By spherical we mean a Gaussian distribution depending only on the mean and variance parameters. A non-parametric estimate of the unknown distribution is then obtained over the sample's domain, by:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{\ell=1}^{L} \mathcal{K}\left(\frac{x - x_\ell}{h}\right). \tag{2.4}$$

Here, h is the *bandwidth* and $\mathcal{K}(x) = e^{-x^2/(2h^2)}$ the Gaussian kernel. In the case of a spherical Gaussian kernel, the bandwidth is usually related to the variance σ. However, particular care must be taken when choosing the bandwidth value. In fact, if h is chosen too small, the density estimate could be noisy and over-fit the data, while if h is too large, the final estimate will be overly smooth, losing most of the fine information in the data. In general, it is arguable that KDE rejects the hypothesis of a model selection, since the global distribution is determined by the selected sample.

A variation of KDE when data is aligned along a lower $j-$dimensional manifold can be found in Manifold-Parzen window Vincent et al., 2003. As in the original Parzen window, all data points are used as centers of the GMM components, however, unlike the seminal work, the covariance matrix of each component is estimated locally. In particular, by decomposing the local covariance matrix of each data point's neighbourhood, the updated covariance ma-

trix is obtained by retaining only the first j eigenvectors (corresponding to the j highest eigenvalues) and computing the sum of their outer products. In order to limit degeneracy of the updated covariance matrix, a non-singular term is added to its diagonal. While mathematically well defined, (Manifold-) Parzen Window is generally computational intensive, in particular for large data sets. Fast-Parzen window X. Wang et al., 2009 aims at solving this issue by segmenting the input data set into hyper-balls, uniformly covering the whole data space. For each hyper-ball, the center of the kernel is overimposed to the local sample mean and the kernel's covariance matrix estimated from the neighborhood. The density estimate is then obtained as a GMM where each component of the mixture is the kernel centered on the corresponding hyper-ball.

Infinite Gaussian Mixture Model Rasmussen, 1999 introduces a Bayesian formulation of Gaussian Mixtures. In this sense, each component's parameters and hyperparameters have an associated prior. Through a Dirichlet process Ferguson, 1973, components of the mixture are automatically created (by sampling from the hyper-parameter's priors) or suppressed (when they are non-representative of a sufficient number of observations). A similar approach is presented in Figueiredo et al., 2002. The drawback of these methodologies is the high computational cost, since multiple Gibbs samplings need to be performed over the parameters' priors.

An alternative way of automatically estimating the number of components in a Mixture Model is given by Hierarchical GMMs. In general, two main strategies are implemented for the identification of the best number of components. The bottom-up (agglomerative) strategy (Goldberger et al., 2005; Garcia et al., 2010) starts by assuming each presented input point is the center of a Gaussian (to speed it up, the number of components can also be initialized by $K-$means with a large number of clusters). Through an iterative process, components of the Mixture are merged based on a dedicated measure (i.e. Kullback-Leibler

divergence, Bregman dvergence). The top-down (divisive) approach (Olech et al., 2016; Li et al., 2018), starts by assuming that all data is represented by a single Gaussian component. The initial component is split in accordance to a dedicated criterion. By applying this procedure recursively, the goodness-of-fit measure generally increases, together with the model complexity, until all data points are centers of a Gaussian component (or a maximum number of components is reached). In both cases, a hierarchy (tree) of GMMs is formed, where each level has the mixture components as nodes. The corresponding dendrogram gives a clear representation of the tree's structure. By evaluating the out-of-sample likelihood recursively at each level of the tree, one can quantitatively estimate the performance of the hierarchy. In the bottom-up procedure, it is often the case that the goodness-of-fit measure has a slow decrease in the initial iterations, however, after a sufficient number of components has been merged, the decrease rate (slope of the function) starts growing, forming a 'knee' on the graph. In the top-down strategy, this behaviour is reversed. The knee is usually considered as the level of the hierarchy where a good compromise between complexity of the model and representation of the data is reached. Based on this metric, the user is able to perform an informed choice on the most suitable level of the hierarchy to retain.

While all these methodologies may reach a good understanding of the global distribution of points in a data set, none of them is able to give a cohesive representation of data if concealed in a low-dimensional structure. In this case, constraining the centers of the components to lie onto a low-dimensional (possibly non-linear) subspace represents a more intuitive solution than estimating the best number of components in the mixture. This approach has been pursued in the Generative Topographic Mapping (GTM, Bishop et al., 1998b).

2.2.4 Generative Topographic Mapping

Generative Topographic Mapping (GTM) was proposed as a probabilistic formulation of the Self-Organizing Map (Kohonen, 1982). Its main advantage is that instead of treating a noisy manifold as a core low-dimensional manifold to be discovered, plus some "noise" around it that somehow needs to be dealt with, it formulates a consistent manifold-aligned density model in the form of a constrained mixture of Gaussians[1]. The original GTM formulation is trained in the maximum likelihood framework using the E-M algorithm (Dempster et al., 1977). Because of the sensitivity to initialization, a suitable initialization is required (e.g, using PCA).

GTM is a latent variable model, meaning that the D dimensional space in which the distribution $p(\mathbf{t})$ lives is represented by a set of L latent variables $\mathbf{x} = (x_1, x_2, ..., x_L)$. The connection between the latent and the data space is performed trough a non-linear mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ where $\mathbf{W}$ are the parameters governing the mapping. If a prior distribution $p(\mathbf{x})$ is defined over the latent-variable space, a distribution $p(\mathbf{y}|\mathbf{W})$ is induced over the data space. The authors chose the conditional distribution over the data space given the latent variables and the parameters $\mathbf{W}$, to be a radially symmetric Gaussian distribution with variance β^{-1}, namely:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \Big(\frac{\beta}{2\pi}\Big)^{D/2} \exp\Big\{ -\frac{\beta}{2}||\mathbf{y}(\mathbf{x}; \mathbf{W}) - \mathbf{t}||^2 \Big\} \tag{2.5}$$

Further developments of GTM (Bishop et al., 1998a) propose a noise model aligned along the principal directions of the manifold. The distribution over the data space for a set of

[1]Location parameters (means) of the Gaussian components are constrained to lie on a smooth manifold - most commonly a smooth image in the data space of a j-dimensional interval (latent space).

parameters $\mathbf{W}$, is obtained by integrating out the dependency on the latent variables:

$$p(\mathbf{t}|\mathbf{W},\beta) = \int p(\mathbf{t}|\mathbf{x},\mathbf{W},\beta)p(\mathbf{x})d\mathbf{x} \tag{2.6}$$

The computation of the best parameters $\mathbf{W}$ over a data set $\mathcal{D} = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$ can then be performed by maximizing the log likelihood:

$$\mathcal{L}(\mathbf{W},\beta) = \ln(\prod_{n=1}^{N} p(\mathbf{t_n}|\mathbf{W},\beta)) \tag{2.7}$$

The prior distribution over the latent space plays an essential role in the computation of the integral in equation 2.6. In general, this integral is not analytically solvable, but if we choose the prior to be a sum of Dirac's delta functions over a regular grid in the latent space,

$$p(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K}\delta(\mathbf{x} - \mathbf{x}_i) \tag{2.8}$$

then the integral in eq. 2.6 can be computed as:

$$p(\mathbf{t}|\mathbf{W},\beta) = \frac{1}{K}\sum_{i=1}^{K} p(\mathbf{t}|\mathbf{x}_i,\mathbf{W},\beta) \tag{2.9}$$

The log likelihood function in 2.7 then becomes:

$$\mathcal{L}(\mathbf{W},\beta) = \sum_{n=1}^{N}\ln\Big\{\frac{1}{K}\sum_{i=1}^{K} p(\mathbf{t}_n|\mathbf{x}_i,\mathbf{W},\beta)\Big\} \tag{2.10}$$

A schematic representation of this method is shown in fig. 2.1 for the case of a two dimensional non-linear manifold embedded in a three dimensional Euclidean space as presented in Bishop et al., 1998b. The assumed noise model in eq. 2.5 implies that the distribution $p(\mathbf{t}|\mathbf{W},\beta)$ is a constrained GMM. Indeed, the points on the latent grid are not able to move independently as they're linked through the function $\mathbf{y}(\mathbf{x};\mathbf{W})$. Furthermore, when this func-

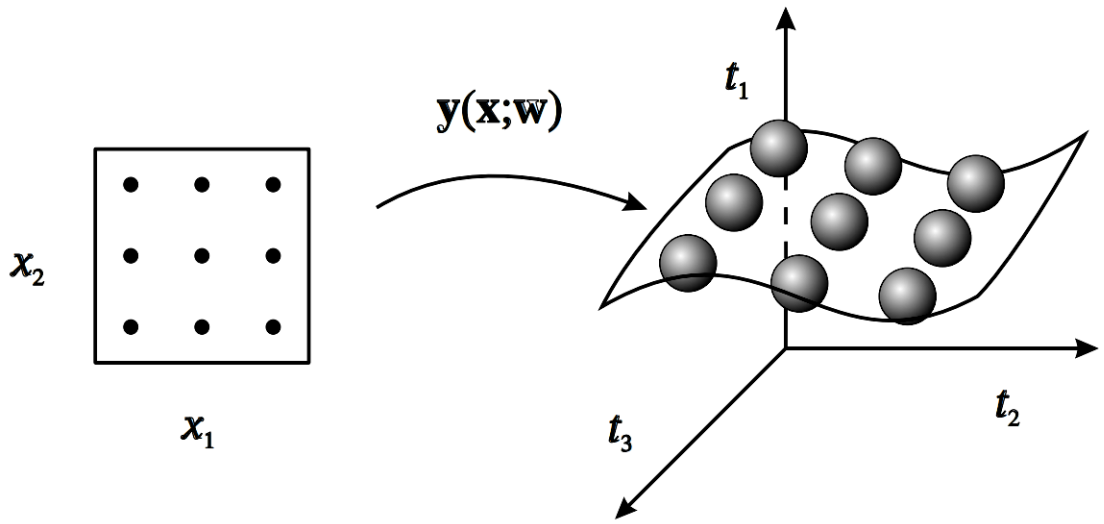


Figure 2.1: Adapted from Bishop et al., 1998b: GTM construction model for a two dimensional non-linear manifold embedded in a three dimensional Euclidean space. The regular grid in the latent space (on the left) is mapped through the mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ onto the data space, where a spherical gaussian noise model is assumed.

tion is chosen to be differentiable and continuous with respect to the parameters $\mathbf{W}$, it maps close points in the latent space to close points in the data space, preserving the topological ordering of data.

The maximization of $\mathcal{L}(\mathbf{W}, \beta)$ is performed through the *Expectation-Maximization* (EM) algorithm, formally introduced in the context of likelihood maximization by Dempster et al., 1977. Being GTM a Gaussian Mixture by nature, this choice is particularly effective, as analytical forms of the parameters updates can be derived. Furthermore, by choosing a set of radial basis M fixed basis functions $\phi_j(\mathbf{x})$ such that

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}\phi(\mathbf{x}) \tag{2.11}$$

$\phi(\mathbf{x})$ being the collection of $\phi_j(\mathbf{x}), j = 1, ..., M$ and $\mathbf{W}$ a $D \times M$ matrix of weights, the M-step in the EM algorithm consists in solving a set of linear equations, given a suitable choice for the basis functions. In this case, the basis functions are radially symmetric Gaussians with centres distributed on a uniform grid on the latent space. The smoothness of the manifold

is then determined by the prior knowledge on the weights in matrix **W** as well as on the number and spacing of the basis functions. Bayesian formulations of GTM have also been proposed Olier et al., 2008.

2.3 Multiple manifolds learning

To deal with complex data sets, where several manifolds of different dimensionalities can co-exist, generalizations of the previous methods have been developed: Multi-Manifold Discriminant Analysis (MMDA, Yang et al., 2011), Multi-Manifold Isomap (M-ISOMAP, Fan et al., 2012), Multi-manifold Proximity Embedding (MPE, Fan et al., 2016), Multi-Manifold LLE (MM-LLE, Hettiarachchi et al., 2015), S-Isomap++, (Mahapatra et al., 2017), Hierarchical GTM (Tino et al., 2002). However, based on the same assumptions as their predecessors, the methods still need to be informed about the dimensionalities of the different manifolds and struggle when dealing with topologically complex, noisy structures.

Other generalizations of PCA that take into consideration multiple coexisting structures over the same space have also been proposed (Generalized PCA, Vidal et al., 2016). However, as per its predecessor, such structures are represented as linear sub-spaces. This inheritance, makes the methodology fail in recovering non-linear low-dimensional manifolds. Sparse-Manifold Clustering and Embedding (SMCE, Elhamifar et al., 2011) aims at forming a local low-dimensional hyper-plane (patch) over the tangential space to a manifold by selecting few points within a fixed size neighbourhood of any given point belonging to the manifold. This is achieved via two optimization processes, both penalizing the contribution of points farther away from the selected center of the hyper-plane. The contribution of every data point to the patch centered on every point in the data are used as entries in a similarity graph. Ideally, the corresponding similarity matrix would have a block-diagonal

form. Different blocks of the matrix can then be associated to different components in the data space, resulting in a clustering of data points based on their mutual inter-dependence. However, as shown later in this work, the similarity matrix quickly loses the block-diagonal shape when even little transverse noise to the manifold is added. Furthermore, due to the imposed sparsity of local neighbourhoods, the global shape of the manifold can be distorted when embedded in its native dimensional space. As a way of solving this problem, Low Rank Neighborhood Embedding (LRNE, Saranathan et al., 2019) aims at preserving both local and global properties of the manifold by minimizing the dimension of the neighborhood of each point. While obtaining better results than SMCE w.r.t. the quality of the recovered embedding, its performance is greatly altered by transverse noise.

The estimation of a correlation dimension as proposed in Grassberger et al., 1983, despite being well rooted in the framework of fractal dimensions (Mandelbrot, 1967), is not suitable in the case at hand. In fact, the intended use of point-estimate dimensionality in this Thesis, is for clustering points according to their "preference" of belonging to a low-dimensional structure of dimension j. As such, the methodology has to provide an integer value representing the dimension of the structure to which an individual point would be assigned to, if the structure had no transverse noise. Since the typical data sets analyzed in this work contain a number of manifolds of different dimensionalities, a local (point-wise) dimensionality estimation is preferred to a global one. The works proposed in Haro et al., 2000 and Allegra et al., 2020, are particularly relevant in terms of dimensionality estimation and clustering. The first methodology relies on the construction of *Translation Poisson Mixture Models* (TPMM) for estimating the dimensionality of local neighbourhoods in the data. However, by design, it is prone to separating a unique manifold if the local density of points changes throughout the manifold. The second methodology, *Hidalgo*, is a Bayesian extension of *TWO-NN*[2] (Facco et al., 2017). The dimensionality of individual

[2]The methodology is based on distances to the two nearest neighbors of each point.

points is obtained by maximization of the posterior distribution over the model's parameters. Despite the appealing formulation, the estimated complexity of $\mathcal{O}(N^2)$, N being the number of points in the data set, is not well suited for applications considered in this study, where N is generally large.

Part I

Methodology

Chapter Three

Adopted data sets

3.1 Astronomical numerical simulations

In the following, a concise description is provided of the scientific data sets where the methodology reported in this work has been successfully applied. Two data sets are analyzed, obtained with different techniques and describing very different objects. Their corresponding numerical schemes are first summarized, highlighting their main properties, and later briefly the simulated objects are described.

3.1.1 Smoothed Particle Hydrodynamics (SPH)

The Smoothed Particle Hydrodynamics (SPH) scheme Gingold et al., 1977 is inspired by the probabilistic modelling technique Parzen window Parzen, 1962. The latter, proposes a methodology for estimating the Probability Density Function (PDF) of a population based on a finite sample. This is achieved by placing a Gaussian smoothing kernel on each sampled point and convolving all kernels on the desired query set.

In the same way, SPH techniques sample the (to be simulated) volume with a finite number

of "particles", each carrying information about many different properties of the surrounding space. Since this methodology is aimed at simulating hydrodynamical systems, the properties sampled by each particle usually describe the state of the fluid (in this case, gas) in its close vicinity. Thus, quantities such as Mass (m), Temperature (T), Density (ρ) and "Metallicity"[1] ($[Fe/H]$) are taken into consideration. While this is a general setup for SPH simulations, the details of the simulation may vary depending on the implementation. In order to keep the work coherent, SPH scheme of the code GADGET2 Springel, 2005 (and its update GADGET-3) is here reported, being the one used for the simulations at hand.

In GADGET2, the location of each particle is identified by Cartesian coordinates $\mathbf{p}_i = (p_i^1, p_i^2, p_i^3)$. A finite-support approximation (spline) to the Gaussian Kernel is introduced at the particle's location:

$$W(r, h_i) = \frac{8}{\pi h_i^3} \begin{cases} 1 - 6\left(\frac{r}{h_i}\right)^2 + 6\left(\frac{r}{h_i}\right)^3 & \text{if} \quad 0 \leq \frac{r}{h_i} \leq \frac{1}{2]} \\ 2\left(1 - \frac{r}{h_i}\right)^3 & \text{if} \quad \frac{1}{2} \leq \frac{r}{h_i} \leq 1 \\ 0 & \text{if} \quad \frac{r}{h_i} > 1 \end{cases} \tag{3.1}$$

where h_i is the distance between the particle and its $50-$th neighbour (*smoothing length*) and r is the distance between the particle and any point in the volume. The continuous quantities of a fluid are then obtained by kernel interpolation over the simulated volume. Given a particle at position $\mathbf{p}_i$, having smoothing length h_i, its density can be defined as:

$$\rho_i = \sum_{i=1}^{N} m_j W(|\mathbf{r}_{i,j}|, h_i). \tag{3.2}$$

[1]Metallicity here is intended as the mass proportion between Hydrogen and heavier elements, with respect to the same quantity measured in the Sun. This over-simplification is generally accepted in astronomy, especially in numerical simulations.

If interested in estimating the value of a certain quantity A in a specific location $\mathbf{x}_j$ within the simulated volume, given the whole set of simulated particles, $A(\mathbf{x}_j)$ is computed as:

$$A(\mathbf{x}_j) = \frac{\sum_{i=1}^{N} A_i \frac{m_i}{\rho_i} W(|\mathbf{r}_{i,j}|, h_i)}{\sum_{i=1}^{N} \frac{m_i}{\rho_i} W(|\mathbf{r}_{i,j}|, h_i)}, \tag{3.3}$$

where $|\mathbf{r}_{i,j}|$ is the distance modulus between particle $\mathbf{p}_i$ and location $\mathbf{x}_j$. A detailed study of interpolated quantities over particular regions of the volume is provided in section 7.

In the SPH formulation, particles are considered to be fluid elements and as such, their motion is described by a Langrangian interpretation of fluids. During the evolution of the system, the particles positions are updated via an equation of motion that takes into consideration the gravitational potential, density, momentum and pressure of each fluid element.

"Jellyfish" galaxy

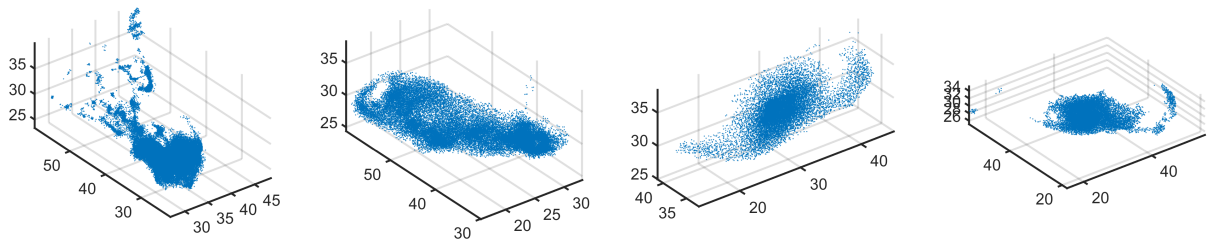


Figure 3.1: Four different orbital stages of a single simulated dwarf galaxy in interaction with a Fornax-like galaxy cluster's halo. Note the typical filamentary structures departing from the "head" and forming the "tail" of the galaxy.

This work includes the analysis of a simulated "Jellyfish" galaxy via the SPH code GADGET2. The name Jellyfish is given to any observed galaxy having extended one-dimensional manifolds (streams) departing from a compact core (head). The streams are usually located within the "tail" of the galaxy, whose origin is still uncertain. Many of these

exotic objects have been associated with "dwarf" galaxies, interacting with the hot gas of their hosting galaxy cluster. The formation of the tail can then be attributed to a combination of hydrodynamical and tidal effects. In order to study the mechanisms underlying the formation of these complicated structures, SPH numerical simulations via GADGET2 have been performed Mastropietro et al., 2021. The simulations initially consider a dwarf galaxy evolved in isolation for a certain amount of time. The term "isolation" refers to the environment in which the galaxy has evolved throughout its life. In this case, during this initial evolutionary stage, the galaxy hasn't been undergoing mergers or any kind of interactions with neighbouring galaxies or hosting gravitational potentials. In this sense, the data set obtained from a single run of the simulation represents a young, fully formed dwarf galaxy. By varying the initial parameters, the population of simulated galaxies closely resembles the statistical properties of the dwarf galaxies observed in the Virgo-cluster. A full catalogue and detailed study of these galaxies can be found in Verbeke et al., 2015; Verbeke et al., 2017.

Taking the end product of this initial evolutionary stage, Mastropietro et al., 2021 study the evolution of these galaxies when shot over different orbits in the gaseous halo of a Fornax-like galaxy cluster. During this stage, the formation of filamentary structures is often found in different orbital epochs (see fig. 3.1 for a few examples). In the following analysis, a single temporal snapshot of a dwarf galaxy evolving through a generic orbit is considered. An example of different orbital stages of the galaxy's evolution as obtained via GADGET2 is shown in figure 3.1. Numerous filamentary structures can be identified departing from the head of the "Jellyfish", forming one or more tails. Given the temporal nature of the data set and the application of the Moving Box technique, the number of particles in the simulation tends to gradually decrease with time. This is due to the fact that particles may get pushed outside of the simulated volume by mostly dynamical processes. The Box is carefully constructed in order to minimize this information loss, while following the trajectory of the simulated object along its orbit. Finally, the particles lost during the evolution are mostly

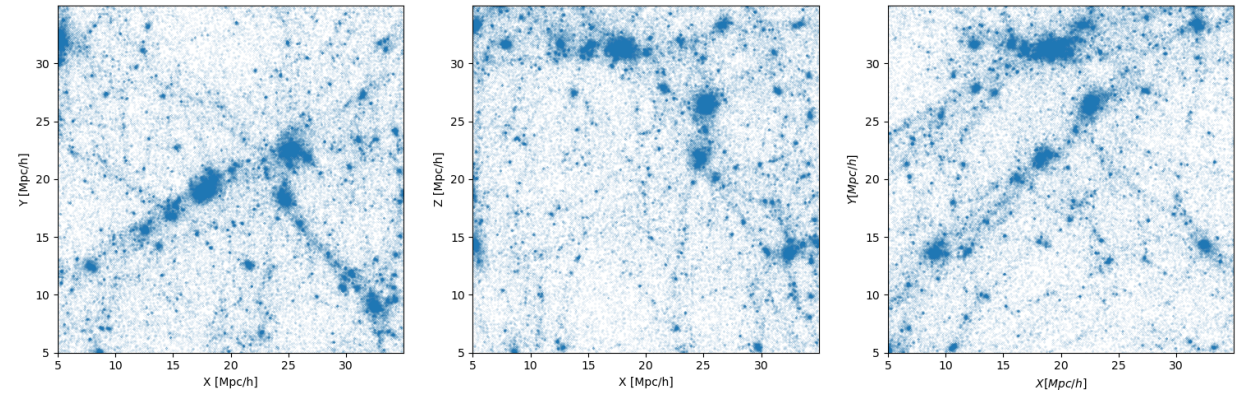
the ones defining the outskirts of the ambient system (gas of the galaxy cluster) and do not affect the properties of the low-dimensional structures. The minimum, maximum and mean number of particles over the whole set of snapshots is given in table 3.1, together with the number of attributes per particles and the specific attributes considered in this work.

Cosmic Web

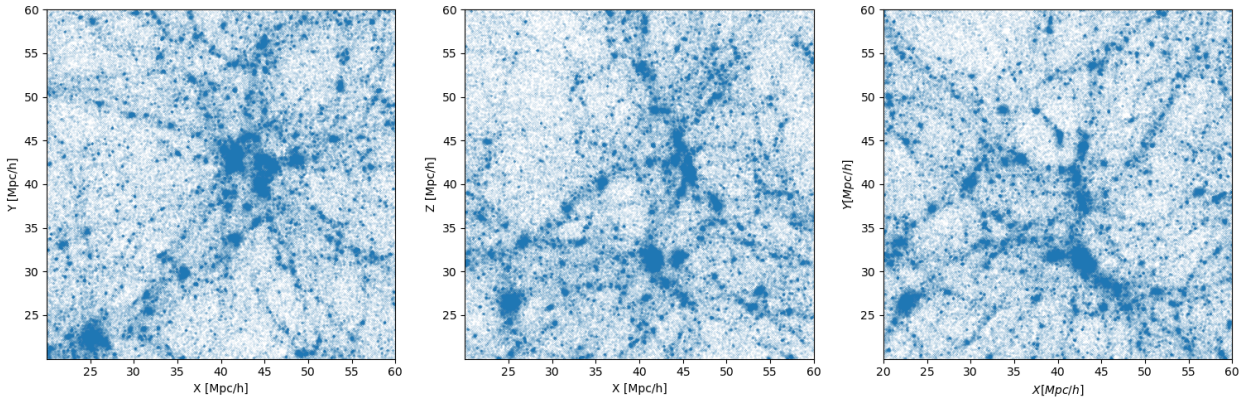
As mentioned earlier, a jellyfish galaxy can be formed in the interaction of a dwarf galaxy with the host galaxy cluster (e.g. the Fornax galaxy cluster). This interaction is due to the strong gravitational pull exercised by the cluster on the surrounding environment. The size of this system can be characterized by the mean orbital radius of the dwarf galaxy, roughly $\sim 10^2$ kpc Mastropietro et al., 2021, or[2] $\sim 3 \times 10^{18}$ km, and the estimated radius of the cluster, ~ 0.7 Mpc Iodice, E. et al., 2019. At larger scales ($\sim 10 - 10^3$Mpc), the distribution of matter is far more complicated. Within the Λ Cold Dark Matter model + inflation, matter is thought to assemble into filaments and walls, delimiting void regions. The distribution of matter at these scales is usually referred to as "Cosmic Web", where the web's threads are the main channels for the accretion of matter into galaxy clusters (and super-clusters) Bond et al., 1996. This highly non-uniform distribution of *visible* matter is supposedly guided by that of Dark Matter. This phenomenon has been tested in dedicated cosmological numerical simulations.

In particular, the analyzed simulation models a $120 \times 120 \times 120$ Mpc^3 volume sampled by only Dark Matter particles, each one having mass $1.072 \times 10^9 M_\odot/h$. The simulation is performed via GADGET-3. The volume is modelled at 5.713 Myr after the Big Bang and evolved until present day, for a total of ~ 13.772 Gyr with the assumed cosmological model. The data set consists of positions and velocities of all the simulated Dark Matter particles at present time. Figure 3.2 shows different samples projected onto pairs of the axis, of different regions

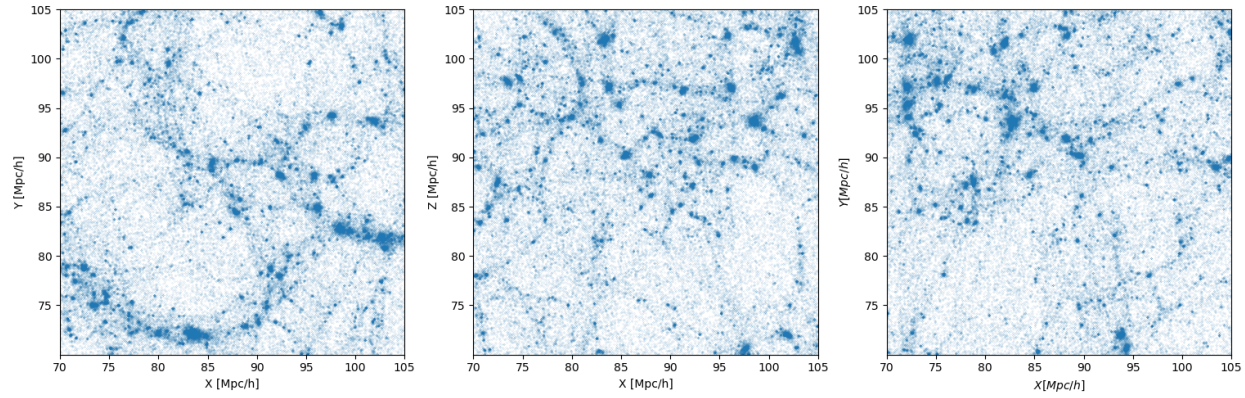
[2]A kpc is an astronomical measure of distance and it corresponds to 3.26 light years or $\sim 3 \times 10^{16}$ km.



(a)



(b)



(c)

Figure 3.2: Three samples from the simulated volume, projected on two of the three axis, showing clear intertwined one-dimensional structures. The snapshots are taken at the final stage of the simulation (current time).

of the same volume. Clear filamentary structures, forming an intricate web, are visible in all images and are the results of the cosmic volume's evolution. A brief summary of the numerical characteristics of the adopted data set is given in table 3.1 where the number of particles in the whole simulation and their attributes are reported. Here, the number of particles in the portion of cosmic volume selected for the analysis is also presented.

Table 3.1: Number of particles and attributes of scientific data sets

Scientific study	**n. points**	**n. attributes**	**Used attributes**
Jellyfish galaxy	min = 56619 max = 130137 mean = 86138	25	$x, y, z, \rho, T, [Fe/H], m, [C\ \text{II}]$
Cosmic Web	Total = 1.34217728×10^8 Sampled = 2.097152×10^6	6	x, y, z, v_x, v_y, v_z

3.2 Synthetic data sets

Throughout this work $\mathcal{Q}^k$ is the $k-$th data set given by the union of the sampled noisy manifolds, whose respective points are collected in data sets $\mathcal{D}^\ell$. Consider a point cloud $\mathcal{Q}^1 = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_L\}$, $\mathbf{t}_i \in \mathbb{R}^d$ containing points sampled from an unknown number of noisy lower-dimensional manifolds embedded in a noisy environment (e.g., points generated from a broad d-dimensional distribution). As an example, fig. 3.3m shows a data set in $\mathbb{R}^3$ obtained from a collection of noisy two-dimensional (central panel) and one-dimensional (left panel, red points) manifolds. All underlying true manifolds are shown in fig. 3.3: the four two-dimensional manifolds (central and bottom row panels: S-shape surface, cap (parabolic surface), toroid and Möbius strip) and the two one-dimensional ones (top row: hyperbolic arm and spiral). The noisy manifolds are embedded in a uniform noise (fig. 3.3m, rightmost panel). A uniformly sampled ball (fig. 3.3m, left (green points) is also included. Characteristics of the manifolds can be found in table 3.2. There, the intrinsic dimension, the generative

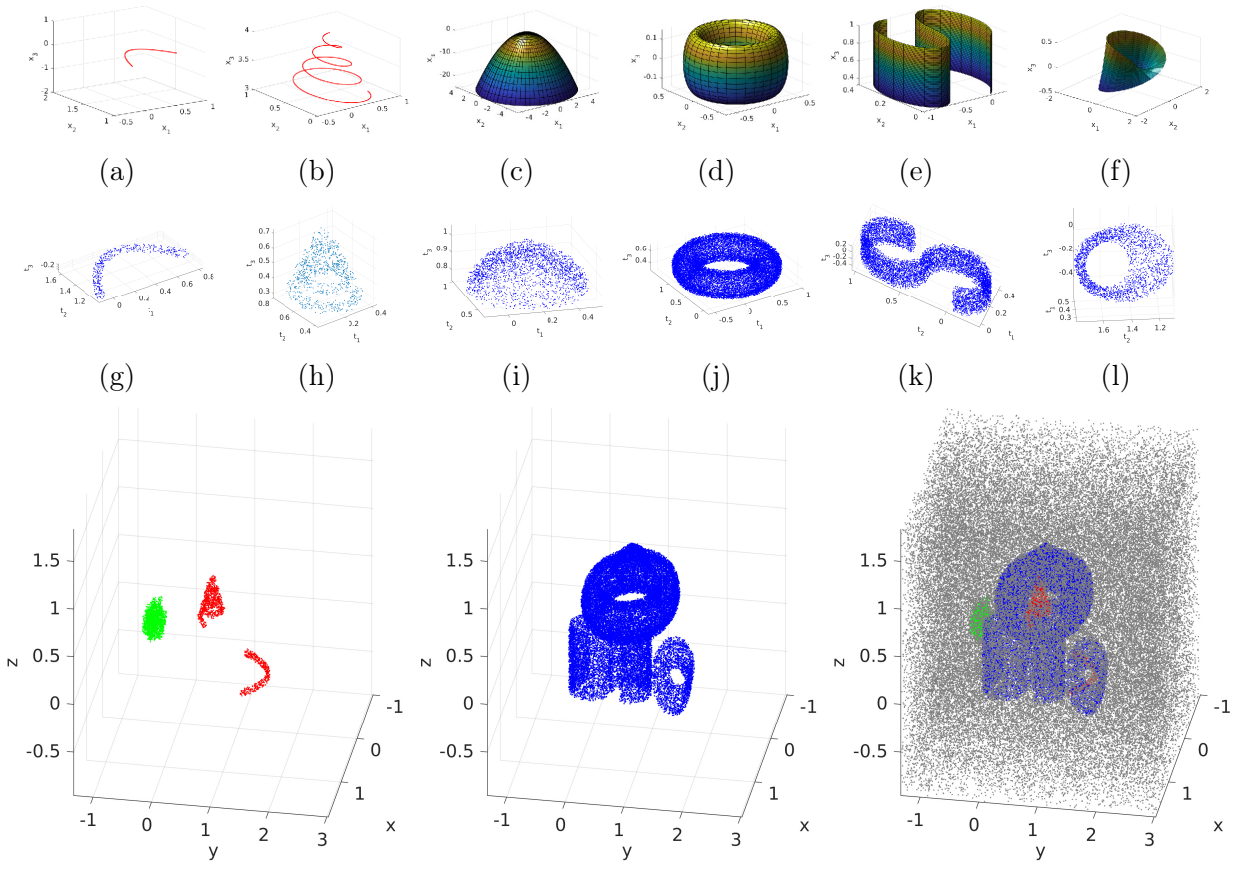


Figure 3.3: Subfigures 5.1b - 5.1f: Ground-truth for each manifold used in dataset $\mathcal{D}$; 3.3h - 3.3l Corresponding noisy manifolds. Subfigure 3.3m: One- and three-dimensional noisy manifolds (left panel), two-dimensional noisy manifolds (central panel) and complete data set $\mathcal{Q}^1$, background noise included (grey points).

clean-distribution and the number of points used for sampling the respective noisy distributions are shown. The union of all points sampling manifolds is $\sim 33\%$ the total number of

Table 3.2: Manifolds' intrinsic dimensionality, underlying distribution and number of points.

Manifold ID	**Dim**	**Distribution**	**# points**
1	1	Spiral Arm	689
2	1	Parabolic arm	424
3	2	"S" surface	8307
4	2	toroidal surface	18959
5	2	Möbius strip	2338
6	2	Hyperbolic surface	2297
7	3	Gaussian point cloud	1958
Summed over manifolds			34972
Background noise			70463
Dataset total			105435

points in the data set. The residual $\sim 67\%$ forms the background noise.

While the purpose of data set $\mathcal{Q}^1$ is mainly to show how to filter out background noise and recover a notion of dimensionality for the residual points, it is not particularly challenging from a topological point of view. In order to verify the capabilities of the proposed methodology when faced with a topologically complicated data set data set $\mathcal{Q}^2$ is constructed by sampling from three interlocked, noisy toroids (figure 3.4). In order to avoid redundancies in the description of the methodology background noise is not considered in data set $\mathcal{Q}^2$,

3.2.1 Manifolds equations and data sets generation

All synthetic data sets documented previously are generated by first defining parametric functions of the "Ground Truth" manifolds, then adding noise to the manifolds, rearranging them via affine operators (scaling, rotation and shift) and adding a background noise. In the following, the parametric equations for each structure is shown and the noise addition process described. The choice of the adopted manifolds is due to particularly interesting

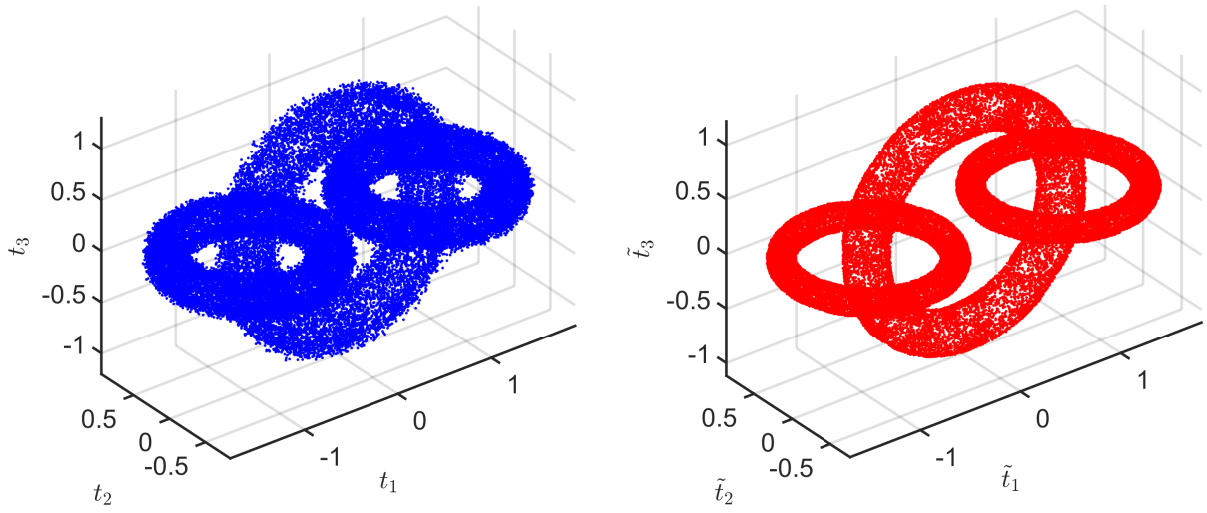


Figure 3.4: Data set $\mathcal{D}^2$ with the three interlocked noisy toroids (right panel) and its diffused and filtered counterpart $\tilde{\mathcal{D}}^2$ (left panel).

features, either geometrical or topological. In particular, the two one-dimensional manifolds have been chosen because of their low ($\mathcal{M}_1$) and increasing ($\mathcal{M}_2$ curvatures respectively). Manifolds $\mathcal{M}_3$ and $\mathcal{M}_4$ are chosen as low- and high- curvature two-dimensional manifolds in the case of a simple topology, since they are both isomorphic to $\mathbb{R}^2$. The choice of manifolds $\mathcal{M}_5$ and $\mathcal{M}_6$ is due to their peculiar topology, being both closed manifolds but respectively orientable and non-orientable. Finally, $\mathcal{M}_7$ is chosen as a prototypical three-dimensional manifold in $\mathbb{R}^3$ and here introduced for completeness of the experiments with respect to the dimensionality index (sec. 5.2). Manifold $\mathcal{M}_8$ is the higher dimensional version of $\mathcal{M}_5$ and it is only used as a demonstrative example of the applicability of the methodology to manifolds with intrinsic dimension $j > 2$ in a higher dimensional space ($\mathbb{R}^4$). The positioning of different manifolds within the synthetic data sets is arbitrary and can be modified, as well as the number and type of manifolds included. The data sets are designed in such a fashion to test the proposed methodology in a general and non-trivial setting.

The functional forms are presented in the following order: Parabolic arm $\mathcal{M}_1$ (1D), Spiral $\mathcal{M}_2$ (1D), S-shape $\mathcal{M}_3$ (2D), Hyperbolic surface $\mathcal{M}_4$ (2D), Toroid $\mathcal{M}_5$ (2D), Möbius strip

$\mathcal{M}_6$ (2D), 3-ball $\mathcal{M}_7$ (3D) and 3-Toroid $\mathcal{M}_8$ ($j = 2$, embedded in $\mathbb{R}^4$).

$$\mathcal{M}_1 : \begin{cases} x \in [0; 1]; \\ y = -2x^2 + 2x + 1; \\ z = -0.2 \end{cases} \quad \mathcal{M}_2(\theta_1, r) : \begin{cases} x = \frac{1}{2}\left[r\cos(\theta_1) + 1\right] - 0.1; \\ y = \frac{1}{2}\left[r\sin(\theta_1) + 1\right] + 0.4 \\ z \in [0.1; 1]; \end{cases} \tag{3.4}$$

where $\theta_1 \in [2\pi; 10\pi]$ and $r \in [0.1; 1]$. In order to obtain the spiral data set, variables θ_1 and r have to be reordered in ascending order, while variable z in descending order.

$$\mathcal{M}_3 : \begin{cases} x = \frac{[\sin(\theta_2)+1]}{5}; \\ y \in [\frac{1}{3}; 1]; \\ z = \frac{\{\mathrm{sgn}(\theta_2)[\cos(\theta_2)+1]\}}{3}. \end{cases} \tag{3.5}$$

Here, $\theta_2 \in [-5; 3\pi - 5]$.

$$\mathcal{M}_4 : \begin{cases} x = R\sin(\theta_3) + 0.5; \\ y = R\cos(\theta_3) - 0.5; \\ z \in [0.4; 0.6] \end{cases} \quad \text{where} \quad \begin{cases} r \in [0; 0.5]; \\ R = -r^2 + 0.25; \\ \theta_3 \in [0; 2\pi]. \end{cases} \tag{3.6}$$

variables r, θ_3 and z must be sorted in ascending order.

$$\mathcal{M}_5 : \begin{cases} x = [R + r\cos(v)]\cos(u); \\ y = [R + r\cos(v)]\sin(u); \\ z = r\sin(v). \end{cases} \tag{3.7}$$

Where $R = 0.5, r = 0.15, (u, v) \in [0; 2\pi] \times [0; 2\pi]$

The Móbius strip is obtained by the equations:

$$\mathcal{M}_6 : \begin{cases} x = \left[1 + \frac{t}{2}\cos(\frac{u}{2})\right]\cos(u); \\ y = \left[1 + \frac{t}{2}\cos(\frac{u}{2})\right]\sin(u); \\ z = \frac{t}{2}\sin(\frac{u}{2}) \end{cases} \tag{3.8}$$

where $t \in [-1; 1]$.

Manifold $\mathcal{M}_7$ is the 3-ball of radius $r = 1$. The 3-torus, embedded in 4 dimensions (discussed in section 5.5) is parameterised by:

$$\mathcal{M}_8 : \begin{cases} x = [3 + \cos(\Phi)]\cos(\theta); \\ y = [3 + \cos(\Phi)]\sin(\theta); \\ z = [3 + \sin(\Phi)]\cos(\Psi); \\ \ell = [3 + \sin(\Phi)]\sin(\Psi); \end{cases} \tag{3.9}$$

where $(\theta, \Phi, \Psi) \in [0; 2\pi] \times [0; 2\pi] \times [0; 2\pi]$.

To generate data set $\mathcal{Q}^1$, a two rotation operators are needed, namely $\mathcal{T}_1$ and $\mathcal{T}_2$:

$$\mathcal{T}_1 : \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathcal{T}_2 : \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0. \end{pmatrix} \tag{3.10}$$

Furthermore, a set of shifting and scaling operators is designed for each manifold, in order to obtain the presented configuration:

$$\hat{\mathcal{M}}_1 = \mathcal{M}_1 + [-0.1; 0.1; 0]; \tag{3.11}$$

$$\hat{\mathcal{M}}_2 = 0.4 \times \mathcal{M}_2 + [0.1; 0.4; -0.9]; \tag{3.12}$$

$$\hat{\mathcal{M}}_3 = \mathcal{T}_1(\mathcal{M}_3) + [0; 0; 0.5]; \tag{3.13}$$

$$\hat{\mathcal{M}}_4 = 1.2 \times \mathcal{M}_4 + [-0.4; 1.3; 0.3]; \tag{3.14}$$

$$\hat{\mathcal{M}}_5 = 1.2 \times \mathcal{M}_5 + [0.2; 0.68; 0.5]; \tag{3.15}$$

$$\hat{\mathcal{M}}_6 = 0.25 \times \mathcal{T}_2(\mathcal{M}_6) + [0.4; 1.5; -0.2]; \tag{3.16}$$

$$\hat{\mathcal{M}}_7 = 0.2 \times \mathcal{M}_7 + [0.2; 0.5; 0]; \tag{3.17}$$

$$\tag{3.18}$$

The construction of data set $\mathcal{Q}^2$ is simpler and relies on specific rotations and shifts of manifold $\mathcal{M}_5$. In particular:

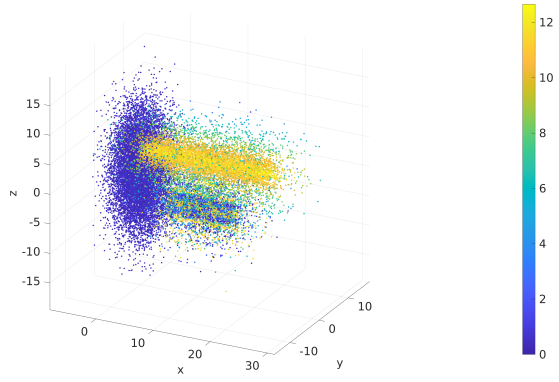
$$\hat{\mathcal{M}}_9 = \mathcal{M}_5 - [0.95; 0; 0]; \tag{3.19}$$

$$\hat{\mathcal{M}}_{10} = 1.4 \times \mathcal{T}_1(\mathcal{M}_5) + [0; 0; 0.12]; \tag{3.20}$$

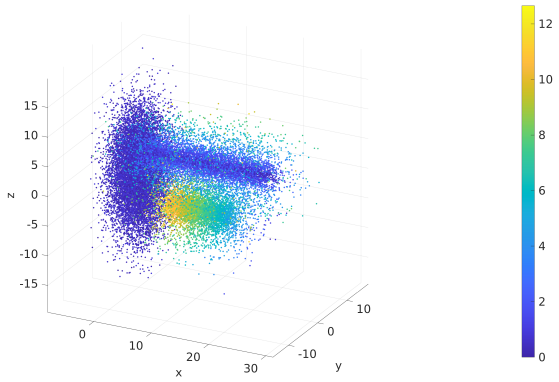
$$\hat{\mathcal{M}}_{11} = \mathcal{M}_5 + [1.9; 0; 0]. \tag{3.21}$$

For each manifold $\hat{\mathcal{M}}_i, i = 1, \ldots, 9$, a large number of $N_{GT} \sim 1e4$ points from its ground truth $\mathcal{M}_i$ is at first sampled, densely covering the whole structure. Around each sample point on the manifold a hyper-ball of fixed radius $r_{GT} = 0.04$ is positioned. Denote the union of the hyper-balls by $\mathcal{B}_{GT}$. To generate a sample from a noisy manifold, the ground truth manifold is enveloped by a cuboid in the ambient space. The cuboid is then covered with a large number $N_{Cub} \sim 1e6$ of points sampled from the uniform distribution and collected in $\mathcal{S}_{Cub}$. The sample from the noisy manifold is then $\mathcal{S}_{Cub} \cap \mathcal{B}_{GT}$. Data set $\mathcal{Q}^3$ is intended to simplify the particle distribution obtained via astronomical numerical simulations while still retaining the properties of the code used to generate and evolve the astronomical system.

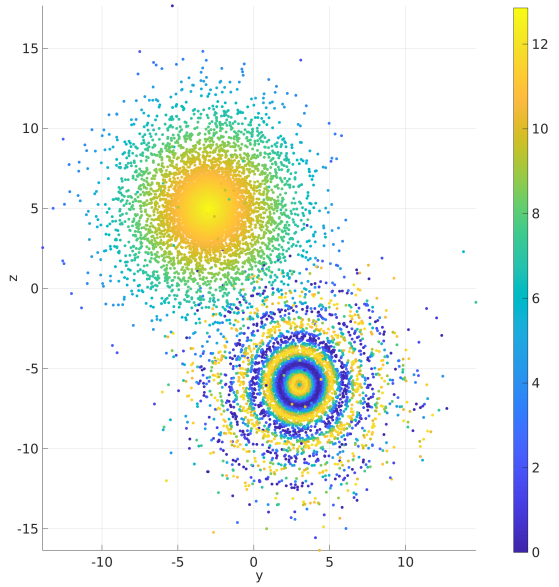
The data set is designed to morphologically resemble a jellyfish galaxy by defining three major manifolds $\mathcal{M}_k,\ k = 12, 13, 14$. $\mathcal{M}_{12}$ is a Gaussian distribution having mean



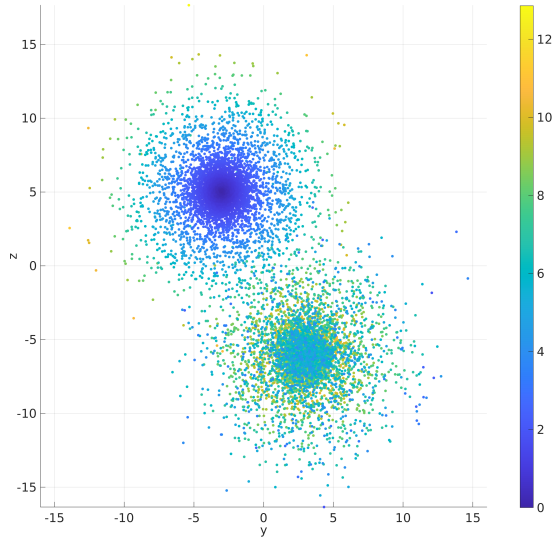
(a) Variable V_1 for total (noisy) toy data set.



(b) Variable V_2 for total (noisy) toy data set.



(c) Variable V_1 for total (noisy) toy data set.



(d) Variable V_2 for total (noisy) toy data set.

Figure 3.5: Distribution of variables V_1 and V_2 for the whole data set (sub-figures 3.5a and 3.5b) mimicking a stage of a simulated Jellyfish galaxy. The variables are designed to have peculiar profiles over the length of the manifolds or over their thickness (see description in the text).

$\mu_1 = (0, 0, 0)$ and covariance matrix

$$\Sigma_1 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 26 \end{pmatrix}$$

Manifolds $\mathcal{M}_{13}$ and $\mathcal{M}_{14}$ are inherently one-dimensional: their underlying true structure is

the unit interval $[-1, 1]$ embedded in $\mathbb{R}^3$ through mappping functions $f_{13} : [-1, 1] \to \mathbb{R}^3$ and $f_{14} : [-1, 1] \to \mathbb{R}^3$. The mapping functions take the form:

$$f_{13}(t) = (3, -3, 5) + \frac{(t+1)}{2}\,(24, -3, 5) \tag{3.22}$$

$$f_{14}(t) = (4, 3, -6) + \frac{(t+1)}{2}\,(14, 3, -6). \tag{3.23}$$

The underlying one-dimensional structure of the manifolds is then convolved with noise in order to obtain a morphological analogy with the structures generally found in astronomical simulations and observations. Both manifolds have an overlapping double noise structure: the thin and dense, inner region (*core*) and the thick, sparser layer (*sparse*), both defined as Gaussian mixtures.

The dense core of manifold $\mathcal{M}_{13}$ is defined as a flat Gaussian mixture whose $N^c_{13} = 43$ centers lie on $L^c_{13} = f_{13}([-1, 1])$ (each distancing from the adjacent by 0.5). The shared covariance matrix for the core Gaussian mixture is the identity matrix $\Sigma^c_{13} = \mathbf{I}$ and the component weights are $p^c_{13} = 1/N^c_{13}$.

The sparse Gaussian mixture has $N^s_{13} = 22$ centers lying on the segment $L^s_{13} = L^c_{13}$ (the distance between adjacent centers is 1). The shared covariance matrix is $\Sigma^s_{13} = 9\mathbf{I}$ and components weights are $p^s_{13} = 1/N^s_{13}$.

Manifold $\mathcal{M}_{14}$ has the same overall structure as manifold $\mathcal{M}_{13}$. The core structure has $N^c_{14} = 21$ centers, regularly sampling segment $L^c_{14} = f_{14}([-1, 1])$, while the sparse structure has $N^s_{14} = 11$ centers on $L^s_{14} = L^c_{14}$. Consequently, the mixture components are $p^c_{14} = 1/N^c_{14}$ for the core and $p^s_{14} = 1/N^s_{14}$. The shared covariance matrices are $\Sigma^c_{14} = \Sigma^c_{13}$ and $\Sigma^s_{14} = \Sigma^s_{13}$ for the core and sparse components respectively.

From each manifold $\mathcal{M}_k, i = 12, 13, 14$ a point cloud in $\mathbb{R}^3$ is generated, obtaining $\mathcal{D}_{12}, \mathcal{D}_{13}$ and $\mathcal{D}_{14}$. The union $\mathcal{Q}^3 = \bigcup_{i=12,13,14} \mathcal{D}_i$ of all point clouds represents the morphology of the synthetic data set (figures 3.5a and 3.5b).

3.2.2 Behaviour of simulated physical properties

Following the SPH formulation (Price, 2012; Cossins, 2010 and citations therein), consider each particle $\mathbf{t}_i$ in data set $\mathcal{Q}^3$ to sample a spherical volume of radius $r = h_i$. Here h_i is the *smoothing length* equal to the distance of particle $\mathbf{t}_i$ to its 50-th neighbour in data set $\mathcal{Q}^3$. Under the assumption of mass preservation Gingold et al., 1977, a constant value $m_i = 1$ is assigned to the mass contained in each volume sampled by particle t_i. The density of the sampled volume can now be defined as:

$$\rho_i = \frac{m_i}{(4/3)\pi h_i^3}. \tag{3.24}$$

Two additional quantities are also defined, having particular behaviours in proximity to the two manifolds in the data set. Quantity V_1 is defined to be uniformly distributed in the interval $[V_{min}, V_{min}+1]$ for particles of manifold $\mathcal{M}_{12}$, decreasing from the center of manifold $\mathcal{M}_{13}$ ($f_{13}([-1,1])$) and sinusoidal depending on the distance to the center of manifold $\mathcal{M}_{14}$ ($f_{14}([-1,1])$):

$$V_1(\mathcal{D}_{12}) = X_{12} \sim \mathcal{U}(V_{min}, V_{min}+1); \tag{3.25}$$

$$V_1(\mathcal{D}_{13}) = \frac{d^M_{13,r} - d(\mathbf{t}_i, L^c_{13})_{\mathcal{D}_{13}}}{4} V_{max}; \tag{3.26}$$

$$V_1(\mathcal{D}_{14}) = 1.5 V_{max} \left\{ 1 + \sin\left[16\pi \frac{d(\mathbf{t}_i, L^c_{14})_{\mathcal{D}_{14}} - d^m_{14}}{d^M_{14,r} - d^m_{14,r}} \right] \right\} \tag{3.27}$$

Where $d(\mathbf{t}_i, L^c_j)_{\mathcal{D}_j}$ is the distance between $\mathbf{t}_i \in \mathcal{D}_j$ from the segment $L^c_j = L^s_j = f_j([-1,1])$, for $j = 13, 14$; $d^M_{j,r} = \max_{\mathbf{t}_i \in \mathcal{D}_j} d(\mathbf{t}_i, L^c_j)$ and $d^m_{j,r} = \min_{\mathbf{t}_i \in \mathcal{D}_j} d(\mathbf{t}_i, L^c_j)$ are the maximum and minimum distances, respectively, within all points $\mathbf{t}_i \in \mathcal{D}_j$ from the core of manifold $\mathcal{M}_j$ (radial distances). Figure 3.5c shows the radial profiles of quantity V_1 for the two manifolds.

Quantity V_2 is uniformly sampled within the interval $[V_{min}, V_{min}+1]$ for all particles

in $\mathcal{D}_{12}$, it has an increasing profile from the center of manifold $\mathcal{M}_{13}$ and it decreases along the longitudinal axis of manifold $\mathcal{M}_{14}$:

$$V_2(\mathcal{D}_{12}) = X_{13} \sim \mathcal{U}(V_{min}, V_{min} + 1); \tag{3.28}$$

$$V_2(\mathcal{D}_{13}) = \frac{d(\mathbf{t}_i, L^c_{13})_{\mathcal{D}_{13}} - d^m_{13,r}}{4} V_{max}; \tag{3.29}$$

$$V_2(\mathcal{D}_{14}) = 3V_{max}\left[\frac{d^M_{14,L} - d(\mathbf{t}_i, f_{14}(-1))}{d^M_{14,L} - d^m_{14,L}}\right], \tag{3.30}$$

where $d^M_{14,L} = \max_{\mathbf{t}_i \in \mathcal{D}_{14}} d(t_i, f_{14}(-1))$ and $d^m_{14,L} = \min_{\mathbf{t}_i \in \mathcal{D}_{14}} d(t_i, f_{14}(-1))$ are the maximum and minimum distances, respectively, within all particles $\mathbf{t}_i \in \mathcal{D}_{14}$ from point $f_{14}(-1)$, representing the head of manifold $\mathcal{M}_{14}$ (longitudinal distances). In figure 3.5d the radial profiles of quantity V_2 for the two manifolds can be found. For completeness the profile defined in equation 3.30 is also included, despite its non-radial nature. This profile is better visualized in figure 3.5b.

Chapter Four

Single noisily sampled manifold

In this section, construction of manifold-aligned density model for a single manifold is described. $\mathcal{D} = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$ is the set of points belonging to a noisy manifold. Set $\tilde{\mathcal{D}} = \{\tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_2, ..., \tilde{\mathbf{t}}_N\}$ contains the points belonging to the noise-free (ground-truth) version of the same manifold. Generally, $\tilde{\mathcal{D}}$ is not available and needs to be approximated. In sec. 5.1, a set of methodologies is proposed, aiming at reducing the amount of transfer noise within a manifold along its perpendicular component, and removing the additional uniform noise that may be present in the whole data set. The first methodology is a physically inspired algorithm that simulates the process of diffusion, locally "pushing" points towards the mean axis of the noisily sampled manifold (Wu et al., 2018). The proposed, subsequent, filtering technique evaluates differences in the local densities of points within the two data sets $\mathcal{D}$ and $\tilde{\mathcal{D}}$. If the local density of a point has increased (and it is large) due to the application of the diffusion algorithm, it is considered to have moved closer to the mean axis of the manifold and therefore kept in $\tilde{\mathcal{D}}$. Otherwise, if its density has not increased it is considered to belong to the uniform noise permeating the analyzed volume, and thus removed from $\tilde{\mathcal{D}}$. A more detailed description is presented in sec. 5.1, also in the case of multi-manifold data sets.

4.1 Probabilistic modelling of a single manifold

Let $\mathcal{M}$ ba a $j-$dimensional manifold embedded in a higher dimensional space $\mathbb{R}^d$, $j < d$. The original GTM Bishop et al., 1998b assumes that the manifold $\mathcal{M}$ is an image of the j-

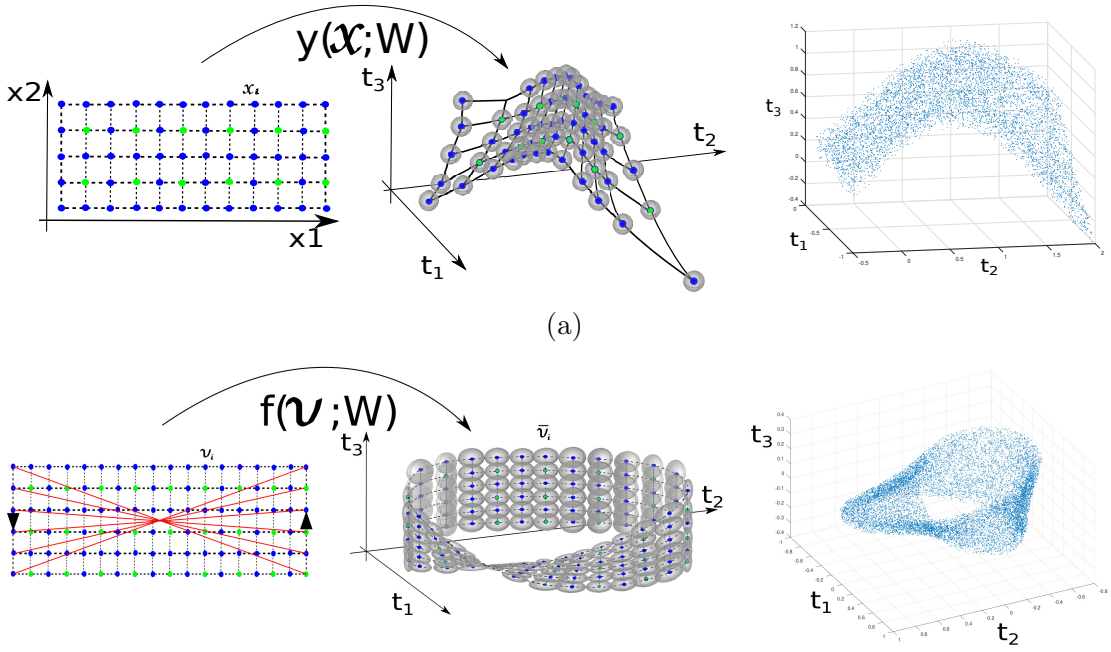
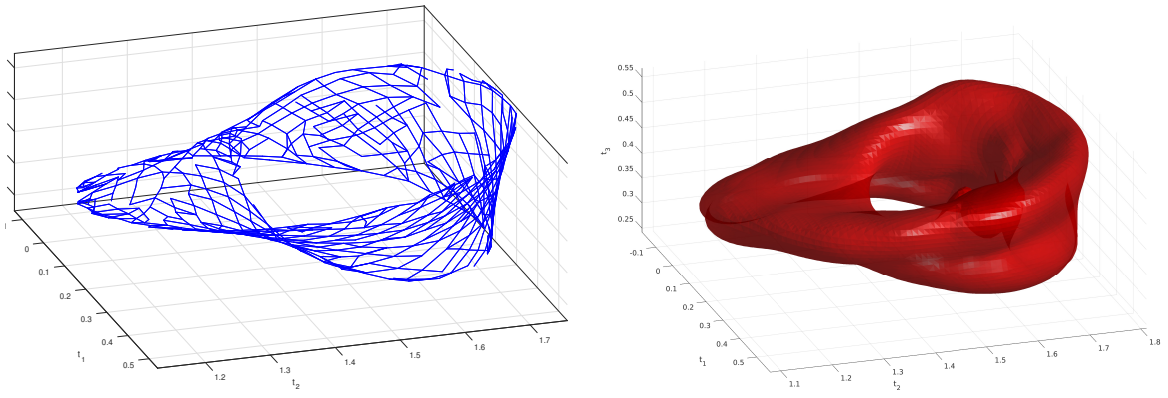


Figure 4.1: Subfig. 4.1a shows the classic GTM setup; the leftmost panel is a representation of the latent space discretized in a grid of points x_i. The j-dimensional interval $[-1, +1]^j$ is mapped through the parametrized function $\mathbf{y}(\mathcal{X}; W)$ onto the data space with a spherical Gaussian centered on each point assumed as noise. The noise model aims at describing the true distribution of points in the data space (rightmost panel). Subfig. 4.1b is a sketch of *AGTM*. The latent space is substituted with an abstract graph, shown here for the Möbius strip, together with its topological representation (the two arrows pointing in opposite directions on the shortest edges of the graph). Function $f(\mathcal{V}; \mathbf{W})$ maps the graph onto the data space and manifold-aligned noise models are estimated on each graph's node $\overline{v}_i$. The true noisy distribution is displayed in the rightmost panel.

dimensional interval $[-1, +1]^j$ (latent space) under a smooth embedding $\mathbf{y} : [-1, +1]^j \to \mathbb{R}^d$. To add noise to $\mathcal{M}$, the latent space $[-1, +1]^j$ is covered by a regular grid $\{\mathbf{x}_i\}_{i=1}^K$ of K points, whose images $\mathbf{y}(\mathbf{x}_i)$ under $\mathbf{y}$ form a "skeleton" of $\mathcal{M}$ in the data space. A spherical Gaussian noise model is then positioned at each skeleton node $\mathbf{y}(\mathbf{x}_i)$, thus representing the

noisy manifold as a mixture of K Gaussians centered at $\mathbf{y}(\mathbf{x}_i)$ (see fig. 4.1a). The latent space grid structure can be represented by an abstract undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each grid point $\mathbf{x}_i$ corresponds to a vertex $v_i \in \mathcal{V}$ and edges $e_{ij} \in \mathcal{E}$ are connecting vertices corresponding to neighbouring grid points $\mathbf{x}_i$ and $\mathbf{x}_j$. The structure of $\mathcal{G}$ is directly reflected in the latent grid shown in fig. 4.1b. The GTM model is generalized in that the manifold skeleton is obtained by direct embedding of the latent graph $\mathcal{G}$ into the data space through a parameterized mapping $\mathbf{f}$. Individual Gaussian noise models are centered at the images $\overline{\mathbf{v}}_i = \mathbf{f}(v_i) \in \mathbb{R}^d$ of the nodes of $\mathcal{G}$. The manifold skeleton is then formed by the embedded graph $\overline{\mathcal{G}} = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$, where $\overline{\mathcal{V}} = \{\overline{\mathbf{v}}_i, ..., \overline{\mathbf{v}}_K\}$ and $\overline{e}_{ij} \in \overline{\mathcal{E}}$ whenever $e_{ij} \in \mathcal{E}$. This enables a natural representation of density models representing noisy manifolds of much more intricate structure than that of a smoothly embedded low dimensional interval. For example, the sample from noisy Möbius strip shown in fig. 4.1b (right panel) is captured by a Gaussian mixture with centers at fixed skeleton points in fig. 4.1b (central panel), The skeleton points themselves are images under $\mathbf{f}$ of vertices $v \in \mathcal{V}$ (fig. 4.1b, left panel). The abstract graph $\mathcal{G}$ is presented with the usual topological convention by identification of opposite sides of a rectangular grid with inversion of direction. The additional edges resulting from such an identification are shown in red.



(a) (b)

Figure 4.2: Left panel: Embedded graph of the Möbius strip used for initialisation of AGTM. Right panel: Probabilistic model obtained after 2 iterations of AGTM when initialised over graphs $\mathcal{G}$ and $\overline{\mathcal{G}}$ with a manifold aligned noise.

Motivated by the original GTM model, the embedding $\mathbf{f} : \mathcal{V} \to \mathbb{R}^d$ is formulated as a nonlinear model, linear in parameters. In particular, a set of M basis functions $\phi_m : \mathcal{V} \to \mathbb{R}$, $m = 1, 2, ..., M$ is here used, operating on the abstract latent space $\mathcal{G}$. The image of a vertex $v \in \mathcal{V}$ under $\Phi = (\phi_1, ..., \phi_M)^T$ is obtained as

$$\overline{\mathbf{v}} = \mathbf{f}(v; \mathbf{W}) = \mathbf{W}\Phi(v), \tag{4.1}$$

where $\mathbf{W}$ is a $d \times M$ matrix of weights. To formulate the basis functions ϕ_m, a regular $\epsilon-$net of the graph $\mathcal{G}$ (see e.g., Boissonnat et al., 2018) is constructed, with M nodes. The nodes $\mathbf{c}_m \in \mathcal{V}$ of the $\epsilon-$net form the centers of the corresponding basis functions:

$$\phi_m(v) = \exp\left\{-\frac{D^2(v, c_m)}{\gamma^2}\right\},$$

where $D(v, c_m)$ is the shortest distance in $\mathcal{G}$ between the vertices v and c_m. Analogously to the original GTM, to ensure smoothness of the embedding $\mathbf{f}$, the scale parameter was set[1] to 2ϵ.

The manifold aligned probabilistic model is a flat mixture model

$$p(\mathbf{t}|\mathbf{W}, \boldsymbol{\zeta}) = \frac{1}{K}\sum_{i=1}^{K} p(\mathbf{t}|v_i, \zeta_i\Sigma_i, \mathbf{W}), \tag{4.2}$$

where the mixture components are locally manifold-aligned multivariate Gaussians centered at the embedded vertices $\overline{\mathbf{v}}_i \in \overline{\mathcal{V}}$:

$$p(\mathbf{t}|v_i, \zeta_i\Sigma_i, \mathbf{W}) = \frac{1}{[(2\pi\zeta_i)^d|\Sigma_i|]^{\frac{1}{2}}} \exp\left\{-\frac{\Delta\mathbf{t}^T\Sigma_i^{-1}\Delta\mathbf{t}}{2\zeta_i}\right\} \tag{4.3}$$

with $\Delta\mathbf{t} = \mathbf{f}(v_i; \mathbf{W}) - \mathbf{t}$. Unlike in Bishop et al., 1998a, the local manifold-aligned covariance matrix is modelled as a scaled version (scaled by $\zeta_i > 0$) of the local covariance matrix

[1] Alternatively, the value of γ can be set according to a suitable criterion through cross-validation.

estimated as

$$\Sigma_i = \frac{1}{\sum_{a=1}^{N} \kappa(\overline{\mathbf{v}}_i, \mathbf{t}_a; \ell_i)} \sum_{n=1}^{N} \kappa(\overline{\mathbf{v}}_i, \mathbf{t}_n; \ell_i)(\mathbf{t}_n - \overline{\mathbf{v}}_i)(\mathbf{t}_n - \overline{\mathbf{v}}_i)^T,$$

with Gaussian smoothing kernel

$$\kappa(\overline{\mathbf{v}}_i, \mathbf{t}_n; \ell_i) = \exp\left\{-\frac{\|\overline{\mathbf{v}}_i - \mathbf{t}_n\|^2}{3\ell_i^2}\right\}.$$

The kernel scale parameter ℓ_i is determined as the average of the distances between the embedded vertex $\overline{\mathbf{v}}_i$ and its neighbouring vertices in $\mathcal{G}$, embedded in $\mathbb{R}^d$,

$$\ell_i = \frac{1}{|\mathcal{N}(v_i)|} \sum_{a \in \mathcal{N}(v_i)} \|\overline{\mathbf{v}}_i - \overline{\mathbf{a}}\|,$$

where $\mathcal{N}(v_i) \subset G$ is the set of neighbours of v_i in the graph $\mathcal{G}$.

As a latent variable model, AGTM can be trained to maximize log-likelihood

$$\mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) = \sum_{n=1}^{N} \ln\left\{\frac{1}{K} \sum_{i=1}^{K} p(\mathbf{t}_n | v_i, \zeta_i \Sigma_i, \mathbf{W})\right\} \tag{4.4}$$

via the E-M algorithm. In the E-step, which is the same as in the original GTM model (in fact, in any mixture model), the responsibilities R_n^i are calculated as the posterior distribution over nodes $v_i \in \mathcal{V}$, given the data points $\mathbf{t}_n \in \mathcal{D}$ Bishop et al., 1998b. The M-step updates for parameters $\mathbf{W}$ and $\boldsymbol{\zeta}$ are obtained by differentiating the expected value of the complete data log-likelihood $< \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) >$ (eq. (4.4)) w.r.t. the corresponding parameters and setting to zero:

$$\nabla_{\mathbf{W}} < \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) > = 0; \tag{4.5}$$

$$\nabla_{\zeta} < \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) > = 0. \tag{4.6}$$

where

$$< \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) >= \sum_{n=1}^{N} \sum_{i=1}^{K} R_n^i(\mathbf{W}^{old}, \zeta^{old}) \ln[p(\mathbf{t}_n | v_i, \zeta_i \Sigma_i, \mathbf{W})] \tag{4.7}$$

Starting from equation 4.5,

$$\begin{aligned} \nabla_{\mathbf{W}} < \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) > &= -\sum_{n=1}^{N} \sum_{i=1}^{K} R_n^i \nabla_{\mathbf{W}} \frac{1}{2\zeta_i} [\mathbf{W}\Phi_i - \mathbf{t}_n]^T \Sigma_i^{-1} [\mathbf{W}\Phi_i - \mathbf{t}_n] \\ &= -\sum_{n=1}^{N} \sum_{i=1}^{K} \frac{R_n^i}{\zeta_i} \Sigma_i^{-1} [\mathbf{W}\Phi_i - \mathbf{t}_n] \Phi_i^T = 0 \end{aligned}$$

the last equality can be rewritten in matrix notation as:

$$\sum_{i=1}^{K} \mathcal{C}_i \mathbf{W} \Psi_i = \sum_{i=1}^{K} \frac{\Sigma_i^{-1}}{\zeta_i} \mathbf{T} \mathbf{R}_i \Phi_i^T, \tag{4.8}$$

where $\mathbf{R}_i$ is the $(N \times 1)$ column vector containing the responsibilities for every point $\mathbf{t}_n$ w.r.t. center $\overline{v}_i$, $\mathbf{T}$ the $D \times N$ matrix having all $\mathbf{t}_n$'s as column vectors, $\Phi_i = \Phi(v_i)$ $(M \times 1)$, $\Psi_i = \Phi_i \Phi_i^T$ and $\mathcal{C}_i = \Big(\sum_{n=1}^{N} R_n^i \Big) \Sigma_i^{-1} / \zeta_i$.

Equation (4.8) can be solved by taking advantage of the properties of the Kronecker product and vec operator (see Petersen et al., 2012, sections 5.1 and 10.2), obtaining:

$$\mathrm{vec}(\mathbf{W}^{new}) = \Big(\sum_{i=1}^{K} \Psi_i^T \otimes \mathcal{C}_i \Big)^{-1} \mathrm{vec}\Big(\sum_{i=1}^{K} \frac{\Sigma_i^{-1}}{\zeta_i} \mathbf{T} \mathbf{R}_i \Phi_i^T \Big), \tag{4.9}$$

Equation (4.6) is now solved by first applying the logarithm and preserving only the terms dependant on ζ_i, leading to:

$$\begin{aligned} \nabla_{\zeta} < \mathcal{L}(\mathbf{W}, \boldsymbol{\zeta}) > &= \sum_{n=1}^{N} \sum_{i=1}^{K} R_n^i \frac{\partial}{\partial \zeta_i} \Big[-\frac{D}{2} \ln \zeta_i - \frac{(\mathbf{W}\Phi_i - \mathbf{t}_n)^T \Sigma_i^{-1} (\mathbf{W}\Phi_i - \mathbf{t}_n)}{2\zeta_i} \Big] \\ &= \sum_{n=1}^{N} \sum_{i=1}^{K} R_n^i \Big[\frac{D}{\zeta_i} - \frac{(\mathbf{t}_n - \overline{\mathbf{v}}_i)^T \Sigma_i^{-1} (\mathbf{t}_n - \overline{\mathbf{v}}_i)}{\zeta_i^2} \Big] = 0. \end{aligned}$$

Solving this for every ζ_i, the scale parameters $\boldsymbol{\zeta}$ of the local covariance matrices are updated as

$$\zeta_i^{new} = \frac{1}{d \sum_{n=1}^{N} R_i^n} \sum_{n=1}^{N} R_n^i D_M^2(\mathbf{t}_n, \overline{\mathbf{v}}_i), \tag{4.10}$$

where $D_M^2(\mathbf{t}_n, \overline{\mathbf{v}}_i)$ is the squared Mahalanobis distance between $\mathbf{t}_n$ and $\overline{\mathbf{v}}_i = \mathbf{W}\Phi_i$:

$$D_M^2(\mathbf{t}_n, \overline{v}_i) = (\mathbf{t}_n - \overline{\mathbf{v}}_i)^T \Sigma_i^{-1} (\mathbf{t}_n - \overline{\mathbf{v}}_i). \tag{4.11}$$

This has an intuitive interpretation: the scale parameter ζ_i of the i-th Gaussian mixture component is obtained as the (effective) squared mean Mahalanobis distance of points around component i to its centre $\overline{\mathbf{v}}_i$, per dimension. As an example, consider the noisy sample around Möbius strip shown in fig. 4.1b (right panel). Fig. 4.2a presents the embedded abstract graph and fig. 4.2b an iso-surface of the density model given by the AGTM model trained on the sample.

4.2 Manifold crawling

The previous section specified how to obtain probabilistic models of noisy manifolds using AGTM. In the following, a recursive algorithm is described that enables recovery of the abstract graph and the embedding function $\mathbf{f}$, given the data set $\tilde{\mathcal{D}}$. The name of this procedure is *Manifold Crawling*, or simply *Crawling*. Recall that the manifold dimensionality is $j < d$ and $\tilde{\mathcal{D}}$ contains points of the sampled manifold after the application of the diffusion algorithm and thus closer to the mean axis of the manifold than in the noisily sampled data set $\mathbf{D}$. The manifold crawling procedure aims at recovering an optimal initialization for AGTM. Other initializations may be utilized (e.g. neighbourhood-graphs) as long as the underlying topological structure of the manifold is preserved by the recovered graph.

4.2.1 Initialisation

Initially, a single "seed" $\tilde{\mathbf{t}}_0 \in \tilde{\mathcal{D}}$ is randomly chosen and Principal Component Analysis (PCA Hotelling, 1933) is applied to its neighbourhood $\mathcal{B}(\tilde{\mathbf{t}}_0, r) \cap \tilde{\mathcal{D}}$. It is assumed that the unit length eigenvectors obtained through PCA are ordered in descending order of their corresponding eigenvalues.

The first j eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_j$ span the tangent space to manifold $\mathcal{M}$ at $\tilde{\mathbf{t}}_0$, $T_{\tilde{\mathbf{t}}_0}\mathcal{M} := \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_j\}$. For every eigenvector $\mathbf{u}_l$, $l = 1, \ldots, j$, two new points are computed at a distance $\eta \cdot r$ from $\tilde{\mathbf{t}}_0$ along directions $\pm\mathbf{u}_l$:

$$\mathbf{z}_l^{\pm} = \tilde{\mathbf{t}}_0 \ \pm \ \eta \cdot r \cdot \mathbf{u}_l^T \tag{4.12}$$

where $0 < \eta < 1$ is a step-size parameter. In order to keep the crawling adherent to the$\tilde{\mathcal{D}}$, the closest neighbour to every $\mathbf{z}_l^{\pm}$ is found:

$$\tilde{\mathbf{t}}_l^{\pm} = \underset{\tilde{\mathbf{t}} \in \tilde{\mathcal{D}}}{\text{argmin}} \, ||\tilde{\mathbf{t}} - \mathbf{z}_l^{\pm}||. \tag{4.13}$$

The construction of the embedded graph $\overline{\mathcal{G}}$ is started by initializing the set of nodes and edges:

$$\overline{\mathcal{V}} = \{\overline{\mathbf{v}}_0, \overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_{2j-1}, \overline{\mathbf{v}}_{2j}\}, \tag{4.14}$$

$$\overline{\mathcal{E}} = \{(\overline{\mathbf{v}}_0, \overline{\mathbf{v}}_1), (\overline{\mathbf{v}}_0, \overline{\mathbf{v}}_2), \ldots, (\overline{\mathbf{v}}_0, \overline{\mathbf{v}}_{2j-1}), (\overline{\mathbf{v}}_0, \overline{\mathbf{v}}_{2j})\}, \tag{4.15}$$

where

$$\overline{\mathbf{v}}_0 = \tilde{\mathbf{t}}_0, \ \overline{\mathbf{v}}_1 = \tilde{\mathbf{t}}_1^+, \ \overline{\mathbf{v}}_2 = \tilde{\mathbf{t}}_1^-, \ \ldots, \ \overline{\mathbf{v}}_{2j-1} = \tilde{\mathbf{t}}_j^+, \ \overline{\mathbf{v}}_{2j} = \tilde{\mathbf{t}}_j^-.$$

At the same time, construction of the abstract graph $\mathcal{G}$ is initialized by imposing

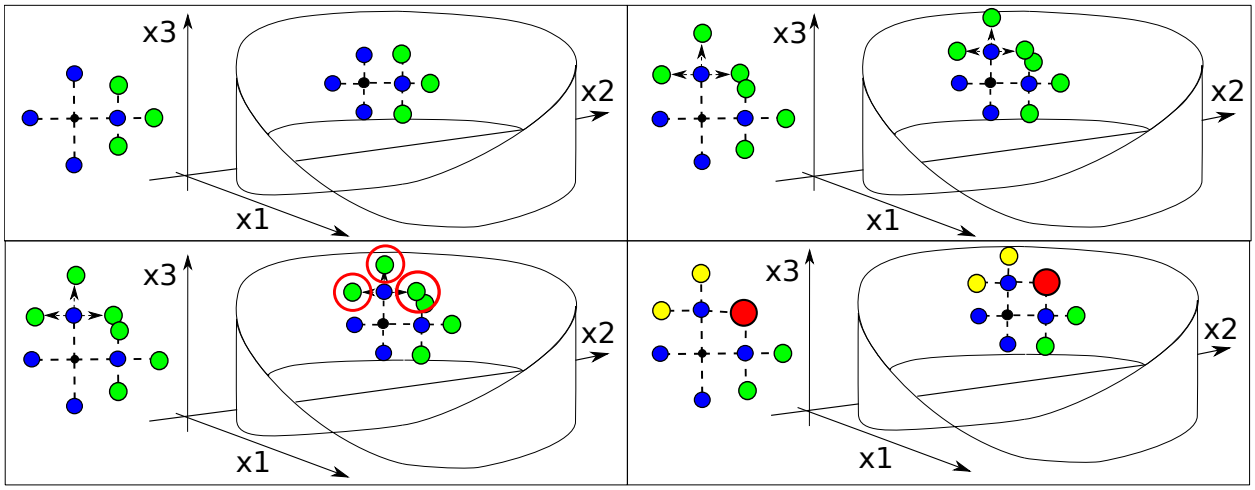


Figure 4.3: Visual representation of the Expansion and Contraction phases in the *Crawling* algorithm. Starting from the initial 1-edge nodes (blue points), the first set of node candidates (green points) are found (top panel) along the directions represented by black dashed lines. For every candidate $\tilde{\mathbf{t}}_c$, construct $\mathcal{B}(\tilde{\mathbf{t}}_c, \beta \cdot r) \cap \overline{\mathcal{V}}$ (red circles in bottom left panel). If the ball is non empty, its only element takes the place of $\tilde{\mathbf{t}}_c$ and is added to the graph (bottom right panel, red point).

that every node $\overline{\mathbf{v}}_l$ in $\overline{\mathcal{V}}$ corresponds to a node $v_l \in \mathcal{V}$ in the abstract counterpart $\mathcal{G}$ of $\overline{\mathcal{G}}$. The node $\overline{\mathbf{v}}_l$ can be considered the image of v_l through an unknown embedding function $\mathbf{f} : \mathcal{V} \longrightarrow \mathbb{R}^d$, $\overline{\mathbf{v}}_l = \mathbf{f}(v_l)$. The pull-back edges in the abstract graph $\mathcal{G}$ are defined as $\mathcal{E} = \{(v_0, v_1), (v_0, v_2), \ldots, (v_0, v_{2j})\}$. The degree of each node in graph $\overline{\mathcal{G}}$ (analogously in $\mathcal{G}$) is then computed. The degree $deg(\overline{\mathbf{v}}_i)$ of node $\overline{\mathbf{v}}_i \in \overline{\mathcal{G}}$ is the number of edges that are incident to that node. The set of vertices having only one edge in $\overline{\mathcal{V}}$ is $\overline{\mathcal{V}} \supset \overline{\mathcal{V}}_1^E = \{\overline{\mathbf{v}} \mid deg(\mathbf{v}) = 1\}$.

After the initialisation phase, each successive iteration is split into two phases: *expansion* and *contraction*. The two phases are alternated for each newly discovered node $\overline{\mathbf{v}}_i \in \overline{\mathcal{V}}_1^E$ (blue dots in fig. 4.3) having only one edge[2] connecting it to its parent node $\overline{\mathbf{v}}_p$.

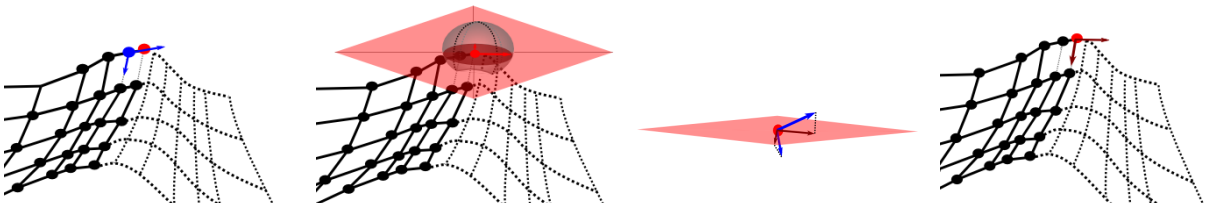
[2]Note that the newly discovered nodes necesserily will be of degree 1.

4.2.2 Expansion Phase

Given a node $\overline{\mathbf{v}}_i$ in $\overline{\mathcal{V}}_1^E$, the eigen-decomposition of its neighbourhood $\mathcal{B}(\overline{\mathbf{v}}_i, r) \cap \tilde{\mathcal{D}}$ is performed, yielding the first j leading eigenvectors $\mathbf{g}_1, \ldots, \mathbf{g}_j$. To preserve the crawling directions, the eigenvectors of the parent's node $\mathbf{u}_1^p, \mathbf{u}_2^p, \ldots, \mathbf{u}_j^p$ are projected onto the tangent space at $\overline{\mathbf{v}}_i$, $T_{\overline{\mathbf{v}}_i}\mathcal{M} := \text{span}\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_j\}$ and renormalized to unit length:

$$\mathbf{u}_i = \frac{1}{||\mathbf{P}\mathbf{u}_i^p||}\mathbf{P}\mathbf{u}_i^p; \qquad i = 1, \ldots, j. \tag{4.16}$$

Here $\mathbf{P} = \mathbf{V}\mathbf{V}^\dagger$ is the orthogonal projection operator onto $T_{\overline{\mathbf{v}}_i}\mathcal{M}$, with $\mathbf{V}$ the matrix containing $\mathbf{g}_l$ as columns and $\mathbf{V}^\dagger$ its pseudo-inverse. (see fig. 4.4). As in the initialization step, new points are then computed using eq. (4.12) and their closest neighbours $\tilde{\mathbf{t}}_l^{\pm}$ in data set $\tilde{\mathcal{D}}$ stored as candidate nodes to be added to $\overline{\mathcal{V}}$. The result of the expansion phase for a given iteration is shown in fig. 4.3. The green points are the new candidates, the dashed black lines represent the principal crawling directions. Construction of the set of new candidate nodes $\mathcal{C}(\overline{\mathbf{v}}_i) = \{\tilde{\mathbf{t}}_l^{\pm}, l = 1, \ldots, j\}$ around $\overline{\mathbf{v}}_i$ terminates the Expansion phase.



(a) (b) (c) (d)

Figure 4.4: Projection of tangent space onto new node's estimate.

4.2.3 Contraction Phase

The *contraction* phase checks if the candidate nodes together with $\overline{\mathcal{V}}$ can be collapsed into a smaller set of nodes.

For every new candidate $\tilde{\mathbf{t}}_c \in \mathcal{C}(\overline{\mathbf{v}}_i)$ from the Expansion phase, it is checked if it lands within a small "tolerance" neighbourhood $\mathcal{B}(\mathbf{v}_e, \beta \cdot r)$ of an already existing node $\overline{\mathbf{v}}_e$ of $\overline{\mathcal{G}}$. Here $0 < \beta < \eta < 1$ is the tolerance scale parameter. If this is the case, the candidate $\tilde{\mathbf{t}}_c$ is identified with $\overline{\mathbf{v}}_e$ and the nodes $\overline{\mathbf{v}}_e$ and $\overline{\mathbf{v}}_i$ are connected by an edge $(\overline{\mathbf{v}}_e, \overline{\mathbf{v}}_i)$ added to $\overline{\mathcal{E}}$ (red point in the lower right plot of fig. 4.3). Otherwise, $\tilde{\mathbf{t}}_c$ is confirmed as a new node and added to $\overline{\mathcal{V}}$ with the edge $(\tilde{\mathbf{t}}_c, \overline{\mathbf{v}}_i)$ added to $\overline{\mathcal{E}}$ (yellow points in the lower right plot of fig. 4.3). This marks the end of the *contraction* phase. The iteration is concluded when *expansion* and *contraction* have been performed for every $\overline{\mathbf{v}}_i$ in the single-edge node set $\overline{\mathcal{V}}_1^E$ from the previous iteration.

After the *expansion* and *contraction* phases, the single-edge node set $\overline{\mathcal{V}}_1^E$ is updated and the new iteration of *expansion-contraction* steps is started. Manifold crawling stops when the number of nodes in $\overline{\mathcal{V}}$ no longer increases in two consecutive iterations. A pseudo-code for the manifold crawling algorithm is provided in alg. 1.

4.2.4 Initialization of the latent graph embedding $\mathbf{f}(v; \mathbf{W})$

The node sets $\overline{\mathcal{V}}$ and $\mathcal{V}$ will be used to initialize the embedding $\mathbf{f}(v; \mathbf{W})$ (equation (4.1)) of the abstract latent space $\mathcal{G}$ into the data space $\mathbb{R}^d$ by setting the weight matrix $\mathbf{W}$ through linear regression so that $\mathbf{f}(v_i; \mathbf{W}) \approx \overline{\mathbf{v}}_i$ for all corresponding couples $(v_i \in \mathcal{V}, \overline{\mathbf{v}}_i \in \overline{\mathcal{V}})$. While $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is unweighted, the edges in $\overline{\mathcal{G}}$ are weighted by the distance between the two connected nodes.

If needed, one can regularize the parameters $\mathbf{W}$ by applying *ridge regression* instead of linear regression, by minimization of the cost function:

$$\mathcal{F} = \sum_{i=1}^{|\mathcal{V}|} (\overline{\mathbf{v}}_i - \mathbf{f}(v_i))^2 + \sum_{j=1}^{M} \chi_j \mathbf{w}_j^2, \tag{4.17}$$

Data: Denoised Single j-dimensional Manifold
Result: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; $\overline{\mathcal{G}} = (\overline{\mathcal{V}}, \overline{\mathcal{E}})$
$A_1 = 1$; $A_2 = 2$; d = degree($\overline{\mathcal{G}}$);
1−Edge_Nodes = $\{\overline{v} \in \overline{\mathcal{G}} | d(\overline{v}) = 1\}$; it = 1;

while $A_2 > A_1$ **do**
  $A_1 = A_2$;
  for $ID \leftarrow 1$ **to** *size(1-Edge_Nodes)* **do**
    EXPANSION
    $\tilde{\mathbf{t}}_0$ = 1-Edge_Nodes(ID);
    Find Graph's Index p corresponding to $\tilde{\mathbf{t}}_0$;
    Compute first j eigen-vectors of $\mathcal{B}(\tilde{\mathbf{t}}_0, r)$: $\mathcal{T}_{\tilde{\mathbf{t}}_0}\mathcal{M}_j = \text{span}\{\mathbf{g}_1, \ldots, \mathbf{g}_j\}$;
    Project parent node's eigen-vectors onto $\mathcal{T}_{\tilde{\mathbf{t}}_0}\mathcal{M}_j$ (eq. 4.16);
    Find new candidate nodes $\mathcal{N} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{2j}\}$ (eq. 4.12): $\forall \mathbf{z}_i \in \mathcal{N}$ compute closest neighbour (eq. 4.13)
    $\mathcal{N}_C = \{\tilde{\mathbf{t}}_1, \ldots \tilde{\mathbf{t}}_{2j}\}$.
    CONTRACTION
    for $l \leftarrow 1$ **to** $|\mathcal{N}|$ **do**
      if $\mathcal{N}_C \cap \overline{\mathcal{V}} \neq \emptyset$ **then**
        $\forall \tilde{\mathbf{t}}_m \in \mathcal{N}_C \cap \overline{\mathcal{V}}$
        Find Graph's Index c corresponding to $\tilde{\mathbf{t}}_m$;
        $\overline{e}_{pc} = (\tilde{\mathbf{t}}_0, \tilde{\mathbf{t}}_m)$; $e_{pc} = (p, c)$;
      else
        Compute $\mathcal{B}(\tilde{\mathbf{t}}_l, \beta r) \cap \overline{\mathcal{V}}$
        if $\mathcal{B}(\tilde{\mathbf{t}}_l, \beta r) \cap \overline{\mathcal{V}} \neq \emptyset$ **then**
          $\forall \tilde{\mathbf{t}}_m \in \mathcal{N}_C \cap \overline{\mathcal{V}}$
          Find Graph's Index c corresponding to $\tilde{\mathbf{t}}_m$;
          $\overline{e}_{pc} = (\tilde{\mathbf{t}}_0, \tilde{\mathbf{t}}_m)$; $e_{pc} = (v_p, v_c)$;
        else
          Confirm candidate as new node;
          $\overline{\mathbf{v}}_l = \tilde{\mathbf{t}}_l$; $\overline{\mathcal{V}} = \overline{\mathcal{V}} \cup \tilde{\mathbf{t}}_l$;
          $\mathcal{V} = \mathcal{V} \cup v_l$;
          Add edge to graphs;
          $\overline{e}_{pl} = (\tilde{\mathbf{t}}_0, \tilde{\mathbf{t}}_l)$;
          $e_{pl} = (v_p, v_l)$;
        end
      end
    end
  end
  d = degree($\overline{\mathcal{G}}$);
  1−Edge_Nodes = $\{\overline{v} \in \overline{\mathcal{G}} | d(\overline{v}) = 1\}$;
  A2 = size(1−Edge_Nodes);
end

Algorithm 1: Crawling algorithm

solving the system of equations:

$$(\Phi^\top \Phi + \mathbf{X})\hat{\mathbf{W}} = \Phi^\top \overline{\mathcal{V}}. \tag{4.18}$$

Here, $\mathbf{X}$ is the diagonal matrix having as diagonal elements the regularization parameters $\{\chi_j\}_{j=1}^{M}$. The solution to the system in equation (4.18) is the *normal equation* Rawlings et al., 2001:

$$\hat{\mathbf{W}} = (\Phi^\top \Phi + \mathbf{X})^{-1}\Phi^\top \overline{\mathcal{V}}. \tag{4.19}$$

In order to recover a smooth mapping function with small optimal weights $\hat{\mathbf{W}}$, throughout the performed tests $\chi_j = \chi = 1e-5$ for all $j = 1\ldots, M$ is chosen.

4.3 Local on-edge curvature

As outlined in Tino et al., 2002, having obtained a smooth representation of a $j-$dimensional manifold via the AGTM formalism, local curvatures along edges of the abstract graph can now be estimated.

Consider $e_{0,1}$ as a segment of unit length and estimate the change in tangent vectors while moving along the embedded image of $e_{0,1}$. Five new nodes are inserted $x_1, \ldots, x_5$ along the edge $e_{0,1}$ and connected in a chain from node v_0 to node v_1, with edges $e_{i,i+1}$ and respective weights:

$$\begin{aligned} &w(v_0, x_1) = \ldots w(x_1, x_2) \cdots = w(x_4, x_5) = h; \\ &w(x_5, v_1) = 1 - 5h, \end{aligned}$$

where, as mentioned above, $h > 0$ is a small perturbation parameter (in this study, $h = 0.01$).

The chain of newly defined nodes, connecting v_0 and v_1, replaces the edge $e_{0,1}$ and acts as a discretization of the continuous segment in a small neighbourhood of the node v_0. A discretized "lifted line" $\omega(x_i) = \mathbf{f}(x_i; \mathbf{W})$ given by the mapping $\mathbf{f}$ through the new nodes $x_i, i = 1 \ldots, 5$ is defined. In the following, two formulations for computing the local curvature on graph's edges are provided, taking advantage of the newly defined nodes, in the vicinity of v_0, along edge $e_{0,1}$.

4.3.1 Algebraic derivation

Given the discrete formulation, it is possible to approximate the first and second order partial derivatives of $\mathbf{f}$ w.r.t. the direction given by $e_{0,1}$, with the central first and second order finite differences $\delta_h[\mathbf{f}]/2h$ and $\delta_h^2[\mathbf{f}]/4h^2$ computed on node x_3:

$$\dot{\omega}(v_0) = \left.\frac{\partial \mathbf{f}}{\partial e_{0,1}}\right|_{v_0} \approx \frac{\delta_h[\mathbf{f}](x_3)}{2h} = \frac{\mathbf{f}(x_4) - \mathbf{f}(x_2)}{2h} \tag{4.20}$$

$$\ddot{\omega}(v_0) = \left.\frac{\partial^2 \mathbf{f}}{\partial^2 e_{0,1}}\right|_{v_0} \approx \frac{\delta_h[\mathbf{f}](x_3)}{4h^2} = \frac{\mathbf{f}(x_5) - 2\mathbf{f}(x_3) + \mathbf{f}(x_1)}{4h^2}. \tag{4.21}$$

Realizing that $\dot{\omega}(v_0) = \mathbf{T}(v_0)_{e0,1}$ is the tangent space of manifold $\mathcal{M}$ at node v_0 on direction $e_{0,1}$, $\ddot{\omega}(v_0)$ is decomposed as the sum of the parallel and perpendicular components to the tangent space $\mathbf{T}(v_0)_{e0,1}$

$$\ddot{\omega}(v_0) = \ddot{\omega}^{\parallel}(v_0) + \ddot{\omega}^{\perp}(v_0); \quad \ddot{\omega}^{\parallel}(v_0) \in \mathbf{T}(v_0)_{e_{0,1}} \quad \text{and} \quad \ddot{\omega}^{\perp}(v_0) \in \mathbf{T}(v_0)^{\perp}_{e_{0,1}}. \tag{4.22}$$

In particular, by defining column vectors $\Gamma^{(1)} = \dot{\omega}(v_0)$ and $\Gamma^{(2)} = \ddot{\omega}(v_0)$, the projector operators over tangent space $\mathbf{T}(v_0)_{e_{0,1}}$ and over the perpendicular space $\mathbf{T}(v_0)^{\perp}_{e_{0,1}}$ are respectively:

$$\Pi_{\|} = \Gamma^{(1)}\left(\Gamma^{(1)}\right)^{+} = \frac{\Gamma^{(1)} \otimes \Gamma^{(1)}}{\|\Gamma^{(1)}\|^2} \tag{4.23}$$

$$\Pi_{\perp} = \mathbf{I} - \Pi_{\|}. \tag{4.24}$$

Using operator $\Pi_{\|}$ and $\Pi_{\perp}$ the parallel and perpendicular components to the tangent space $\mathbf{T}(v_0)_{e_{0,1}}$ of $\ddot{\omega}(v_0)$ can be rewritten as:

$$\ddot{\omega}^{\|}(v_0) = \Pi_{\|}\ddot{\omega}(v_0); \tag{4.25}$$

$$\ddot{\omega}^{\perp}(v_0) = \Pi_{\perp}\ddot{\omega}(v_0). \tag{4.26}$$

While equation (4.25) describes the local "deformatio" of the mapping along the tangent space (magnification factor), equation (4.26) presents the variation in the perpendicular direction to tangent space $\mathbf{T}(v_0)_{e_{0,1}}$ of the mapping function $\mathbf{f}$, which is easily interpretable as the curvature of $\mathbf{f}$ in the vicinity of node v_0 along edge $e_{0,1}$.

4.3.2 Geometric derivation

Given the discretization of the lifted line $\omega(x_i)$, an approximation of the unit tangent vectors at $\omega(x_2)$ and $\omega(x_4)$ is achievable via central differences:

$$\vartheta(x_2) \approx \frac{\omega(x_3) - \omega(x_1)}{\|\omega(x_3) - \omega(x_1)\|} \tag{4.27}$$

$$\vartheta(x_4) \approx \frac{\omega(x_5) - \omega(x_3)}{\|\omega(x_5) - \omega(x_3)\|}. \tag{4.28}$$

The curvature at $\omega(x_3)$ can be evaluated through the difference of the unit tangent vectors at $\omega(x_2)$ and $\omega(x_4)$, relative to the geodesic distance from $\omega(x_2)$ to $\omega(x_4)$ (approximated

through $\|\omega(x_3) - \omega(x_2)\| + \|\omega(x_4) - \omega(x_3)\|$):

$$\Upsilon(x_3) \approx \frac{\vartheta(x_4) - \vartheta(x_2)}{\|\omega(x_3) - \omega(x_2)\| + \|\omega(x_4) - \omega(x_3)\|},$$

which can be written as

$$\Upsilon(x_3) \approx \frac{\frac{\omega(x_3)-\omega(x_1)}{\|\omega(x_3)-\omega(x_1)\|} - \frac{\omega(x_5)-\omega(x_3)}{\|\omega(x_5)-\omega(x_3)\|}}{\|\omega(x_3) - \omega(x_2)\| + \|\omega(x_4) - \omega(x_3)\|}. \tag{4.29}$$

This perturbation based estimation can be regularised by observing that for sufficiently small h, the distances between the embedded neighbouring nodes, $\|\omega(x_2)-\omega(x_1)\|$, $\|\omega(x_3)-\omega(x_2)\|$, $\|\omega(x_4)-\omega(x_3)\|$ and $\|\omega(x_5)-\omega(x_4)\|$, are approximately equal and can be represented by $\tilde{h}$, and $\|\omega(x_3) - \omega(x_1)\| \approx |\omega(x_5) - \omega(x_3)\| \approx 2\tilde{h}$. Thus, the final result for $\Upsilon(x_3)$ is:

$$\Upsilon(x_3) \approx \frac{\omega(x_5) - 2\omega(x_3) + \omega(x_1)}{4\tilde{h}^2}. \tag{4.30}$$

In practice, the mean length $\tilde{\ell}$ of edges of the abstract latent graph embedded in the data space is calculated and $\tilde{h} = h \cdot \tilde{\ell}$. The norm $\|\Upsilon(x_3)\|$ quantifies the directional curvature of $\mathbf{f}$ in the vicinity of node v_0, along edge $e_{0,1}$.

4.3.3 Demonstrative applications

To demonstrate the workings of the local curvature estimations a set of controlled experiments on concentric circles with known radii is first performed. The geometrical estimation of local curvature is expected to approximate the inverse radius of the circles, while the algebraic formulation estimates the curvature up to a scalar. After generating points on circles of radii $r_0 = 1, 2, 4$, crawling is performed and a latent graph constructed on each circle manifold. Curvatures on all edges of the embedded graphs are then evaluated follow-

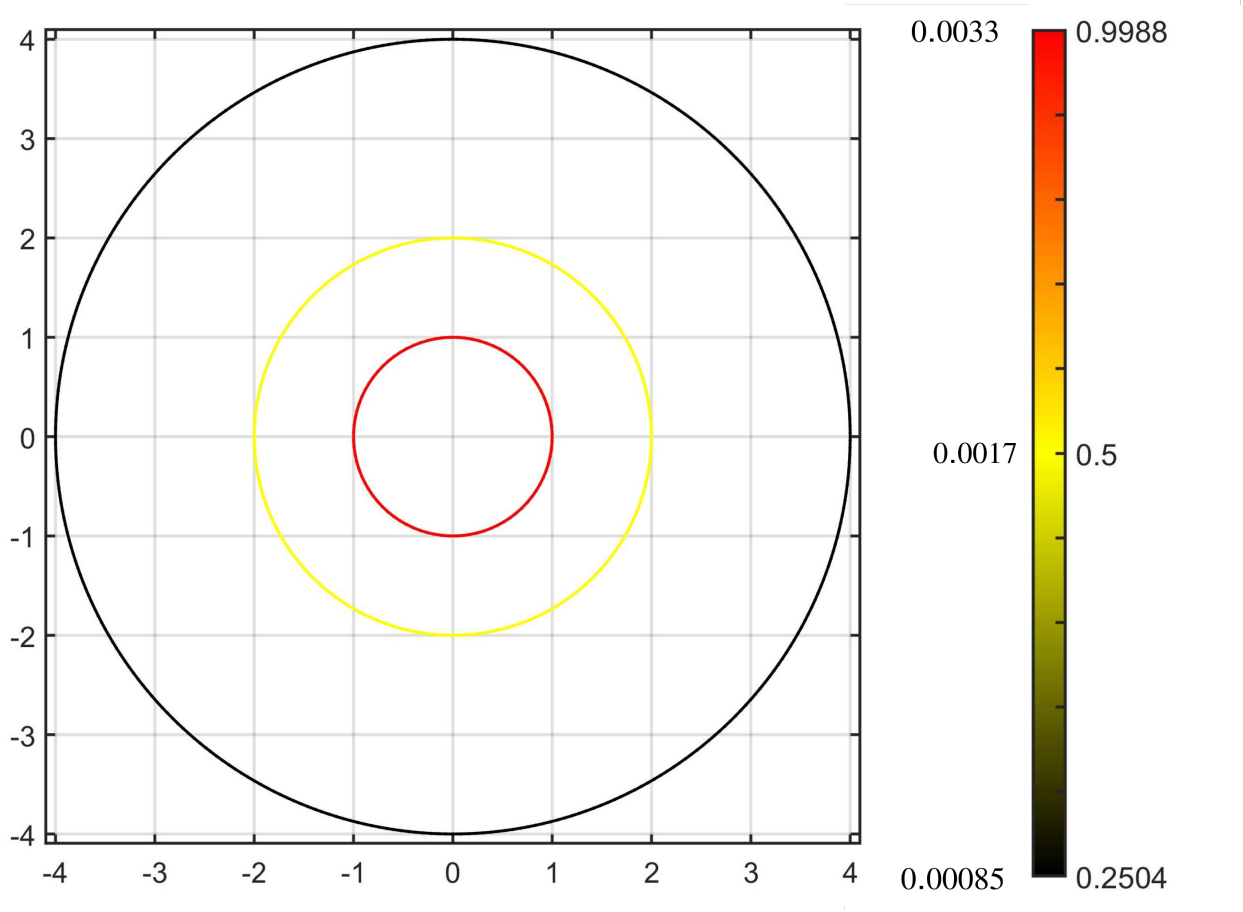


Figure 4.5: Curvature computed on the recovered graphs of three concentric circles data sets. the radii of the circles are $1, 2$ and 4, respectively. The curvature values for the corresponding circles are shown in the colormap (algebraic and geometric on its left and right side respectively). Note the relation to the inverse of the radius.

ing the two described formulations. The results are presented in figure 4.5. As expected, the geometrically estimated curvatures (right column of color map) closely follow the r_0^{-1} relation - the curvatures on circles of radii 1, 2 and 4 are approximately 1, 0.5 and 0.25, respectively. The curvatures estimated via the algebraic formulation (left column of color map) are still related to the inverse of the circles' radii. In this particular case, the normalizing scalar is the curvature of the smallest circle, since in this case it has to be equal to 1 to coincide with the geometric interpretation. Dividing the algebraic curvatures of the outer circles by the normalizing factor, they read ~ 0.5 and ~ 0.25, in agreement with the geometric formulation.

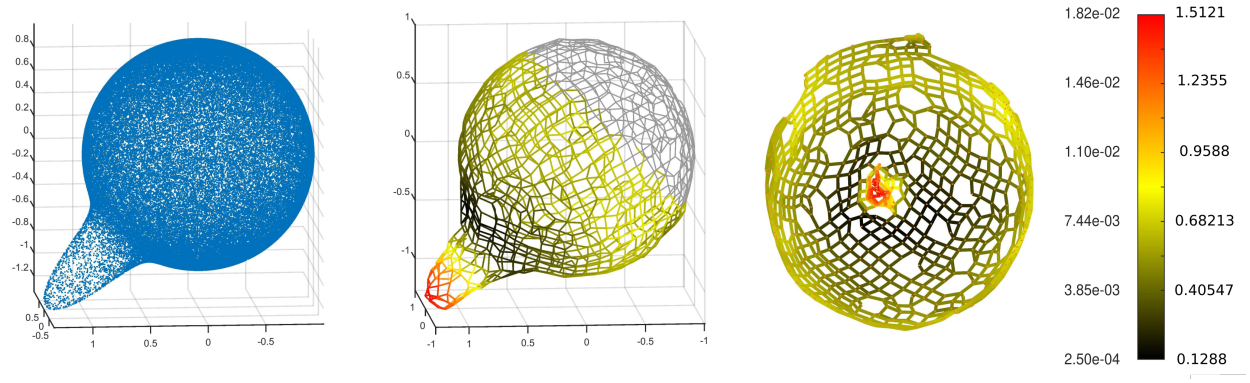


Figure 4.6: Data set (left panel), extracted graph (central panel) and planar representation (right panel) for manifold $\mathcal{M}(x, y, z)$, and its on-edge curvature distribution in the two formulations: algebraic and geometric (left and right respectively).

As a further illustrative example, consider the case of a 2-dimensional manifold $\mathcal{C}$ embedded in $\mathbb{R}^3$, represented in figure 4.6, left panel. Manifold $\mathcal{C}$ can be expressed in terms of a "smooth" and a "peaked" component, respectively $\mathcal{C}_S$ and $\mathcal{C}_P$. The "smooth" component $\mathcal{C}_S$ is generated by a uniform distribution $\mathcal{U}(\theta, \phi)$ over the angular components of the spherical coordinate system $(\theta, \phi) \in \mathcal{I}_\theta \times \mathcal{I}_\phi = [0, 2\pi] \times [-\pi/2, \pi/2]$ with fixed radius $r = 1$. The "peaked" component is an additive term to coordinate r generated by a Gaussian distribution $\mathcal{N}(\mu_P, \mathbf{C}_P)$, where $\mu_P = (\theta_P, \phi_P) = (\pi/2, -\pi/4)$ and covariance matrix $\mathbf{C}_P = \text{diag}(\pi/90, \pi/180)$. In other words, a varying radius is defined over the angular component of the coordinates as:

$$r_{\mathcal{C}} \sim 1 + 1.5 \times \mathcal{N}(\mu_P, \mathbf{C}_P); \tag{4.31}$$

Manifold $\mathcal{C}$ is then obtained by converting the spherical coordinate system to the cartesian coordinates x, y, z using:

$$x = r \sin\theta \cos\phi; \quad y = r \sin\theta \sin\phi; \quad z = r \cos\theta.$$

By applying manifold crawling and AGTM to this sample a smooth manifold is recovered through the basis functions formulation and, for every edge in the abstract latent graph a

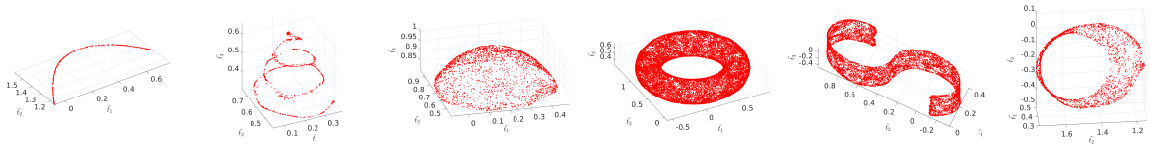
value for the curvature as defined in equation (4.30) is computed. Figure 4.6 (central panel) presents the extracted graph where edges are color-coded by their curvature (only curvature for the portion of the graph containing the peak and the adjacent spherical structure is shown). The visualization in figure 4.6, right panel presents the same portion of the graph as in the central one, but unfolded on a plane. Here the central region captures the extreme curvature of the culminating part of the peak. The curvature decreases to a minimum where the Gaussian peak smoothly approaches the sphere's surface, but steadily converges to the constant curvature of the spherical surface in the outer edges of the planar representation. This behaviour can also be verified in the central panel and is to be expected by construction of the data set. Once again, the two formulations are in agreement, up to a scalar, in the description of the curvature over the edges of the manifold.

Chapter Five

Multiple noisily sampled manifolds

5.1 Diffusion and background noise removal

The previously described methodology for modelling a single noisy manifold is now extended to the case where multiple noisy manifolds of different dimensionalities coexist in the same data set $\mathcal{Q}$. First, the extended methodology for a general data set $\mathcal{Q}$ is described, for later focusing on the tests performed on $\mathcal{Q}^1$ and $\mathcal{Q}^2$, described in chapter 3. At first, a physics-based diffusion method is applied, Structure-Aware Filtering technique (SAF, Wu et al., 2018), that collapses points in close vicinity of dense structures onto them, resulting in a diffused data set $\tilde{\mathcal{Q}} = \{\tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_2, ..., \tilde{\mathbf{t}}_L\}$, $\tilde{\mathbf{t}}_i \in \mathbb{R}^d$. The SAF method moves points towards high density regions, enabling points in the vicinity of a noisy manifold to migrate towards its "spine" or "mean surface". Figure 5.1 shows the noisy manifolds described previously (subfigures 3.3g-3.3l) together with the corresponding recovered mean manifolds via SAF (subfigures 5.1a-5.1f). Despite a few imperfections when compared to their ground-truths (e.g. deformation of the spiral in the second ring from the bottom), the method generally reduces transverse noise to the manifolds. Assuming that the data structures to be modeled are more densely sampled than the noisy environment, the data sets $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ is filtered



(a) (b) (c) (d) (e) (f)

Figure 5.1: Diffused manifolds using SAF technique.

by removing point couples $(\mathbf{t}_i, \tilde{\mathbf{t}}_i)$ that have sparse neighbourhood in *both* $\mathcal{Q}$ *and* $\tilde{\mathcal{Q}}$. In particular, around each $\mathbf{t}_i \in \mathcal{Q}$ and $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ a hyperball $\mathcal{B}(\mathbf{t}_i; r)$ and $\mathcal{B}(\tilde{\mathbf{t}}_i; r)$ is constructed, respectively, in $\mathbb{R}^d$ of radius $r > 0$. In case a point $\mathbf{t}_i \in \mathcal{Q}$ lies further apart from a manifold, both $\mathcal{B}(\mathbf{t}_i; r)$ and $\mathcal{B}(\tilde{\mathbf{t}}_i; r)$ will be sparsely populated. Hence, if both $\mathcal{B}(\mathbf{t}_i; r)$ and $\mathcal{B}(\tilde{\mathbf{t}}_i; r)$ contain less points from $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$, respectively, than a pre-specified threshold $\tau > 0$, the points $\mathbf{t}_i$ and $\tilde{\mathbf{t}}_i$ are removed from their corresponding data sets.

Following this, the first task in capturing the multi-manifold structure in $\tilde{\mathcal{Q}}$ is to estimate local dimensionality of the cloud point around each $\tilde{\mathbf{t}}_i$ in the form of a *dimensionality index* δ_i (section 5.2). Using the dimensionality indices, partitions of the data into subsets $\mathcal{Q}_j, \tilde{\mathcal{Q}}_j$ are created according to the local dimensionalities $j = 1, 2, ..., d$. Since $\mathcal{Q}_j, \tilde{\mathcal{Q}}_j$ can contain several distinct sampled manifolds of dimensionality j, a dedicated "manifold crawling" procedure operating on $\tilde{\mathcal{Q}}$ is used to separate the individual manifolds (section 5.3). Moreover, the crawling also produces for each manifold a graph structure embedded in $\mathbb{R}^d$ representing a piece-wise linear "skeleton" approximation of the spine of the noisy manifold (section 4.2). For every manifold, the associated graph will then function as an abstract latent space of a generalized form of Generative Topographic Mapping (GTM Bishop et al., 1998b) called *Abstract GTM* (AGTM). Using points in $\mathcal{Q}$, the generalized GTM produces manifold aligned density models (section 4.1). Finally, if a single summary model is required, density models of all detected manifolds (AGTMs) can be grouped together in a hierarchical mixture model Tino et al., 2002 representing in a concise manner the global density of the low-dimesional structures in dataset $\mathcal{Q}$. An alternative approach to produce a de-

noised version of the noisy data set can be achieved by first recovering noisy low-dimensional structures with the aid of an ant-colony inspired algorithm, LAAT: Locally Aligned Ant Technique (Taghribi et al., 2020). This methodology lets an ant-colony jump between neighboring points, while leaving a trace of pheromone on the visited points. The pheromone is an additive scalar value that evaporates through time (jump iterations). The jumping probability of each ant is dictated by the local density in the surrounding of the point and the current value of pheromone in the surrounding points. After a certain number of iterations the pheromone distribution on the data set tends to highlight the low-dimensional structures (independently on their dimensionalities) so that denser regions have higher pheromone. By thresholding the pheromone value, the uniform noise can be discarded, while the noisy low-dimensional structures preserved. Taking the end result of LAAT, a different ant-colony inspired algorithm can be applied to reduce the transverse noise on each structure. EM3A: Evolutionary Manifold Alignment Aware Agents (Mohammadi et al., 2020), uses local information to reconstruct the manifolds and then move data points close to them. Moreover, principles of evolutionary game theory are used to encourage ants to select better strategies and hence optimize the hyper-parameters automatically. While the effect is similar to SAF, its implementation and fundamental approach stems from a completely different idea.

5.2 Intrinsic dimensionality of manifolds

Around each $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ local Principal Component Analysis (PCA, alternatively one could use Singular Value Decomposition, SVD) is performed using points from $\mathcal{B}(\tilde{\mathbf{t}}_i; r) \cap \tilde{\mathcal{Q}}$, obtaining eigenspectrum $\lambda_{i,1} \geq \lambda_{i,2} \geq ... \geq \lambda_{i,d}$.

The dimensionality index of $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$ used in Xiaoxia Wang et al., 2008 (limited to

3-dimensional data and derived from Mordohai et al., 2005) was obtained as

$$\delta_i^O = \underset{j}{\operatorname{argmax}}\, S_{i,j} \tag{5.1}$$

where $S_{i,1} = \lambda_{i,1} - \lambda_{i,2}$, $S_{i,2} = 2(\lambda_{i,2} - \lambda_{i,3})$ and $S_{i,3} = 3\lambda_{i,3}$.

Here, a general method is proposed for computing dimensionality index of points distributed in spaces of arbitrary finite dimension d, based on renormalized eigenvalues,

$$\tilde{\lambda}_{i,j} = \frac{\lambda_{i,j}}{\sum_{k=1}^{d} \lambda_{i,k}},$$

viewed as "likelihoods" of different dimensionalities j of the cloud of points around $\tilde{\mathbf{t}}_i$. Note that $\tilde{\Lambda}_i = (\tilde{\lambda}_{i,1}, \tilde{\lambda}_{i,2}, ..., \tilde{\lambda}_{i,d})$ lies in the $(d-1)$-dimensional simplex $\mathcal{S}_0$ with vertices $\mathbf{s}_1 = (1, 0, 0, \ldots, 0)$, $\mathbf{s}_2 = (1/2, 1/2, 0, \ldots, 0)$, $\mathbf{s}_3 = (1/3, 1/3, 1/3, \ldots, 0)$, $\ldots \mathbf{s}_d = (1/d, 1/d, \ldots, 1/d)$. The simplex $\mathcal{S}_0$ is a subset of the standard simplex $\mathcal{S}$ with vertices equal to the standard basis $\mathbf{e}_1 = (1, 0, 0, \ldots, 0); \mathbf{e}_2 = (0, 1, 0, \ldots, 0); \mathbf{e}_d = (0, 0, \ldots, 1)$. Considering $\mathcal{S}$ as the simplex of multinomial distributions, the appropriate Riemannian geodesic distance is the Fisher distance Lebanon, 2005. In particular, for any $\tilde{\Lambda}_k, \tilde{\Lambda}_l \in \mathcal{S}$,

$$d_J(\tilde{\Lambda}_k, \tilde{\Lambda}_l) = 2 \arccos\left(\sum_{i=1}^{d} \sqrt{(\tilde{\lambda}_{ki} \cdot \tilde{\lambda}_{li})} \right). \tag{5.2}$$

Note that the vertex $\mathbf{s}_j$ of $\mathcal{S}_0$ corresponds to the ideal normalized eigen-spectrum of a j-dimensional neighbourhood. Hence, it is possible to quantify the degree to which the neighbourhood of $\tilde{\mathbf{t}}_i$ resembles a j-dimensional hyperplane in $\mathbb{R}^d$ by the closeness of $\tilde{\Lambda}_i$ to $\mathbf{s}_j$ in terms of the geodesic distance (5.2),

$$\delta_i^G = \underset{j}{\operatorname{argmin}}\, d_J(\tilde{\Lambda}_i, \mathbf{s}_j). \tag{5.3}$$

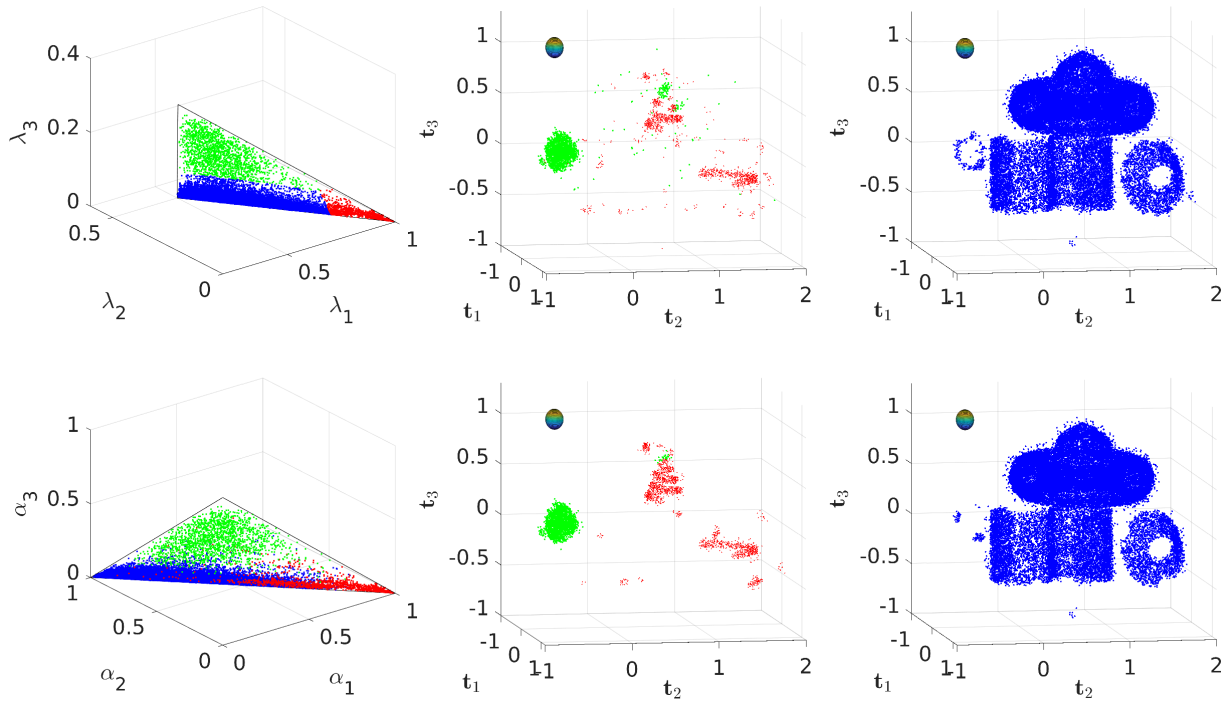


Figure 5.2: Eigenvalue distribution on simplex and dataset labelled using the dimensionality index δ^G (upper row) and δ^S (lower row).

A 'soft' and spatially smoothed version of dimensionality index is also here proposed. This is done by positioning an isotropic kernel

$$K(\alpha; \mathbf{e}_j) = \exp\left[-\frac{d_J(\alpha, \mathbf{e}_j)^2}{2\kappa^2}\right], \quad \alpha \in \mathcal{S},$$

on top of each vertex $\mathbf{e}_j$ of $\mathcal{S}$ with kernel scale κ set to the geodesic distance between the equidistant point $\mathbf{s}_d \in \mathcal{S}$ to every vertex of $\mathcal{S}$ (the 'center' of $\mathcal{S}$), i.e. $\kappa = d_J(\mathbf{e}_j, \mathbf{s}_d)$, for any j. As explained above, due to the imposed eigenvalue order, the normalized eigen-spectra $\tilde{\Lambda}_i$ live in the sub-simplex $\mathcal{S}_0$ of the standard simplex $\mathcal{S}$. $\mathcal{S}_0$ is transformed into $\mathcal{S}$, via the barycentric coordinates Möbius, 1827 $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,d})$ of $\tilde{\Lambda}_i$ by solving,

$$\tilde{\Lambda}_i = \sum_{j=1}^{d} \alpha_{i,j}\mathbf{s}_j, \quad \alpha_{i,j} \geq 0, \sum_{j=1}^{d} \alpha_{i,j} = 1.$$

The eigen-spectrum $\tilde{\Lambda}_i$ is thus mapped to the point in $\mathcal{S}$ with barycentric coordinates α_i.

This leads to a soft kernel-based dimensionality distribution

$$P_i(j) = \frac{K(\alpha_i; \mathbf{e}_j)}{\sum_{k=1}^{d} K(\alpha_i; \mathbf{e}_k)}.$$

It may happen that the normalized index distribution P_i is abruptly changing across neighbouring points due to low sample density (e.g., at the manifold edges). Therefore spatially smoothing operator P_i is implemented, using a Gaussian filter[1]

$$c(i,l) = \exp\left[-\|\tilde{\mathbf{t}}_i - \tilde{\mathbf{t}}_l\|^2/(2r^2)\right].$$

The smoothed normalized index distribution reads:

$$P_i^S(j) = \frac{1}{\sum_{\tilde{\mathbf{t}}_l \in \mathcal{B}(\tilde{\mathbf{t}}_i, r)} c(i,l)} \sum_{\tilde{\mathbf{t}}_l \in \mathcal{B}(\tilde{\mathbf{t}}_i, r)} c(i,l) \cdot P_l(j), \tag{5.4}$$

where the sum is taken over diffused points $\tilde{\mathbf{t}}_l$ in the spherical neighbourhood of $\tilde{\mathbf{t}}_i$ of radius r. The smoothed dimensionality index of $\tilde{\mathbf{t}}_i$ is then

$$\delta_i^S = \operatorname*{argmax}_j P_i^S(j). \tag{5.5}$$

Fig. 5.2 shows the results obtained using dimensionality indices δ^G (top row) and δ^S (bottom row). Each column presents the results of the two indices on the toy data set. The first column shows the 1- (red), 2- (blue) and 3-dimensional (green) eigenvalue-related distributions on the simplices $\mathcal{S}_0$ and $\mathcal{S}$ for δ_i^G and δ_i^S, respectively, for all points in the data set (after removing the low-density background noise). Central column shows points from $\mathcal{Q}$ with dimensionality index equal to 1 (red) and 3 (green). Points with dimensionality index 2 (blue) are shown in the last column. The central column clearly shows the beneficial effect of spatial smoothing of the dimensionality index. In both columns, a small ball of radius

[1] r is the radius used for the local PCA estimation

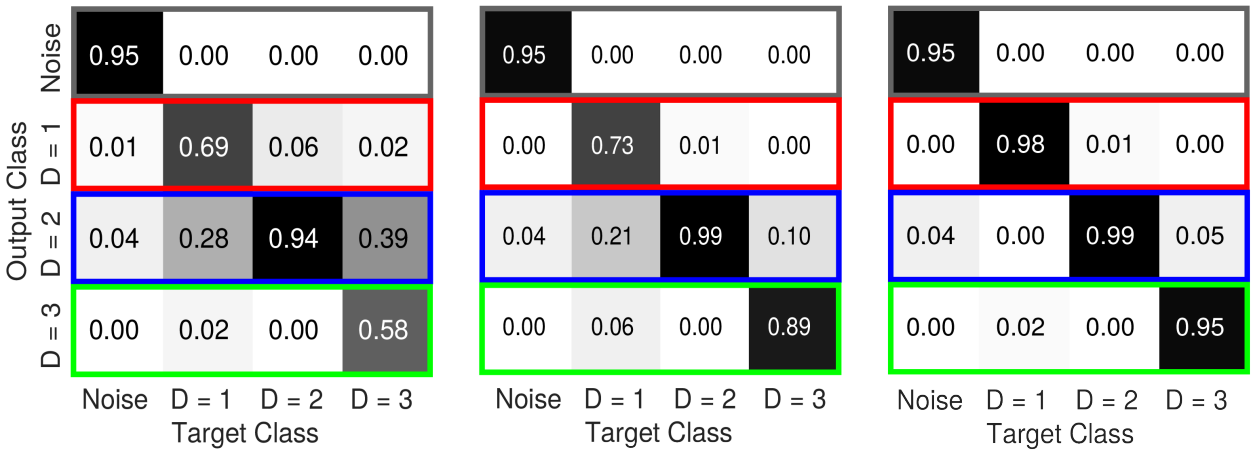


Figure 5.3: Confusion matrices for dimensionality indices δ_i^O (left), δ_i^G (center) and δ_i^S (right). The superior performance of δ_i^S is confirmed in fig. 5.2.

$r = 0.1$ is also shown. The same value of r is used as the characteristic scale in local manifold linearisations employed in dimensionality index calculations and manifold crawling on this data set.

Knowing a-priori the true dimensionality of each point, a qualitative analysis of the different dimensionality index formulations is provided in the form of confusion tables shown in fig. 5.3. The confusion tables present percentages of correctly and incorrectly classified points depending on their true dimensionality and their corresponding dimensionality indices. Also included, the "Noise" class containing points generated as the lower density background noise and supposed to be filtered out in the initial pre-processing step outlined in section 5.1. Left, middle and right panels show confusion matrices of indices δ_i^O, δ_i^G and δ_i^S, respectively. Compared with δ_i^O, reformulation of dimensionality index as proximity of renormalized eigen-spectrum to the "ideal" eigen-spectrum for each dimensionality $1 \leq j \leq d$ (index δ_i^G) stabilizes the estimation. The index is further stabilized by spatial smoothing of the soft version of δ_i^G, provided by index δ_i^S.

5.3 Multi-Manifold Crawling

Having obtained dimensionality index δ_i^S for each point $\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}}$, and hence also for the corresponding point $\mathbf{t}_i \in \mathcal{Q}$, a partition of the two data sets (noisy and diffused) is performed according to their dimensionality indices, $\mathcal{Q}_j = \{\mathbf{t}_i \in \mathcal{Q} |\ \delta_i^S = j\}$ and $\tilde{\mathcal{Q}}_j = \{\tilde{\mathbf{t}}_i \in \tilde{\mathcal{Q}} |\ \delta_i^S = j\}$. For a given $1 \leq j < d$, the data sets $\mathcal{Q}_j$ and $\tilde{\mathcal{Q}}_j$ contain the noisy and "crisp" samples, respectively, along manifolds of intrinsic dimension j.

While the Manifold Crawling algorithm of section 4.2 can be used to crawl and assist model building of a single manifold, the method can be extended to crawl and isolate multiple manifolds of a given dimensionality j. The "to be processed" data set $\mathcal{R}$ is initialized as $\tilde{\mathcal{Q}}_j$. Then the Manifold Crawling algorithm is run on $\mathcal{R}$ with a scale parameter r. This results in graphs $\mathcal{G}^1$ and $\overline{\mathcal{G}}^1$, forming skeletons of individual manifolds of dimensionality j. The underlying assumption is that the separation between manifolds in $\tilde{\mathcal{Q}}$ is greater than r.

For every node $\overline{\mathbf{v}}_l$ of $\overline{\mathcal{G}}^1$ its neighbourhood in $\mathcal{R}$ is computed as $\overline{\mathcal{O}}_l = \mathcal{B}(\overline{\mathbf{v}}_l, r) \cap \mathcal{R}$. The neighbourhoods of all nodes of $\overline{\mathcal{G}}^1$, $\bigcup_l \overline{\mathcal{O}}_l$, are then subtracted from $\mathcal{R}$. At the same time, by computing $\mathcal{O}_l = \mathcal{B}(\overline{\mathbf{v}}_l, r) \cap \mathcal{Q}_j$ for every $\overline{\mathbf{v}}_l \in \overline{\mathcal{V}}^1$, the "noisy" data set $\mathcal{A}^1 = \bigcup_l \mathcal{O}_l$ aligned along the 1-st manifold of dimensionality j is recovered.

The manifolds sampled in $\tilde{\mathcal{Q}}_j$ could be separated by larger distances, so that the previous crawling run may not have reached them. Therefore, while $\mathcal{R} \neq \emptyset$, another seed is picked at random from $\mathcal{R}$ and a new crawling procedure is initiated.This produces graphs $\overline{\mathcal{G}}^2$ and $\mathcal{G}^2$, accompanied by the corresponding manifold-aligned data $\mathcal{A}^2$. After reduction of $\mathcal{R}$ as described above, while $\mathcal{R} \neq \emptyset$, the whole process is repeated.

Finally, assuming the processing of $\tilde{\mathcal{Q}}_j$ took H runs, the procedure results in a collection of H extracted j-dimensional manifolds represented by the graphs $\left\{\overline{\mathcal{G}}^\ell\right\}_{\ell=1}^H$, $\left\{\mathcal{G}^\ell\right\}_{\ell=1}^H$ and the associated manifold specific data sets $\left\{\mathcal{A}^\ell\right\}_{\ell=1}^H$. The individual graphs $\overline{\mathcal{G}}^\ell$, $\mathcal{G}^\ell$ and data

sets $\mathcal{A}^\ell$ can now be used by the AGTM algorithm (section 4.1) to construct the manifold-aligned density models for the manifolds of intrinsic dimensionality j.

To guarantee smoothness of the AGTM density estimates, the ϵ-net parameter for basis function positioning in the abstract graphs (see section 4.1) is conservatively set to $\epsilon = 1$. Due to relatively modest curvature in the low-dimensional structures, the crawling step size parameter η (see section 4.2) was set to $\eta = 0.75$. The tolerance scale parameter in the contraction phase of the crawling procedure is set to $\beta = 0.4$. As mentioned earlier, the characteristic scale parameter is set to $r = 0.1$. Fig. 5.4 shows the embedded graphs $\overline{\mathcal{G}}^\ell$ (black lines) and overlaid AGTMs (red and blue pdf isosurfaces) of manifolds in the example data set described in section 3.2, plotted by intrinsic dimensionality of the corresponding manifolds ($j = 1$, left; $j = 2$, right). The results were insensitive to small variations in the parameters r, η and β. Larger values of η can cause "overshooting" of highly curved local manifold structures, while smaller values of η are harmless, modulo increased computational complexity. Likewise, larger values of β can result in neglecting important details of the manifold structure, while smaller values are harmless, but can lead to more mixture components in the final probabilistic model. The most important parameter is the characteristic scale r governing local manifold linearisations used in determination of dimensionality indices and manifold crawling. This can be set based on prior knowledge, or after inspection of the data set. Alternatively, the multimanifold learning method can be used as a semi-automatic exploratory tool by the domain experts, where the focus and characteristic scale r of the structures to be mined can be varied continuously, starting from capturing coarse structures in the data (larger r), and proceeding with more detailed analysis, as potentially more and more low-dimensional manifolds start to appear with decreasing r.

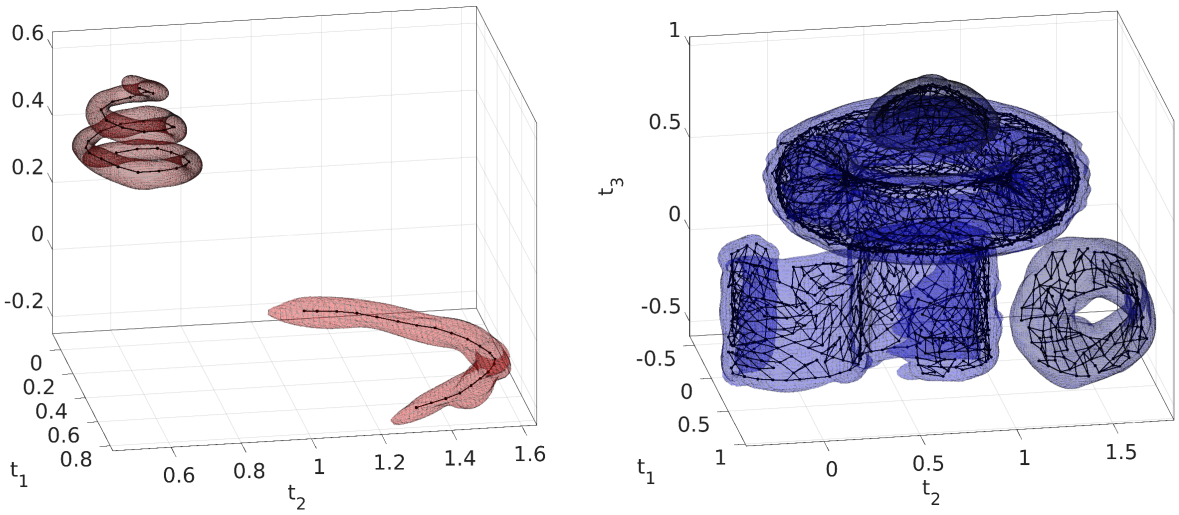


Figure 5.4: Both panels show the probabilistic model of each manifold in data set $\mathcal{D}$, recovered through AGTM, highlighted as an isosurface of the respective Probability Density Function. The black skeletons in each panel are the embedded graphs, as obtained after MultiManifold Crawling and training of AGTM.

5.3.1 An efficient alternative to Multi-Manifold Crawling

One can approximate the results of Multi-manifold crawling by constructing an ϵ-neighborhood graph on specific samples of the data set $\tilde{\mathcal{Q}}_j$, with ϵ set to r, the radius used for local PCA in the crawling algorithm.

First, $\tilde{\mathcal{Q}}_j$ is covered via L overlapping hyper-balls $\mathcal{B}(\tilde{\mathbf{t}}_k, c)$, $k = 1, 2, ..., L$, of radius $c = r/2$, centered at $\tilde{\mathbf{t}}_k \in \tilde{\mathcal{Q}}_j$ located using the recursive algorithm of X. Wang et al., 2009. Hence, $\tilde{\mathcal{Q}}_j = \cup_{k=1}^{L} \mathcal{S}_k$, where $\mathcal{S}_k \subseteq \mathcal{B}(\tilde{\mathbf{t}}_k, c) \cap \tilde{\mathcal{Q}}_j$. For every set $\mathcal{S}_k$, its center $\mathbf{s}_k$ is defined as its sample mean, obtaining the set of centers $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_L\}$. By design, the minimum Euclidean distance between any pair of centers $\mathbf{s}_m, \mathbf{s}_n \in \mathcal{S}$, $m \neq n$ is:

$$\frac{r}{2} < \min_{m \neq n} d_E(\mathbf{s}_m, \mathbf{s}_n) < r.$$

The distance between all centers is now computed, obtaining a square, symmetric matrix $D_{m,n} = d_E(\mathbf{s}_m, \mathbf{s}_n), \forall m \neq n, m, n = 1, \ldots, L$, from which an initial estimate of the adjacency

matrix $\mathbf{A}$ of graph $\overline{\mathcal{G}}$ is recovered. The adjacency matrix has non-zero elements only for elements of $\mathbf{D}$ smaller than r:

$$A_{m,n} = \begin{cases} 1 & \text{if } D_{m,n} \leq r; \\ 0 & \text{Otherwise.} \end{cases} \tag{5.6}$$

This (initial) adjacency matrix $\mathbf{A}$ tends to be overly connected, because it does not take into consideration the alignment of the tangent spaces of neighbouring partition centers. In order to include this geometrical information in $\mathbf{A}$, for every set $\mathcal{S}_k$its eigen-decomposition is computed so that an approximation of the local tangent space to the unknown manifold $\mathcal{M}_\ell$ at center $\mathbf{s}_k$ is given by the first j eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_j$ collected in matrix $\mathbf{U}_k = [\mathbf{u}_1, \ldots, \mathbf{u}_j]$. For each pair of connected points (whose corresponding entry in the initial adjacency matrix is 1), the angle between the corresponding tangent spaces Gunawan et al., 2005 is computed as $|\Theta_{m,n}| = \arccos(|\det \mathbf{M}_{m,n}|)$, where $\mathbf{M}_{m,n} = \mathbf{U}_m^\top \mathbf{U}_n \in \mathbb{R}^{j \times j}$. The existing edge is retained only if the angle between the connected points is $|\Theta_{m,n}| \leq \pi/3$. Alternatively, one could define an additional hyper-parameter $0 < T_\theta < 1$ so that smaller or larger angles are tolerated in the construction of the graph. The condition on the final adjacency matrix would then be:

$$A_{m,n} = \begin{cases} 1 & \text{if } D_{m,n} \leq r \;\; \& \;\; |\Theta_{m,n}| < T_\theta \cdot \pi/2; \\ 0 & \text{Otherwise.} \end{cases} \tag{5.7}$$

In this work, $T_\theta = \frac{2}{3}$.

Having constructed the adjacency matrix $\mathbf{A}$ for a set $\tilde{\mathcal{Q}}_j$, graph $\overline{\mathcal{G}}$, corresponding to $\mathbf{A}$, is split into its maximal connected components $\mathcal{K}_c = (\overline{\mathcal{V}}_c, \overline{\mathcal{E}}_c)$, $c = 1, \ldots, N_G$, Here, $\overline{\mathcal{V}}_c$ and $\overline{\mathcal{E}}_c$ are the sets of centres (nodes) and edges, respectively, of the component $\mathcal{K}_c$. For each component the corresponding abstract graph is defined as $\mathcal{K}_c = (\mathcal{V}_c, \mathcal{E}_c)$ as in the

previous section. Now, the N_G extracted pairs of embedded and abstract graphs can be used as initialization to AGTM, as an alternative to the Multi-manifold crawling initialization presented in the last section.

5.4 Comparison to existing methods

While multi-manifold AGTM covers both multi-manifold learning and probabilistic modelling simultaneously, most methodologies in the literature focus on one or the other aspect of capturing spatial structures in point clouds. The evaluation of AGTM is therefore split in comparative experiments along two lines: (1) recovery of a collection of smooth manifolds of possibly different dimensionalities and (2) full probabilistic modelling of the data distribution.

Two main data sets are used. The first one is data set $\mathcal{Q}^1$ presented in section 3.2 composed of six sampled manifolds of different dimensionalities embedded in $\mathbb{R}^3$. Recall the presence of two non-linear $1-$dimensional (curved spiral and hyperbolic arm) and four $2-$dimensional manifolds (toroid, Möbius strip, parabolic cap, S-shape). The second data set $\mathcal{Q}^2$ consists of three interlocked noisy toroids (figure 3.4). Both data sets are embedded in uniformly distributed noise over the domain. All methods for this comparison are applied to the data sets after the initial filtering methodology (section 5.1), obtaining both a diffused $\tilde{\mathcal{Q}}^i$ and noisy $\mathcal{Q}^i$ version of each data set.

5.4.1 Multi-manifold learning

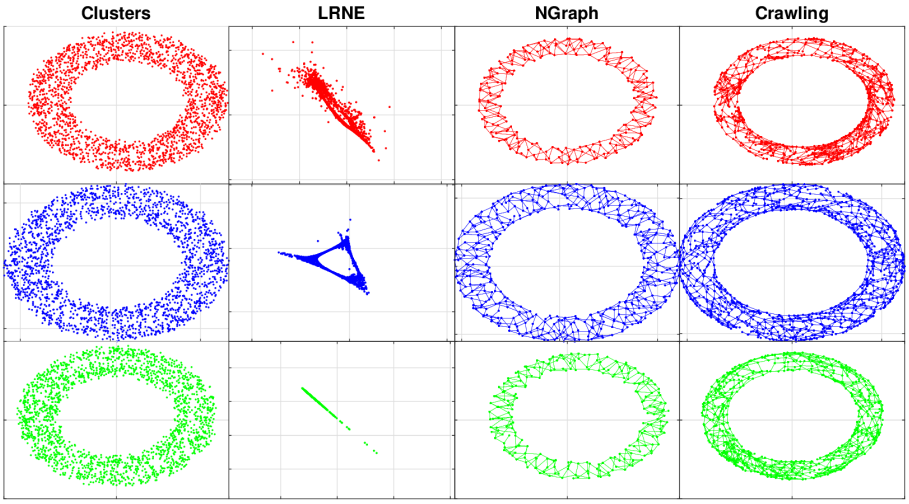
In this section manifold crawling is compared with the alternative efficient graph construction presented in section 5.3.1, as well as with two main algorithms found in the literature, for which the codes are publicly available: *Sparse Manifold Clustering and Embedding* (SMCE

Elhamifar et al., 2011) and *Low Rank Neighborhood Embedding* (LNRE Saranathan et al., 2019). The four methodologies are tested on data sets $\tilde{\mathcal{Q}}^1$ and $\tilde{\mathcal{Q}}^2$ in order to verify their ability to deal with multiple manifolds of different dimensionalities and topologies. Both SMCE and LRNE are designed to cluster a mixture of low-dimensional manifolds embedded in a higher dimensional space, while recovering a low-dimensional embedding for each manifold. In both methods the number of manifolds in the data set has to be provided by the user. In contrast, Multi-manifold crawling and its alternative graph construction recover the number of manifolds in the data set automatically, once the characteristic size of the manifolds is provided. Figure 5.5a, presents a qualitative comparison[2] of the embeddings recovered by LRNE (second column), the efficient alternative to Manifold Crawling (third column) and Manifold Crawling (fourth column) on data sets $\tilde{\mathcal{Q}}^2$. While the clustering performance is similar for all methods (first row), the embeddings recovered by LRNE do not explicitly display the properties of the three toroids. The graphs recovered both by Manifold Crawling and its alternative graph construction provide much more insight into the topological structures underlying the sampled noisy manifolds. The stability of Multi-manifold crawling and the $\epsilon-$graph alternative with respect to (hyper-)parameter settings is also tested. To that end, the number of recovered manifolds is evaluated over a range of hyper-parameter values (figure 5.5b, upper row). The blue box-plots show the statistics over the number of manifolds recovered by the $\epsilon-$graph construction, for 30 different samplings of data set $\mathcal{Q}^2$ at a given radius. The black compact box-plots[3] show the same measure for 30 iterations

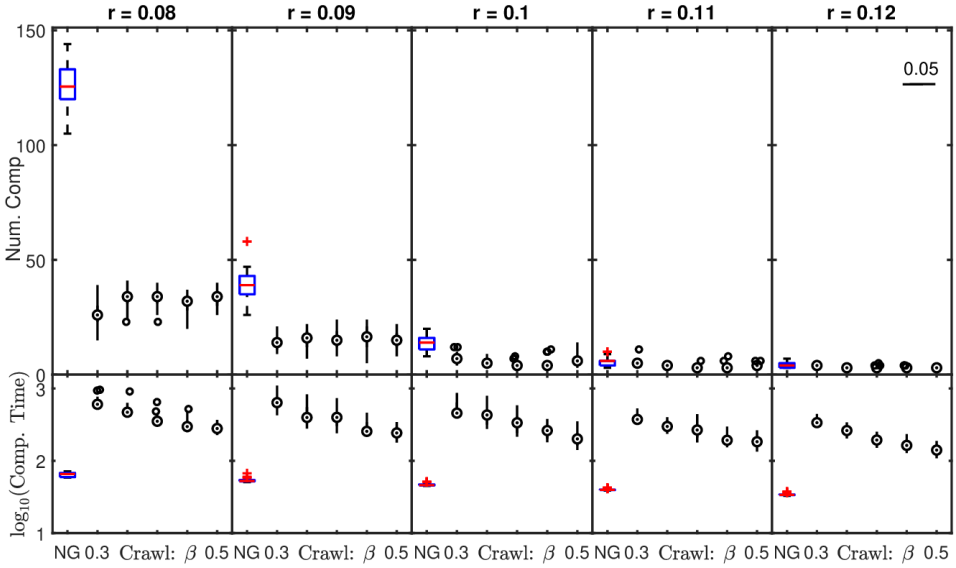
Methods	Parameters	
$\epsilon-$graph	$r \in [0.08, 0.09, \ldots, 0.17]$	
Crawling	$r \in \{0.08, 0.09, \ldots, 0.17\}$	$\beta \in \{0.3, 0.35, \ldots, 0.5\}$
LRNE	$k \in \{25, 30, \ldots, 50\}$	$l_{LRNE} \in \{1e-3, 5e-3, 1e-2, \ldots, 10\}$
SMCE	$k \in \{30, 35, 40, 45, 50\}$	$l_{SMCE} \in \{1e-3, 1e-2, \ldots, 1e2\}$

Table 5.1: Parameters used for all Manifold Learning methods in this analysis for both data sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$.

[2]Parameters of each method were tuned to optimize performance on $\tilde{\mathcal{Q}}^2$.



(a)



(b)

Figure 5.5: Comparison of embeddings obtained for data set $\mathcal{Q}^2$ via LRNE (second column) Multi-Manifold Crawling (rightmost column) and its $\epsilon-$mesh alternative (third column). The clusters shown in the leftmost column are equally recovered by the three methods (figure 5.5a). Figure 5.5b shows the number of components (top row) and computational time (bottom row) for Multi-Manifold Crawling (black boxplots) and its $\epsilon-$mesh alternative (blue boxplots, identified by label "NG") evaluated over 30 samplings of synthetic data set $\mathcal{Q}^2$.

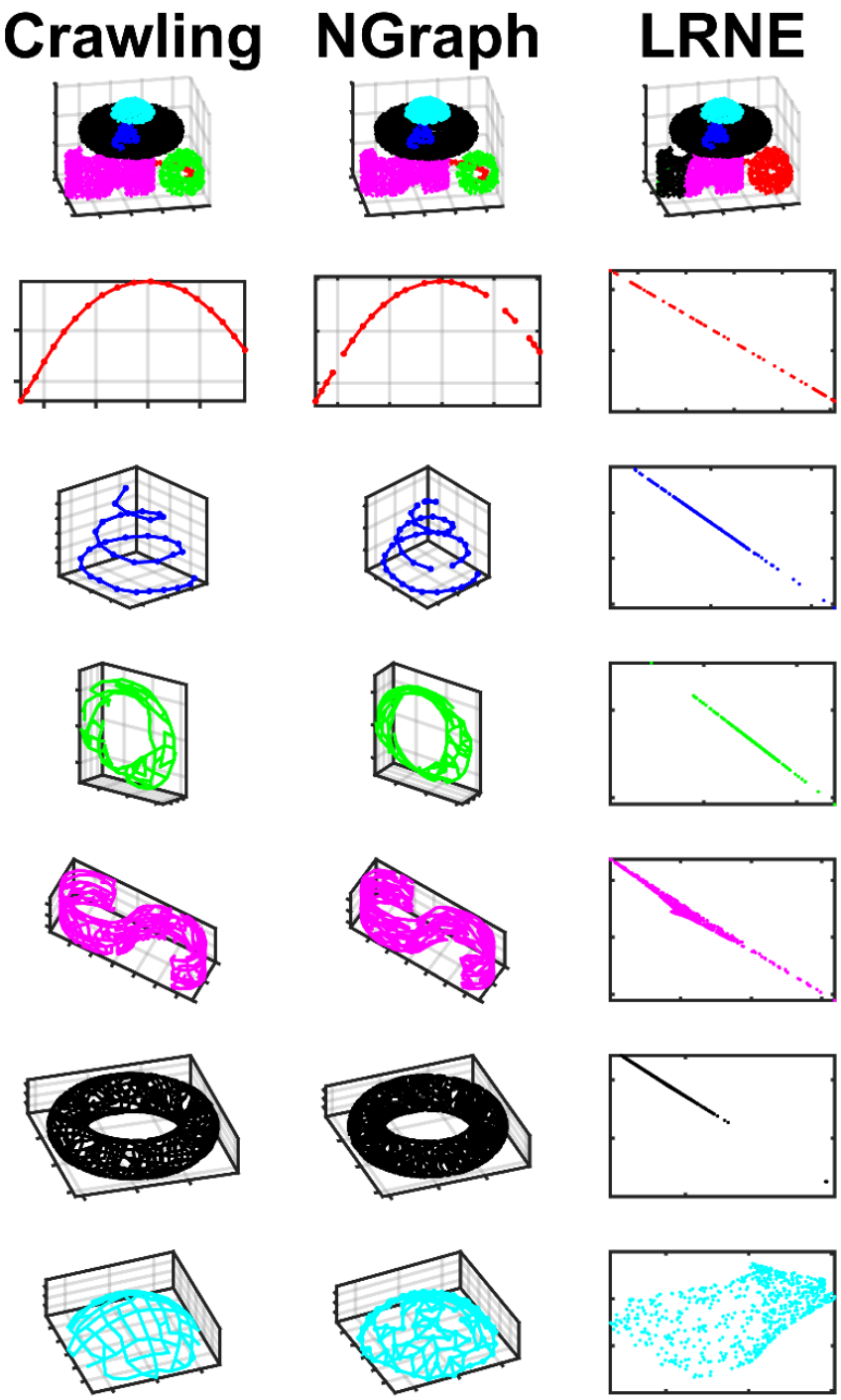


Figure 5.6: Same as figure 5.5a, but for data set $\mathcal{Q}^1$.

of Multi-Manifold crawling over the same data set at the same radius, for different values of the parameter β. Not surprisingly, Multi-manifold crawling displays a higher stability then its alternative, over a range of values of the parameter r, while little change is due to the parameter β for a given radius. While both methods converge to the correct number of manifolds (three in this case) starting at $r = 0.1$, the efficient $\epsilon-$graph method vastly overestimates the number of manifolds when the radius is too small. The computational times are shown in figure 5.5b, lower row. This discrepancy in the quality of the results stems from the imposition, in the alternative methodology to crawling, of a "hard" angular threshold that controls variability of the angles between neighbouring tangent spaces. Even a slight deviation from this threshold may prevent two nodes belonging to the same manifold from being connected. The efficient alternative critically relies on angle relations between local PCA subspaces. The subspace decompositions become increasingly unreliable with decreasing radius, while no such reliance on subspace angles is needed in the crawling procedure.

The same comparison is shown in figures 5.6 and 5.7 for data set $\mathcal{Q}^1$. For both Crawling and its approximation, the clusters are obtained by applying the procedure outlined in section 5.3 and 5.3.1 respectively. While in the case of data set $\mathcal{Q}^2$, LRNE was able to successfully recover the manifolds for a set of carefully tuned parameter configurations, the results proved unsatisfactory when applied to data set $\mathcal{Q}^1$. In particular, no parameter setting was able to distinguish between the Möbius strip and the parabolic arm, as shown in figure 5.6, top right panel, red points. All the resulting embeddings (fig. 5.6, rightmost column), with the exception for the hyperbolic surface (cyan points, bottom panel) and the spiral (third top panel, blue points), fail in capturing the low-dimensional structure of the corresponding manifold.

[3]For visualization purposes, the black compact boxplots are shown as hollow dots for the means and black vertical lines for the variances, plus smaller black dots for outliers where present.

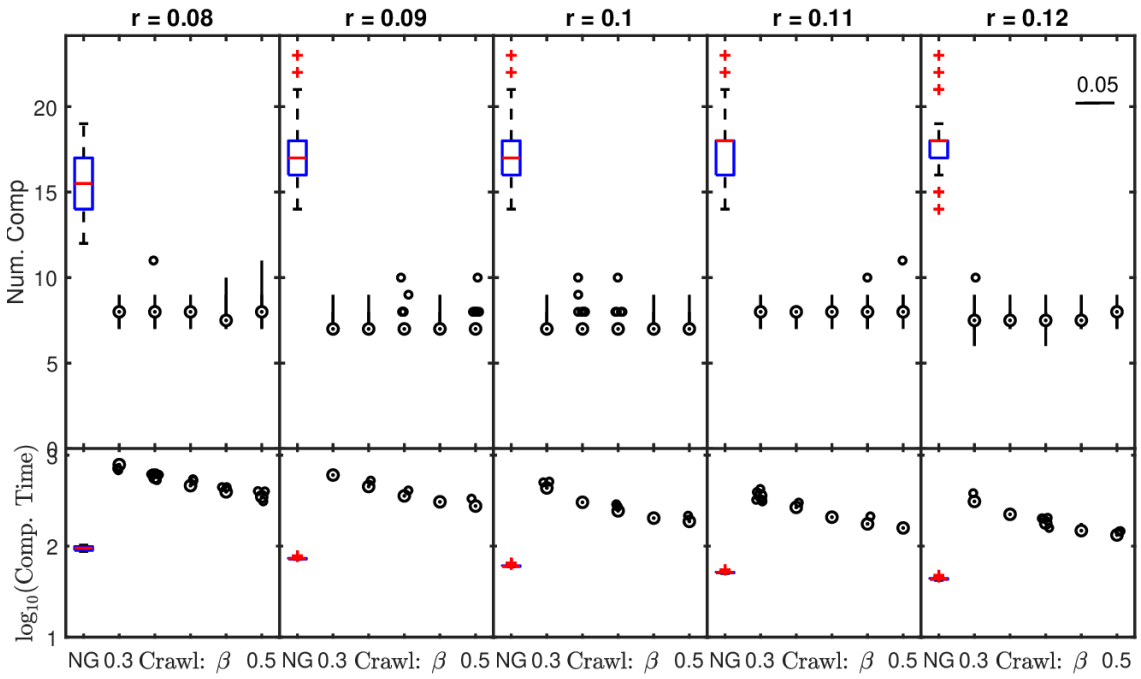


Figure 5.7: Same as figure 5.5b, but for data set $\mathcal{Q}^1$.

The first two columns of figure 5.6 illustrate the performance of Multi-manifold crawling and the $\epsilon-$graph alternative on $\mathcal{Q}^1$. Although the representation of the 2-dimensional manifolds is coherent over the two methodologies, the $\epsilon-$graph tends to break prematurely the 1-dimensional structures into smaller segments. This behaviour is confirmed over various values for the radius, as shown in figure 5.7, top row. Here, in contrast to the data set $\mathcal{Q}^2$, convergence to the correct number of manifolds is achieved only by Multi-Manifold Crawling, while the efficient $\epsilon-$graph alternative, consistently overestimates this measure. This clearly demonstrates the enhanced accuracy and stability of Multi-manifold crawling, of course, at the price of a higher computational time (figure 5.7, bottom row).

The application of SMCE with the set of parameters in table 5.1 has proven unsuccessful on both data sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$. In order to perform at least a qualitative comparison with the methodology proposed in the current work, manifold crawling is applied to the data set described in Elhamifar et al., 2011. The data set consists of two noise-less, 3-foiled in-

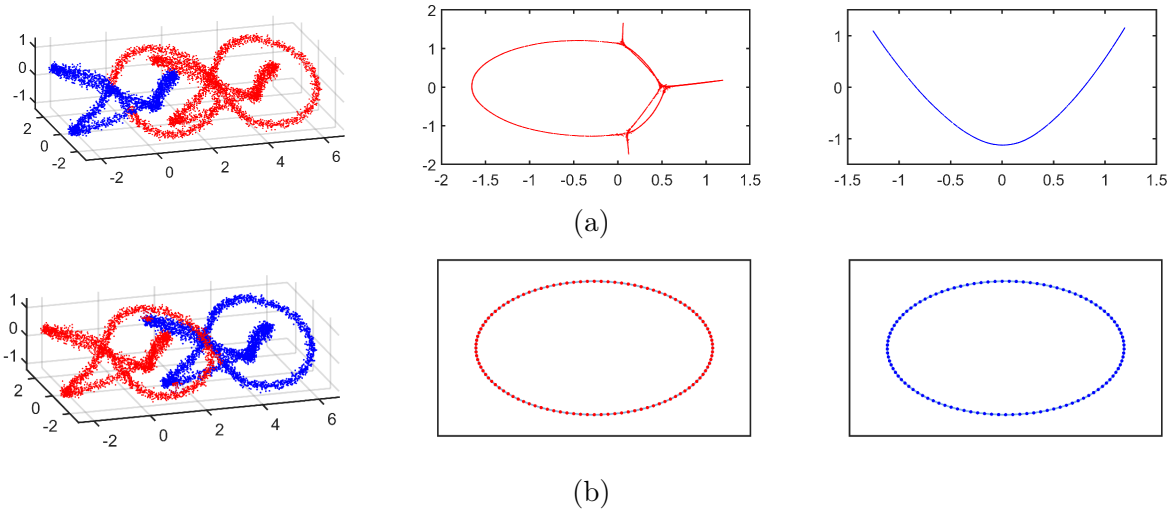


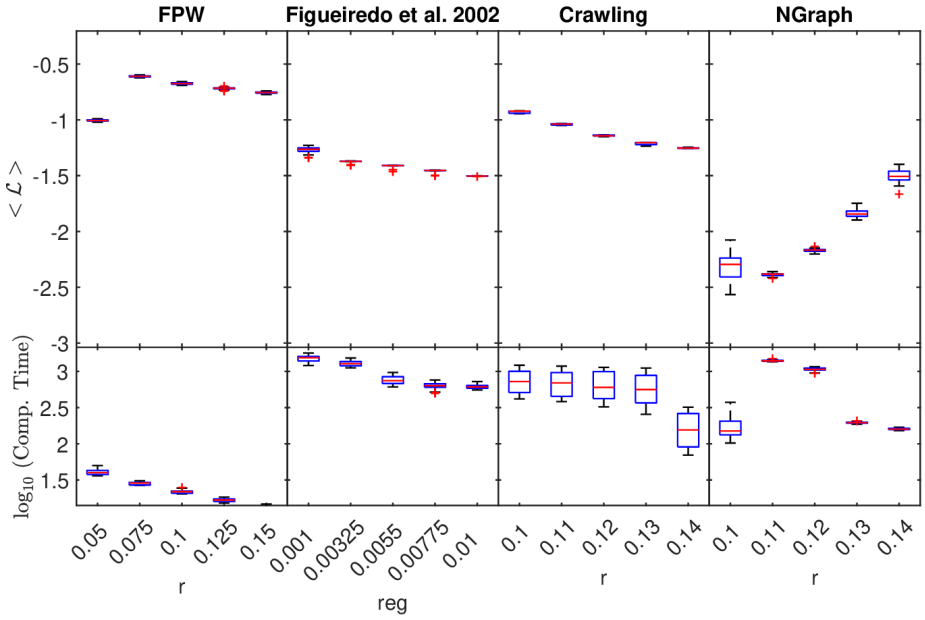
Figure 5.8: Top row shows the results obtained with SMCE on the data set described in Elhamifar et al., 2011, while bottom row shows the results obtained with Multi-Manifold Crawling. From left to right: two obtained clusters, embedding (bottom: graph) of the first cluster and embedding (bottom: graph) of the second cluster.

terlocked knots. The application of both methods on the presented data set proved equally satisfying for various parameter settings. However, when introducing a small amount on noise along the manifold, the performance of SMCE greatly deteriorates, both in terms of clustering accuracy and quality of the embedding (figures 5.8a, top row). The results do not improve when the data set is diffused and filtered beforehand, for any parameter setting. In contrast, Manifold crawling recovers the underlying ground truth by building two 1-dimensional, closed graphs and clustering the noisy data set accordingly (figure 5.8b).

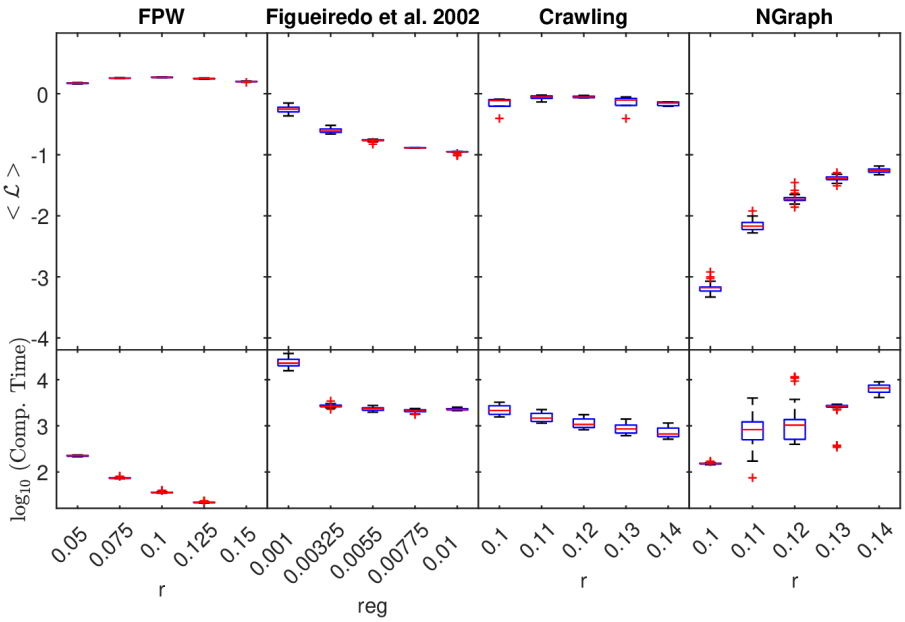
5.4.2 Probabilistic Modelling

The last subsection evaluated the construction of embedded latent space of the AGTM model as a manifold detection and approximation method, specifically designed to handle general low-dimensional manifolds embedded in finite dimensional Euclidean spaces. In this section AGTM (section 4.1) is treated as a generative probabilistic model of data aligned along such (noisy) manifolds. As explained above, the model evaluation is split into these two distinct

strands as competitors naturally spanning general manifold learning and probabilistic modelling are still missing. AGTM (initialized with Manifold crawling or the efficient ϵ-graph



(a)



(b)

Figure 5.9: Subfigures 5.9a and 5.9b present the average log-likelihood per point of the probabilistic models generated by the four methodologies with different parameter settings (top row) and the computational time required to train the models (bottom row) for data sets $\mathcal{Q}^2$ and $\mathcal{Q}^1$ respectively.

alternative) is compared with two probabilistic modelling techniques, namely *Fast Parzen Window* (FPW, X. Wang et al., 2009) and another unsupervised algorithm for learning *finite mixture models from multivariate data* (Figueiredo et al., 2002).

For both data sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$ a 5-fold cross-validation scheme is applied. Each method (except initialization via Multi-manifold crawling) is applied thirty times on training folds for a given parameter setting and the average log-likelihood per point of the generated model is evaluated on the corresponding hold-out fold. The choice of log-likelihood as a comparative measure for different probabilistic modelling techniques is motivated by the fact that it can be considered as an upper limit for the cross-entropy $H(P, Q)$, when the comparison is performed with respect to the common underlying, unknown distribution. Consider data set $\{x_i\}_{i=1}^N \sim P(x)$ as a i.i.d. sample of the underlying distribution $P(x)$. Suppose the final model obtained via any of the proposed methodologies is $Q(x)$, then the cross-entropy between $Q(x)$ and $P(x)$ is the expectation of the relative entropy with respect to distribution $P(x)$ is:

$$H(P, Q) = -\mathrm{E}_{P(x)}\left[\log\left(Q(x)\right)\right] = -\int_{-\infty}^{+\infty} P(x)\log\left(Q(x)\right) \simeq -\frac{1}{N}\sum_{i=1}^{N}\log\left(Q(x_i)\right). \qquad (5.8)$$

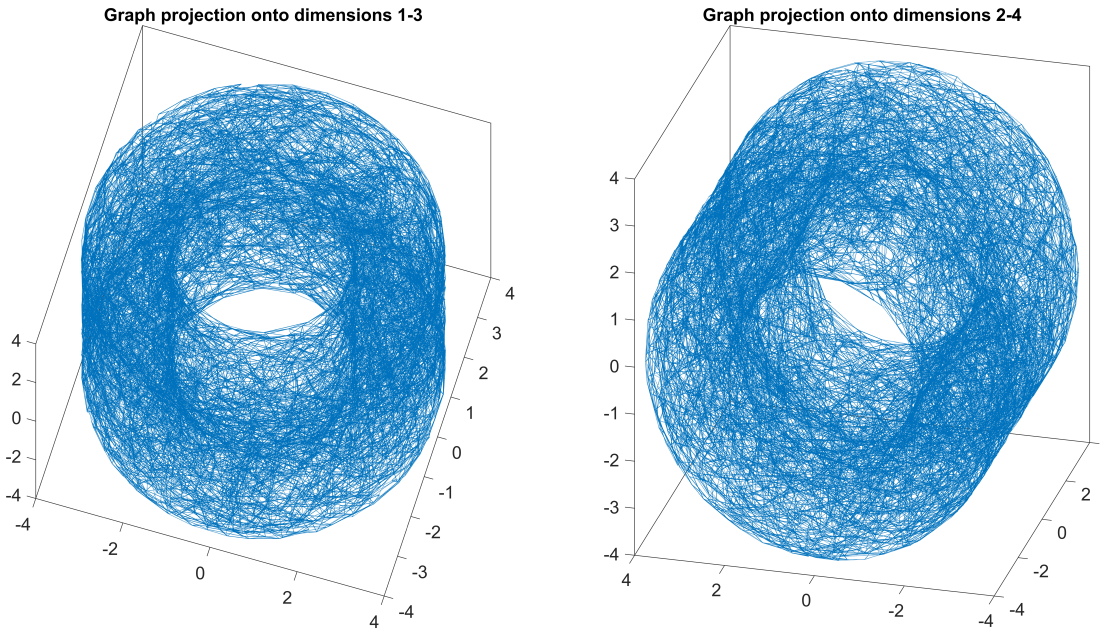
In the case of sampled distributions, equation (5.8) can be approximated by the sum over the sample and further manipulated, yielding the last term. This term is non other than the expected negative log-likelihood of distribution $Q(x)$ for the sample generated from $P(x)$: the empirical average negative log-likelihood. Since, in every experiment performed in this work, methodologies are compared on individual data sets and the underlying distribution for each specific data set is common to all methodologies, the empirical average negative log-likelihood is an approximation to the cross entropy and as such a quantitative measure for comparison.

The Multi-manifold crawling initialization is repeated five times on each training

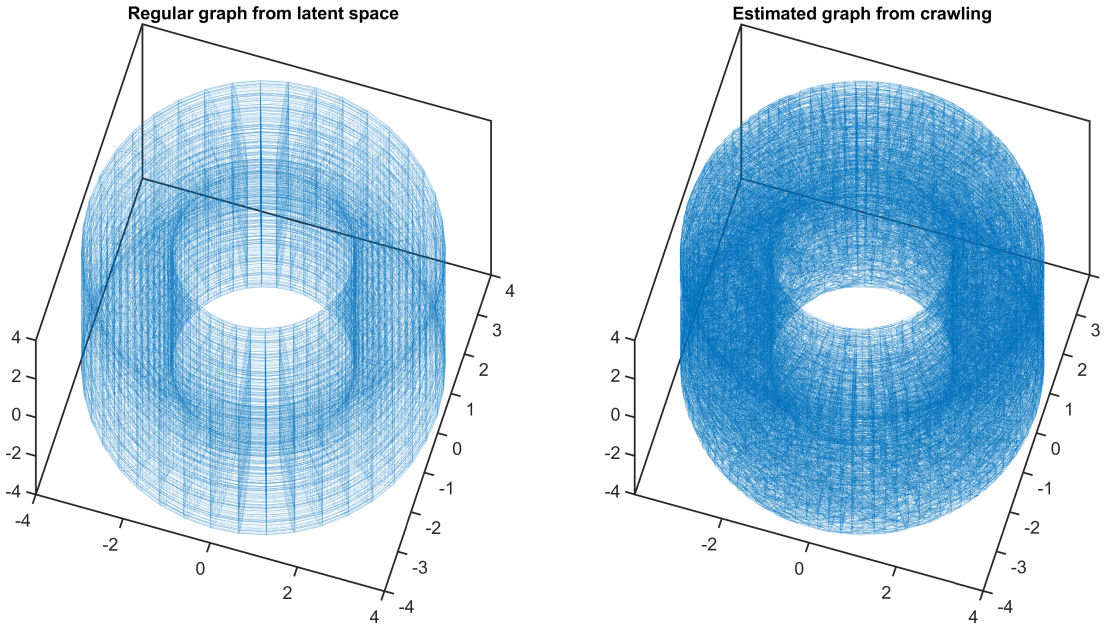
set, each run with a different β parameter. The results for the two data sets, multiple manifolds of different dimensionalities ($\mathcal{Q}^1$) and three interlocked toroids ($\mathcal{Q}^2$), are presented in figures 5.9b and 5.9a, respectively. Top row in each plot shows the average log-likelihood for different parameters as a boxplot over the initialization repetitions. Bottom row shows the computational time required for training of each model. The performance of FPW is generally higher and faster than all other methods, although this is a natural consequence of it being an efficient unconstrained probabilistic modelling algorithm. The manifold structure was often lost in the set of mixture components. However, for both data sets, AGTM initialized with Multi-Manifold crawling is comparable to FPW and outperforms both its alternative initialization and the algorithm proposed in Figueiredo et al., 2002 in terms of average log-likelihood. It is also worth noting that the computational time of AGTM is comparable to, if not lower than, the finite mixture method of Figueiredo et al., 2002. In conclusion, AGTM provides faithful probabilistic model of the data, while providing a clear interpretation of the low-dimensional general manifold structures along which the data is organized.

5.5 Application to higher dimensional manifold

An additional experiment on a 3−dimensional torus embedded in four dimensions (manifold $\mathcal{M}_8$ as sampled by data set $\mathcal{D}^8$) is here proposed. The torus is parameterized in the angle space $(\theta, \phi, \psi) \in [0, 2\pi] \times [0, 2\pi] \times [0, 2\pi]$ by equation 3.9. The 3−torus is covered by sampling the angle parameters with a large number of uniformly distributed values. the same procedure for producing a sample from the noisy manifold as used in the previous synthetic data experiments is then applied. After pre-processing the noisy data set by SAF and filtering, the estimation of dimensionality index on every remaining point is performed, finding that most of the structure is retained as made of 3-dimensional points. The Crawling algorithm



(a)



(b)

Figure 5.10: Results of crawling on sampled 3−torus (subfig. 5.10a) and comparison between ground truth graph (subfig. 5.10b, left) and crawled regularly sampled 3−torus (subfig. 5.10b, right).

is then applied, constructing an abstract-embedded pair of graphs. The embedded latent space is shown in figure 5.10a, projected onto two different triplets of dimensions. The global structure is clearly recovered through crawling. However, given that the abstract latent space was induced from a diffused sample of points not regularly covering the $3-$torus, the internal structure of the manifold is not easily identifiable. To demonstrate more clearly the ability of the crawling algorithm to recover the underlying 3-dimensional manifold structure, the crawling is operated operate on a data set created as a regular grid of points lying on the 3-torus. This way, a direct comparison between the ground-truth embedded latent space graph (obtained by connecting neighbouring grid points in the 3-torus topology, shown in a sub-sampled version in figure 5.10b, left panel) and the one recovered by crawling (figure 5.10b, right panel) is achievable. The two embedded graphs (shown here in the projection corresponding to that of figure 5.10a, left panel) clearly sample the same manifold structure. As shown in this example, the methodology could in principle be applied to higher dimensional manifolds embedded in even higher dimensional space. This is also true if the ambient dimension increases, provided the manifold is well sampled and local neighbourhoods are populated enough to perform PCA. However, the computation time of the crawling algorithm is bound to increase, since at each expansion phase, $2j$ new children node estimates are created per parent node.

Part II

Applications

Chapter Six

Spatial distribution along manifolds

6.1 Experiments on a jellyfish galaxy

Thew proposed methodology is demonstrated on the analysis of formation of a curious astronomical object, "jellyfish galaxy", through a detailed astrophysical simulation of its formation. The term "jellyfish galaxy" refers to an observed galaxy showing signs of gas stripping (Poggianti et al., 2017), whose signatures are a dense "head" of mainly gas and stars and an elongated gaseous, star-forming tail. These galaxies are usually observed when falling into large clusters of other galaxies, where the hot ionized gas filling the cluster is able to strip away "tentacles" of relatively cold gas from the galactic body.

The technique described above can be a valuable tool to investigate the behaviour of physical quantities in the head and in the gaseous tail. This study is focused on identifying star formation regions in the head and in the tail of the galaxy for two main reasons:

i) presence of new stars born in the head is an important indicator of how much the galactic gas is affected by the stripping pressure;

ii) since stars are created from dense and cold regions of gas, the presence (or lack thereof)

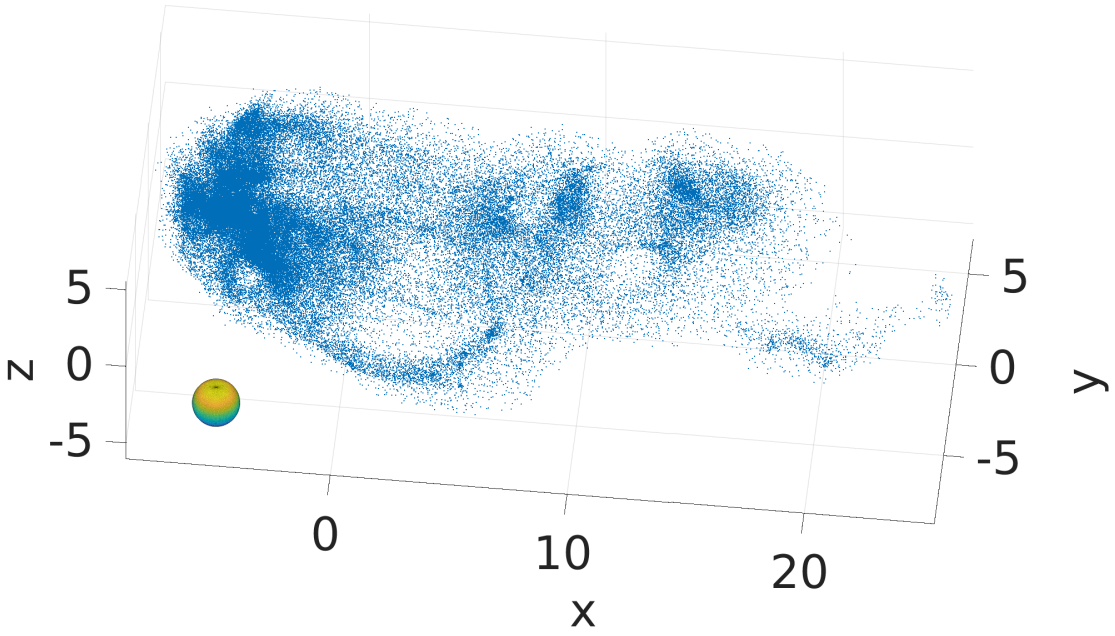


Figure 6.1: Gas particles of the simulated dwarf galaxy falling into the halo of the Fornax Galaxy Cluster.

of stars formed in the gaseous tail carries information about how much the galaxy gas is mixed with the hot gas in the surrounding environment.

To quantify the star formation the intensity of the [C II] emission line is measured. Recent observations with the Herschel Space Observatory showed a tight correlation between the intensity of [C II] and other well known tracers of Star Formation Rate (SFR, De Looze et al., 2011; Herrera-Camus et al., 2015). The study can provide useful insights on the formation scenario of galaxies infalling in a cluster (Ebeling et al., 2014). As an example, dwarf galaxy NGC 1427A in the Fornax cluster provides an interesting case of still unclear formation scenario and a generally accepted common interpretation is still lacking (Mora et al., 2015; Lee-Waddell et al., 2017; Mastropietro et al., 2021).

Starting from an existing suite of simulations of dwarf galaxies evolving in a Fornax-like cluster environment (Mastropietro et al., 2021), a single simulated snapshot is chosen representing an irregular, gas rich galaxy, exposing an elongated star forming gaseous tail

during intense ram pressure stripping. The simulation is performed using a modified version of the mixed N-body/Smoothed Particle Hydrodynamics (SPH) code GADGET-2 Volker Springel, 2005; De Rijcke et al., 2013.

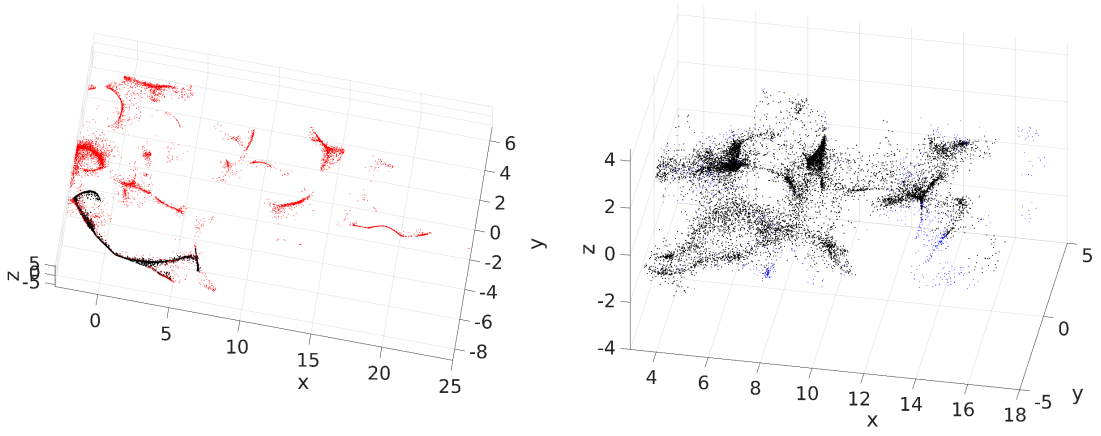


Figure 6.2: 1-D (left) and 2-D (right) distributions of the diffused particles in the tail of the jellyfish structure. Highlighted in black are the points belonging to two distinct 1-D and 2-D structures.

In order to simulate the evolution of the galaxy with high resolution and to reduce the computational cost of simulating the whole galaxy cluster, the *moving box* technique (Nichols et al., 2015b) is used. In this formulation, the simulation follows closely the evolution of the galaxy by immersing a cubic volume ("box") enveloping the galaxy, in the gravitational potential of the galaxy cluster. The gravitational potential is modelled via a spherically symmetric, static, Navarro-Frenk-White profile (NFW, Navarro et al., 1996). The box position and orientation is updated at every time-step so that its $x-$axis is always tangential to the orbit of the galaxy at every moment. Also at each time-step, new gas particles are injected into the simulation box from its open "front" face and existing ones removed when close to the opposite side of the box[1]. This injection-ejection technique is used to simulate the motion of the galaxy through the cluster's gas and it is crucial in studying processes such as gas stripping. The density and temperature of the injected particles change

[1]the particles "falling out of the box"

according to the position of the box within the cluster, following the radial profiles described in Paolillo et al., 2002 and assuming hydrostatic equilibrium of the gas.
This setup allows the study of the evolution of the galaxy in high detail, while being consistent with the characteristics of the ambient space (physical properties of the galaxy cluster) at a comparatively lower computational cost.

The computation model in SPH simulations is based on a particle formulation of hydrodynamics where each particle samples physical properties (mass, temperature, density etc.) of a volume of radius r_N - radius of the sphere containing N neighbouring particles (smoothing length). A continuous distribution of the physical variables over the full domain is then obtained by spatial smoothing with Gaussian kernels centered on each particle (Gingold et al., 1977). Associated with each gas particle (representing the corresponding volume) are values of physical quantities such as density, temperature, pressure etc.

An estimate of [C II] emission in this kind of simulations is obtained by using evolved quantities of the gas (metallicity, density and temperature) as inputs of chemical evolution models of the radiating gas, taking into account its ionization equilibrium and ion level occupation model (Maio et al., 2007; De Rijcke et al., 2013). This study, focuses on the state of the galaxy after 2.5 Gyrs of evolution since its injection in the galaxy cluster. The data set consists of 135530 gas particles (observations), each having 11 observed variables[2] (dimensions): r_N (smoothing length); $x-, y-, z-$ coordinates, representing the position in the moving box; v_x, v_y, v_z, the velocities along the three axis of the box; m, ρ, T and [C II]: the mass, density, temperature and [C II] emission for the volume of gas sampled by the particle.
Since the particles, by construction of the SPH methodology, fill the whole volume of the moving box, a physically motivated filtering is at first performed by removing all particles

[2]The number of observed variables is larger than the one presented here. The selected ones were deemed useful for the presented analysis.

whose density is $\rho \leq 10^{-3}$ atom/cm^3. This is a lower limit in density for the particles to be able to emit in the radio band, and so to be observed as part of the galaxy.
After this initial filtering, the remaining data set consists of 83772 particles. Their distribution over the box's volume is shown in fig. 6.1.
Several low dimensional structures are clearly visible, for example the long 1-dimensional manifold departing from the head and elongating along the x-axis of the simulation box. Visually inspecting the data set, radius $r = 1$ kpc is chosen as the characteristic scale parameter for the manifolds (shown as the small sphere in fig. 6.1). The chosen radius is in agreement with the spatial resolution of recent observations for galaxies as distant as NGC 1427A. This parameter is fixed for preprocessing through diffusion and filtering (section 5.1), local dimensionality estimation (section 5.2) and manifold crawling (sections 4.2 and 5.3). The other parameters are set as in the synthetic data experiments of section 5.3: $\epsilon = 1$, $\eta = 0.75$ and $\beta = 0.4$.

As mentioned earlier, the main body of the data set can be visually divided into head and tail parts. However, from the topological standpoint there is no justification for clear segregation into 1-D manifolds in the tail and 2-D manifolds in the head. Dimensionality index estimation clearly identified 1-D structures in the tail, such as the elongated stream of particles starting from the head. However, points in the head were also predominantly identified as[3] 1-D, due to its complicated, intertwined filamentary structure. On the other hand, distribution of 2-D points was more localized in the main body of the tail.

6.1.1 A Multi Manifold analysis of a dwarf jellyfish galaxy

The performance of AGTM is evaluated with both its initializations and compared with the probabilistic modelling methods introduced above using 5-fold cross-validation scheme

[3]Loose terminology referring to points with dimensionality index j as j-D points.

on the 1− and 2−dimensional gas particles. The average out-of-sample log-likelihood per point and computational times are reported in figure 6.3 for all methods and their parameter settings. From the probabilistic modelling standpoint, all methods perform similarly, with FPW providing for the best score. Both initializations of AGTM are comparable to the results obtained by the method described in Figueiredo et al., 2002, with lower computational effort. Initialization by crawling also shows a high stability of the results when compared to both its alternative initialization and the competing methodology. The $\epsilon-$graph initialization, despite its larger sensitivity to the randomness of the sampling process, reaches higher values of the average log-likelihood than the initialization obtained via crawling. This behaviour is explained by the tendency of the former to break manifolds into sub-structures. Having more structures to cover the same spatial region, leads to a better performance of the probabilistic modelling algorithm. However, the information about low-dimensional structures retained by the sub-structures, gradually deteriorates as this phenomenon occurs more frequently. The limiting case of this information loss is represented by FPW, where the distribution of the data set is globally captured with a high efficiency, but no additional information on the low-dimensional local structures can be inferred. Having obtained the multi-manifold probabilistic profile of the gas particles in the tail of the jellyfish galaxy, it is possible to perform various kinds of detailed analysis of how physical properties vary along the manifolds.

The following discussion focuses on the curvature (figure 6.4), as defined in section 4.3, and the star formation potential by analysing the behaviour of emission line [C II] over the 1-D and 2-D structures in the jellyfish tail shown as black dots in fig. 6.2 (left) and (right), respectively - the gaseous stream of gas particles departing from the head and reaching half way through the tail and the predominant 2-D structure in the tail. Figure 6.4, left panel shows two regions of high curvature, for the 2-dimensional manifold presented in figure 6.2, right panel. The far right region of intense curvature, is located towards the end of the tail,

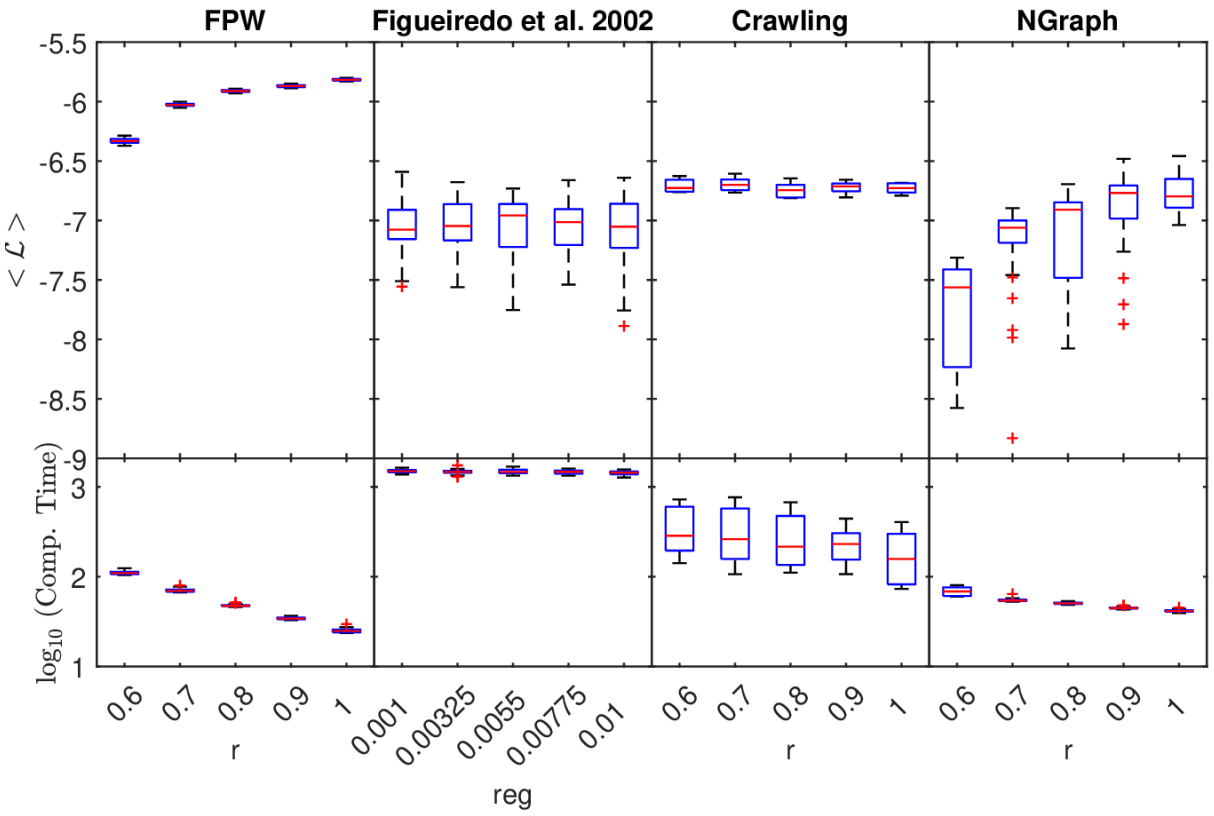


Figure 6.3: Average log-likelihood (top row) and log computational time of the probabilistic modelling techniques described in section 5.4.2 applied to the jellyfish data set.

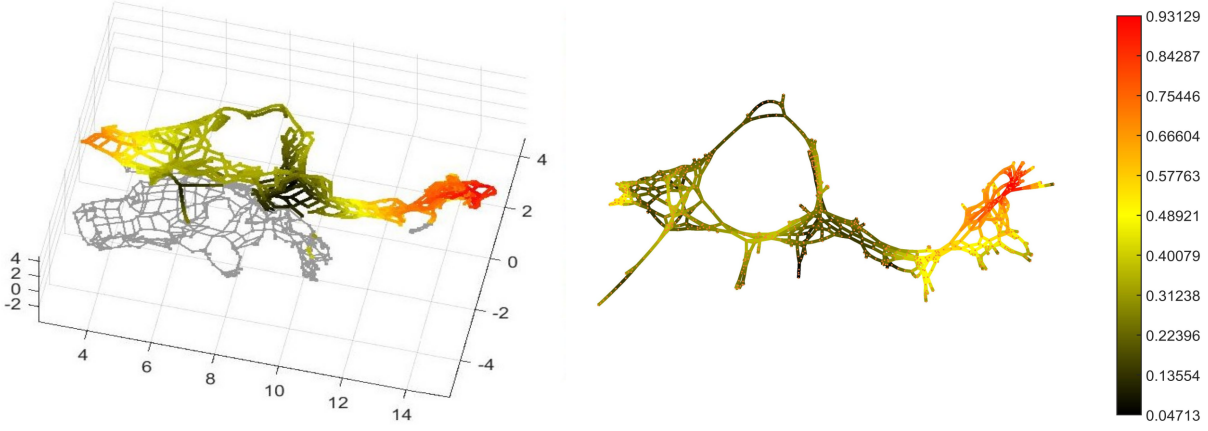


Figure 6.4: Embedded graph (left panel) and planar representation (right panel) of a 2-dimensional manifold extracted from the jellyfish data set, and its on-edge "geometric" curvature distribution.

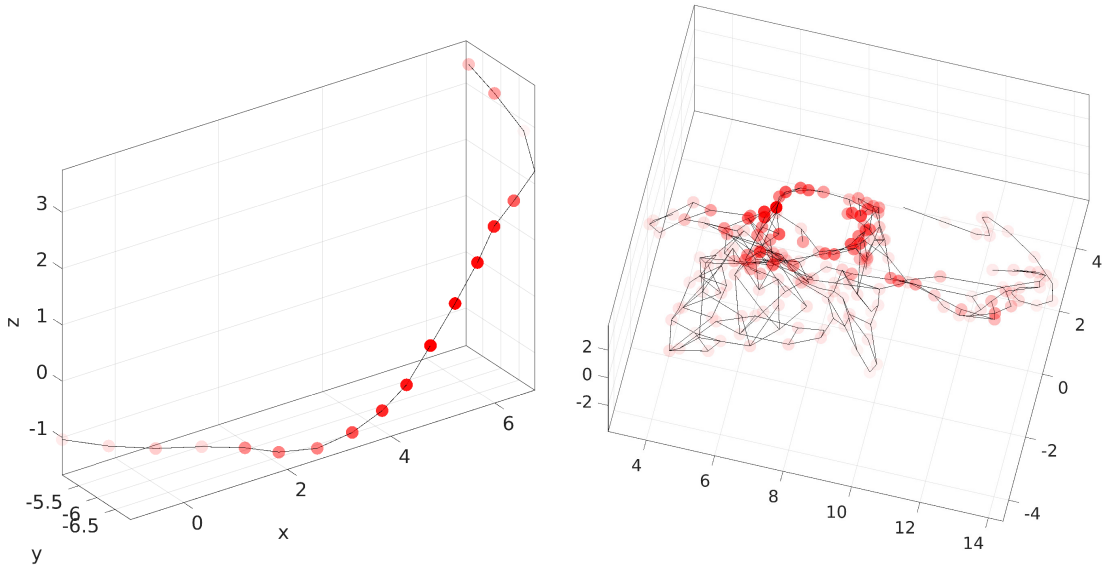
where the gas motion is more chaotic due to the motion of the galaxy through the halo of the galaxy cluster. However, the spherical region on the left side of the manifold, presenting a coherent curvature throughout its elongation (top circle of figure 6.4, left panel) suggests, as a possible cause of formation, an isotropic expansion, typical of Supernova remnants.

Taking advantage of the probabilistic nature of the AGTM model, figure 6.5a shows the embedded vertices $\overline{\mathbf{v}}_j$ of the graph $\overline{\mathcal{G}}$ for the stream model, with intensity modulated by the weighted mean of [C II] values $\mathcal{I}_i^{[\text{C II}]}$ of particles $\mathbf{t}_i$ in the manifold, where the weights are the posterior probabilities of the node v_j, given particles $\mathbf{t}_i$:

$$\overline{\mathcal{I}}_j^{[\text{C II}]} = \frac{\sum_{i=1}^{N_{\mathcal{M}}} p(v_j|\mathbf{t}_i, \zeta_j\Sigma_j, \mathbf{W}_j)\ \mathcal{I}_i^{[\text{C II}]}}{\sum_{k=1}^{N_{\mathcal{M}}} p(v_j|\mathbf{t}_k, \zeta_j\Sigma_j, \mathbf{W}_j)} \tag{6.1}$$

Analogous figure for the 2-D structure is presented in fig. 6.5b.

In the 1-D case, the manifold is located at the outskirts of the jellyfish (fig. 6.1, left panel), meaning that it is more exposed during the evolution to the surrounding gas of the galaxy cluster. This implies that the manifold is subject to a higher ram pressure than the tail, leading to a higher density and lower temperature of the gas - necessary conditions for the formation of new stars. These conditions are reflected in an increase of the [C II] emission line over the middle section of the manifold, $3 < x < 6$, thus informing of an enhanced Star Formation Rate, compared to the rest of the manifold. The 2-D structure shows an overall constant [C II] intensity whereas the region at $5 < x < 9$ presents sharply higher values. The shape of this region is particularly interesting. It is, in fact, a hole with an almost spherical section. This structure detected with AGTM and Manifold Crawling, with potent [C II] emission at its boundary, is the remnant of a supernova explosion. This process is modeled in the simulation via an injection of 10^{51} erg of energy for a short amount of time and a transfer of metallic elements (in the case of the analyzed simulation, the model tracks iron and magnesium, De Rijcke et al., 2013) to the neighbouring gas particles. The



(a) (b)

Figure 6.5: 1D (6.5a) and 2D (6.5b) manifolds extracted from the data set in fig. 6.1. The graphs' nodes are colored based on the value of [C II] of the surrounding particles lying not further than 1 kpc from the local tangent space of the manifold, weighted by the nodes' responsibilities.

metal-enriched gas particles (like in the surrounding of a supernova explosion) are then able to cool down more efficiently and show strong [C II] emission line.

Our methodology provides a strong tool for extracting such an information from the morphology of gas particles and can be used to effectively calibrate feedback models[4] in simulations.

Such a detailed analysis of low dimensional structures (remnants of galaxy interactions) is not currently possible with tools routinely used to calibrate and analyse astrophysical simulations of galaxy evolution. As mentioned in section 5.3, the technique presented in this paper can be used as a semi-automatic exploratory tool by the domain experts, where the focus and characteristic scale of the structures to be mined can be varied continuously with analysis of their physical properties of interest (after necessary computations) performed and

[4]A feedback mechanisms, is any process that allows to exchange energy, matter and/or momentum among galaxy components.

studied on the fly.

Chapter Seven

Moving through manifolds

T following sections are focused on one-dimensional manifolds. This is due to the particular shape of Jellyfish galaxies and cosmic web structure, where information lies mainly onto streams. In order to maintain coherence throughout this work, the following methodologies are described referring to the pair of abstract and embedding graphs described in Part I. However, since the low-dimensional representation of a one-dimensional manifold is the unit interval, it is always possible to revert back to the original formulation of GTM. Nonetheless, an estimation of dimensionality index per point, filtering of noise and initialization of the mapping function parameters is necessary in order to obtain an accurate model of the embedded streams. Since the proposed methodology is a generalization of GTM to abstract manifolds, this process is already accounted for by design. All the results pertaining to the studied cases, plus additional analysis of the extracted manifolds can be found at https://github.com/MarcoCanducci/general-noisy-multi-manifold-learning.

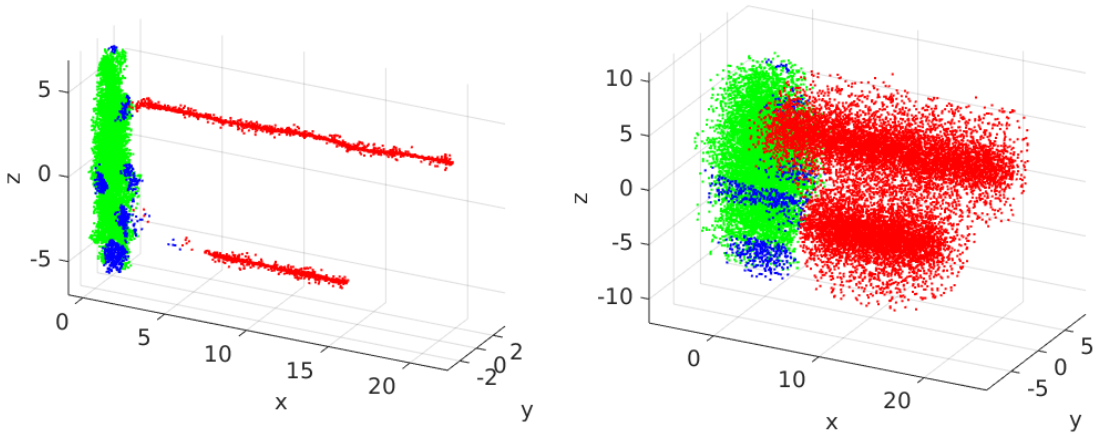


Figure 7.1: Partition of data sets $\tilde{\mathcal{Q}}^3$ (left) and $\mathcal{Q}^3$ (right) into their corresponding low dimensional subsets: $\tilde{\mathcal{Q}}^3_1$ ($\mathcal{Q}^3_1$, red), $\tilde{\mathcal{Q}}^3_2$ ($\mathcal{Q}^3_2$, blue), $\tilde{\mathcal{Q}}^3_3$ ($\mathcal{Q}^3_3$, green).

7.1 Bi-dimensional profiles

Take into consideration data set $\mathcal{Q}^3$ described in sec.3.2. After application of filtering, dimensionality index and manifold crawling the latent and embedded graphs representing manifolds $\mathcal{M}_{12}$ and $\mathcal{M}_{13}$ are recovered as outlined in part I of this work. The results of filtering and dimensionality index are presented for both the noisy and diffused data set in figure 7.1, left and right panel respectively, with the same color-coding used throughout the previous chapters. AGTM is then applied on the individual manifolds so that each one is represented as a Gaussian mixture with manifold aligned noise, whose updated centers $\{\overline{\mathbf{v}}_\ell; \ell = 1, \ldots, L^k\}$ are constrained to lie on a one-dimensional latent abstract graph $\mathcal{G}^k = \left(\mathcal{V}^k, \mathcal{E}^k\right)$. Taking full advantage of the probabilistic nature of AGTM the behaviour of simulated properties are recovered along the manifold's elongation and its thickness within the simulated volume. The centers obtained at the end of optimization have most likely shifted along the manifold, due to local variations of point density within the manifold itself. In order to take this shift into consideration the tangent bundle to manifold $\mathcal{M}_k$ is updated by computing the deriva-

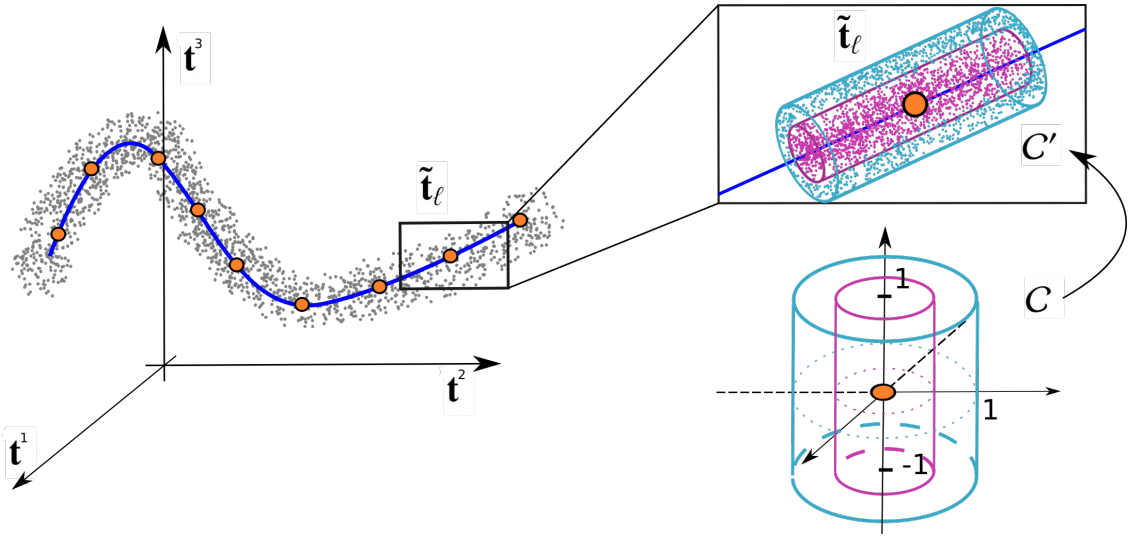


Figure 7.2: Sketch depicting the formation of the uniformly sampled concentric cylindrical volumes, aligned with the local tangent space of manifold $\mathcal{M}_k$ on the AGTM center $\overline{\mathbf{v}}_\ell$.

tive of the (trained) mapping function w.r.t. latent center v_ℓ: $\hat{\boldsymbol{\xi}}_\ell = \theta(v_\ell) = \dot{\omega}(v_\ell)$, where $\theta(v_\ell)$ is derived in equation (4.20). Doing this for every center $v_\ell \in \mathcal{V}^k$ the tangent bundle $TM^k = \{\hat{\boldsymbol{\xi}}_\ell;\ \ell = 1, \ldots, L^k - 1\}$ of manifold $\mathcal{M}_k$ is updated.

Consider now a point cloud $\mathcal{C}$ containing points as row vectors, uniformly sampling a cylindrical volume of radius 2, aligned along the $z-$axis and centered at the origin $\mathcal{O} = (0,\ 0,\ 0)$. In order to radially partition the point cloud $\mathcal{C}$ into concentric cylindrical shells (bottom right panel of figure 7.2, cyan and magenta cylinders), first, $\overline{\mathcal{C}}$ is created by projecting $\mathcal{C}$ onto the $x - y$ plane:

$$\overline{\mathcal{C}} = \mathcal{C} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{7.1}$$

For every point $\overline{\mathbf{p}} \in \overline{\mathcal{C}}$ its distance to the projected origin $\overline{d}(\overline{\mathbf{p}}, \overline{\mathcal{O}}) = ||\overline{\mathbf{p}} - [0, 0]||$ is computed. Points in $\overline{\mathcal{C}}$ are now grouped so that

$$\mathcal{I}_i = \left\{ j \mid d^r_{i-1} \leq d(\overline{\mathbf{p}}_j, \overline{\mathcal{O}}) < d^r_i \right\}, \quad \forall \overline{\mathbf{p}}_j \in \overline{\mathcal{C}}, \tag{7.2}$$

where $d^r = \{0, 0.2, 0.4, \ldots, 2\}$ is a discrete collection of radial distances and $i = 1, \ldots, N^r = |r|$. The sets $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_{N^r}$ contain all indices of points in $\overline{\mathcal{C}}$ (and thus in $\mathcal{C}$) belonging to specific cylindrical shells concentric w.r.t. the origin $\overline{\mathcal{O}}$ ($\mathcal{O}$) and the $z-$axis.

It is always possible to scale, translate and rotate the point cloud so that the cylindrical axis is oriented as vector $\hat{\boldsymbol{\xi}}_\ell$, the origin over-posed to center $\overline{\mathbf{v}}_\ell$ and the axis length equal to the distance between two adjacent centers $d_{\ell,\ \ell+1} = ||\overline{\mathbf{v}}_{\ell+1} - \overline{\mathbf{v}}_{\ell+1}||$ (as shown in figure 7.2, top, right panel).

Scaling: In matrix notation, the scaling operator is S

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d_{\ell,\ \ell+1} \end{pmatrix}$$

Rotation Tthe quaternion $\mathbf{q} = (q_1, q_2, q_3, q_4)$ Hamilton, 1866 is computed, where the first three components are given by $\hat{\mathbf{v}}_\ell \times (0,\ 0,\ 1)$ and the $4-$th component by $\hat{\mathbf{v}}_\ell \cdot (0,\ 0,\ 1)$. Its matrix representation is given by R:

$$R = \begin{pmatrix} 1 - 2q_2^2 - 2q_3^2 & 2q_1q_2 - 2q_3q_4 & 2q_1q_3 + 2q_2q_4 \\ 2q_1q_2 + 2q_3q_4 & 1 - 2q_1^2 - 2q_3^2 & 2q_2q_3 - 2q_1q_4 \\ 2q_1q_3 - 2q_2q_4 & 2q_2q_3 + 2q_1q_4 & 1 - 2q_1^2 - 2q_2^2 \end{pmatrix}$$

Shift The scaled and rotated point cloud can then be shifted so that its origin is on center $\overline{\mathbf{v}}_\ell$.

Any point $\mathbf{p} \in \mathcal{C}$ is then mapped to $\mathbf{p}' \in \mathcal{C}'$ under the combined operator as:

$$\mathbf{p}' = (\mathbf{p}\ S)\ R + \tilde{\mathbf{t}}_\ell \tag{7.3}$$

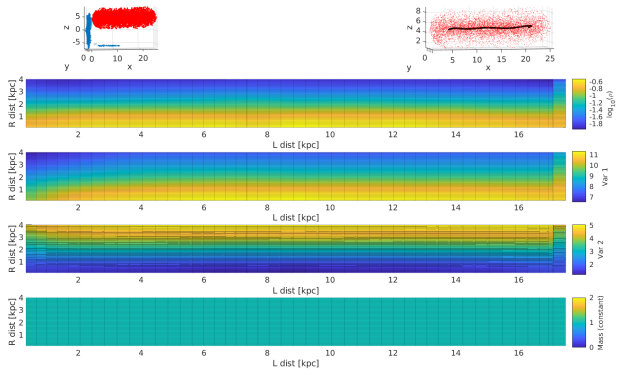
Having obtained a point cloud $\mathcal{C}'$ uniformly sampling a thick cylindrical volume with axis tangential to manifold $\mathcal{M}_k$ on point $\overline{\mathbf{v}}_\ell$, the weighted mean of any quantity contained in

the data set, over the volume sampled by $\mathcal{C}'$ can be computed. Note that the families of indices $\mathcal{I}_1, \ldots, \mathcal{I}_{N^r}$, when applied to $\mathcal{C}'$, contain all indices of points in cylindrical shells concentric w.r.t. point $\overline{\mathbf{v}}_\ell$ and vector $\hat{\boldsymbol{\xi}}_\ell$, radially partitioning the cylindrical volume sampled by $\mathcal{C}'$. For each linear segment of manifold $\mathcal{M}_k$, parameterized by its corresponding AGTM, a uniformly sampled cylindrical volume aligned along its corresponding local tangent space is then obtained. Computing $\langle V_m(\mathbf{p}')\rangle$ for every $\mathbf{p}' \in \mathcal{C}'$ the mean value of V_m over the concentric rings defined by the index families $\mathcal{I}_1, \ldots, \mathcal{I}_{N^r}$ can now be evaluated as:

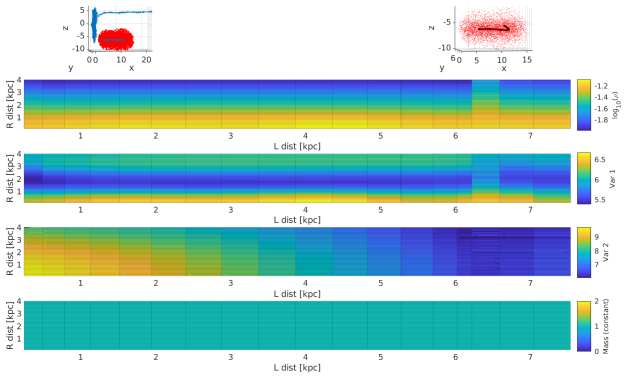
$$\overline{V_{m,i}} = \langle V_m(d^r_{i-1}, d^r_i)\rangle = \frac{\sum_{j\in\mathcal{I}_i}\langle V_m(p'_j)\rangle}{|\mathcal{I}_i|}, \tag{7.4}$$

obtaining the mean of V_m over the cylindrical shell between (d^r_{i-1}, d^r_i) for every $i = 1, \ldots, N^r$.

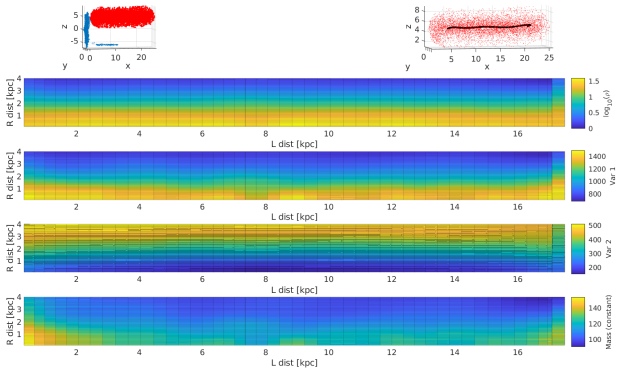
The whole process is iterated for every center of AGTM, obtaining for each linear segment, the distribution of V_m in concentric cylindrical shells centered on the current center (figure 7.2, left panel). The plots shown in figures 7.3a and 7.3b are obtained by considering both longitudinal (defined recursively by the centers of AGTM) and radial (obtained by the linear operator defined in eq. 7.3 on point cloud $\mathcal{C}$) profiles. In both panels, the vertical axis of each plot contains the radius of the cylindrical shells d^r and the horizontal axis the approximated geodesic distance (computed by summation of the lengths of the individual linear segments) from the head of the manifold. From these plots it can be verified that the behaviour of quantities V_1 (second panel from top) and V_2 (third panel from top) are in agreement with how they were designed when constructing the data set (sec. 3.2). The decreasing profile for quantity V_1 and increasing for V_2 along manifold $\mathcal{M}_{13}$ is detected, as well as the sinusoidal behaviour of quantity V_1 along manifold $\mathcal{M}_{14}$'s thickness and V_2's decreasing profile along its longitudinal elongation. The bottom plot in each panel presents the "mass" distribution over the radial and longitudinal dimensions of the manifolds. As expected, the mass is constant throughout the sampled volumes and it is everywhere $\overline{m}_i = 1$.



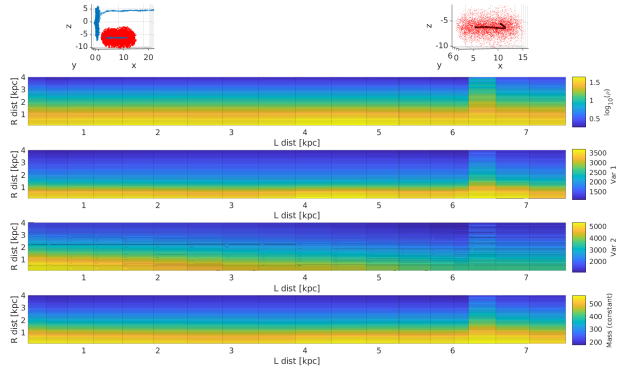
(a) Manifold $\mathcal{M}_{13}$, normalized weighted mean.



(b) Manifold $\mathcal{M}_{14}$, normalized weighted mean.



(c) Manifold $\mathcal{M}_{13}$, weighted mean, no normalization.



(d) Manifold $\mathcal{M}_{14}$, weighted mean, no normalization.

Figure 7.3: Bi-dimensional profiles for manifolds $\mathcal{M}_{13}$ (left column) and $\mathcal{M}_{14}$ (right column) with (top row) and without (bottom row) the normalization term in equation (6.1), for variables $\overline{\rho}_i, \overline{V}_{1,i}, \overline{V}_{2,i}$, and $\overline{m}_i$ (left and right panel respectively).

The uniformity of a constant scalar is guaranteed by the normalization term in equation 3.3. Not to including this term, a constant scalar would still be showing variations over the bi-dimensional plot, due to the non-uniform distribution of particles in the volume. An example over data set $\mathcal{Q}^3$ is given in the bottom panel of figures 7.3c and 7.3d. As clearly visible in this controlled experiment, the true behaviour of the mass parameter is distorted when the normalization term is removed from the computation of the weighted mean. The mass at low radial distance from the manifolds' cores is higher, tracing the particle's distribution along each manifold: by design, $\mathcal{M}_{13}$ and $\mathcal{M}_{14}$ are denser along their core while increasingly sparse along the radial direction. This behaviour is reflected on the estimation of the weighted means of variables V_1 and V_2.

7.1.1 Application to Jellyfish data set

The methodology proposed in Part I of this work is now applied to the case of the simulated Jellyfish galaxy, described in section 3.1.1. The same setup as in section 6.1 is followed, disregarding Dark Matter and Stellar particles and only considering the distribution of Gas particles and their properties: Density (ρ), Neutral Fraction, Metallicity ($[Fe/H]$) and Temperature (T). The term Neutral Fraction, refers to the proportion between the mass of neutral (non-ionized) gas and the total mass of the gas. The study of Neutral gas distribution over the simulated galaxy is essential in the identification of possible Star Formation loci, since generally stars are more likely to form in regions where gas is colder, denser and more neutral.

Applying manifold crawling and AGTM to the Jellyfish data set results in multiple manifolds being detected and modelled. In particular, 15 streams of different elongations are recovered. In the following, only the most elongated manifolds recovered through the methodology are discussed, since they are more likely to be clearly visible in astronomical observations. Figure 7.4, shows four different manifolds recovered via the proposed methodology, residing heterogeneously throughout the simulated Jellyfish galaxy. While the elongation of individual manifolds vary, the inspected properties behave similarly for all detected structures. In particular, Neutral fraction (second bi-plot from the top), along with Density (top bi-plot), is generally higher in the inner regions of the manifolds, across roughly their whole elongation. At the same time Temperature (bottom bi-plot, in logarithmic scale) is consistently lower in the same regions. A similar behaviour can not be found for the Metallicty (third bi-plot from the top). It can be argued the this quantity is generally higher in the inner regions as well, however the non-uniformity of its distribution discourages an accurate inference on the causes of this anomaly. It is possible that, being the manifolds from different regions of the galaxy, their interaction with the gas of the galaxy cluster's halo in previous epochs heavily influenced their evolution. Nonetheless, since the gas' metallicity does not directly affect the

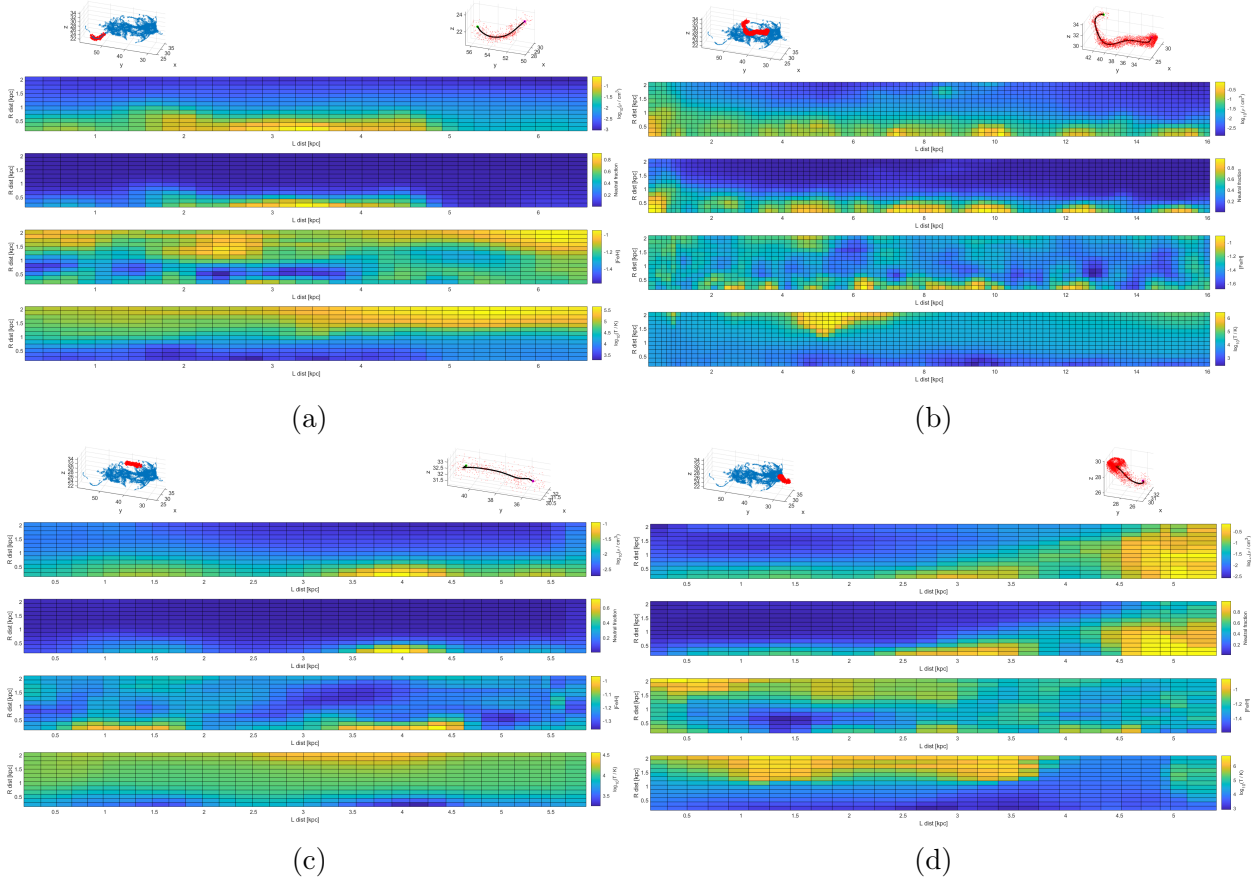


Figure 7.4: Bi-dimensional profiles of four elongated manifolds recovered via the methodology discussed in this work. The manifolds are located in different regions of the simulated Jellyfish galaxy, but show very similar behaviours w.r.t. the analyzed properties.

chances of its collapse,it is arguable that the streams in Jellyfish galaxies are effective loci of Star Formation. Furthermore, the core of these streams are more likely to contain newly born stars, and thus be observed via optical observatories.

7.2 Co-Moving orthonormal coordinate frames

For each manifold $\mathcal{M}_k$, a better discretization of its "spine" can be obtained by introducing more points in the latent space and propagating them in the ambient space through the mapping function $\mathbf{f}(x; \mathbf{W}) : \mathcal{V}^k \longrightarrow \overline{\mathcal{V}}^k$. This is achieved by regularly introducing N_ℓ new

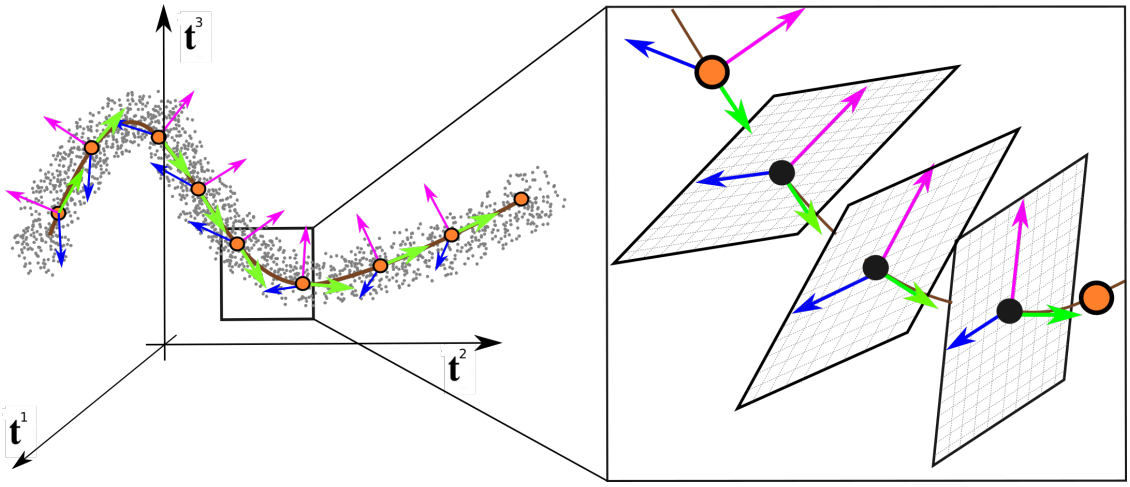


Figure 7.5: Sketch depicting the formation of co-moving orthonormal planes along a portion of manifold $\mathcal{M}_k$, enclosed within two centers of the corresponding AGTM.

nodes on the abstract latent graph within edge $e_{\ell,\ell+1}$ connecting original nodes v_ℓ and $v_{\ell+1}$. Recall that the weight of edge $e_{\ell,\ell+1}$ is $w(v_\ell, v_{\ell+1}) = 1$. The new nodes $x_\ell^1, \ldots, \mathbf{x}_\ell^{N_\ell}$ are introduced as a chain of connected nodes, where

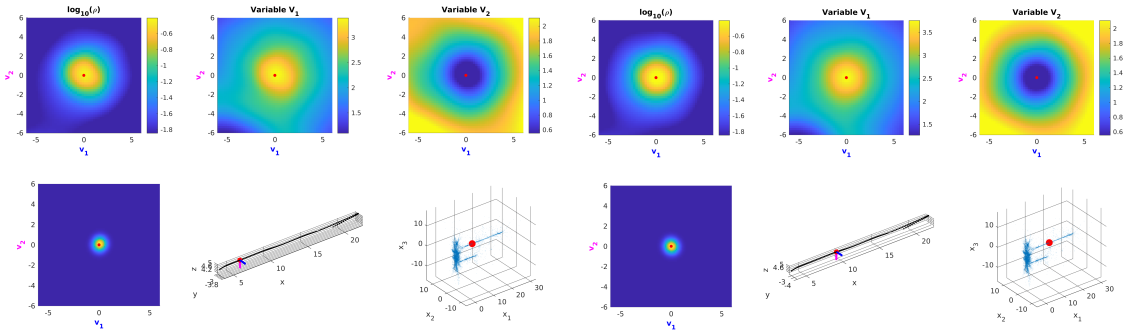
$$w(v_\ell, x_\ell^1) = w(x_\ell^1, x_\ell^2) = \cdots = w(x_\ell^{N_\ell}, v_{\ell+1}) = 1/(N_\ell + 1), \tag{7.5}$$

for any $i = 1, \ldots, N_\ell$. As in section 4.3, the "lifted line" $\omega(x_i) = \mathbf{f}(x_i; \mathbf{W})$ is defined.

Applying this procedure to points $v_\ell, \forall \ell = 1, \ldots, L^k - 1$, the up-sampled set of abstract nodes $\mathcal{V}_\uparrow^k = \left\{v_0, x_0^1, \ldots, x_0^{N_\ell}, v_1, x_1^1, \ldots, x_1^{N_\ell}, \ldots, v_{L^k-1}, x_{L^k-1}^1, \ldots, x_{L^k-1}^{N_\ell}, v_{L^k}\right\}$ is obtained, having size $\left|\mathcal{V}_\uparrow^k\right| = N^\ell \left(L^k - 1\right) + L^k$.

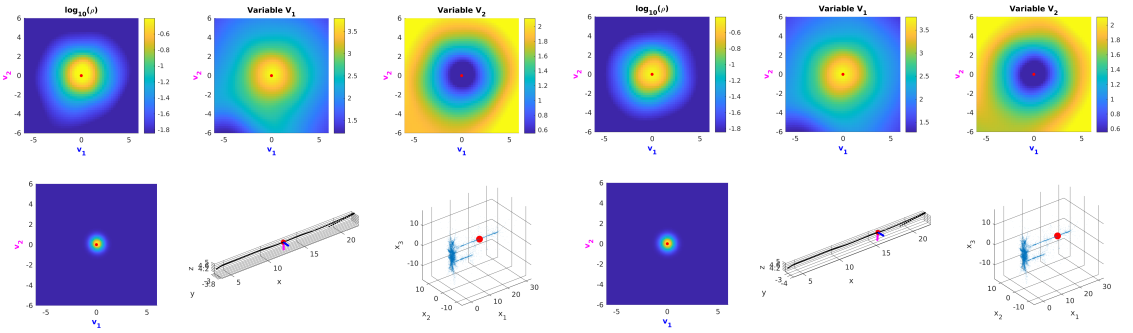
Propagation of latent abstract graph $\mathcal{G}_\uparrow^k = \left(\mathcal{V}_\uparrow^k, \mathcal{E}_\uparrow^k\right)$ into the ambient space through mapping function $\mathbf{f}(x_\ell; \mathbf{W})$ leads to the up-sampled embedded graph $\overline{\mathcal{G}_\uparrow^k} = \left(\overline{\mathcal{V}_\uparrow^k}, \overline{\mathcal{E}_\uparrow^k}\right)$, as depicted in figure 7.5, right panel, black dots. The tangent bundle TM^k is updated on every point in $\overline{\mathcal{V}_\uparrow^k}$, by computing the first-order derivative of the lifted line along edge $e_{\ell,\ell+1}, \forall \ell = 1, \ldots, L^k - 1$, as in section 7.1. In this section, slightly abusing mathematical notation, index ℓ is used for any latent (and corresponding embedded) point belonging to $\mathcal{V}_\ell^k$ and drop

subscript $\uparrow$, for readability purposes. For every $\hat{\boldsymbol{\xi}}_\ell \in TM^k$ a set of two vectors is recovered,



(a) Manifold $\mathcal{M}_{13}$, snapshot n.1 of associated movie clip.

(b) Manifold $\mathcal{M}_{13}$, snapshot n.2 of associated movie clip.



(c) Manifold $\mathcal{M}_{13}$, snapshot n.3 of associated movie clip.

(d) Manifold $\mathcal{M}_{13}$, snapshot n.3 of associated movie clip.

Figure 7.6: Selected snapshots of the video clip generated by considering co-moving coordinate frames on the elongation of manifold $\mathcal{M}_{13}$. The top row panels show the distribution of variables ρ, V_1 and V_2 over the coordinate frame $\mathcal{T}^\ell_\perp$ defined on $\overline{\mathbf{v}}_\ell$, shown in bottom central and right panels as a red sphere. Bottom left panel presents the Probability Density Function obtained through AGTM on the same plane.

$\mathbf{u}_1 = (\mathbf{u}_1^1, \mathbf{u}_1^2, \mathbf{u}_1^3)$ and $\mathbf{u}_2 = (\mathbf{u}_2^1, \mathbf{u}_2^2, \mathbf{u}_2^3)$, perpendicular to $\hat{\boldsymbol{\xi}}_\ell$, spanning the perpendicular plane $\mathcal{T}^\ell_\perp$ to manifold $\mathcal{M}_k$ on center $\overline{\mathbf{v}}_\ell$.

This is achieved by solving the system of linear equations given by:

$$\begin{cases} \mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1^1\mathbf{u}_2^1 + \mathbf{u}_1^2\mathbf{u}_2^2 + \mathbf{u}_1^3\mathbf{u}_2^3 = 0 \\ \mathbf{u}_1 \cdot \hat{\boldsymbol{\xi}}_\ell = \mathbf{u}_1^1\hat{\boldsymbol{\xi}}_\ell^1 + \mathbf{u}_1^2\hat{\boldsymbol{\xi}}_\ell^2 + \mathbf{u}_1^3\hat{\boldsymbol{\xi}}_\ell^3 = 0 \\ \mathbf{u}_2 \cdot \hat{\boldsymbol{\xi}}_\ell = \mathbf{u}_2^1\hat{\boldsymbol{\xi}}_\ell^1 + \mathbf{u}_2^2\hat{\boldsymbol{\xi}}_\ell^2 + \mathbf{u}_2^3\hat{\boldsymbol{\xi}}_\ell^3 = 0 \end{cases} \tag{7.6}$$

Being a degenerate system of linear equations there can be an infinite number of solutions giving infinite pairs of vectors spanning the perpendicular plane $\mathcal{T}_{\perp}^{\ell}$. In order to maintain consistency throughout the manifold's elongation, the solution is chosen to be:

$$\mathbf{u}_1 = (\hat{\boldsymbol{\xi}}_{\ell}^2,\ -\hat{\boldsymbol{\xi}}_{\ell}^1,\ 0), \tag{7.7}$$

$$\mathbf{u}_2 = (\hat{\boldsymbol{\xi}}_{\ell}^1\hat{\boldsymbol{\xi}}_{\ell}^3,\ \hat{\boldsymbol{\xi}}_{\ell}^2\hat{\boldsymbol{\xi}}_{\ell}^3,\ -[(\hat{\boldsymbol{\xi}}_{\ell}^1)^2 + (\hat{\boldsymbol{\xi}}_{\ell}^2)^2]), \tag{7.8}$$

so that $\mathcal{T}_{\perp}^{\ell} = \text{span}(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2,)$, where $\hat{\mathbf{u}}_1 = \mathbf{u}_1/||\mathbf{u}_1||$ and $\hat{\mathbf{u}}_2 = \mathbf{u}_2/||\mathbf{u}_2||$. Under this scheme, the two vectors form an orthonormal coordinate frame for the plane locally perpendicular to center $\overline{\mathbf{v}}_{\ell}$. The two vectors are shown in figure 7.5 as the magenta and blue arrows, changing direction slightly, between any pair of adjacent centers in $\overline{\mathcal{V}}^k$. The tangent bundle is here also shown (sampled on points in $\overline{\mathcal{V}}^k$) as a green arrow.

Taking advantage of the local coordinate frame given by unit vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ (black gridded rotated squares in figure 7.5), a regular $M \times M$ square grid of side a can now be imposed on plane $\mathcal{T}_{\perp}^{\ell}$. Define the set $\mathcal{Y}_{\ell} = \{\mathbf{y}_{11}, \ldots, \mathbf{y}_{1M}, \mathbf{y}_{21}, \ldots, \mathbf{y}_{2M}, \ldots, \mathbf{y}_{M1}, \ldots, \mathbf{y}_{MM}\}$, where

$$\mathbf{y}_{ij} = \mathcal{O}_{\ell} + i\ \delta\hat{\mathbf{u}}_1 + j\ \delta\hat{\mathbf{u}}_2; \qquad \mathcal{O}_{\ell} = \mathbf{t}_{\ell} - 2(\hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2)$$

and the increments along vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ are given by $\delta\hat{\mathbf{u}}_1 = (a/M)\hat{\mathbf{u}}_1$ and $\delta\hat{\mathbf{u}}_2 = (a/M)\hat{\mathbf{u}}_2$ respectively. The number of bins M must be chosen as an odd integer in order for $\overline{\mathbf{v}}_{\ell}$ to be on the origin of the local coordinate frame. In order to represent only local properties of the manifold a selection of relevant particles in the data set is performed based on their position w.r.t. the new reference frame. First, the projection $\mathbf{t}_m^{\parallel}$ of all particle's original positions $\mathbf{t}_m$ onto the tangent vector $\hat{\boldsymbol{\xi}}_{\ell}$ (note that $||\hat{\boldsymbol{\xi}}_{\ell}|| = 1$) to manifold $\mathcal{M}_k$ at point $\overline{\mathbf{v}}_{\ell}$ is computed: $\mathbf{t}_m^{\parallel} = \left[(\mathbf{t}_m - \overline{\mathbf{v}}_{\ell}) \cdot \hat{\boldsymbol{\xi}}_{\ell}\right] \hat{\boldsymbol{\xi}}_{\ell}$. The distance $\overline{d}_{\ell}$ between adjacent points in $\overline{\mathcal{V}}_{\ell}^k$ is hereon always assumed to be proportional to the distance d_{ℓ} between corresponding adjacent points in $\mathcal{V}_{\ell}^k$

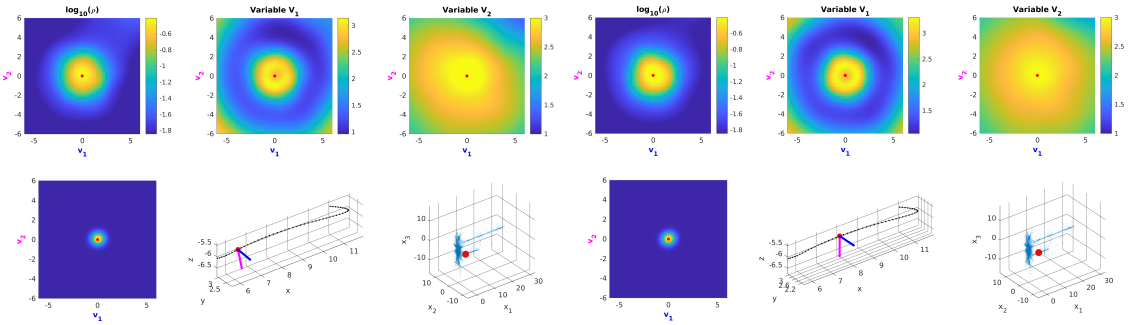
and constant[1]. Only those particles such that $||\mathbf{t}_m^{\parallel}|| \leq \overline{d}_\ell/2$ are then selected. Additionally, the perpendicular component $\mathbf{t}_m^{\perp}$ of position $\mathbf{t}_m$ is computed, by building the projection operator onto $\mathcal{T}_\perp^\ell$ as in section 4.2: $\mathbf{P} = \mathbf{V}\mathbf{V}^\dagger$, where $\mathbf{V}$ is the matrix having $\mathbf{u}_1$ and $\mathbf{u}_2$ as column vectors: $\mathbf{t}_m^{\perp} = \mathbf{P}(\mathbf{t}_m - \overline{\mathbf{v}}_\ell)$. In this analysis only particles lying within the sphere of radius b, centered on $\mathbf{t}_\ell$: $||\mathbf{t}_m^{\perp}|| <= b$ are considered, where b can be imposed by the user or automatically selected as $b = a\sqrt{2}/2$: the half-diagonal of the gridded plane $\mathcal{T}_\perp^\ell$. Identifying with $\mathcal{J}$ the index set of all particles satisfying these two conditions, the weighted mean of variables $V_1(\mathcal{J}, \mathbf{y}_{ij})$ and $V_2(\mathcal{J}, \mathbf{y}_{ij})$ w.r.t. points $\mathbf{y}_{ij} \in \mathcal{Y}_\ell$ can now be computed, following the SPH formulation presented in eq. 3.3. Top row of figures 7.6a-7.6d and 7.7a-7.7d present the behaviours of variables V_1, V_2 and density ρ respectively, for both manifolds detected in data set $\mathbf{Q}_1$, computed at position $\overline{\mathbf{v}}_\ell^k$. Having the probabilistic model for manifold $\mathcal{M}_k$ as AGTMs, the Probability Density Function of the Mixture on plane $\mathcal{T}_\perp^\ell$ is:

$$p^k(y_{ij}) = \sum_{v_\ell \in \mathcal{V}^k} \pi_\ell p(y_{ij}|v_\ell, \hat{\Sigma}_\ell, \hat{\mathbf{W}}), \tag{7.9}$$

where $\hat{\Sigma}_\ell$ and $\hat{\mathbf{W}}$ are the parameters found through optimization with the EM algorithm (see sec. 4.1) and π_ℓ the component proportion given by $\pi_\ell = 1/|\overline{\mathcal{P}}^k|$. The PDFs for manifolds $\mathcal{M}_k$ on plane $\mathcal{T}_\perp^\ell$ at position $\overline{\mathbf{v}}_\ell^k$ are shown in the bottom left panel of figures 7.6a-7.7d, while central and right panels show the current position on the manifold and w.r.t. the whole data set respectively. By applying recursively this procedure to every $\overline{\mathbf{v}}_\ell \in \overline{\mathcal{V}}^k$ a representation of the variables of interest is obtained, on co-moving orthonormal reference frames along each manifold, showing the clear radial behaviour of these variables.

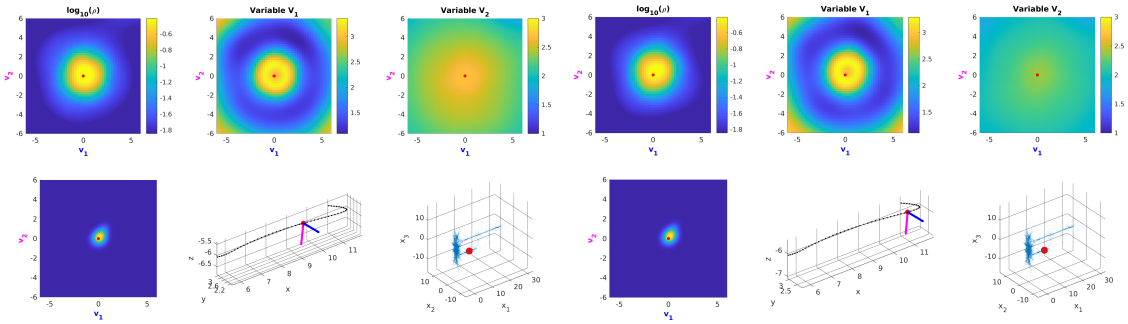
Figures 7.6a-7.6d present the results for selected points on manifold $\mathcal{M}_{13}$. The top central and right panels in each figure clearly manifest the radial behaviour of variable V_1 and V_2 respectively. While the first one decreases linearly from the center outwards, the second

[1]This might not always be the case, but present small variations. However, in this case this effect is usually negligible, as it can be regularized by up-sampling more densely $\mathcal{V}_\ell^k$ (and thus its embedding $\overline{\mathcal{V}}_\ell^k$).



(a) Manifold $\mathcal{M}_{14}$, snapshot n.1 of associated movie clip.

(b) Manifold $\mathcal{M}_{14}$, snapshot n.2 of associated movie clip.



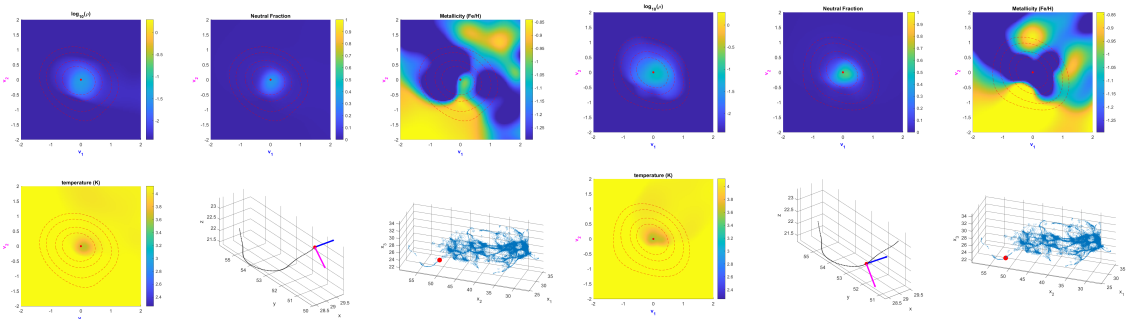
(c) Manifold $\mathcal{M}_{14}$, snapshot n.3 of associated movie clip.

(d) Manifold $\mathcal{M}_{14}$, snapshot n.4 of associated movie clip.

Figure 7.7: As in figure 7.6 but representing properties of manifold $\mathcal{M}_{14}$.

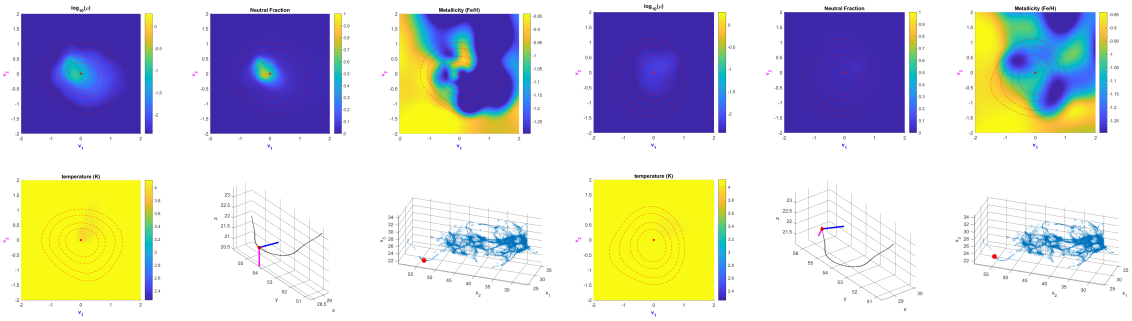
one increases linearly. This behaviour is present throughout the whole elongation of the manifold.

Figures 7.7a-7.7d, top central and left panels show variables V_1 and V_2 for manifold $\mathcal{M}_{14}$. Particularly interesting is the case of variable V_1, where the sinusoidal profile can be seen to be captured. This implies that the center of the manifold has been correctly identified by AGTM and its centers lie closely to its underlying nature (see eq. 3.22). Comparing top right panel in figure 7.7a and the one in figure 7.7d, the longitudinal dimming of variable V_2 can be verified, showing that its true nature has been correctly recovered.



(a) Manifold n. 1, snapshot n.1

(b) Manifold n. 1, snapshot n.30



(c) Manifold n. 1, snapshot n.60

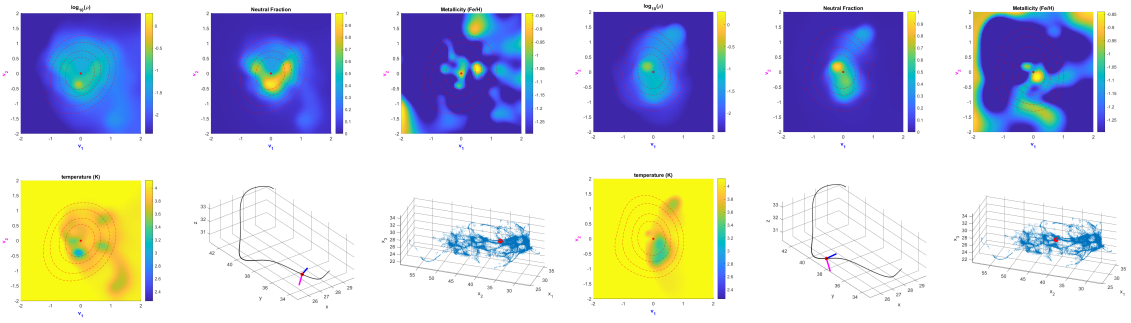
(d) Manifold n. 1, snapshot n.89

Figure 7.8: Equally spaced snapshots of manifold 1 recovered from Jellyfish data set.

7.2.1 Application to Jellyfish

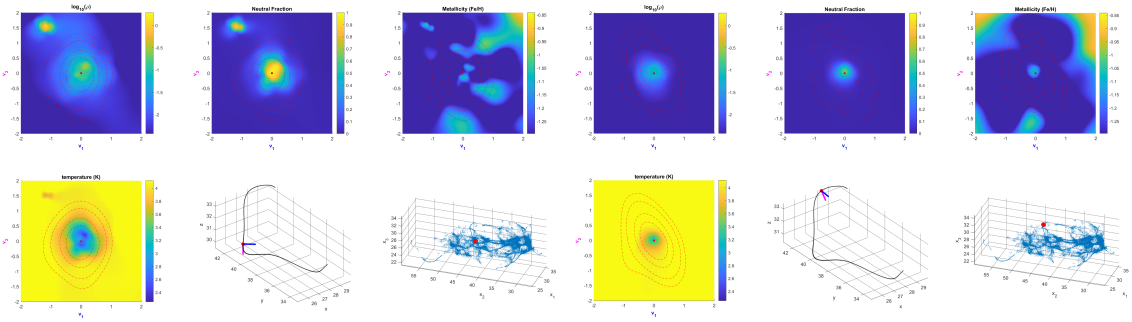
The same analysis has been performed on the manifolds extracted from the Jellyfish data set, applied to the quantities ρ, Neutral Fraction, $[Fe/H]$ and T. The three panels on top (left to right) and the one on the bottom left, show the distribution of these quantities on the perpendicular planes to the tangent bundle the manifold. Over-plotted to the quantities distribution, in every panel PDF of the probabilistic model obtained via AGTM is shown as (red) iso-curves. The red dot represents the center of the plane, corresponding to the embedded skeleton of the manifold. The bottom central and right panel show the location of the current plane w.r.t. the manifold and the whole data set respectively. For consistency, the same manifolds presented in section 7.1.1 are shown here.

Throughout the manifold's elongation, via the snapshots presented in figures 7.8 and 7.9, the centering of the recovered skeleton w.r.t. the regions having highest Density is verified. This



(a) Manifold n. 2, snapshot n.34

(b) Manifold n. 2, snapshot n.90



(c) Manifold n. 2, snapshot n.136

(d) Manifold n. 2, snapshot n.199

Figure 7.9: Snapshots of manifold 2 recovered from Jellyfish data set.

region is also always associated to a higher Neutral fraction and lower Temperature. In the case of manifold 1, as studied via the corresponding 2-dimensional profile (see sec. 7.1.1), little can be said about the behaviour of metallicity, despite the tendency of creating cores with local peaks surrounded by lower metallicity regions. The second manifold, shown in figures 7.9a-7.9d, presents a similar behaviour of the analysed quantities to the one found in manifold n.1. However, the metallicity peak in the core, corresponding to density and Neutral Fraction increase, entails a different evolutionary scenario than the previous stream. When stars are formed in the head of the Jellyfish, their life-span is very short and they terminate their evolution (if massive enough) in a Supernova stage, where the elements synthesized in the star's interior are ejected into the surrounding environment, enriching it of heavier elements and increasing its overall metallicity. If the stream was formed from the metal poor gas of the galaxy cluster's halo, it would still be possible for it to contain metal rich gas, if Star Formation had happened in the past. However, the metallicity distribution would be

heavily influenced by local variations in T, ρ and Neutral Fraction, in a similar fashion to manifold 1 (fig. 7.8). The consistency found throughout the whole elongation of manifold 2, in terms of metallicity distribution in its core, suggests that the gas composing the stream belonged originally to the head, only for it to be stripped away via hydrodynamical and gravitational interactions with the galaxy cluster's halo.

7.2.2 Application to Cosmic Web

As briefly described in section 3.1, the mass distribution over the Large Scale Structure (LSS) of the Universe is highly inhomogeneous. Elongated curvilinear filaments form an intricate "cosmic web", delimiting large empty regions called "voids". Since the visible mass distribution is guided by that of Dark Matter, characterizing the Dark Matter properties of the filaments may shed new light on the mechanisms responsible for mass accretion onto nodes of the cosmic web. Furthermore, linking the results obtained by simulations to the observed kinematic properties of galaxies and (super-)clusters may aid the fine-tuning of the cosmological parameters adopted by the ΛCDM formation scenario.

Adopting the methodology presented in Part I, two manifolds are isolated in the simulated volume, connected by a node of the Cosmic Web. By building orthonormal co-moving frames over the skeletons of these manifolds, their Dark Matter's distribution and kinematic properties are studied. In particular, for each particle in the simulation, at each location over the manifold's skeleton, its tangential and orthogonal velocity w.r.t. the local reference frame is computed. This is achieved by considering that the simulation is modelling Dark Matter. In GADGET-3, Dark Matter is modelled as a collision-less fluid, so that the only force driving DM's motion is (self-) gravity. In this, disregarding the SPH formulation of particles being samples of their local volume, smoothed out by a finite-support kernel of specific (variable) smoothing lengths (see sec. 3.1 for more details) does not cause any harm

in the analysis. Instead, the DM particles' distribution is considered a discrete sample of the actual Dark Matter, where each particle is representative of the kinematic state of a fixed size neighbourhood.

Given a specific point $\overline{\mathbf{v}}_\ell \in \overline{\mathcal{V}_\uparrow}^k$ belonging to manifold $\mathcal{M}_k$'s up-sampled skeleton, as a product of the proposed methodology, the corresponding tangent vector $\hat{\boldsymbol{\xi}}_\ell$ and perpendicular plane $\mathcal{T}_\ell^\perp = \text{span}\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2\}$ are recovered. As discussed in the previous section, the index set $\mathcal{J}$ is defined as the set of all particles falling within the box of size $\overline{d}_\ell \times 2b$ centered on $\overline{\mathbf{v}}_\ell$. Each $\mathbf{t}_m$ such that $m \in \mathcal{J}$, has a velocity $\boldsymbol{\nu}_m \in \mathbb{R}^3$ describing its motion. The velocity vector is projected onto $\mathcal{T}_\ell^\perp$ and along the tangent vector $\hat{\boldsymbol{\xi}}_\ell$ to manifold $\mathcal{M}_k$ on $\overline{\mathbf{v}}_\ell$:

$$\boldsymbol{\nu}_m^\perp = \mathbf{P}\boldsymbol{\nu}_m; \qquad \boldsymbol{\nu}_m^\parallel = (\boldsymbol{\nu}_m \cdot \hat{\boldsymbol{\xi}}_\ell)\hat{\boldsymbol{\xi}}_\ell, \tag{7.10}$$

for every $m \in \mathcal{J}$. On each point in $\mathcal{Y}$ an isotropic Gaussian kernel can now be placed $\mathcal{G}(\mathbf{t}^\perp, \mathbf{y}_{ij})$, with fixed variance κ, in order to obtain a weighted mean of the two projected velocities:

$$\overline{\boldsymbol{\nu}}(\mathbf{y}_{ij})^\perp = \frac{\sum_{m\in\mathcal{J}} \mathcal{G}(\mathbf{t}_m^\perp, \mathbf{y}_{ij})\boldsymbol{\nu}_m^\perp}{\sum_{q\in\mathcal{J}} \mathcal{G}(\mathbf{t}_q^\perp, \mathbf{y}_{ij})} \tag{7.11}$$

$$\overline{\boldsymbol{\nu}}(\mathbf{y}_{ij})^\parallel = \frac{\sum_{m\in\mathcal{J}} \mathcal{G}(\mathbf{t}_m^\perp, \mathbf{y}_{ij})\boldsymbol{\nu}_m^\parallel}{\sum_{q\in\mathcal{J}} \mathcal{G}(\mathbf{t}_q^\perp, \mathbf{y}_{ij})} \tag{7.12}$$

$$\mathcal{G}(\mathbf{t}_m^\perp, \mathbf{y}_{ij}) = \exp \frac{\|\mathbf{t}_m^\perp - \mathbf{y}_{ij}\|^2}{\kappa}. \tag{7.13}$$

Note that parameter κ can be fixed by the user to match a desired smoothing, however in the experiments proposed in this work it is fixed to $\kappa = \frac{\sqrt{a/M}}{4}$: the half-diagonal of the square formed by four adjacent points sampling $\mathcal{T}_\ell^\perp$ in $\mathcal{Y}_\ell$. Having obtained a velocity field on $\mathcal{T}_\ell^\perp$, it is now possible to compute its rotor and divergence w.r.t. the local coordinate frame. Define the ∇ operator on the local reference frame given by $\text{span}\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \hat{\xi}_\ell\}$ as $\nabla = (\partial/\partial u_1, \partial/\partial u_2, \partial/\partial \xi_\ell)$, so that the rotor and divergence of vector field $\overline{\boldsymbol{\nu}}$ can be written

as:

$$\begin{aligned}\nabla \times \overline{\boldsymbol{\nu}}^{\perp} &= \left(\frac{\partial \overline{\boldsymbol{\nu}}^{\perp,3}}{\partial u_1} - \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial \xi_\ell}\right)\hat{\mathbf{u}}_1 + \left(\frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial \xi_\ell} - \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,3}}{\partial u_1}\right)\hat{\mathbf{u}}_2 + \left(\frac{\partial \overline{\boldsymbol{\nu}}^{\perp,2}}{\partial u_1} - \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial u_2}\right)\hat{\boldsymbol{\xi}}_\ell \\ &= \left(\frac{\partial \overline{\boldsymbol{\nu}}^{\perp,2}}{\partial u_1} - \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial u_2}\right)\hat{\boldsymbol{\xi}}_\ell; \qquad (7.14)\end{aligned}$$

$$\nabla \cdot \boldsymbol{\nu}^{\perp} = \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial u_1} + \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,2}}{\partial u_2} + \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,3}}{\partial \xi_\ell} = \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,1}}{\partial u_1} + \frac{\partial \overline{\boldsymbol{\nu}}^{\perp,2}}{\partial u_2}. \qquad (7.15)$$

The final forms of the rotor and divergence are obtained by noting that only the third term in the expansion is non-null in the first case, while it is the only null element in the second (being the vector field over the plane).

The results for the two manifolds selected from the simulated volume are shown in figures 7.10 and 7.11. In each sub-figure, information is presented by the following scheme:

Top left corner: Skeleton (solid, black line) recovered by up-sampling the embedded graph obtained via AGTM and isosurface of the corresponding model's PDF (red, opaque surface). Extracted simulated particles obtained via soft assignment on the whole data set, given the model (grey dots). Current center $\overline{\mathbf{v}}_\ell$ (red sphere) of the perpendicular plane $\mathcal{T}_\ell^{\perp}$ and vectors $\hat{\mathbf{u}}_1$ $\hat{\mathbf{u}}_2$ spanning it (blue and magenta arrows respectively).

Top central panel: Heatmap of the weighted mean projected tangential velocity $||\overline{\boldsymbol{\nu}}^{\parallel}||$, at rest frame: the velocity of the central point $\overline{\boldsymbol{\nu}}^{\parallel}(\overline{\mathbf{v}}_\ell)$ has been removed in order to represent local velocities w.r.t. the local frame. Over-plotted are the iso-contours of the model's PDF (dashed black lines) and a mask hiding sparsely populated regions of plane $\mathcal{T}_\ell^{\perp}$.

Top right panel: Quiver plot of the weighted mean velocity field (blue arrows) over the plane with over-plotted iso-contours of the model's PDF. Again the central velocity $\overline{\boldsymbol{\nu}}^{\perp}(\overline{\mathbf{v}}_\ell)$ has been subtracted to the velocity field.

Bottom left panel: Number of particles within the sphere of radius κ for each point $\mathbf{y}_{ij} \in$

$\mathcal{Y}_\ell$, sampling regularly plane $\mathcal{T}_\ell^\perp$, with over-plotted mask.

Bottom central panel: Tangential component of the rotor (also called "curl") of the weighted mean velocity field computed over the perpendicular plane.

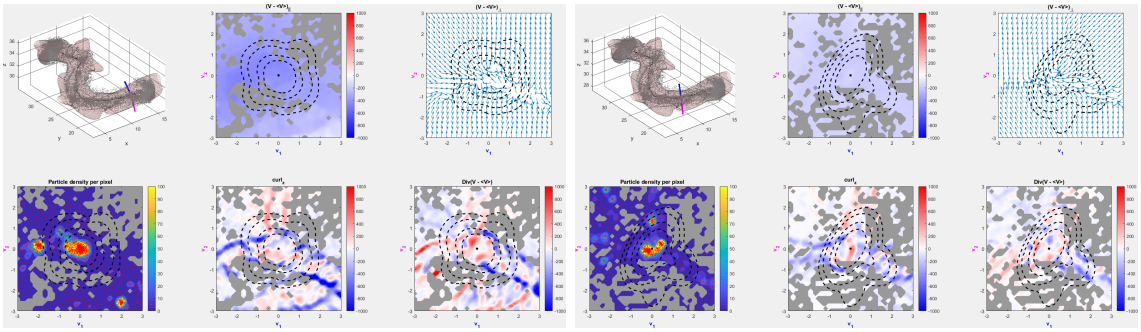
Bottom right panel: Divergence of the weighted mean velocity field computed over the perpendicular plane.

A few main results can be drawn from the visualization of the kinematic properties of the two manifolds using the proposed technique:

- The manifolds' skeletons are consistently aligned with their corresponding densest regions. This is visible in the bottom left panel of each figure. The regions containing the largest amount of particles are consistently in the center of the plane, for every point on the skeleton. Indeed, while the distribution of points is sparser in the outer regions of the plane, its center is always densely populated.

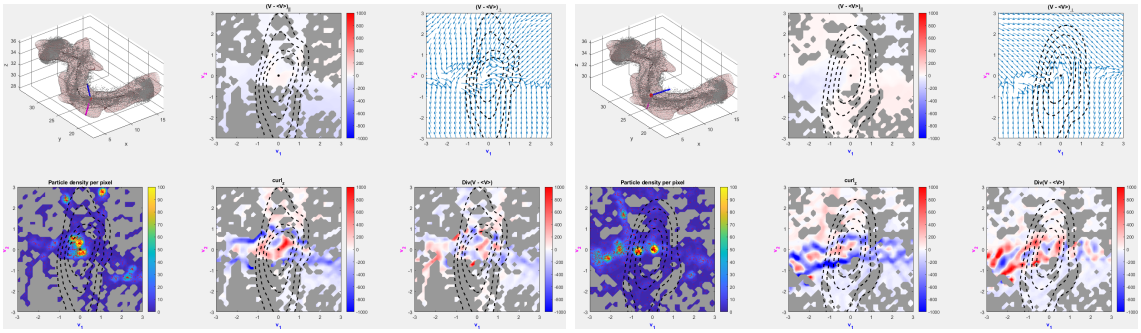
- The tangential velocity of particles on the manifold (top central panel) has the tendency to change sign at a certain at a certain distance from the two extremities. The extremities of each manifold are the nodes of the cosmic web. In particular, starting from one node and moving towards the other, the tangential velocity has a negative sign at first (overall blue color), meaning that particles are pulled towards the starting node. Moving towards the second node, the tangential velocity's module tends to decrease until it reaches zero. After this "breaking" point, the tangential velocity is aligned with the crawling direction, meaning that the particles tend to be pushed towards this second node. This "breaking" point is clearly visible in figure 7.10c for manifold n.1 and 7.11e for manifold n.2.

- The parallel curl (perpendicular to the plane) of the velocity field shows clear islands of opposite signs. This result confirms what found in Laigle et al., 2014. However,



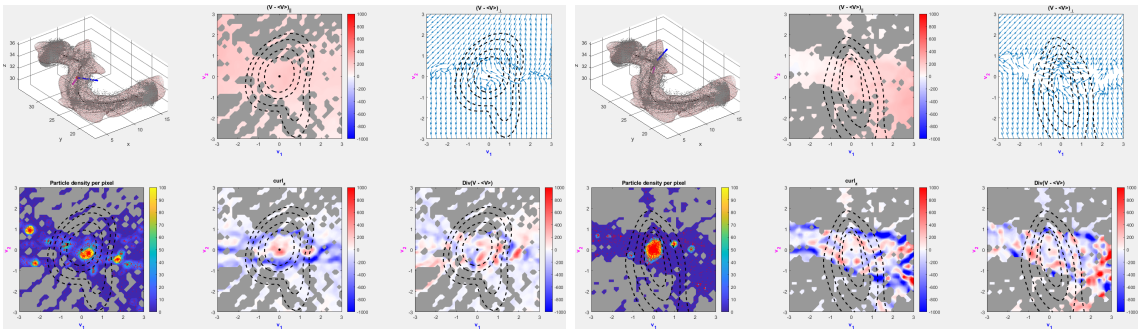
(a) Manifold n. 1, snapshot n.8

(b) Manifold n. 1, snapshot n.16



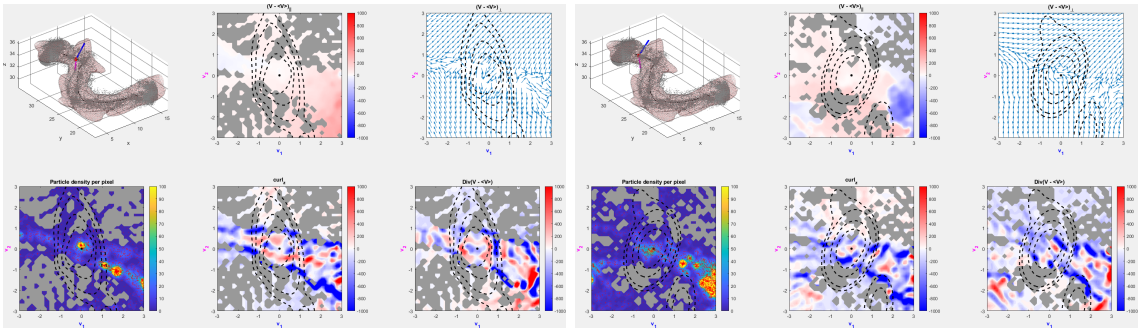
(c) Manifold n. 1, snapshot n.24

(d) Manifold n. 1, snapshot n.32



(e) Manifold n. 1, snapshot n.40

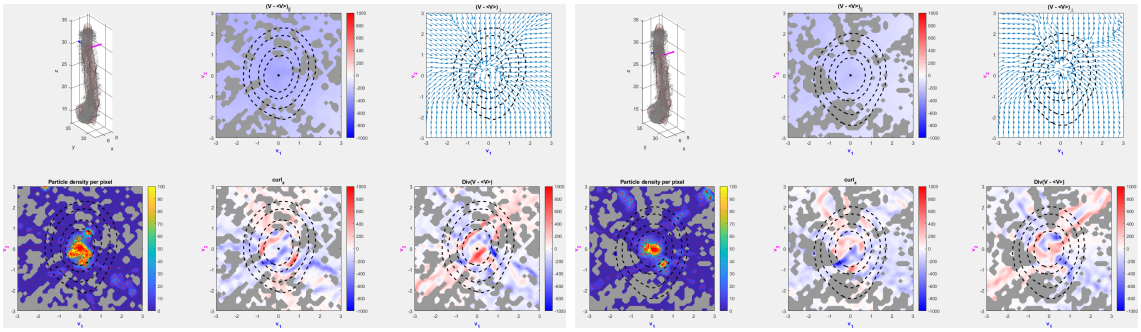
(f) Manifold n. 1, snapshot n.48



(g) Manifold n. 1, snapshot n.52

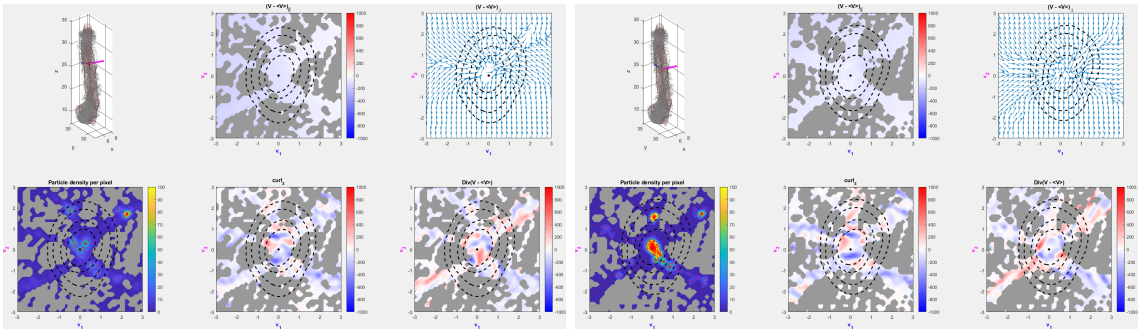
(h) Manifold n. 1, snapshot n.56

Figure 7.10: Snapshots of manifold 1 recovered from Cosmic Web data set.



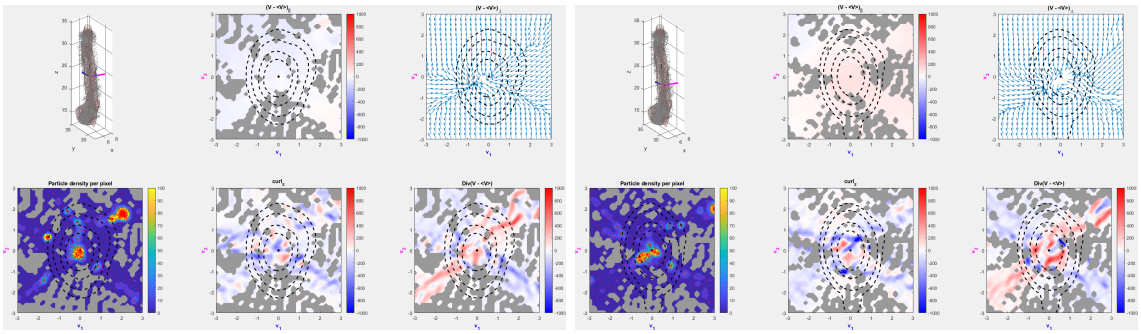
(a) Manifold n. 2, snapshot n.8

(b) Manifold n. 2, snapshot n.12



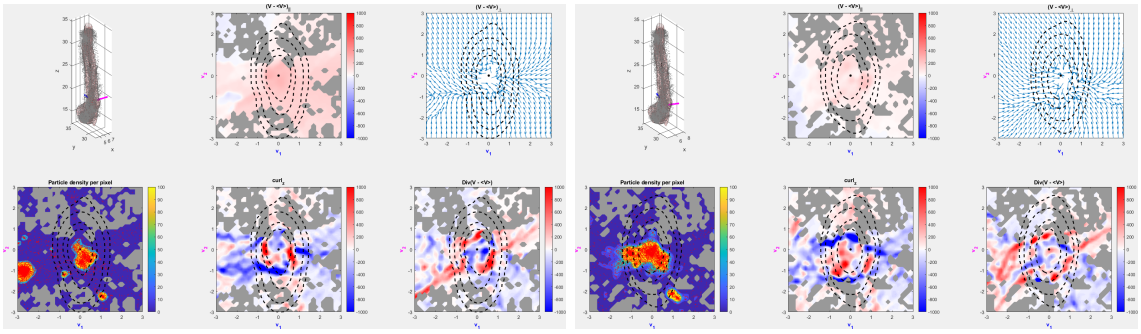
(c) Manifold n. 2, snapshot n.16

(d) Manifold n. 2, snapshot n.20



(e) Manifold n. 2, snapshot n.24

(f) Manifold n. 2, snapshot n.28



(g) Manifold n. 2, snapshot n.36

(h) Manifold n. 2, snapshot n.40

Figure 7.11: Snapshots of manifold 2 recovered from Cosmic Web data set.

contrary to the findings in the presented work, it is generally assumed that the core of the streams can be divided into only four distinct vorticity regions. Although in some iterations an agreement can be found with the predictions (e.g. figures 7.10c, 7.10g, 7.11c,7.11d and 7.11e), that is generally not the case, as the different vorticity regions may have a more irregular distribution.

- Particles in the outskirts of the manifolds are attracted towards their cores. This results transpires from a visual inspection of the top right panels of all figures, for both manifolds. The weighted mean perpendicular velocity field is generally oriented towards the center of the plane (the manifold's core). Except for the central region of the planes, this conclusion is also confirmed by the divergence plot (bottom right panel), presenting generally uniform behaviour in the outskirts.

Chapter Eight

Modelling time-evolving manifolds

In the previous chapters a description of how to extract and model all of the structures hidden within a data set, without having prior information on their nature has been proposed. However, when such information is available, its incorporation into the model leads to a more reliable interpretation of the results. Furthermore, in several situations the nature of the manifold is dynamic and it evolves through time, e.g. astronomical numerical simulations. In these cases, a joint spatio-temporal representation of the manifold is needed, in order to capture its underlying nature. It has been shown how GTM can be generalized to model time-series (GTM Through Time, C.M. Bishop, 1997), by imposing a Hidden Markov Model (HMM) structure on latent space nodes Rabiner, 1989. This enables one to project whole trajectories evolving around a lower-dimensional manifold onto the latent space of a single GTM. Here a different problem is tackled - a methodology for tracking the evolution of a time series of low-dimensional manifolds based on the first-order Markovian assumption Gagniuc, 2017, where the spatial structure of the manifold at each time-step is modelled via GTM. The model is propagated through adjacent time-steps by minimization of a set of cost functions, aimed at preserving the position of the Gaussian Mixture's centres, while still fitting the evolved data set. The proposed regularizations are based on L_1- and L_2-norms on changes of model parameters.

The methodology is applied to a particle simulation of a dwarf galaxy interacting with its host galaxy cluster. The manifold of interest is a cavity ("bubble") generated by a Supernova and detected via the Dionysus toolbox Morozov, 2007 (Persistent Homology, PH) at a given temporal stage. The probabilistic model is obtained via Geodesic GTM A. Taghribi et al., 2021, a reformulation of GTM where the linear latent space is replaced by a spherical one.

Consider a sequence $\mathcal{S} = (\mathcal{Q}^0, ..., \mathcal{Q}^\tau, ..., \mathcal{Q}^T)$ of point clouds $\mathcal{Q}^\tau = \left\{\mathbf{t}_1^\tau, \mathbf{t}_2^\tau, ..., \mathbf{t}_{N_\tau}^\tau\right\} \subset \mathbb{R}^d$, sampling a noisy, time-dependent, low-dimensional manifold $\mathcal{M}(\tau)$ of dimension ℓ throughout its evolution, $\tau = 1, 2, ..., T$. A naive approach to produce a probabilistic model of the sequence $\mathcal{S}$ is to model each temporal realization of the manifold individually, disregarding the mutual temporal dependencies of its components. However, it is sometimes useful to include the time-dependency of data sets when interested in joint space-time variations of the manifold's properties. The chosen methodology for modelling the spatial state of manifold $\mathcal{M}(\tau)$ is first described for two topologically different cases. Afterwards, different measures for the propagation of such models through adjacent temporal stages are investigated.

8.1 One-dimensional GTM

Assume that the manifolds $\mathcal{M}(\tau)$ are open, one-dimensional and non-linear. As previously mentioned, each $\mathcal{M}(\tau)$ is sampled by $\mathcal{Q}^\tau$. Having the information about the manifold's dimensionality ($\ell = 1$), the Generative Topographic Mapping (GTM) formulation can be used in order to define the proposed model as a constrained mixture of Gaussians, where the latent space representation of centres is the interval $\mathrm{I} = [-1, 1]$. The interval I is sampled regularly with M centres $\mathbf{X} = \{x_1, \ldots, x_M\}$ and define a set of $K < M$ Radial Basis Functions (RBFs) $\phi(\mu_k, x_m)$ centred on uniformly spaced points $\mathbf{K} = \{\mu_1, \ldots, \mu_K\} \in \mathrm{I}$.

RBFs are chosen to be spherical Gaussians:

$$\phi(\mu_k, x_m) = \exp\left[-\frac{(\mu_k - x_m)^2}{2\sigma^2}\right] \ , \tag{8.1}$$

where σ is a scale parameter chosen to be a multiple of the distance between neighboring RBFs centres. The set of latent centres $\mathbf{X}$ is then mapped onto the ambient space of manifold $\mathcal{M}(\tau)|_{\tau=0}$ through the non-linear mapping function $\mathbf{y}(\mathbf{X};\mathbf{W})$ defined as:

$$\mathbf{y}(\mathbf{X};\mathbf{W}) = \mathbf{W}\Phi(\mathbf{X},\mathbf{K}) \ . \tag{8.2}$$

Here, Φ is the matrix with $\phi(\mu_k, x_m)$ being its $(k,m)-$th element. Each $x_m \in \mathbf{X}$ is thus mapped, through eq. (8.2), to point $\mathbf{c}_m \in \mathbb{R}^d$, which acts as a centre of the Gaussian mixture. On each mapped centre $\mathbf{c}_m$ a manifold-aligned noise model is imposed as obtained by differentiation of the mapping function $\mathbf{y}(\mathbf{X};\mathbf{W})$ w.r.t. the latent space coordinates, as described in Christopher M. Bishop et al., 1998a:

$$\mathbf{C}_m = \frac{1}{\beta}\mathbf{I} + \eta_m \sum_{p=1}^{\ell} \left[\left(\frac{\partial \mathbf{y}}{\partial x_p}\right)^\top \left(\frac{\partial \mathbf{y}}{\partial x_p}\right)\right]\Bigg|_{x_m} \ . \tag{8.3}$$

Here, x_p denotes the latent coordinate of dimension p. Where the derivative of the mapping function is given by

$$\left.\frac{\partial \mathbf{y}}{\partial x}\right|_{x_m} = \mathbf{W}\left.\frac{\partial \Phi}{\partial x}\right|_{x_m} = -\frac{\mu_k - x_m}{\sigma^2}\Phi(x_m). \tag{8.4}$$

With suitable values for parameters β and η_m $(m = 1, \ldots, M)$ the density of each component of the mixture reads:

$$p(\mathbf{t}|x_m, \mathbf{W}) = \left(\frac{1}{2\pi}\right)^{d/2} |\mathbf{C}_m|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{c}_m - \mathbf{t})^\top \mathbf{C}_m^{-1}(\mathbf{c}_m - \mathbf{t})\right] \ . \tag{8.5}$$

The overall GTM model is a flat mixture of Gaussians,

$$p(\mathbf{t}|\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{t}|x_m, \mathbf{W}) \ . \tag{8.6}$$

whose parameters can be trained via a maximum-likelihood optimization, e.g. through the Expectation-Maximization (EM) algorithm Christopher M. Bishop et al., 1998a.

8.2 Geodesic GTM for cavities

In A. Taghribi et al., 2021, *Geodesic-GTM* (gGTM) has been proposed, where the linear latent space of classical GTM is replaced by a spherical one with dedicated geodesic distance. While classical GTM can be applied in principle to spherical manifolds, its performance is largely guided by its initialization. Furthermore, the low-dimensional representation of the manifold would still be inherently incorrect, since a sphere is not homeomorphic to the linear subspace of the same dimension. Thus, in order to correctly capture the topology of the manifold and systematically solve the initialization issue, gGTM is here adopted as the preferential methodology for modelling cavities. Hence, every point $\boldsymbol{\xi}$ on the surface of the sphere is uniquely determined by a pair of angular coordinates: θ and λ where $(\theta, \lambda) \in \mathrm{I}^{\ell}_{\angle} = [-\pi; \pi] \times [-\pi/2; \pi/2]$ and $\mathrm{I}^{\ell}_{\angle}$ denotes the ℓ-dimensional, angular interval ($\ell = 2$). Assuming a radius $r = 1$ the geodesic distance between any pair of points $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ can be rewritten:

$$d_{\Omega}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = 2\arcsin\left(\frac{\Delta\Omega}{2}\right), \qquad \Delta\Omega = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} \ , \tag{8.7}$$

where $\Delta X = \sin\theta_i \cos\lambda_i - \sin\theta_j \cos\lambda_j$, $\Delta Y = \sin\theta_i \sin\lambda_i - \sin\theta_j \sin\lambda_j$ and $\Delta Z = \cos\theta_i - \cos\theta_j$. As in the previous section, a set $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subset \mathrm{I}^{\ell}_{\angle}$ of latent points and $\mathbf{K} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\} \subset \mathrm{I}^{\ell}_{\angle}$ of centres for the RBFs $\boldsymbol{\phi} = \{\phi(\boldsymbol{\mu}_1), \ldots, \phi(\boldsymbol{\mu}_K)\}$ is defined on the

latent space. Using this formulation each RBF takes the form:

$$\phi(\boldsymbol{\mu}_k, \boldsymbol{\xi}_m) = \exp\left[-\frac{d_\Omega(\boldsymbol{\mu}_k, \boldsymbol{\xi}_m)^2}{2\sigma^2}\right] \quad . \tag{8.8}$$

The probabilistic model is then formed analogously to the one-dimensional case, with the only exception for the estimation of the covariance matrix of the centres, that is now dependant on the geodesic distance on a sphere (eq. (8.7)). In particular, the derivative of the RBFs w.r.t. to latent variables θ and λ are:

$$\left.\frac{\partial \Phi}{\partial \theta}\right|_{\boldsymbol{\xi}_m} = -\frac{2\arcsin\left(\frac{\Delta\Omega}{2}\right)}{\sigma^2}\Phi_m \left.\frac{\partial d_\Omega}{\partial \theta}\right|_{\boldsymbol{\xi}_m} \tag{8.9}$$

$$\left.\frac{\partial \Phi}{\partial \lambda}\right|_{\boldsymbol{\xi}_m} = -\frac{2\arcsin\left(\frac{\Delta\Omega}{2}\right)}{\sigma^2}\Phi_m \left.\frac{\partial d_\Omega}{\partial \lambda}\right|_{\boldsymbol{\xi}_m} \tag{8.10}$$

and the derivatives of the geodesic distance w.r.t. to the latent variables:

$$\left.\frac{\partial d_\Omega}{\partial \theta}\right|_{\boldsymbol{\xi}_m} = -\omega\left(\Delta X \cos\theta_m \cos\lambda_m + \Delta Y \cos\theta_m \sin\lambda_m - \Delta Z \sin\theta_m\right) \tag{8.11}$$

$$\left.\frac{\partial d_\Omega}{\partial \theta}\right|_{\boldsymbol{\xi}_m} = \omega\left(\Delta X \sin\theta_m \sin\lambda_m - \Delta Y \sin\theta_m \cos\lambda_m\right), \tag{8.12}$$

where $\omega = 2/\left(\Delta\Omega\sqrt{4+\Delta\Omega^2}\right)$. The covariance matrix of each centre $\mathbf{y}(\boldsymbol{\xi_m}; \mathbf{W})$ of the Gaussian mixture model is obtained via eq. (8.3), where $\ell = 2$, $x_1 = \theta$ and $x_2 = \lambda$.

8.3 Temporal evolution

Given a data set $\mathcal{Q}^\tau = \left\{\mathbf{t}_1^\tau, \mathbf{t}_2^\tau, \ldots, \mathbf{t}_{N_\tau}^\tau\right\} \subset \mathbb{R}^d$ sampling the $\ell-$dimensional manifold $\mathcal{M}(\tau)$ at time τ, its evolution in time can be modelled by propagating the model created at time $\tau - 1$ through the sequence $\mathcal{S}(\tau)$. The temporal smoothness of model structure is imposed

by chaining the model parameters in the first-order Markov dependency structure through regularization terms. Once the probabilistic model of data set $\mathcal{Q}^{\tau-1}$ is obtained, when presented with the next data set $\mathcal{Q}^{\tau}$, the minimization of a cost function is required for the recovery of the updated parameters $\mathbf{W}^{\tau}$. Multiple choices could be made for the cost function; in the following four options are presented and compared, that behave differently when applied iteratively on each pair of adjacent data sets in $\mathcal{S}$:

$$\mathrm{l}_{2\mathrm{W}}(\mathbf{W}^{\tau};\Lambda) = -\mathcal{L}(\mathcal{Q}^{\tau}|\mathbf{X},\mathbf{W}^{\tau}) + \Lambda\|\mathbf{W}^{\tau} - \mathbf{W}^{\tau-1}\|_2^2 \tag{8.13}$$

$$\mathrm{l}_{2\mathrm{C}}(\mathbf{W}^{\tau};\Lambda) = -\mathcal{L}(\mathcal{Q}^{\tau}|\mathbf{X},\mathbf{W}^{\tau}) + \Lambda\|\mathbf{W}^{\tau}\Phi - \mathbf{W}^{\tau-1}\Phi\|_2^2 \tag{8.14}$$

$$\mathrm{l}_{1\mathrm{W}}(\mathbf{W}^{\tau};\Lambda) = -\mathcal{L}(\mathcal{Q}^{\tau}|\mathbf{X},\mathbf{W}^{\tau}) + \Lambda\|\mathbf{W}^{\tau} - \mathbf{W}^{\tau-1}\|_1 \tag{8.15}$$

$$\mathrm{l}_{1\mathrm{C}}(\mathbf{W}^{\tau};\Lambda) = -\mathcal{L}(\mathcal{Q}^{\tau}|\mathbf{X},\mathbf{W}^{\tau}) + \Lambda\|\mathbf{W}^{\tau}\Phi - \mathbf{W}^{\tau-1}\Phi\|_1 \ . \tag{8.16}$$

Here $\mathcal{L}(\mathcal{Q}|\mathbf{X},\mathbf{W})$ is the log likelihood of the GTM model with latent structure $\mathbf{X}$ and parameters $\mathbf{W}$, given a data set $\mathcal{Q}$. The cost functions are based on the L_1 and L_2-norms ($\|\cdot\|_1$ and $\|\cdot\|_2$) of the (vectorized) differences between weights or centres at times $\tau-1$ and τ and are minimized following the simplex search method presented in Lagarias et al., 1998. At each time cost functions evaluation, the covariance matrices of the components are also updated according to eq. (8.3), keeping fixed parameters β and η_m for $m = 1,\ldots,M$.

8.4 Experiments on a synthetic data set

In order to test the performance of the minimization procedure when presented with the four cost functions, a controlled experiment is constructed where the Ground Truth is known. A synthetic sequence $\mathcal{S}(\tau)$ is presented in figure 8.1 (top row). Each panel shows the spatial distribution of points at the corresponding time (grey points) and a pre-defined selection of centres considered optimal (Ground Truth: $C_{GT}(\tau)$) for the desription and modelling of

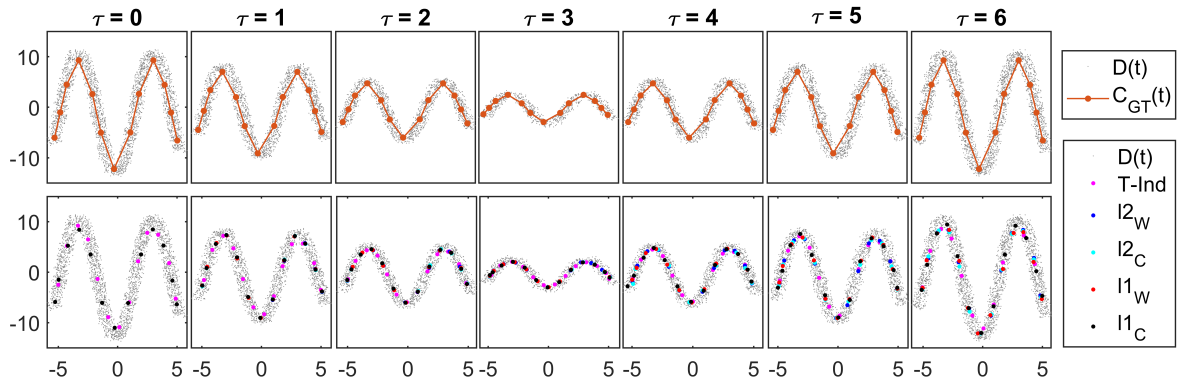


Figure 8.1: Top row: 1D time-evolving data set with corresponding "Ground Truth". Bottom: centres obtained via independent (magenta line) or time-dependent modelling.

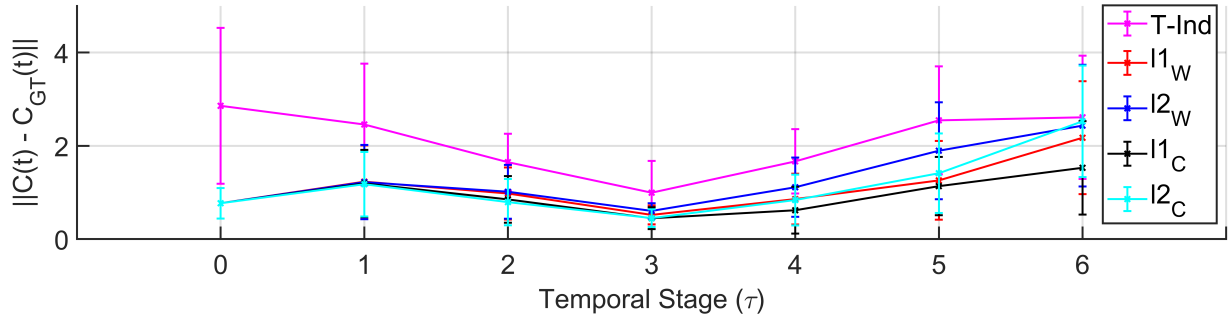


Figure 8.2: Mean and standard deviation of distances between updated centres and corresponding Ground Truth at time τ.

the temporal evolotion. For example, to understand the evolution of such "wave shapes", one would need to consistently track the extremal "knot points" and the inner nodes in between them. The one-dimensional GTM presented in section 8.1 is set with 13 centres and 7 RBFs. The weights of the mapping function $\mathbf{y}(\mathbf{X}, \mathbf{W})$ are initialized via linear regression imposing congruence between the mapped centres and $C_{GT}(\tau = 0)$ at stage $\tau = 0$. The Λ hyper-parameter was estimated through $5-$fold cross-validation. For comparison, each temporal stage has been modelled individually with the same initialization, but additionally trained with the EM-algorithm. The centres recovered via both time-independent (T-Ind) and dependent modelling are shown in the bottom row of figure 8.1. At each temporal stage, the distance between the Ground Truth centres at that time and the centres obtained by each modelling approach is computed, with results presented in figure 8.2. Here shown are

the mean and standard deviation of this measure at each temporal stage. The minimum distance is consistently achieved by the l1-regularization on centres (l_{1C}), followed in order by l_{1W}, l_{2C}, l_{2W} and finally time-independent modelling.

Given the results on this simple synthetic experiment, $l1-$regularization over the mixture's centres results particularly convenient when identifiability of local properties of a manifold through time is required. The proposed methodology is now demonstrated in the context of a simulated dwarf galaxy, orbiting around the halo of its host galaxy cluster. Taking advantage of both the 1D and Geodesic formulations of GTM, the evolution of the dwarf galaxy's tail (what makes it a JellyFish galaxy) is tracked, together with detected cavities in the point cloud.

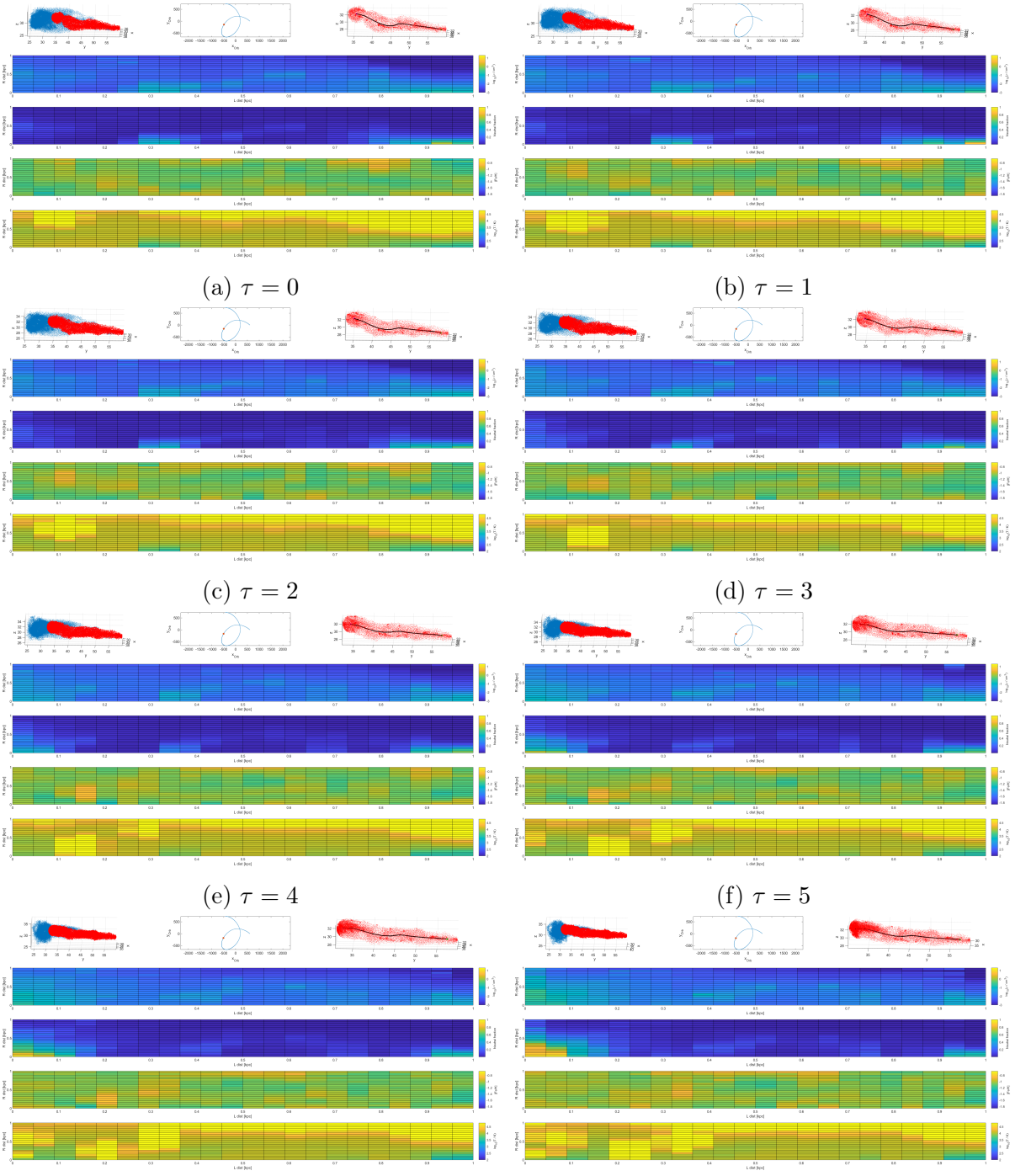
8.5 Tracking the JellyFish galaxy's tail through time

Having already tested the performance of temporal 1D GTM on the synthetic data set in fig. 8.1, in this section, only results of the l_{1C} regularization on Centres are taken into consideration. Through its interaction with the galaxy cluster, the dwarf galaxy may form elongated streams departing from its core. While there are evolutionary stages where multiple streams may appear, when approaching the orbital pericenter (closest orbital point to the galaxy cluster's center) the galaxy often shows a single elongated tail. In order to extract information from the tail's evolution, 1D GTM, described in section 8.1, is applied with the temporal evolution obtained by l_{1C} regularization to these stages. Since this manifold is again one-dimensional, the visualization technique proposed in sec. 7.1 can be used, creating a sequence of bi-plots summarizing the stream's properties through its temporal evolution. The results are shown in figures 8.3-8.4. In addition to the properties described in section 7.1, at each temporal snapshot, the current position of the galaxy on its orbit

around the galaxy cluster's center is shown. The same behaviour found for singular streams w.r.t. Neutral Fraction, Temperature and Density can be found on the tidal tail as well. Furthermore, these properties are preserved through the modelled evolutionary stages. In particular, Density and Neutral Fraction (first and second bi-plots from top) have a tendency of being higher in the central region of the streams (closer to the manifold's skeleton), while Temperature (third bi-plot from top) is generally lower there and tends to increase with the distance to the manifold's core. It is arguable that the tail of JellyFish galaxies (interpreted as evolving dwarf galaxies), is a possible locus of star formation, throughout its evolution. Indeed, later stages show how the tail can be broken in detached pieces, after a Supernova explosion (or multiple SNe explosions) takes place. SNe (especially Type II SNe) are the final evolutionary stage of highly massive, young stars. The formation of such massive stars is preferentially located in the core of the streams, as seen both in this case and in the ones treated in sections 7.1.1 and 7.2. The SNe explosion is a highly energetic process, releasing in the ambient space energies up to 10^{53} ergs. The kinetic energy of the material ejected from the outer shells of the starts is $\sim 10^5 1$ ergs. So it is plausible that the material in the surroundings of the SN explosion is swept away from the the centre of expansion (the stellar Iron core), which is preferentially located on the internal regions of the stream. This violent energy release may cause the disruption of the tail at certain evolutionary stages of the galaxy through its orbit.

8.6 Tracking supernovae explosions through time

By increasing local densities the interaction between the gas of the galaxy and that of the cluster may enforce the formation of massive stars which later approach the phase of Supernovae. When a star reaches this stage it expands outwards in a short time, sweeping away the surrounding gas. The "explosion" of the Supernova generates a cavity in the ambient



(a) $\tau = 0$ (b) $\tau = 1$

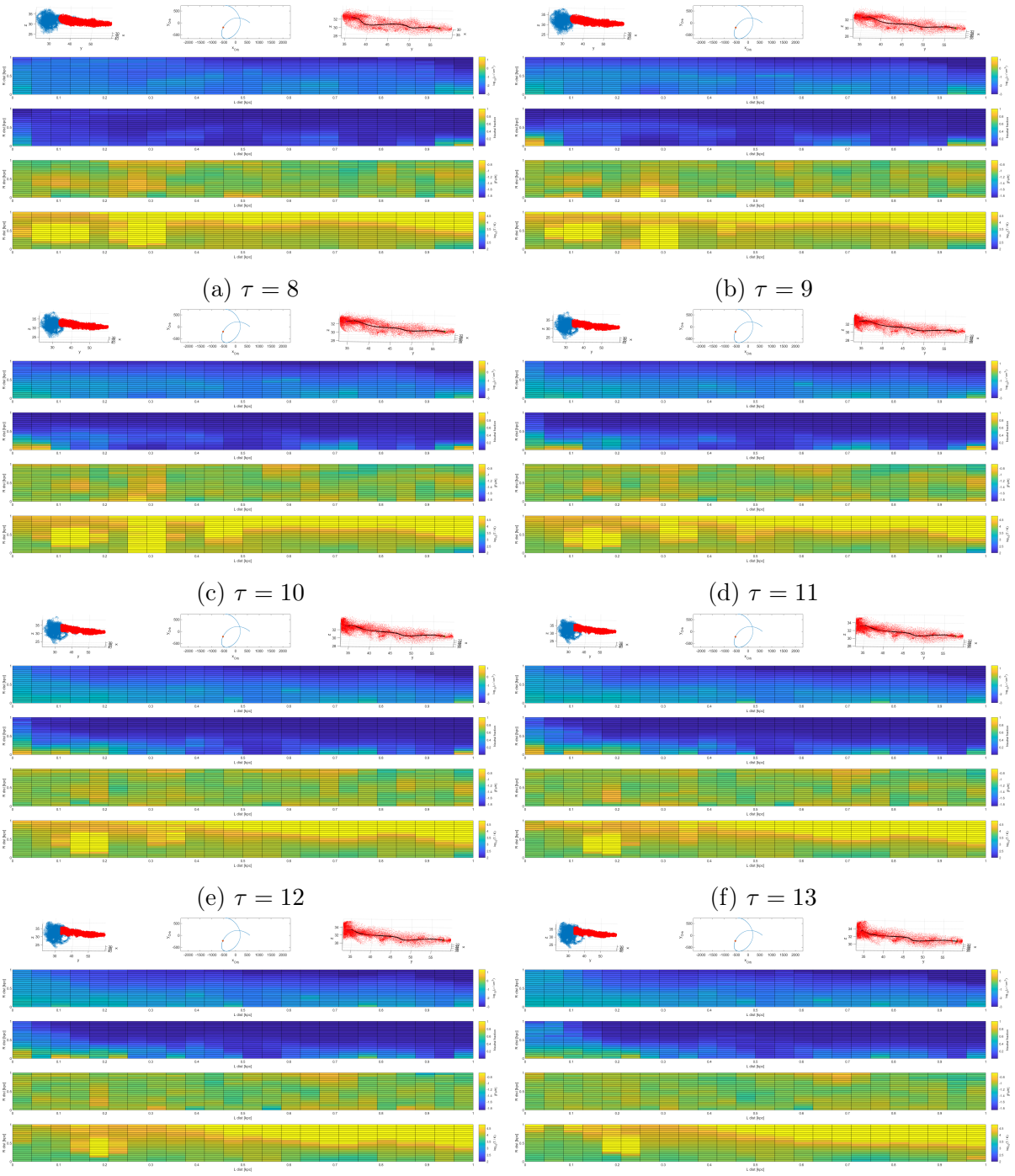
(c) $\tau = 2$ (d) $\tau = 3$

(e) $\tau = 4$ (f) $\tau = 5$

(g) $\tau = 6$ (h) $\tau = 7$

Figure 8.3: Bi-dimensional profiles of the JellyFish's tidal tail tracked thtough 8 consecutive temporal stages.

gas, usually referred to as a "bubble". In Morozov, 2007, a methodology for detecting cavities has been proposed using PH. In A. Taghribi et al., 2021 a sub-sampling procedure (ASAP)



(a) $\tau = 8$ (b) $\tau = 9$

(c) $\tau = 10$ (d) $\tau = 11$

(e) $\tau = 12$ (f) $\tau = 13$

(g) $\tau = 14$ (h) $\tau = 15$

Figure 8.4: Continuation of figure 8.3.

has been shown to regularize the detection through PH and applied to a similar astronomical simulation. The aim of this section is to track the evolution of a detected bubble through

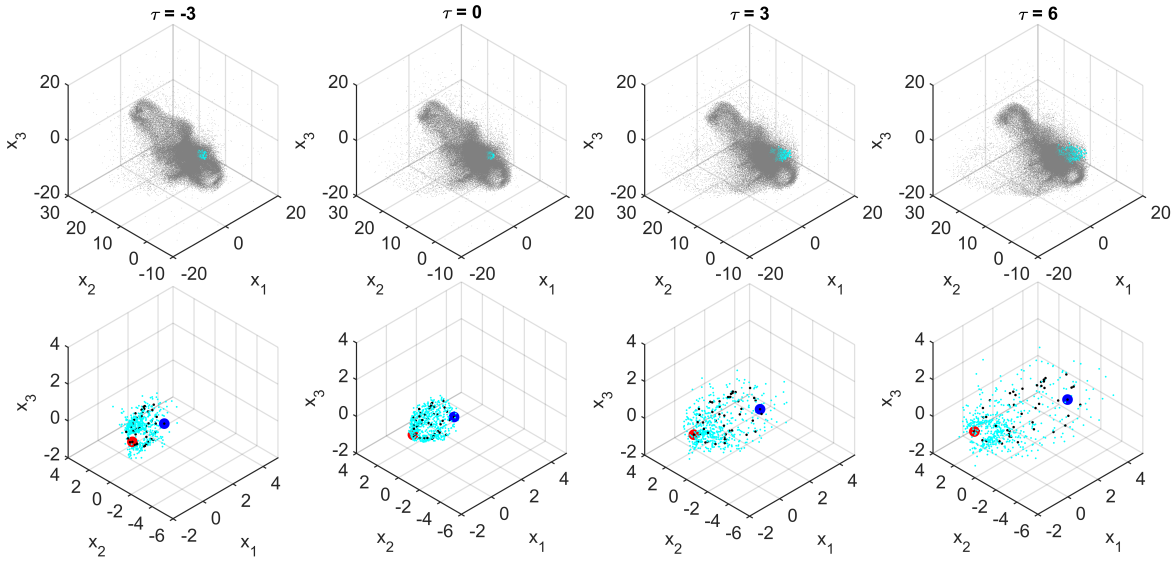


Figure 8.5: Top row: Temporal stages of simulated dwarf galaxy. Grey dots are galaxy's point distribution, in cyan the recovered bubble propagated through time. Bottom row: Zoom in on the bubble with centres (black) representing opposite lobes (red and blue).

adjacent temporal-stages of the particle simulation. The probabilistic modelling technique proposed in section 8.2 is particularly convenient for the preservation of the position of centres in adjacent temporal stages while capturing the topological properties of the bubble. All four cost functions tested in the 1D case are adopted. However, since the $l1-$regularization on centres proved to be the optimal choice in the previous analysis, the detailed comparison will be performed w.r.t. this approach only. The latent space $\mathrm{I}_{\angle}^{\ell}$ is regularly sampled with 45 centres and 25 RBFs, and the weights of the mapping function initialized by assuming a spherical distribution of the embedded centres[1]. The model is initialized at time $\tau = 0$ and EM-algorithm run in order to obtain an accurate model at fixed time of the cavity. The model is them propagated from time $\tau = 0$ to the following and preceding stages, obtaining the spatio-temporal model of the bubble throughout 11 consecutive snapshots. A selection of snapshots equally spaced in time is shown in figure 8.5, top row. Grey dots represent the position of the simulated particles and cyan ones show the bubble's position w.r.t. the galaxy, at each temporal stage. The bubble's particles, detected at $\tau = 0$ via the

[1]The radius of the embedded sphere is estimated from the data similar to A. Taghribi et al., 2021.

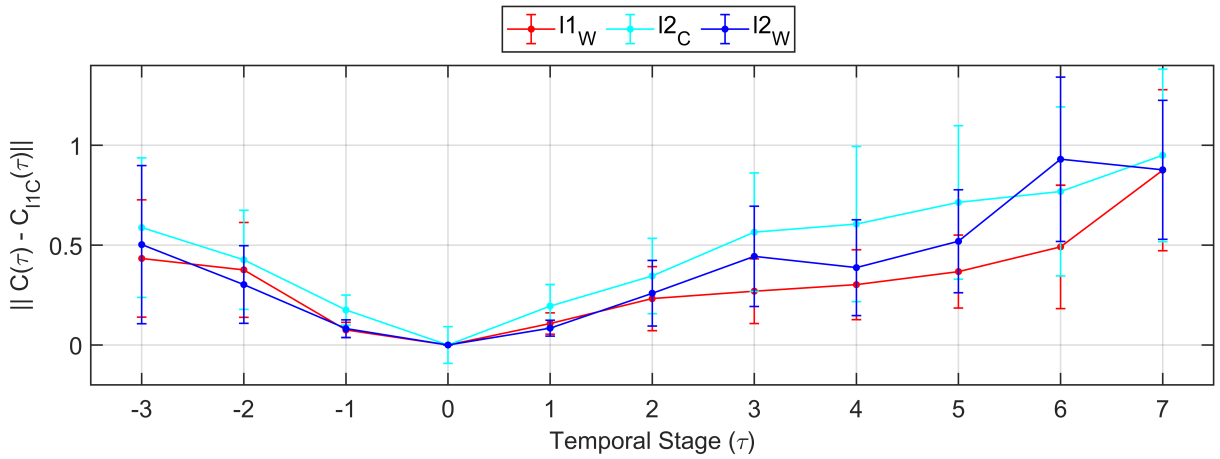


Figure 8.6: As in figure 8.2 but w.r.t. the centres obtained via $l1-$regularization "on centres", for the remaining optimization formulations.

Dionysus toolbox and ASAP, have been tracked through time using their unique identifiers assigned by the simulation. The bottom row shows a zoom in on the bubble's particles at the corresponding stage with the centres of the respective models in black. Additionally, two centres are highlighted, lying on opposite lobes of the cavity. The red one is closest to the galaxy while the blue one points outwards.

The model's centres have been propagated through adjacent temporal snapshots using the four cost functions defined in eqs. (8.13)-(8.16). Figure 8.6 shows the mean and standard deviation, at each temporal stage (τ), of the distances from all centres recovered by eqs. (8.15) (red), (8.14) (cyan) and (8.13) (blue), to the corresponding ones obtained via eq. (8.16). Except for a general agreement at temporal stage $\tau = 0$, when the models are initialized, the centres obtained via the three remaining cost functions increasingly diverge with time (both forward and backward). For each selected centre, the change in the properties of the particles through time nearby the two selected centres can now be studied, for all adopted cost functions. In particular, density ρ, temperature T and Neutral Fraction (NF) are considered. The Neutral Fraction is the ratio of the neutral Hydrogen mass to the total particle mass. For each temporal stage (and each propagated model), the weighted mean of

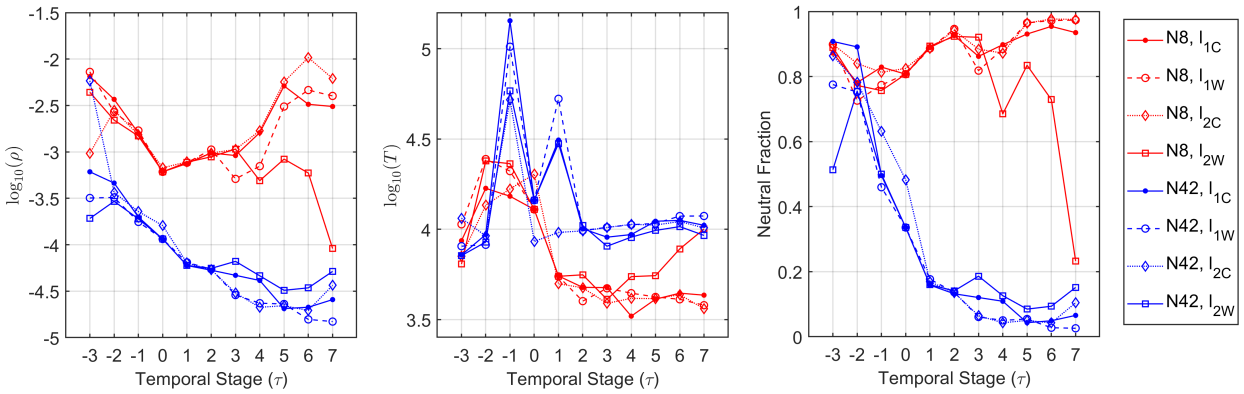


Figure 8.7: Weighted mean of simulated quantities ($\log_{10}(\rho)$, T and Neutral Fraction (NF)), on centres $N8$ and $N42$ of the models propagated through time using costs (8.13)-(8.16).

quantity a at centre c_m is computed as:

$$\langle a \rangle|_{\mathbf{c}_m} = \frac{\sum_{n=1}^{N_\tau} a_n R_m^n}{\sum_{n=1}^{N_\tau} R_m^n} \quad , \tag{8.17}$$

where R_m^n is the *responsibility* of the m-th Gaussian component (around $\mathbf{c}_m$) for the n-th particle and N_τ the number of particles belonging to the bubble at time τ. When applied to properties $a \in \{\rho, T, NF\}$, at each temporal stage for both centres, the profiles shown in fig. 8.7 are obtained. The results confirm the intuitive expectation of $l1-$regularization to be preferable to the $l2$ one. In addition, direct $l2-$regularization of the centers' position stabilizes the results to a greater degree than $l2-$regularization on $\mathbf{W}$. Centre 8 being closer to the galaxy, has a consistently higher ρ and NF, while retaining lower T w.r.t. to centre 42 that is further away. The cluster's gas is generally hotter and less dense than the gas of the galaxy. The hot gas connected with the blue centre being pushed out can escape the galaxy, keeping a low neutral fraction and high temperature (the UV background is shining on it, keeping it hot and ionized). Gas at the red centre is inside the galaxy's interstellar medium, sheltered from the external UV and can cool and condense to higher densities, which is, indeed, favorable for enhancing star formation.

Chapter Nine

Conclusions

A novel, semi-automated framework for denoising, dimensionality estimation, multi-manifold extraction and manifold aligned density estimation from complex data sets containing samples from noisy manifolds of diverse dimensionalities embedded in a noisy environment has been presented in this work. The framework uses only a few easy-to-understand hyper-parameters, such as characteristic scale, that can be manipulated to obtain an emerging picture of the multi-manifold structure of the data. The workings of the methodology are illustrated on different synthetic data sets, presenting peculiar challenges for the ours and the existing methodologies in the field. The first data set contains a number of manifolds, including a toroid and a Möbius strip, embedded in an extremely noisy environment. The second data set is a topologically challenging example of three interlocked noisy toroids. The third data set is aimed at reproducing properties of the astronomical cases treated in this work. For each synthetic data set manifolds having the same dimensionality have been isolated (via the proposed dimensionality index estimation), and a representation of each manifold as a pair of abstract and embedded graphs constructed via Multi-Manifold crawling. Furthermore the corresponding probabilistic model, describing its density, has also been built via Abstract GTM (AGTM). An efficient alternative to the Crawling algorithm has also been proposed. The alternative is inherently less computationally intensive and can

be considered a "batch" version of Crawling. While competitive with other manifold learning algorithms, the alternative to Crawling proved less accurate than its slower, but more stable, graph construction method. Methodologies for the computation of the curvature on graph's edges have also been proposed, that take advantage of the smooth mapping function formulation of AGTM.

A detailed comparison of the proposed methodology has been proposed, with both Multi-Manifold learning and Probabilistic modelling algorithms, widely used in the literature The Multi-Manifold learning alternatives proved sensitive to noise, or impractical when dealing with multiple manifolds of different dimensionalities. The probabilistic modelling techniques, although capable of representing globally the data sets, are not designed to explicitly capture general noisy low-dimensional structures along which the data can be organized. AGTM proved equally capable of modelling the global distribution of the data sets, while still carrying meaningful information about their underlying low-dimensional manifolds.

The methodology has then been applied to two complex data sets, each one being the output of numerically simulated astronomical objects. The first data set contained simulated gas volume particles from a numerical simulation of a dwarf galaxy interacting with its host galaxy cluster. The extraction of 1-D and 2-D manifolds has proven advantageous in understanding the nature of star formation inside such disrupted dwarf galaxies. This data set, and its temporal evolution, have been also analyzed in chapter 8. In particular, a first-order Markov dependency structure has been imposed by introducing penalization terms in the estimation of the log-likelihood of the model in nearby stages. Modelling cavities through a dedicated formulation of Geodesic GTM, has allowed for a description of their expansion and measurement of important simulated quantities. Their behaviour hinted to a Supernova explosion as their formation scenario. The second data set contained numerical simulated gas volume particles, simulating a single temporal stage of the cosmic volume. The Large Scale Structure (LSS) of the Universe as simulated under a specific cosmological theory,

presented intertwined filaments connecting in nodes and forming an overall web structure. By extracting two of these streams via the proposed complete methodology, their kinematic properties were studied.The nodes enclosing a filament, resulted being actively pulling particles towards themselves. This pull is reflected on the particles' tangential velocities along the stream, as their sign changes in proximity to its half length. The rotor of the velocities' perpendicular components has also been studied, noting how vortices can be formed within the central region of the streams. These vortices may be driving the kinematic properties of the host galaxies (in particular their spin).

The study over different astronomical cases has proven insightful in many aspects. In particular, the role of dynamical aspects of a galaxy's evolution (at scales of kpc) over time allows for a detailed study of its Star Formation. As the radiation coming from stars (scales vary within fractions of a pc) is mostly within the visible light wavebands and radiation at these wavelengths is observable, via either ground or space based observatories, with higher resolution (w.r.t. radio wave lengths for instance), it is essential to understand the processes regulating the formation of new stars. As discussed in sections 7.1.1, 7.2 and 8.5, simulated data hints at a high Star Formation Rate in the streams of Jellyfish galaxies, generated through the interaction with the gravitational field of the hosting galaxy cluster and its dynamic properties. A high Star Formation rate implies a high number of stars being born over a specific time interval. The evolution of an individual star is determined by its mass at its formation time. In first approximation, the larger the initial mass of a star is, the shorter its lifespan. In particular, stars having initial mass larger than $10M_{\odot}$ (solar masses) undergo a series of thermonuclear reactions in their core that produce increasingly heavier elements. When the central core (and the first to be formed) of Hydrogen is exhausted, the pressure support provided by its combustion is lost and the star is able to rapidly collapse onto its core. The rebound of the external combustion shells over the core generates a rapid expansion outward, which is usually referred to as a Supernova stage. The external shells

get dispersed into the ambient space of the star (interstellar medium), enriching it with heavy elements (up to Iron) and the eventual remnant of the star is either a Black Hole or a White Dwarf (when a remnant exists). While binary systems of stars may keep interacting by orbiting around their center of mass, and generating gravitational waves, the expanding outer shells, expelled into the interstellar medium swipe the surrounding gas. This gas, that has a low percentage of heavy elements, pushed by the Supernova explosion is able to get denser, while converting part of the heat generated by the Supernova into radiation (in the infrared waveband) and thus cooling down. The dense, cold and neutral Hydrogen gas, swept by the outer shells of the Supernova is then able to locally collapse, under the action of its self-gravity, generating new stars, whose radiation becomes observable. Given that the main natural force operating at the astronomical scales is gravitation, and that by nature gravitation is only attractive, the galaxies undergoing Star Formation (in fact, all galaxies) often reside in much larger regions of the Universe, populated by other galaxies. These groups or clusters (depending on the number of members and size) of galaxies (typical scales of hundreds of kpc), are permeated by hot gas (intergalactic medium, mostly composed of Hydrogen), which is able to cool down via mixing with the gas of the hosted galaxies. Thus, the intergalactic medium offers a large reservoir of gas, necessary for replenishing galaxies with material for the formation of new stars. The dynamical and gravitational interaction (sometimes referred to as a "merger") between galaxies in clusters, may form gaseous or stellar structures such as the galactic streams shown in this work, or the two-dimensional manifold described in section 6.1, providing favorable conditions to Star Formation. However, the dynamical properties of individual galaxies (or populations of them) are not only driven by their initial properties at their formation, but also from the characteristics of the larger scale structure of the Universe as predicted by numerical simulations; what is often referred to as the Cosmic Web (scales of Mpc). The Large Scale Structure (LSS) of the Universe is thought to be composed of filaments and walls: regions where dark matter is preferentially distributed and thus have a large gravitational pull over the surrounding visible matter (section 7.2.2).

These low-dimensional structures are delimited by "voids", where matter (both visible and Dark) is scarce. Filaments and walls, containing numerous galaxy clusters, connect in nodes, towards which matter is constantly accreted. From this stems the belief (driven by simulation evidence) that nodes of the Cosmic Web host superclusters of galaxies. The Cosmic Web results from the propagation through time of anisotropies in the dark matter distribution after the inflation period following the Big Bang. The existence of these initial anisotropies is supported by the map of the Cosmic Microwave Background (CMB) radiation: a radiation observed today in the microwave band, which permeates the entire Universe. The CMB radiation has been linked, via General Relativity and the consequent cosmological model of the Universe, to the high energy, temperature and density plasma forming entirely the Universe 4×10^5 yrs after the Big Bang.

Besides many other possible applications, the proposed methodology offers new exciting ways of analysing details emerging from astrophysical simulations that are not possible with the current tools commonly used in astronomy. The semi-supervised nature of this methodology entails its large applicability to many different cases. Ongoing work is dedicated to the extraction of structures in observations. Low-dimensional structures can be identified both in the parameter space and the positional space of astronomical observations. For instance, gravitational waves are generated by the gravitational interaction of two massive systems. The detected gravitational wave can thus be characterized by the two masses of the binary system, but also by their spins (the spin is here referred to the rotation on an axis of individual objects). Since the instrumentation dedicated to gravitational waves detection is inherently subject to noise, the masses and spins associated to a detection are not point-wise, but are distributed according to the instrumental noise and the uncertainty on the binary object's model fitting. Hence, each gravitational event is identified by a posterior distribution over the four-dimensional parameter space (two masses and two spins). This can have a complicated shape, making the decision process about the identity of the two objects

forming the binary system non trivial. However, modelling each posterior (observation) via AGTM, it is possible to combine all the models and obtain the probability mass distribution over the parameter space for the collection of observations. This is the process behind Probabilistic Hough Transform as described in Tino et al., 2010. The mass distribution over the parameter space can then be used to estimate peaks of probability. The location of these peaks in the parameter space correspond to prototypical binary systems (e.g. Black Hole - Black Hole, Neutron Star - Black Hole, Neutron Star - Neutron Star). Having the locations of the prototypical systems, individual observations can be assigned to either of them based on their posterior probability.

While gravitational waves are generated by highly massive binary systems and they are a recent type of observations, interesting results may be achieved via the analysis of fairly known objects: stars, and in particular their grouping via mutual gravitational interaction, globular clusters. These are clusters of stars, tied by mutual gravitational interaction, that due to their small size when compared to galaxies can only be efficiently observed in the vicinity of the Milky Way (the galaxy hosting the Solar System), but are supposed to populate every disc galaxy. They are believed to be remnants of dwarf galaxies that, affected by tidal and gravitational interactions with a more massive galaxy, were depleted of their gas content while still preserving their star distribution. In this sense, the stars within a globular cluster have a specific distribution in terms of their magnitudes (measured luminosities in different electromagnetic bands). Furthermore, since observations of globular clusters are carried out by looking at a square region in the sky, numerous other sources are detected that pollute the globular cluster's distribution over the luminosity space (roughly uniformly). However, by restricting the selection of stars within a small circle in the observed region, centered around the cluster's core, it is possible to obtain a refined selection of stars characterizing the prototypical luminosity distribution of the cluster. Obtaining a model of both the cluster (through AGTM) and a selection of the stellar pollutant component (in regions further away

from the cluster's core but within the observed square) allows to refine the selection of stars belonging to the cluster (e.g. via a simple Bayesian classifier). Selecting stars throughout the whole observational luminosity space, according to the results provided by the models, and bringing them onto the positional space (observed region of sky), highlights the true positional distribution of cluster's stars. In particular, the positional distribution of stars belonging to the cluster (in terms of their luminosity properties), highlights streams that were previously hidden by the pollutant component. The streams can further be modelled via AGTM and provide insights on their thickness (distribution along the local perpendicular component) or significance of detection (signal to noise ratio). The same methodologies presented in sections 7 can be applied to the filaments with little to none modifications.

In conclusion, the Universe has a high degree of self-similiarity, where structures of similar shapes and topologies may be seen at different length scales, ranging from fractions of a pc (globular clusters) to hundreds of Mpcs (Cosmic Web filaments). Having a consistent methodology that is able to extract low-dimensional, noisily sampled manifolds while capturing their transverse noise distribution, is extremely important in understanding the formation and evolution through time of these low-dimensional structures. Enabling this kind of study might reveal hidden information and aid scientific inference at multiple levels, ranging from the characterization of Star Formation in galaxies, up to the description of the evolution of the LSS of the Universe. Furthermore, the proposed methodology might open the possibility to multiple different kinds of structures, even higher dimensional, that might have been overseen so far due to unavailable detection and modelling schemes. A significant amount of work is still due in order to improve the available knowledge of the Universe, but this methodology might represent a stepping stone from which to gaze onto new horizons.

References

Aarseth, S.J. (1963). "Dynamical evolution of clusters of galaxies, I". In: *MNRAS* 126, p. 223. DOI: 10.1093/mnras/126.3.223.

Abbott, B. P. et al. (Feb. 2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Physical Review Letters* 116.6, 061102, p. 061102. DOI: 10.1103/PhysRevLett.116.061102. arXiv: 1602.03837 [gr-qc].

Alahakoon, D. et al. (2000). "Dynamic self-organizing maps with controlled growth for knowledge discovery". In: *IEEE Transactions on Neural Networks* 11.3, pp. 601–614. DOI: 10.1109/72.846732.

Alahmadi, Hanin H. et al. (2016). "Classifying Cognitive Profiles Using Machine Learning with Privileged Information in Mild Cognitive Impairment". In: *Frontiers in Computational Neuroscience* 10, p. 117. ISSN: 1662-5188. DOI: 10.3389/fncom.2016.00117. URL: https://www.frontiersin.org/article/10.3389/fncom.2016.00117.

Allegra, Michele et al. (Oct. 2020). "Data segmentation based on the local intrinsic dimension". In: *Scientific Reports* 10, 16449, p. 16449. DOI: 10.1038/s41598-020-72222-0. arXiv: 1902.10459 [stat.ML].

Bacciu, D. et al. (2013). "Compositional Generative Mapping for Tree-Structured Data—Part II: Topographic Projection Model". In: *IEEE Transactions on Neural Networks and Learning Systems* 24.2, pp. 231–247.

Bacon, R. et al. (Nov. 2017). "The MUSE Hubble Ultra Deep Field Survey. I. Survey description, data reduction, and source detection". In: *A&A* 608, A1, A1. DOI: 10.1051/0004-6361/201730833. arXiv: 1710.03002.

Becher, Erich (1905). "The Philosophical Views of Ernst Mach". In: *The Philosophical Review* 14.5, pp. 535–562. ISSN: 00318108, 15581470. URL: http://www.jstor.org/stable/2177489.

Belkin, Mikhail et al. (2001). "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering". In: *Advances in Neural Information Processing Systems 14*. MIT Press, pp. 585–591.

Ben-Israel, Adi et al. (2003). *Generalized inverses: theory and applications*. Vol. 15. Springer Science & Business Media.

Bishop, C. M. et al. (Mar. 1998). "A hierarchical latent variable model for data visualization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.3, pp. 281–293. ISSN: 0162-8828. DOI: 10.1109/34.667885.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Bishop, Christopher M. et al. (1998a). "Developments of the generative topographic mapping". In: *Neurocomputing* 21.1, pp. 203–224. ISSN: 0925-2312. DOI: 10.1016/S0925-2312(98)00043-5.

— (1998b). "GTM: The Generative Topographic Mapping". In: *Neural Computation* 10, pp. 215–234.

Boissonnat, Jean-Daniel et al. (2018). *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press. URL: https://hal.inria.fr/hal-01615863.

Bond, J. Richard et al. (Apr. 1996). "How filaments of galaxies are woven into the cosmic web". In: *Nature* 380.6575, pp. 603–606. DOI: 10.1038/380603a0. arXiv: astro-ph/9512141 [astro-ph].

Brinchmann, J. et al. (Nov. 2017). "The MUSE Hubble Ultra Deep Field Survey. III. Testing photometric redshifts to 30th magnitude". In: *A&A* 608, A3, A3. DOI: 10.1051/0004-6361/201731351. arXiv: 1710.05062.

Bryan, G. L. et al. (Apr. 2014). "ENZO: An Adaptive Mesh Refinement Code for Astrophysics". In: *ApJs* 211, 19, p. 19. DOI: 10.1088/0067-0049/211/2/19. arXiv: 1307.2265 [astro-ph.IM].

C.M. Bishop G.E. Hinton, I.G.D. Strachan (Jan. 1997). "GTM through time". English. In: *IET Conference Proceedings*, 111–116(5).

Camastra, Francesco et al. (2016). "Intrinsic dimension estimation: Advances and open problems". In: *Information Sciences* 328, pp. 26–41. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2015.08.029. URL: http://www.sciencedirect.com/science/article/pii/S0020025515006179.

Canducci, M. et al. (2021). "Tracking the Temporal-Evolution of Supernova Bubbles in Numerical Simulations". In: ed. by Hujun Yin et al., pp. 493–501.

Canducci, M. et al. (2022). "Probabilistic modelling of general noisy multi-manifold data sets". In: *Artificial Intelligence* 302, p. 103579. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103579. URL: https://www.sciencedirect.com/science/article/pii/S0004370221001302.

Chazal, Frédéric et al. (Nov. 2013). "Persistence-Based Clustering in Riemannian Manifolds". In: *J. ACM* 60.6, 41:1–41:38. ISSN: 0004-5411. DOI: 10.1145/2535927. URL: http://doi.acm.org/10.1145/2535927.

Ching, Travers et al. (2018). "Opportunities and obstacles for deep learning in biology and medicine". In: *Journal of The Royal Society Interface* 15.141. ISSN: 1742-5689. DOI: 10.1098/rsif.2017.0387. URL: http://rsif.royalsocietypublishing.org/content/15/141/20170387.

Chung, F. R. K. (1997). *Spectral Graph Theory*. American Mathematical Society.

Comaniciu, D. et al. (May 2002). "Mean shift: a robust approach toward feature space analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5, pp. 603–619. ISSN: 0162-8828. DOI: 10.1109/34.1000236.

Comaniciu, Dorin et al. (1999). "Mean Shift Analysis and Applications". In: *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*. ICCV '99. Washington, DC, USA: IEEE Computer Society, pp. 1197–. ISBN: 0-7695-0164-8. URL: http://dl.acm.org/citation.cfm?id=850924.851593.

Cossins, Peter J. (July 2010). "Smoothed Particle Hydrodynamics". PhD thesis. -.

Cottrell, Marie et al. (2006). "Batch and median neural gas". In: *Neural Networks* 19.6. Advances in Self Organising Maps - WSOM'05, pp. 762–771. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2006.05.018. URL: https://www.sciencedirect.com/science/article/pii/S0893608006000815.

Couchman, H. M. P. et al. (Oct. 1995). "Hydra: an Adaptive-Mesh Implementation of P 3M-SPH". In: *ApJ* 452, p. 797. DOI: 10.1086/176348. eprint: astro-ph/9409058.

Cox, Michael A. A. et al. (2008). "Multidimensional Scaling". In: *Handbook of Data Visualization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 315–347. ISBN: 978-3-540-33037-0. DOI: 10.1007/978-3-540-33037-0_14. URL: https://doi.org/10.1007/978-3-540-33037-0_14.

De Looze, Ilse et al. (Oct. 2011). "The reliability of [C II] as an indicator of the star formation rate". In: *MNRAS* 416.4, pp. 2712–2724. DOI: 10.1111/j.1365-2966.2011.19223.x. arXiv: 1106.1643 [astro-ph.CO].

De Rijcke, S. et al. (Aug. 2013). "New composition-dependent cooling and heating curves for galaxy evolution simulations". In: *MNRAS* 433.4, pp. 3005–3016. ISSN: 00358711. DOI: 10.1093/mnras/stt942. eprint: 1306.4860 (astro-ph.CO). URL: http://academic.oup.com/mnras/article/433/4/3005/1749217/New-compositiondependent-cooling-and-heating.

Delingette (1994). "Simplex meshes: a general representation for 3D shape reconstruction". In: *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 856–859. DOI: 10.1109/CVPR.1994.323913.

Delingette, Hervé (Feb. 1997). *General Object Reconstruction based on Simplex Meshes*. Tech. rep. RR-3111. INRIA. URL: https://hal.inria.fr/inria-00073579.

Dempster, A. P. et al. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.

Deutsch, Shay et al. (2015). "Intersecting Manifolds: Detection, Segmentation, and Labeling". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, pp. 3445–3452. ISBN: 978-1-57735-738-4. URL: http://dl.acm.org/citation.cfm?id=2832581.2832729.

Dial, Robert B. (Nov. 1969). "Algorithm 360: Shortest-path Forest with Topological Ordering [H]". In: *Commun. ACM* 12.11, pp. 632–633. ISSN: 0001-0782. DOI: 10.1145/363269.363610. URL: http://doi.acm.org/10.1145/363269.363610.

Dijkstra, E. W. (Dec. 1959). "A note on two problems in connexion with graphs". In: *Numerische Mathematik* 1.1, pp. 269–271. ISSN: 0945-3245. DOI: 10.1007/BF01386390. URL: https://doi.org/10.1007/BF01386390.

Ding, Chris et al. (2017). "L1-norm Error Function Robustness and Outlier Regularization". In: *CoRR* abs/1705.09954.

Donoho, David L. et al. (2003). "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". In: *Proceedings of the National Academy of Sciences* 100.10, pp. 5591–5596. ISSN: 0027-8424. DOI: 10.1073/pnas.1031596100. eprint: http://www.pnas.org/content/100/10/5591.full.pdf. URL: http://www.pnas.org/content/100/10/5591.

Dunn, J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3, pp. 32–57. DOI:

10.1080/01969727308546046. eprint: https://doi.org/10.1080/01969727308546046. URL: https://doi.org/10.1080/01969727308546046.

Ebeling, H. et al. (Jan. 2014). "JELLYFISH: EVIDENCE OF EXTREME RAM-PRESSURE STRIPPING IN MASSIVE GALAXY CLUSTERS". In: *The Astrophysical Journal* 781.2, p. L40. ISSN: 2041-8213. DOI: 10.1088/2041-8205/781/2/l40. URL: http://dx.doi.org/10.1088/2041-8205/781/2/L40.

Edelsbrunner, H. et al. (Nov. 2000). "Topological persistence and simplification". In: *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pp. 454–463. DOI: 10.1109/SFCS.2000.892133.

Einstein, A. (Jan. 1916). "Die Grundlage der allgemeinen Relativitätstheorie". In: *Annalen der Physik* 354.7, pp. 769–822. DOI: 10.1002/andp.19163540702.

Einstein, Albert (Jan. 1908). "Über das Relativitätsprinzip und die aus demselben gezogenen Folgerungen". In: *Jahrbuch der Radioaktivität und Elektronik* 4, pp. 411–462.

Elhamifar, Ehsan et al. (2011a). "Sparse Manifold Clustering and Embedding". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., pp. 55–63. URL: http://papers.nips.cc/paper/4246-sparse-manifold-clustering-and-embedding.pdf.

— (2011b). "Sparse Manifold Clustering and Embedding". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., pp. 55–63. URL: https://proceedings.neurips.cc/paper/2011/file/fc490ca45c00b1249bbe3554a4fdf6fb-Paper.pdf.

Facco, E. et al. (2017). "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *Scientific Reports* 7.

Fan, Mingyu et al. (2012). "Isometric Multi-manifold Learning for Feature Extraction". In: *2012 IEEE 12th International Conference on Data Mining*, pp. 241–250.

Fan, Mingyu et al. (June 2016). "Efficient Isometric Multi-manifold Learning Based on the Self-organizing Method". In: *Inf. Sci.* 345.C, pp. 325–339. ISSN: 0020-0255. DOI: 10.1016/j.ins.2016.01.069. URL: https://doi.org/10.1016/j.ins.2016.01.069.

Ferguson, Thomas S. (1973). "A Bayesian Analysis of Some Nonparametric Problems". In: *The Annals of Statistics* 1.2, pp. 209–230. DOI: 10.1214/aos/1176342360. URL: https://doi.org/10.1214/aos/1176342360.

Figueiredo, M. A. T. et al. (2002). "Unsupervised learning of finite mixture models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.3, pp. 381–396. DOI: 10.1109/34.990138.

Fredman, M. L. et al. (Oct. 1984). "Fibonacci Heaps And Their Uses In Improved Network Optimization Algorithms". In: *25th Annual Symposium onFoundations of Computer Science, 1984.* Pp. 338–346. DOI: 10.1109/SFCS.1984.715934.

Gagniuc, Paul A. (May 2017). *Markov Chains: From Theory to Implementation and Experimentation.* WILEY. ISBN: 978-1-119-38755-8. DOI: 10.1002/9781119387596.

Garcia, Vincent et al. (Jan. 2010). "Hierarchical Gaussian Mixture Model." In: pp. 4070–4073. DOI: 10.1109/ICASSP.2010.5495750.

Gauss, Carl Friedrich (1981). "Disquisitiones generales circa superficies curvas". In.

Gianniotis, N. et al. (Aug. 2008). "Visualization of Tree-Structured Data Through Generative Topographic Mapping". In: *IEEE Transactions on Neural Networks* 19.8, pp. 1468–1493. ISSN: 1045-9227. DOI: 10.1109/TNN.2008.2001000.

Gingold, R.A. et al. (Nov. 1977). "Smoothed particle hydrodynamics - Theory and application to non-spherical stars". In: *MNRAS* 181, pp. 375–389. DOI: 10.1093/mnras/181.3.375.

Goldberger, Jacob et al. (2005). "Hierarchical Clustering of a Mixture Model". In: *Advances in Neural Information Processing Systems.* Ed. by L. Saul et al. Vol. 17. MIT Press. URL: https://proceedings.neurips.cc/paper/2004/file/36e729ec173b94133d8fa552e4029f8b-Paper.pdf.

Grassberger, Peter et al. (Oct. 1983). "Measuring the strangeness of strange attractors". In: *Physica D Nonlinear Phenomena* 9.1-2, pp. 189–208. DOI: 10.1016/0167-2789(83)90298-1.

Guérou, A. et al. (Nov. 2017). "The MUSE Hubble Ultra Deep Field Survey. V. Spatially resolved stellar kinematics of galaxies at redshift $0.2 < z > 0.8$". In: *A&A* 608, A5, A5. DOI: 10.1051/0004-6361/201730905. arXiv: 1710.07694.

Guhathakurta, P. et al. (May 2006). "Dynamics and Stellar Content of the Giant Southern Stream in M31. I. Keck Spectroscopy of Red Giant Stars". In: *The Astronomical Journal* 131, pp. 2497–2513. DOI: 10.1086/499562. eprint: astro-ph/0406145.

Gunawan, Hendra et al. (Jan. 2005). "A Formula for Angles between Subspaces of Inner Product Spaces". In: *Beiträge zur Algebra und Geometrie* 46.

Hamilton, William Rowan (1866). *Elements of quaternions*. Longmans, Green, & Company.

Haro, Gloria et al. (2000). "Translated Poisson mixture model for stratification learning". In: *Int. J. Comput. Vision*.

Hastie, Trevor et al. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Herrera-Camus, R. et al. (Feb. 2015). "[C II] 158 μ m EMISSION AS A STAR FORMATION TRACER". In: *The Astrophysical Journal* 800.1, p. 1. ISSN: 1538-4357. DOI: 10.1088/0004-637x/800/1/1. URL: http://dx.doi.org/10.1088/0004-637X/800/1/1.

Hettiarachchi, R. et al. (Sept. 2015). "Multi-manifold LLE Learning in Pattern Recognition". In: *Pattern Recogn.* 48.9, pp. 2947–2960. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2015.04.003. URL: http://dx.doi.org/10.1016/j.patcog.2015.04.003.

Hinton, Geoffrey et al. (2003). "Stochastic Neighbor Embedding". In: *Advances in neural information processing systems* 15. Ed. by S Thrun S Becker et al., pp. 833–840. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.7959&rep=rep1&type=pdf.

Hoerner S., von (1960). "Die numerische Integration des n-Körper-Problemes für Sternhaufen. I". In: *Z. Astrophys.* 50.

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables with Principal Components". In: *Journal of Educational Psychology* 24, pp. 417–441.

Hunter, Deidre A. et al. (Oct. 2012). "LITTLE THINGS". In: *The Astronomical Journal* 144.5, p. 134. ISSN: 1538-3881. DOI: 10.1088/0004-6256/144/5/134. URL: http://dx.doi.org/10.1088/0004-6256/144/5/134.

Huo, Xiaoming et al. (2008). "A Survey of Manifold-Based Learning Methods". In: *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*, pp. 691–745. DOI: 10.1142/9789812779861_0015. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789812779861_0015. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812779861_0015.

Ibata, R. et al. (July 2001). "A giant stream of metal-rich stars in the halo of the galaxy M31". In: *Nature* 412, pp. 49–52. eprint: astro-ph/0107090.

Iodice, E. et al. (2019). "The Fornax Deep Survey with the VST - V. Exploring the faintest regions of the bright early-type galaxies inside the virial radius". In: *A&A* 623, A1. DOI: 10.1051/0004-6361/201833741. URL: https://doi.org/10.1051/0004-6361/201833741.

Israel, F. P. (1998). "Centaurus A - NGC 5128". In: *A&A Rev.* 8, pp. 237–278. DOI: 10.1007/s001590050011. eprint: astro-ph/9811051.

Izenman, Alan Julian (Sept. 2012). "Introduction to Manifold Learning". In: *WIREs Comput. Stat.* 4.5, pp. 439–446. ISSN: 1939-5108. DOI: 10.1002/wics.1222. URL: https://doi.org/10.1002/wics.1222.

Jain, Anil K. (2010). "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR), pp. 651–666. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2009.09.011. URL: http://www.sciencedirect.com/science/article/pii/S0167865509002323.

Kaufman, Leonard et al. (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Kingma, Diederik P. et al. (2014). "Auto-Encoding Variational Bayes". In: *CoRR* abs/1312.6114.

Kohonen, Teuvo (Jan. 1982). "Kohonen, T.: Self-Organized Formation of Topologically Correct Feature Maps." In: *Biological Cybernetics 43(1), 59-69* 43, pp. 59–69.

Kriegel, Hans-Peter et al. (Mar. 2009). "Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering". In: *ACM Trans. Knowl. Discov. Data* 3.1, 1:1–1:58. ISSN: 1556-4681. DOI: 10.1145/1497577.1497578. URL: http://doi.acm.org/10.1145/1497577.1497578.

Kruskal, J. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1, pp. 1–27. URL: https://EconPapers.repec.org/RePEc:spr:psycho:v:29:y:1964:i:1:p:1-27.

Kruskal, J. B. (June 1964). "Nonmetric multidimensional scaling: A numerical method". In: *Psychometrika* 29.2, pp. 115–129. ISSN: 1860-0980. DOI: 10.1007/BF02289694. URL: https://doi.org/10.1007/BF02289694.

Kullback, S. et al. (1951a). "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1, pp. 79–86.

— (1951b). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

Lagarias, Jeffrey C. et al. (1998). "Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions". In: *SIAM Journal on Optimization* 9.1, pp. 112–147. DOI: 10.1137/S1052623496303470.

Laigle, C. et al. (Nov. 2014). " Swirling around filaments: are large-scale structure vortices spinning up dark haloes?" In: *Monthly Notices of the Royal Astronomical Society* 446.3, pp. 2744–2759. ISSN: 0035-8711. DOI: 10.1093/mnras/stu2289. eprint: https:

//academic.oup.com/mnras/article-pdf/446/3/2744/10448057/stu2289.pdf. URL: https://doi.org/10.1093/mnras/stu2289.

Laparra, Valero et al. (Jan. 2016). "Principal Polynomial Analysis". In: *arXiv e-prints*, arXiv:1602.00221, arXiv:1602.00221. arXiv: 1602.00221 [stat.ML].

Larsen, Anders Boesen Lindbo et al. (2015). "Autoencoding beyond pixels using a learned similarity metric". In: *CoRR* abs/1512.09300. arXiv: 1512.09300. URL: http://arxiv.org/abs/1512.09300.

Larson, R. et al. (2003). *Elementary Linear Algebra*. Houghton Mifflin. ISBN: 9780618400508. URL: https://books.google.co.uk/books?id=zeXpAgAACAAJ.

Lebanon, Guy (2005). "Riemannian Geometry and Statistical Machine Learning". AAI3159986. PhD thesis. Pittsburgh, PA, USA: Carnegie Mellon University. ISBN: 0-496-93472-4.

Lee, Chan-Su et al. (Feb. 2016). "Learning Representations from Multiple Manifolds". In: *Pattern Recogn.* 50.C, pp. 74–87. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2015.08.024. URL: http://dx.doi.org/10.1016/j.patcog.2015.08.024.

Lee-Waddell, K. et al. (Oct. 2017). "Tidal origin of NGC 1427A in the Fornax cluster". In: *Monthly Notices of the Royal Astronomical Society* 474.1, pp. 1108–1115. ISSN: 0035-8711. DOI: 10.1093/mnras/stx2808.

Li, Jichuan et al. (2018). "Gaussian mixture learning via adaptive hierarchical clustering". In: *Signal Process.* 150, pp. 116–121.

Lin, Chen-Hsuan et al. (June 2019). "Photometric Mesh Optimization for Video-Aligned 3D Object Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, Tong et al. (May 2008). "Riemannian Manifold Learning". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 30.5, pp. 796–809. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.70735. URL: https://doi.org/10.1109/TPAMI.2007.70735.

Lin, Tony et al. (2006). "Riemannian Manifold Learning for Nonlinear Dimensionality Reduction". In: *Computer Vision – ECCV 2006*. Ed. by Ales Leonardis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 44–55. ISBN: 978-3-540-33833-8.

Luxburg, Ulrike von (Dec. 2007). "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4, pp. 395–416. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9033-z. URL: https://doi.org/10.1007/s11222-007-9033-z.

Mahapatra, S. et al. (Oct. 2017). *S-Isomap++: Multi Manifold Learning from Streaming Data*. arXiv: 1710.06462 [stat.ML].

Maio, U. et al. (Aug. 2007). "Metal and molecule cooling in simulations of structure formation". In: *MNRAS* 379.3, pp. 963–973. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2007.12016.x. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2007.12016.x.

Mandelbrot, Benoit (1967). "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension". In: *Science* 156.3775, pp. 636–638. DOI: 10.1126/science.156.3775.636. eprint: https://www.science.org/doi/pdf/10.1126/science.156.3775.636. URL: https://www.science.org/doi/abs/10.1126/science.156.3775.636.

Martinetz, T.M. et al. (1993). "'Neural-gas' network for vector quantization and its application to time-series prediction". In: *IEEE Transactions on Neural Networks* 4.4, pp. 558–569. DOI: 10.1109/72.238311.

Masci, Jonathan et al. (2011). "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction". In: *Artificial Neural Networks and Machine Learning – ICANN 2011*. Ed. by Timo Honkela et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 52–59. ISBN: 978-3-642-21735-7.

Mastropietro, M et al. (Apr. 2021). "A tale of two tails: insights from simulations into the formation of the peculiar dwarf galaxy NGC 1427A". In: *MNRAS* 504.3, pp. 3387–3398. ISSN: 0035-8711. DOI: 10.1093/mnras/stab1091. eprint: 2104.07671. URL: https://academic.oup.com/mnras/advance-article/doi/10.1093/mnras/stab1091/6237530.

Maxwell, J.C. (1873). *A Treatise on Electricity and Magnetism*. Clarendon Press series. Clarendon Press. URL: https://books.google.co.uk/books?id=OcS4BEiMNwoC.

McLachlan, Geoffrey J et al. (1988). *Mixture models: Inference and applications to clustering*. Vol. 38. M. Dekker New York.

McLachlan, Geoffrey J et al. (2019). "Finite mixture models". In: *Annual review of statistics and its application* 6, pp. 355–378.

Möbius, A. F. (1827). "Der barycentrische Calcul: ein neues Hülfsmittel zur analytischen Behandlung der Geometrie". In: Barth.

Mohammadi, Mohammad et al. (2020). "Multi-agent Based Manifold Denoising". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2020*. Ed. by Cesar Analide et al. Cham: Springer International Publishing, pp. 12–24. ISBN: 978-3-030-62365-4.

Monti, Stefano et al. (July 2003). "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". In: 52, pp. 91–118.

Mora, Marcelo D. et al. (Sept. 2015). "A Starburst in the Core of a Galaxy Cluster: the Dwarf Irregular NGC 1427A in Fornax". In: *The Astronomical Journal* 150.3, 93, p. 93. DOI: 10.1088/0004-6256/150/3/93. arXiv: 1411.7314 [astro-ph.GA].

Mordohai, Philippos et al. (2005). "Unsupervised Dimensionality Estimation and Manifold Learning in High-dimensional Spaces by Tensor Voting". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., pp. 798–803. URL: http://dl.acm.org/citation.cfm?id=1642293.1642421.

Morozov, Dmitriy (2007). *Dionysus, a C++ library for computing persistent homology*. en. URL: https://mrzv.org/software/dionysus/ (visited on 2007).

Nabney, Ian T. (2002). *NETLAB. Algorithms for Pattern Recognition*. Advances in Pattern Recognition. London et. al.: Springer, p. 420. ISBN: 1852334401.

Navarro, Julio F. et al. (May 1996). "The Structure of Cold Dark Matter Halos". In: *ApJ* 462, p. 563. DOI: 10.1086/177173. arXiv: astro-ph/9508025 [astro-ph].

Nguyen, Nam et al. (2007). "Consensus Clusterings". In: *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. ICDM '07. Washington, DC, USA: IEEE Computer Society, pp. 607–612. ISBN: 0-7695-3018-4. DOI: 10.1109/ICDM.2007.73. URL: https://doi.org/10.1109/ICDM.2007.73.

Nichols, Matthew et al. (Sept. 2015a). "The post-infall evolution of a satellite galaxy". In: *Astronomy & Astrophysics* 582, A23. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201526113. URL: http://dx.doi.org/10.1051/0004-6361/201526113.

— (Oct. 2015b). "The post-infall evolution of a satellite galaxy". In: *A&A* 582, A23, A23. DOI: 10.1051/0004-6361/201526113. arXiv: 1503.05190 [astro-ph.GA].

Olech, Łukasz P. et al. (2016). "Hierarchical Gaussian Mixture Model with Objects Attached to Terminal and Non-terminal Dendrogram Nodes". In: *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*. Ed. by Robert Burduk et al. Cham: Springer International Publishing, pp. 191–201. ISBN: 978-3-319-26227-7.

Olier, Iván et al. (2008). "Variational Bayesian Generative Topographic Mapping". In: *Journal of Mathematical Modelling and Algorithms* 7, pp. 371–387.

Paolillo, M. et al. (Feb. 2002). "DeepROSATHRI Observations of the NGC 1399/NGC 1404 Region: Morphology and Structure of the X-Ray Halo". In: *The Astrophysical Journal* 565.2, pp. 883–907. DOI: 10.1086/337919. URL: https://doi.org/10.1086/337919.

Parzen, Emanuel (1962). "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3, pp. 1065–1076.

Pearson, K. (1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11. PCA beginnings, pp. 559–572. DOI: 10.1080/14786440109462720. eprint:

https://doi.org/10.1080/14786440109462720. URL: https://doi.org/10.1080/14786440109462720.

Petersen, K. B. et al. (Nov. 2012). *The Matrix Cookbook*. Version 20121115. Technical University of Denmark. URL: http://localhost/pubdb/p.php?3274.

Pidopryhora, Y. et al. (Feb. 2007). "The Ophiuchus Superbubble: A Gigantic Eruption from the Inner Disk of the Milky Way". In: *ApJ* 656, pp. 928–942. DOI: 10.1086/510521. eprint: astro-ph/0610894.

Poggianti, Bianca M. et al. (July 2017). "GASP. I. Gas Stripping Phenomena in Galaxies with MUSE". In: *The Astrophysical Journal* 844.1, p. 48. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aa78ed. URL: http://dx.doi.org/10.3847/1538-4357/aa78ed.

Price, Daniel J. (2012). "Smoothed particle hydrodynamics and magnetohydrodynamics". In: *Journal of Computational Physics* 231.3. Special Issue: Computational Plasma Physics, pp. 759–794. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2010.12.011. URL: http://www.sciencedirect.com/science/article/pii/S0021999110006753.

Qiu, Teng et al. (2014). "A Physically Inspired Clustering Algorithm: to Evolve Like Particles". In: *CoRR* abs/1412.5902.

Qiu, Teng et al. (2015a). "A general framework for the IT-based clustering methods". In: *CoRR* abs/1506.06068. URL: http://arxiv.org/abs/1506.06068.

— (2015b). "IT-map: an Effective Nonlinear Dimensionality Reduction Method for Interactive Clustering". In: *CoRR* abs/1501.06450. URL: http://arxiv.org/abs/1501.06450.

Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626.

Rahmani, M. et al. (Dec. 2017). "Coherence Pursuit: Fast, Simple, and Robust Principal Component Analysis". In: *IEEE Transactions on Signal Processing* 65, pp. 6260–6275. DOI: 10.1109/TSP.2017.2749215. arXiv: 1609.04789 [cs.LG].

Rasmussen, Carl Edward (1999). "The Infinite Gaussian Mixture Model". In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*. NIPS'99. Denver, CO: MIT Press, pp. 554–560.

Rawlings, John O et al. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

Rodriguez, Alex et al. (2014). "Clustering by fast search and find of density peaks". In: *Science* 344.6191, pp. 1492–1496. ISSN: 0036-8075. DOI: 10.1126/science.1242072. eprint: http://science.sciencemag.org/content/344/6191/1492.full.pdf. URL: http://science.sciencemag.org/content/344/6191/1492.

Rokach, Lior et al. (2005). "Clustering Methods". In: *Data Mining and Knowledge Discovery Handbook*. Ed. by Oded Maimon et al. Boston, MA: Springer US, pp. 321–352. ISBN: 978-0-387-25465-4. DOI: 10.1007/0-387-25465-X_15. URL: https://doi.org/10.1007/0-387-25465-X_15.

Rosenblatt, Murray (1956). "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3, pp. 832–837. DOI: 10.1214/aoms/1177728190. URL: https://doi.org/10.1214/aoms/1177728190.

Roweis, Sam T. et al. (2000). "Nonlinear dimensionality reduction by locally linear embedding". In: *SCIENCE* 290, pp. 2323–2326.

Rumelhart, David E. et al. (1986). "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Ed. by David E. Rumelhart et al. Cambridge, MA: MIT Press, pp. 318–362.

Saranathan, A. M. et al. (2019). "On Clustering and Embedding Mixture Manifolds Using a Low Rank Neighborhood Approach". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.6, pp. 3890–3903. DOI: 10.1109/TGRS.2018.2888983.

Schölkopf, Bernhard et al. (1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10.5, pp. 1299–1319. DOI: 10.1162/089976698300017467.

Shepard, Roger N. (June 1962). "The analysis of proximities: Multidimensional scaling with an unknown distance function. I." In: *Psychometrika* 27.2, pp. 125–140. ISSN: 1860-0980. DOI: 10.1007/BF02289630. URL: https://doi.org/10.1007/BF02289630.

Silva, Vin de et al. (2002). "Global Versus Local Methods in Nonlinear Dimensionality Reduction". In: *Proceedings of the 15th International Conference on Neural Information Processing Systems*. NIPS'02. Cambridge, MA, USA: MIT Press, pp. 721–728. URL: http://dl.acm.org/citation.cfm?id=2968618.2968708.

Sivanandam, S. et al. (Dec. 2014). "Tracing Ram-pressure Stripping with Warm Molecular Hydrogen Emission". In: *ApJ* 796, 89, p. 89. DOI: 10.1088/0004-637X/796/2/89. arXiv: 1410.0688.

Springel, V. (Dec. 2005). "The cosmological simulation code GADGET-2". In: *MNRAS* 364, pp. 1105–1134. DOI: 10.1111/j.1365-2966.2005.09655.x. eprint: astro-ph/0505010.

Springel, Volker (Dec. 2005). "The cosmological simulation code gadget-2". In: *Monthly Notices of the Royal Astronomical Society* 364.4, pp. 1105–1134. ISSN: 1365-2966. DOI: 10.1111/j.1365-2966.2005.09655.x. URL: http://dx.doi.org/10.1111/j.1365-2966.2005.09655.x.

Stone, J. M. et al. (Sept. 2008). "Athena: A New Code for Astrophysical MHD". In: *ApJS* 178, 137-177, pp. 137–177. DOI: 10.1086/588755. arXiv: 0804.0402.

Student (1908). "The Probable Error of a Mean". In: *Biometrika* 6.1, pp. 1–25. ISSN: 00063444. URL: http://www.jstor.org/stable/2331554.

Taghribi, A. et al. (2021). "ASAP – A Sub-sampling Approach for Preserving Topological Structures Modeled with Geodesic Topographic Mapping". In: *Neurocomputing*. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.05.108.

Taghribi, Abolfazl et al. (2020). "LAAT: Locally Aligned Ant Technique for detecting manifolds of varying density". In: *ArXiv* abs/2009.08326.

Taghribi, Abolfazl et al. (2022). "ASAP – A sub-sampling approach for preserving topological structures modeled with geodesic topographic mapping". In: *Neurocomputing* 470,

pp. 376–388. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.05.108. URL: https://www.sciencedirect.com/science/article/pii/S0925231221011139.

Tenenbaum, Joshua B. et al. (2000). "A Global Geometric Framework for Nonlinear Dimensionality Reduction". In: *Science* 290.5500, pp. 2319–2323. ISSN: 0036-8075. DOI: 10.1126/science.290.5500.2319. eprint: http://science.sciencemag.org/content/290/5500/2319.full.pdf. URL: http://science.sciencemag.org/content/290/5500/2319.

Tino, P. et al. (May 2002). "Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5, pp. 639–656. ISSN: 0162-8828. DOI: 10.1109/34.1000238.

Tino, P. et al. (Nov. 2010). "Searching for Coexpressed Genes in Three-Color cDNA Microarray Data Using a Probabilistic Model-Based Hough Transform". In: *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 8, pp. 1093–107. DOI: 10.1109/TCBB.2010.120.

Tipping, Michael E. et al. (1999). "Probabilistic Principal Component Analysis". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 61.3, pp. 611–622.

Torgerson, Warren (Dec. 1952). "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4, pp. 401–419. DOI: 10.1007/BF02288916. URL: https://ideas.repec.org/a/spr/psycho/v17y1952i4p401-419.html.

Torgerson, Warren S. (1958). "Warren S. Torgerson, Theory and methods of scaling. New York: John Wiley and Sons, Inc., 1958. Pp. 460". In: *Behavioral Science* 4.3, pp. 245–247. DOI: 10.1002/bs.3830040308. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830040308. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830040308.

Tosi, Monica (Apr. 2003). "Understanding dwarf galaxies as galactic building blocks". In: *Astrophysics and Space Science* 284.2, pp. 651–660. DOI: 10.1023/A:1024038913272. arXiv: astro-ph/0210200 [astro-ph].

Tu, L.W. (2010). *An Introduction to Manifolds*. Universitext. Springer New York. ISBN: 9781441973993. URL: https://books.google.co.uk/books?id=br1KngEACAAJ.

Vahdat, Arash et al. (2020). *NVAE: A Deep Hierarchical Variational Autoencoder*. DOI: 10.48550/ARXIV.2007.03898. URL: https://arxiv.org/abs/2007.03898.

Van der Maaten, Laurens et al. (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

Vancura, Vojtech et al. (2021). "Deep Variational Autoencoder with Shallow Parallel Path for Top-N Recommendation (VASP)". In: *CoRR* abs/2102.05774. arXiv: 2102.05774. URL: https://arxiv.org/abs/2102.05774.

Vaswani, N. et al. (Nov. 2017). "Robust PCA, Subspace Learning, and Tracking". In: *ArXiv e-prints*.

Verbeke, R. et al. (Dec. 2015). "How the First Stars Shaped the Faintest Gas-dominated Dwarf Galaxies". In: *ApJ* 815, 85, p. 85. DOI: 10.1088/0004-637X/815/2/85. arXiv: 1511.01484.

Verbeke, R. et al. (Oct. 2017). "A new astrophysical solution to the Too Big To Fail problem. Insights from the moria simulations". In: *A&A* 607, A13, A13. DOI: 10.1051/0004-6361/201730758. arXiv: 1703.03810 [astro-ph.GA].

Vidal, Ren et al. (2016). *Generalized Principal Component Analysis*. 1st. Springer Publishing Company, Incorporated. ISBN: 0387878106.

Vincent, Luc (1989). "Graphs and mathematical morphology". In: *Signal Processing* 16.4. Special Issue on Advances in Mathematical Morphology, pp. 365–388. ISSN: 0165-1684. DOI: https://doi.org/10.1016/0165-1684(89)90031-5. URL: http://www.sciencedirect.com/science/article/pii/0165168489900315.

Vincent, Pascal et al. (2003). "Manifold Parzen Windows". In: *Advances in Neural Information Processing Systems 15*. Ed. by S. Becker et al. MIT Press, pp. 849–856. URL: http://papers.nips.cc/paper/2203-manifold-parzen-windows.pdf.

Wang, X. et al. (June 2009). "Fast parzen window density estimator". In: *2009 International Joint Conference on Neural Networks*, pp. 3267–3274. DOI: 10.1109/IJCNN.2009.5178637.

Wang, Xiaoxia et al. (2008). "Multiple Manifolds Learning Framework Based on Hierarchical Mixture Density Model". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Walter Daelemans et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 566–581. ISBN: 978-3-540-87481-2.

Wedin, P. (1983). "On angles between subspaces of a finite dimensional inner product space". In: *Lecture Notes in Mathematics* 973, pp. 263–285.

Wu, S. et al. (2018). "Structure-Aware Data Consolidation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.10, pp. 2529–2537. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2017.2754254.

Wu, Yiming et al. (2004). "An extended Isomap algorithm for learning multi-class manifold". In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)* 6, 3429–3433 vol.6.

Yang, Wankou et al. (Aug. 2011). "A Multi-manifold Discriminant Analysis Method for Image Feature Extraction". In: *Pattern Recogn.* 44.8, pp. 1649–1657. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.01.019. URL: http://dx.doi.org/10.1016/j.patcog.2011.01.019.

Yin, Hujun et al. (Nov. 1995). "On the Distribution and Convergence of Feature Space in Self-Organizing Maps". In: *Neural Computation* 7.6, pp. 1178–1187. ISSN: 0899-7667. DOI: 10.1162/neco.1995.7.6.1178. eprint: https://direct.mit.edu/neco/article-pdf/7/6/1178/813050/neco.1995.7.6.1178.pdf. URL: https://doi.org/10.1162/neco.1995.7.6.1178.

You, Chong et al. (2017). "Provable Self-Representation Based Outlier Detection in a Union of Subspaces". In: *CVPR*. IEEE Computer Society, pp. 4323–4332.

Zabel, Nikki et al. (Nov. 2018). "The ALMA Fornax Cluster Survey I: stirring and stripping of the molecular gas in cluster galaxies". In: *Monthly Notices of the Royal Astronomical Society* 483.2, pp. 2251–2268. ISSN: 1365-2966. DOI: 10.1093/mnras/sty3234. URL: http://dx.doi.org/10.1093/mnras/sty3234.

Zhang, Zhenyue et al. (2004). "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment". In: *SIAM Journal on Scientific Computing* 26.1, pp. 313–338. DOI: 10.1137/S1064827502419154. eprint: https://doi.org/10.1137/S1064827502419154. URL: https://doi.org/10.1137/S1064827502419154.

Zibulevsky, Michael et al. (2010). "L1-L2 Optimization in Signal and Image Processing". In: *IEEE Signal Processing Magazine* 27.3, pp. 76–88. DOI: 10.1109/MSP.2010.936023.

Zollhöfer, M. et al. (2018). "State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications". In: *Computer Graphics Forum* 37.2, pp. 523–550. DOI: https://doi.org/10.1111/cgf.13382. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13382.