Kernel Smoothing and Mean Shift Theories with Applications to Cosmic Web Detection

Yikun Zhang*

Joint work with *Yen-Chi Chen** and *Rafael S. de Souza*†

* *Department of Statistics, University of Washington*
† *School of Physics, Engineering and Computer Science, University of Hertfordshire*

5th SIAM PNWS Biennial Meeting
October 4, 2025

Department of STATISTICS

UNIVERSITY of WASHINGTON

What Is Cosmic Web?

Cosmic Web is a large-scale network structure revealing that the matter in our Universe is not uniformly distributed (Bond et al., 1996).

Cosmic Web is a large-scale network structure revealing that the matter in our Universe is not uniformly distributed (Bond et al., 1996).

- **Large scale**: 1 Mpc $\approx$ 3.26 light-years.
- **Cause**: the anisotropic collapse of matter in gravitational instability scenarios at the early stage of the Universe (Zel'Dovich, 1970).

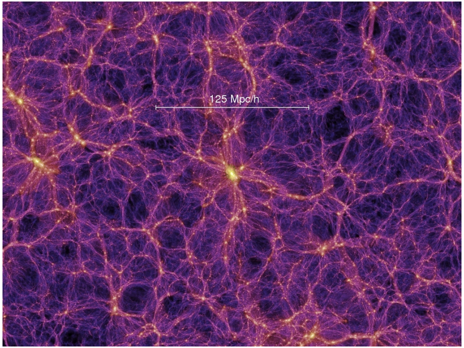


Figure: Visualization of *Cosmic Web* (credited to the millennium simulation project (Springel et al., 2005)).

Key Characteristics of Cosmic Web

Cosmic web consists of four distinct components (Libeskind et al., 2018):

- Massive galaxy *clusters* (*or nodes*),
- Interconnected **filaments**,
- Two-dimensional tenuous *sheets/walls*,

} on which matter concentrates.

around • Vast and near-empty *voids*.

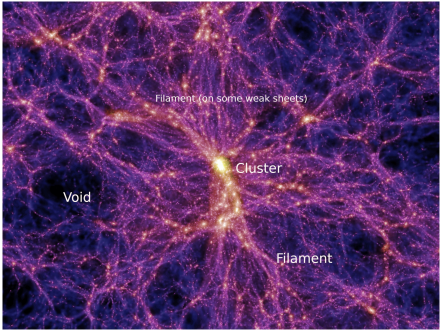


Figure: Characteristics of *Cosmic Web* (credited to the millennium simulation).

Significance of Cosmic Filaments

We will focus on detecting the (one-dimensional) **cosmic filaments**, because

- They connect complexes of super-clusters (Lynden-Bell et al., 1988).
- They contain information about the global cosmology and the nature of dark matter (Zhang et al., 2009; Tempel et al., 2014).

We will focus on detecting the (one-dimensional) **cosmic filaments**, because

- They connect complexes of super-clusters (Lynden-Bell et al., 1988).
- They contain information about the global cosmology and the nature of dark matter (Zhang et al., 2009; Tempel et al., 2014).
- The trajectory of cosmic microwave background light can be distorted due to cosmic filaments, creating the weak lensing effect.

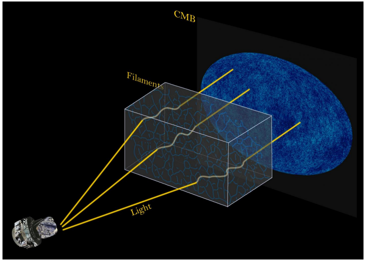


Figure: Illustration of the bending trajectory of CMB lights (credit to Siyu He, Shadab Alam, Wei Chen, and Planck/ESA; see He et al. (2018) for details).

- The filamentary structures are overwhelmingly complex (Cautun et al., 2013):
  - Lack of structural symmetries,
  - Uncertainty in measuring its connectivity,
  - Intrinsic multi-scale nature, etc.

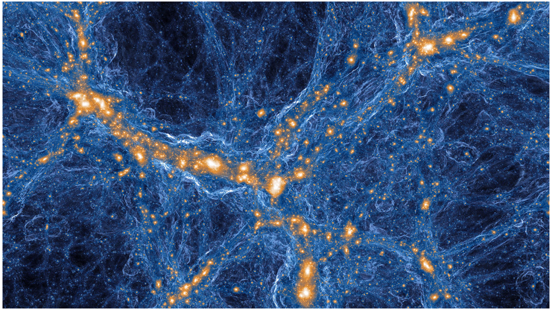


Figure: A view of the present-day cosmic web 300 million light-years across, as modeled by IllustrisTNG (Vogelsberger et al., 2014).

Challenges in Detecting Cosmic Filaments

- The filamentary structures are overwhelmingly complex (Cautun et al., 2013):
  - Lack of structural symmetries,
  - Uncertainty in measuring its connectivity,
  - Intrinsic multi-scale nature, etc.

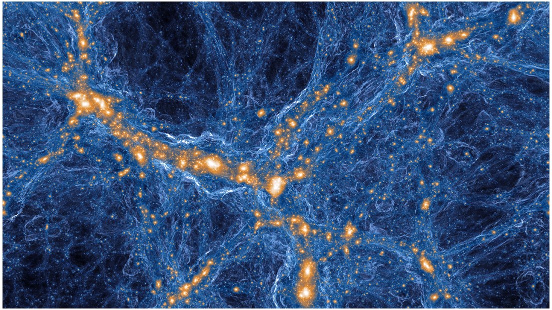


Figure: A view of the present-day cosmic web 300 million light-years across, as modeled by IllustrisTNG (Vogelsberger et al., 2014).

▶ There are no universal and mathematically rigorous definitions of cosmic filament!

Outline of Today's Talk

1. Review existing approaches for cosmic web detection and discuss their drawbacks.

Outline of Today's Talk

1. Review existing approaches for cosmic web detection and discuss their drawbacks.
2. Introduce our cosmic filament model, which formulates cosmic filaments as the ***directional density ridges*** on a celestial sphere.

Outline of Today's Talk

1. Review existing approaches for cosmic web detection and discuss their drawbacks.
2. Introduce our cosmic filament model, which formulates cosmic filaments as the ***directional density ridges*** on a celestial sphere.
3. Discuss ***directional density ridges*** from both statistical and computational perspectives:
   - Establish the statistical consistency of estimating the true density ridges with directional kernel density estimator (KDE).
   - Estimate the directional density ridges via our proposed ***Directional Subspace Constrained Mean Shift*** (DirSCMS) algorithm.
   - Establish the linear convergence properties of our DirSCMS algorithm.

Outline of Today's Talk

1. Review existing approaches for cosmic web detection and discuss their drawbacks.
2. Introduce our cosmic filament model, which formulates cosmic filaments as the ***directional density ridges*** on a celestial sphere.
3. Discuss ***directional density ridges*** from both statistical and computational perspectives:
   - Establish the statistical consistency of estimating the true density ridges with directional kernel density estimator (KDE).
   - Estimate the directional density ridges via our proposed ***Directional Subspace Constrained Mean Shift*** (DirSCMS) algorithm.
   - Establish the linear convergence properties of our DirSCMS algorithm.
4. Apply our DirSCMS algorithm to the galaxy observations in the Sloan Digital Sky Survey (SDSS-IV; Ahumada et al. 2020) and construct a cosmic web catalog.

Previous Works on Cosmic Filament Detection

Observational Data for Filament Detection

In astronomical survey data, the observed objects (or galaxies) are recorded as:

$$\{(\alpha_1, \delta_1, Z_1), \ldots, (\alpha_n, \delta_n, Z_n)\},$$

where, for $i = 1, \ldots, n$,

- $\alpha_i \in [0, 360°)$ is the *right ascension* (RA), i.e., celestial longitude,
- $\eta_i \in [-90°, 90°]$ is the *declination* (DEC), i.e., celestial latitude,
- $Z_i \in (0, \infty)$ is the *redshift* value, i.e., measuring its distance to the Earth.

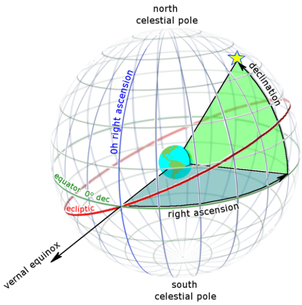


Figure: Illustration of RA and DEC (Image Courtesy of Wikipedia).

In astronomical survey data, the observed objects (or galaxies) are recorded as $\{(\alpha_1, \delta_1, Z_1), ..., (\alpha_n, \delta_n, Z_n)\}$.

The existing methods for detecting cosmic filaments from survey data can be classified into the following two categories:

In astronomical survey data, the observed objects (or galaxies) are recorded as $\{(\alpha_1, \delta_1, Z_1), ..., (\alpha_n, \delta_n, Z_n)\}$.

The existing methods for detecting cosmic filaments from survey data can be classified into the following two categories:

- **3D methods:** Convert $\{(\alpha_i, \delta_i, Z_i)\}_{i=1}^n$ to their Cartesian coordinates as

$$X_i = d(Z_i)\cos\alpha_i\cos\delta_i, \quad Y_i = d(Z_i)\sin\alpha_i\cos\delta_i, \quad Z_i = d(Z_i)\sin\delta_i,$$

  where $d(\cdot)$ is a distance transforming function; see Tempel et al. (2014) for details.

- **2D methods:** Slice the Universe into thin redshift slices (Chen et al., 2015b; Duque et al., 2022).

In astronomical survey data, the observed objects (or galaxies) are recorded as $\{(\alpha_1, \delta_1, Z_1), ..., (\alpha_n, \delta_n, Z_n)\}$.

The existing methods for detecting cosmic filaments from survey data can be classified into the following two categories:

- **3D methods:** Convert $\{(\alpha_i, \delta_i, Z_i)\}_{i=1}^n$ to their Cartesian coordinates as

$$X_i = d(Z_i) \cos\alpha_i \cos\delta_i, \quad Y_i = d(Z_i) \sin\alpha_i \cos\delta_i, \quad Z_i = d(Z_i) \sin\delta_i,$$

  where $d(\cdot)$ is a distance transforming function; see Tempel et al. (2014) for details.
- **2D methods:** Slice the Universe into thin redshift slices (Chen et al., 2015b; Duque et al., 2022).

▶ **Note:** Our method can easily switch between the above two categories.

2D Methods: Slicing the Universe

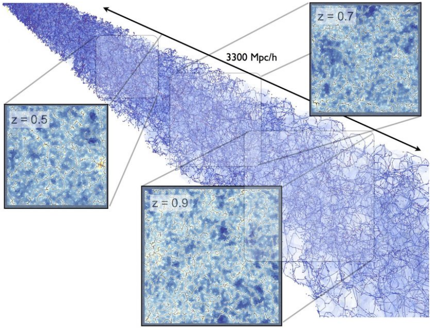


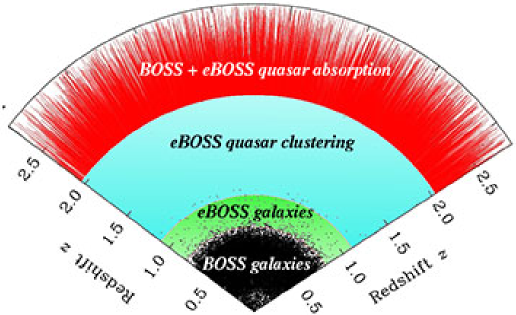
Figure: Illustration of slicing the Universe (credit to Laigle et al. 2018).

The tomographic filament detection has its own advantages over 3D methods:

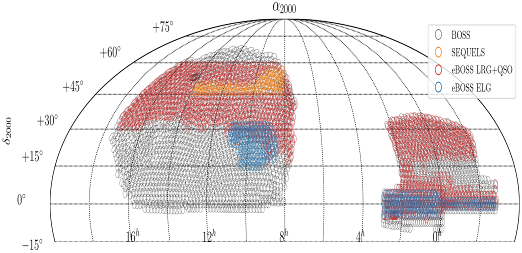
- It controls the redshift distortions along the line-of-sight direction (i.e., the *finger-of-god* effect).
- The measurement error in one slice will not propagate to other slices.
- It helps reduce computational cost...

The slices ($\Delta Z = 0.005$) in the survey data are not some flat 2D planes, but some **spherical shells**, which have a *nonlinear* curvature!

- Recall that the locations of astronomical objects in a slice are recorded by $\{(\alpha_i, \delta_i)\}_{i=1}^n$ on a celestial sphere $\Omega_2 = \left\{ \boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x}||_2 = 1 \right\}$.



(a) Planned eBOSS coverage of the Universe (credit to M. Blanton and SDSS)



(b) BOSS/eBOSS Spectroscopic Footprint as of DR16 (credit to SDSS)

Why can't we ignore the spherical geometry?

Setup: Suppose that we want to recover the true ring/filament structure across the North and South pole of a unit sphere given some noisy data points from it.

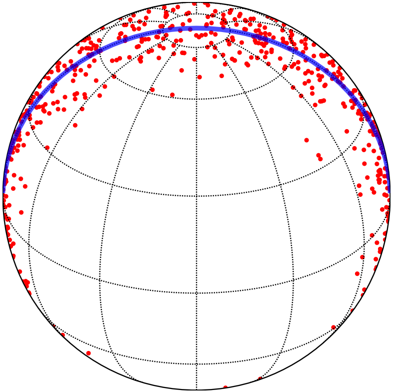
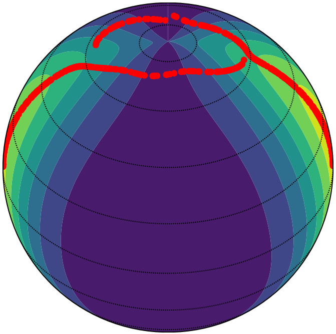
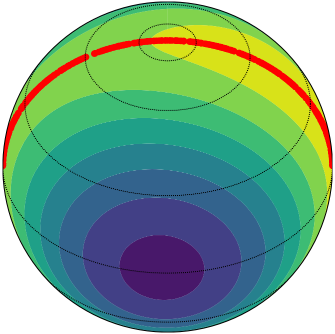


Figure: Noisy observations (red points) and the underlying true ring/filament structure (blue line).

The background contour plots are kernel density estimators on the flat plane $[-90°, 90°] \times [0°, 360°)$ and unit sphere $\Omega_2 = \left\{ \boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x}||_2 = 1 \right\}$, respectively.



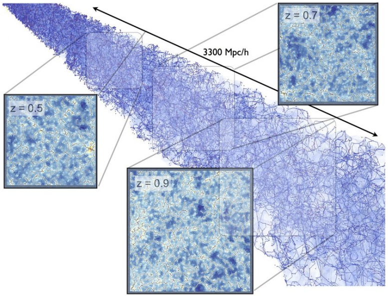
(a) Euclidean SCMS Method (ignoring the spherical geometry).



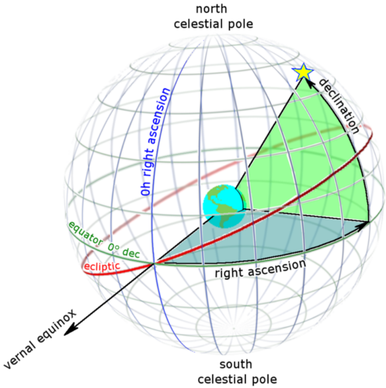
(b) Directional SCMS Method.

∗ SCMS: subspace constrained mean shift (Ozertem and Erdogmus, 2011).

Importance of Modeling Cosmic Web Under Spherical Geometry



(a) Slicing the Universe.



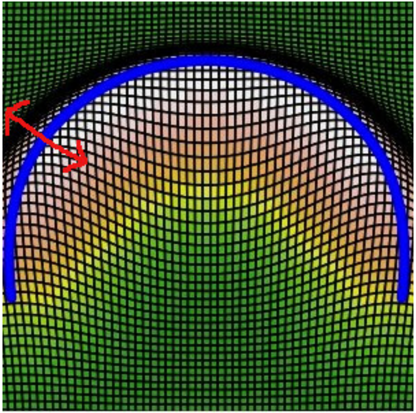
(b) Positioning of the observed galaxy.

▶ **Research Question:** How do we model and estimate the cosmic filaments based on the observed galaxies in each (redshift) **spherical** slice?

Cosmic Filament Model: Directional Density Ridges

- **Fact:** The cosmic filaments are 1D curves tracing over the high-density regions of matter (or galaxy) density field.

- **Fact:** The cosmic filaments are 1D curves tracing over the high-density regions of matter (or galaxy) density field.
- **Our Model:** (Directional) density ridges are generalized local maxima (within some subspaces) of the underlying density function (on $\Omega_q = \left\{ \boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1 \right\}$).



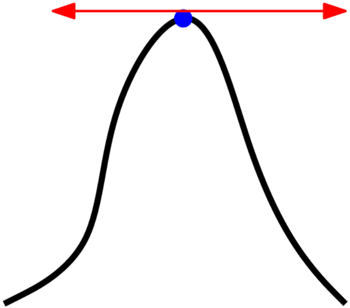


Figure: Density ridge (lifted onto the underlying density function; Chen et al. 2015a)

- **Local Modes/Maxima of f on $\Omega_q = \left\{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2 = 1\right\}$:**

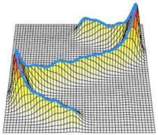
$$\mathcal{M} \equiv \texttt{Mode}(f) = \left\{\boldsymbol{x} \in \Omega_q : \texttt{grad} f(\boldsymbol{x}) = \mathbf{0}, \lambda_1(\boldsymbol{x}) < 0\right\}.$$

- $\texttt{grad} f(\boldsymbol{x})$ is the Riemannian gradient and $\mathcal{H}f(\boldsymbol{x})$ is the Riemannian Hessian on Ω_q.
- $\lambda_1(\boldsymbol{x}) \geq \cdots \geq \lambda_q(\boldsymbol{x})$ are (descending) eigenvalues of $\mathcal{H}f(\boldsymbol{x})$ associated with eigenvectors $\boldsymbol{v}_1(\boldsymbol{x}), ..., \boldsymbol{v}_q(\boldsymbol{x})$ that lies within the tangent space $T_{\boldsymbol{x}}$.

- **Local Modes/Maxima of f on $\Omega_q = \left\{x \in \mathbb{R}^{q+1} : ||x||_2 = 1\right\}$:**

$$\mathcal{M} \equiv \mathtt{Mode}(f) = \{x \in \Omega_q : \mathtt{grad}\, f(x) = \mathbf{0}, \lambda_1(x) < 0\}.$$

- $\mathtt{grad}\, f(x)$ is the Riemannian gradient and $\mathcal{H}f(x)$ is the Riemannian Hessian on Ω_q.
- $\lambda_1(x) \geq \cdots \geq \lambda_q(x)$ are (descending) eigenvalues of $\mathcal{H}f(x)$ associated with eigenvectors $v_1(x), ..., v_q(x)$ that lies within the tangent space T_x.



- **Density ridge on Ω_q (or *directional density ridge*) of f:**

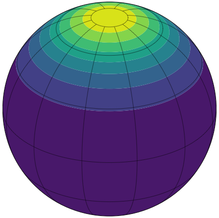
$$\mathcal{R}_d \equiv \mathtt{Ridge}(f) = \left\{x \in \Omega_q : V_d(x)V_d(x)^T \mathtt{grad}\, f(x) = \mathbf{0}, \lambda_{d+1}(x) < 0\right\},$$

where $V_d(x) = [v_{d+1}(x), ..., v_q(x)] \in \mathbb{R}^{(q+1)\times(q-d)}$ consists of the last $q-d$ eigenvectors of $\mathcal{H}f(x)$ within T_x.

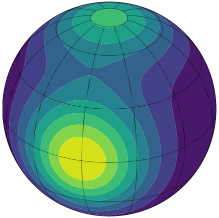
We first estimate the density function f on Ω_q via the directional KDE (Hall et al., 1987; Bai et al., 1988; García-Portugués, 2013) as:

$$\widehat{f}_h(\boldsymbol{x}) = \frac{C_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right),$$

- $L : [0, \infty) \to [0, \infty)$ is a directional kernel, *i.e.*, a rapidly decaying nonnegative function. (Example: von Mises kernel $L(r) = e^{-r}$.)
- $h > 0$ is the bandwidth parameter, and $C_{L,q}(h)$ is a normalizing term.



(a) $f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}, \nu)$ with $\boldsymbol{\mu} = (0, 0, 1)$ and $\nu = 4.0$.



(b) $\frac{2}{5} \cdot f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}_1, 5) + \frac{3}{5} \cdot f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}_2, 5)$ with $\boldsymbol{\mu}_1 = (0, 0, 1), \boldsymbol{\mu}_2 = (1, 0, 0)$.

The directional KDE $\widehat{f}_h$ is useful because its plug-in estimators

$$\widehat{\mathcal{M}} = \left\{ \boldsymbol{x} \in \Omega_q : \texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) = \mathbf{0}, \widehat{\lambda}_1(\boldsymbol{x}) < 0 \right\}$$

and

$$\widehat{\mathcal{R}}_d = \left\{ \boldsymbol{x} \in \Omega_q : \widehat{V}_d(\boldsymbol{x}) \widehat{V}_d(\boldsymbol{x})^T \texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) = \mathbf{0}, \widehat{\lambda}_{d+1}(\boldsymbol{x}) < 0 \right\}$$

The directional KDE $\widehat{f}_h$ is useful because its plug-in estimators

$$\widehat{\mathcal{M}} = \left\{ \boldsymbol{x} \in \Omega_q : \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) = \boldsymbol{0}, \widehat{\lambda}_1(\boldsymbol{x}) < 0 \right\}$$

and

$$\widehat{\mathcal{R}}_d = \left\{ \boldsymbol{x} \in \Omega_q : \widehat{V}_d(\boldsymbol{x}) \widehat{V}_d(\boldsymbol{x})^T \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) = \boldsymbol{0}, \widehat{\lambda}_{d+1}(\boldsymbol{x}) < 0 \right\}$$

approach $\mathcal{M}$ and $\mathcal{R}_d$ in a statistically consistent way (Theorem 6 in Zhang and Chen 2021 and Theorem 4.1 in Zhang and Chen 2022):

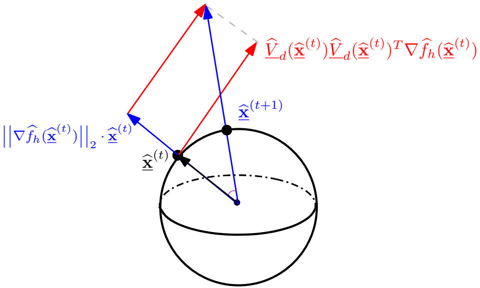
- $\mathtt{Haus}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$, as $h \to 0$ and $nh^{q+2} \to \infty$,
- $\mathtt{Haus}\left(\mathcal{R}_d, \widehat{\mathcal{R}}_d\right) = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right)$, as $h \to 0$ and $\frac{nh^{q+6}}{|\log h|} \to \infty$,

where $\mathtt{Haus}(A, B) = \max \left\{ r > 0 : \sup_{\boldsymbol{x} \in A} d(\boldsymbol{x}, B), \sup_{\boldsymbol{y} \in B} d(\boldsymbol{y}, A) \right\}$.

We generalize the traditional subspace constrained mean shift algorithm Ozertem and Erdogmus (2011) in $\mathbb{R}^D$ to estimate $\widehat{\mathcal{R}}_d$ in practice as the **directional subspace constrained mean shift (DirSCMS)** algorithm (Section 4.2 in Zhang and Chen 2022):

$$\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \widehat{\boldsymbol{x}}^{(t)} + \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T \cdot \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{x}}^{(t)})}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{x}}^{(t)})\right|\right|_2} \quad \text{and} \quad \widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \frac{\widehat{\boldsymbol{x}}^{(t+1)}}{\left|\left|\widehat{\boldsymbol{x}}^{(t+1)}\right|\right|_2},$$

for $t = 0, 1, \ldots$.



Essentially, this is a subspace constrained gradient ascent algorithm on Ω_q, for which we establish the linear convergence results in Section 4.3 of Zhang and Chen (2022).

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

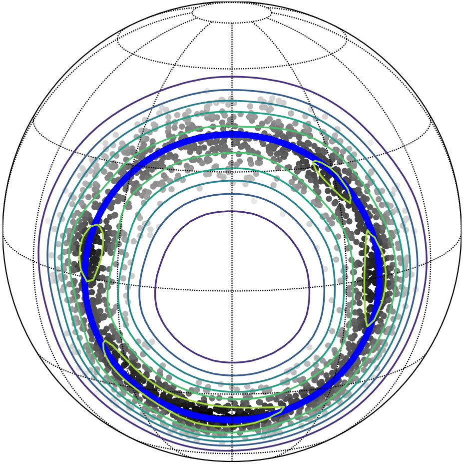


Figure: The underlying circle (blue curve) and sampled points (gray dots) on Ω_2.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

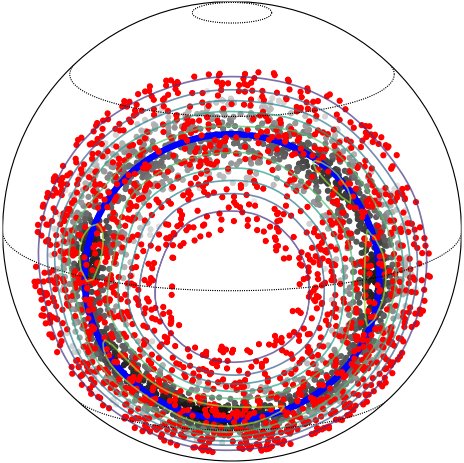


Figure: Directional SCMS at Step 0.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

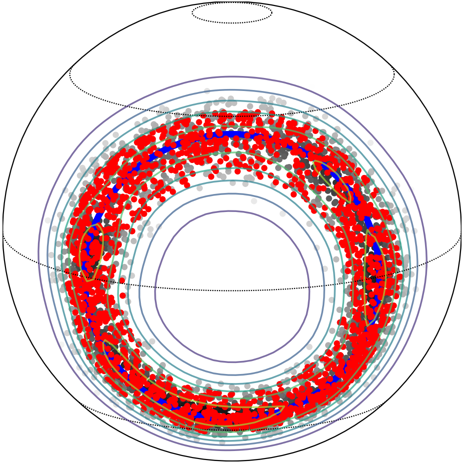


Figure: Directional SCMS at Step 1.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

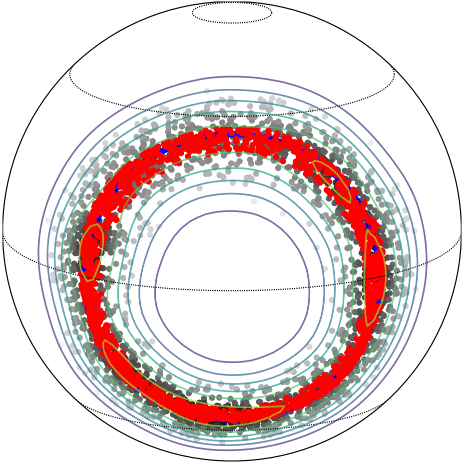


Figure: Directional SCMS at Step 2.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

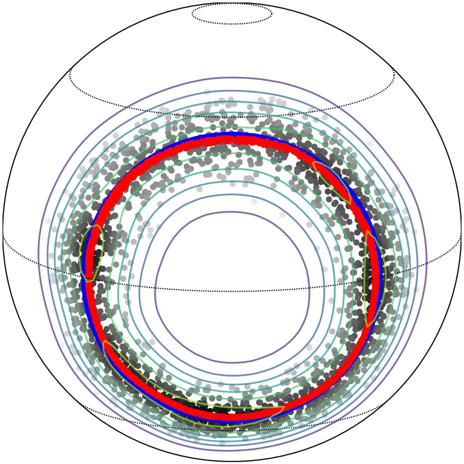


Figure: Directional SCMS at Step 4.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

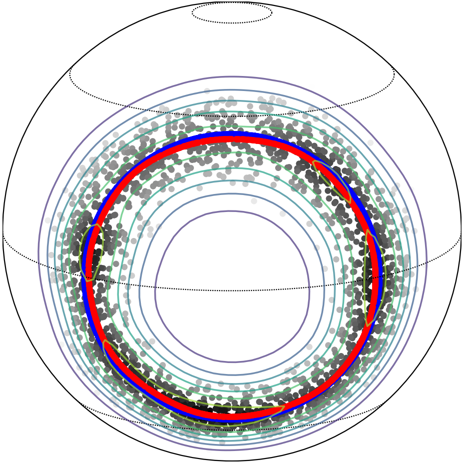


Figure: Directional SCMS at Step 8.

We simulate 2000 data points from a circle on Ω_2 with additive Gaussian noises $\mathcal{N}(0, 0.1^2)$ on their Cartesian coordinates and L_2 normalization.

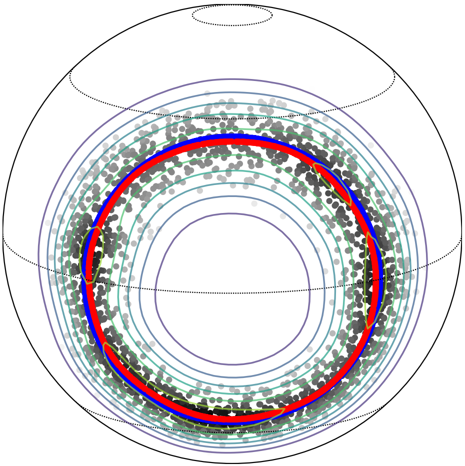


Figure: Directional SCMS at Step 24 (converged).

The observed galactic data $\{(\phi_i, \eta_i, z_i)\}_{i=1}^n \subset \Omega_2 \times \mathbb{R}^+$ are directional-linear, and the density ridges in $\Omega_2 \times \mathbb{R}^+$ (Zhang and Chen, 2025) can also be estimated as:

- Density estimation at $(\boldsymbol{x}, z) \in \Omega_q \times \mathbb{R}$ (García-Portugués et al., 2015):

$$\widehat{f}_{\boldsymbol{h}}(\boldsymbol{x}, z) = \frac{C_L(h_1)}{nh_2} \sum_{i=1}^n L\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h_1^2}\right) K\left(\frac{z - Z_i}{h_2}\right).$$

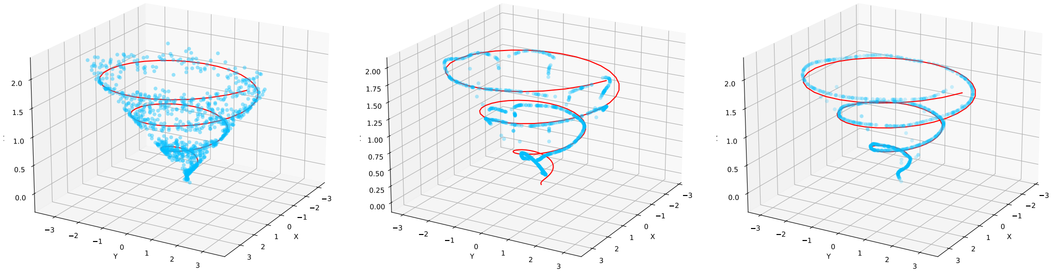
- Ridge-Finding via SCMS algorithm on $\boldsymbol{y}^{(t)} = (\boldsymbol{x}^{(t)}, z^{(t)})$ as:

$$\boldsymbol{y}^{(t+1)} \leftarrow \boldsymbol{y}^{(t)} + \eta \cdot \widehat{V}_d(\boldsymbol{y}^{(t)}) \widehat{V}_d(\boldsymbol{y}^{(t)})^T \boldsymbol{H}^{-1} \Xi(\boldsymbol{y}^{(t)}) \quad \text{with}$$

$$\Xi(\boldsymbol{y}) = (\Xi_x(\boldsymbol{x}, z), \Xi_z(\boldsymbol{x}, z))^T = \left(\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{X}_i^T \boldsymbol{x}^{(t)}}{h_1^2}\right) K\left(\frac{z^{(t)} - Z_i}{h_2}\right)}{\sum_{i=1}^n L'\left(\frac{1 - \boldsymbol{X}_i^T \boldsymbol{x}^{(t)}}{h_1^2}\right) K\left(\frac{z^{(t)} - Z_i}{h_2}\right)} - \boldsymbol{x}, \frac{\sum_{i=1}^n z_i L\left(\frac{1 - \boldsymbol{X}_i^T \boldsymbol{x}^{(t)}}{h_1^2}\right) K'\left(\frac{z^{(t)} - Z_i}{h_2}\right)}{\sum_{i=1}^n L\left(\frac{1 - \boldsymbol{X}_i^T \boldsymbol{x}^{(t)}}{h_1^2}\right) K'\left(\frac{z^{(t)} - Z_i}{h_2}\right)} - z \right).$$

▶ **Note:** A naive generalization of SCMS algorithm $\boldsymbol{z}^{(t+1)} \leftarrow \boldsymbol{z}^{(t)} + \widehat{V}_d(\boldsymbol{z}^{(t)}) \widehat{V}_d(\boldsymbol{z}^{(t)})^T \Xi(\boldsymbol{z}^{(t)})$ plus standardization as with pure Euclidean/directional data does not work (Zhang and Chen, 2025)!

We sample 1000 observations on a spiral curve with additive Gaussian noises $\mathcal{N}(0, 0.2^2)$ to their angular-linear coordinates.



(a) Simulated data points. (b) Euclidean SCMS. (c) Directional-linear SCMS.

- Our directional-linear SCMS algorithm is stabler than its Euclidean prototype.

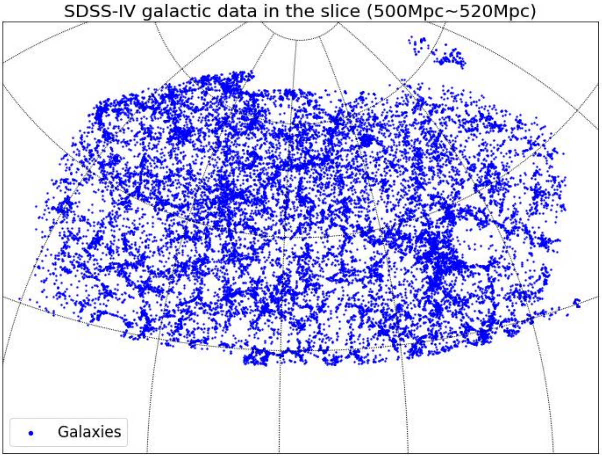
All of our proposed methods are encapsulated in a Python package called **SCONCE-SCMS** (**S**pherical and **CON**ic **C**osmic w**E**b finder with the extended **SCMS** algorithms; Zhang et al. 2022).

- **Python Package Index**: https://pypi.org/project/sconce-scms/.
- **Documentation**: https://sconce-scms.readthedocs.io/en/latest/.

SDSS-IV Cosmic Web Catalog

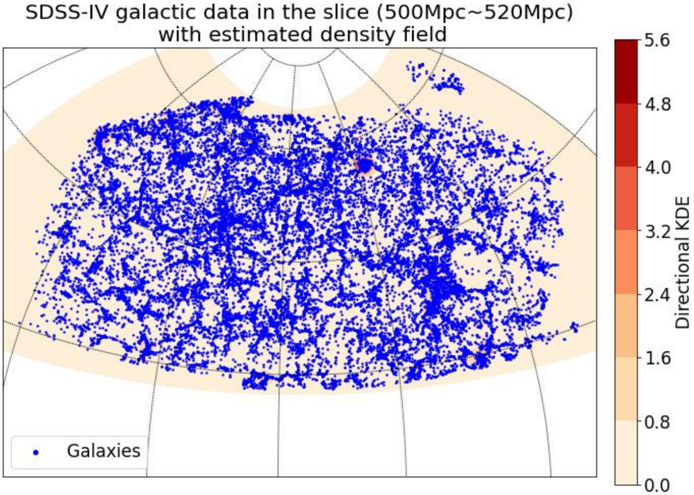
Step 1 (Slicing the Universe): Partition the redshift range into 325 spherical slices based on the comoving distance $\Delta L = 20\,\text{Mpc}$.

- Within each slice, we consider the redshifts of galaxies to be the same so that the galaxies are located on Ω_2.



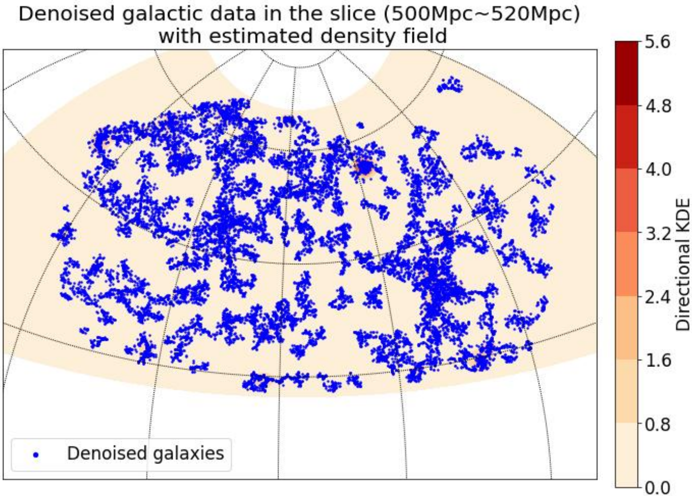
Step 2 (Density Estimation): Estimate the galaxy density field within each spherical slice by directional KDE.

- The bandwidth parameter is selected via a data-adaptive approach.

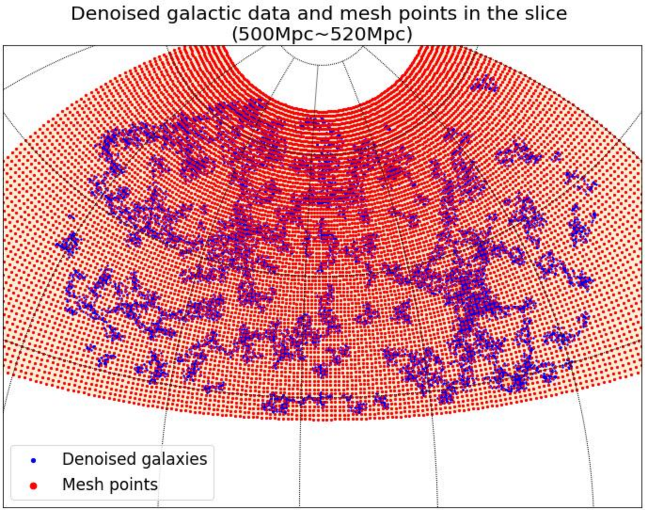


Step 3 (Denoising): Remove the observations with low-density values.

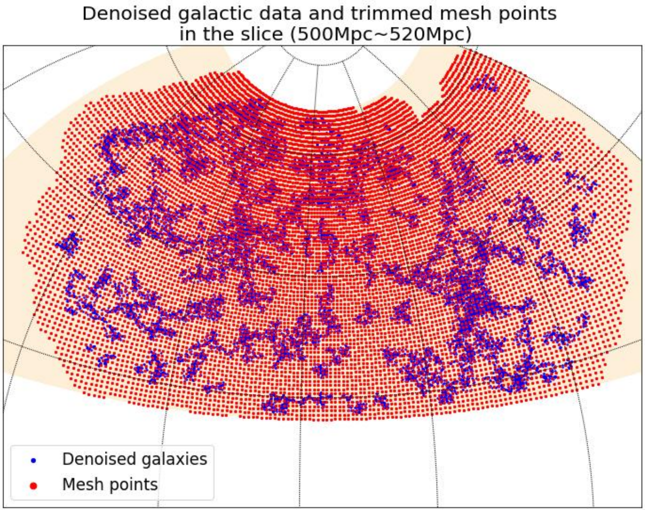
- We keep at least 80% of the original galaxy data in the slice.



Step 4 (Laying Down the Mesh Points): We place a set of dense mesh points on the interested region, which are the initial points of our DirSCMS iterations.



Step 5 (Thresholding the Mesh Points): We discard those mesh points with low-density values and keep 85% of the original mesh points.



Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

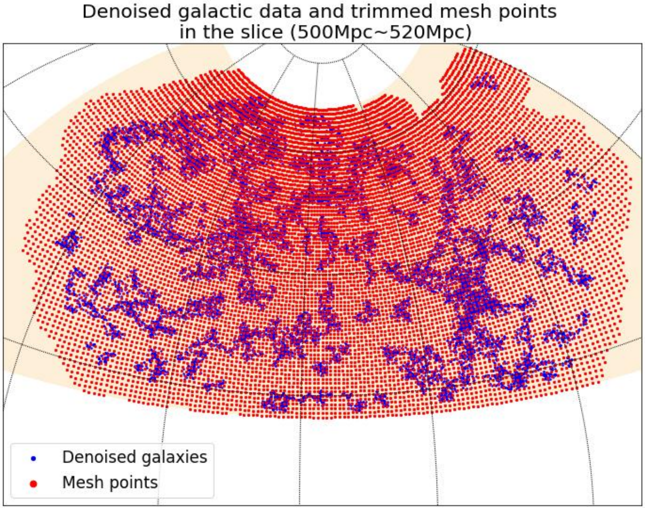


Figure: DirSCMS Iterations (Step 0).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

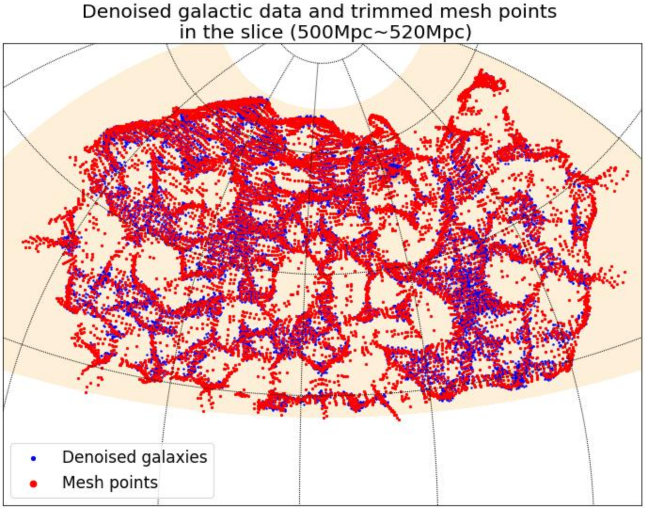


Figure: DirSCMS Iterations (Step 1).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

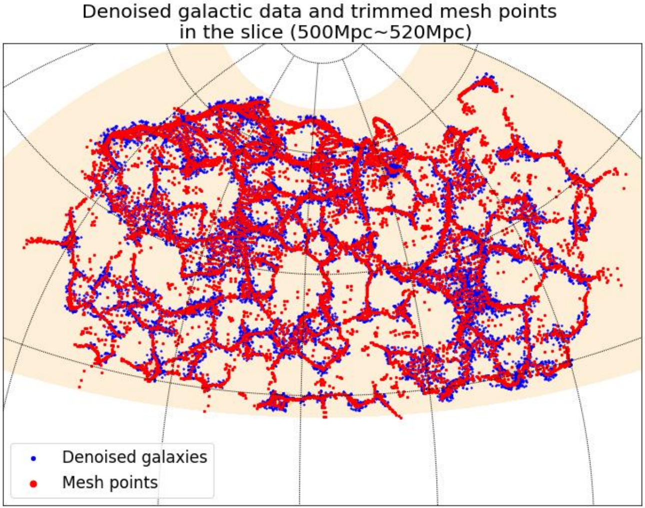


Figure: DirSCMS Iterations (Step 2).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

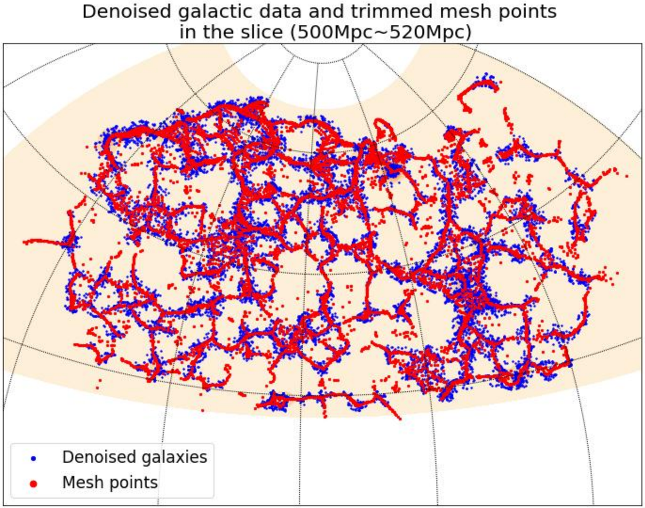


Figure: DirSCMS Iterations (Step 3).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

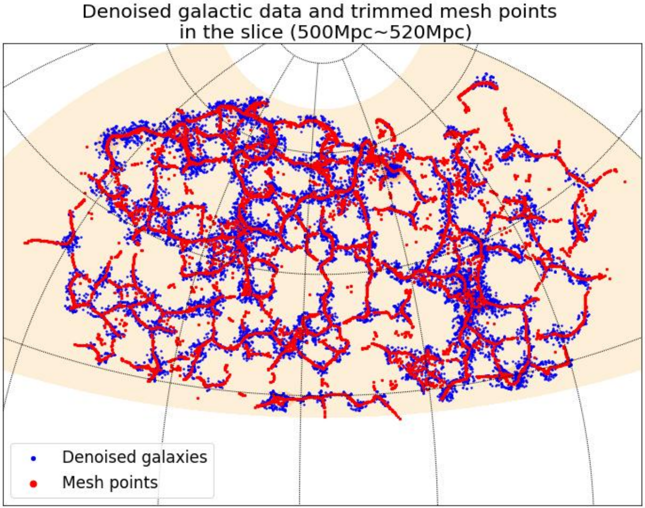


Figure: DirSCMS Iterations (Step 5).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

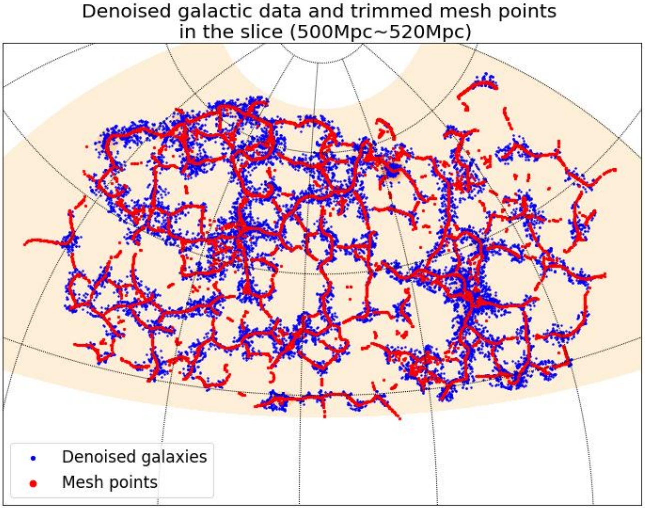


Figure: DirSCMS Iterations (Step 8).

Step 6 (DirSCMS Iterations): We iterate our DirSCMS algorithm on each remaining mesh point until convergence.

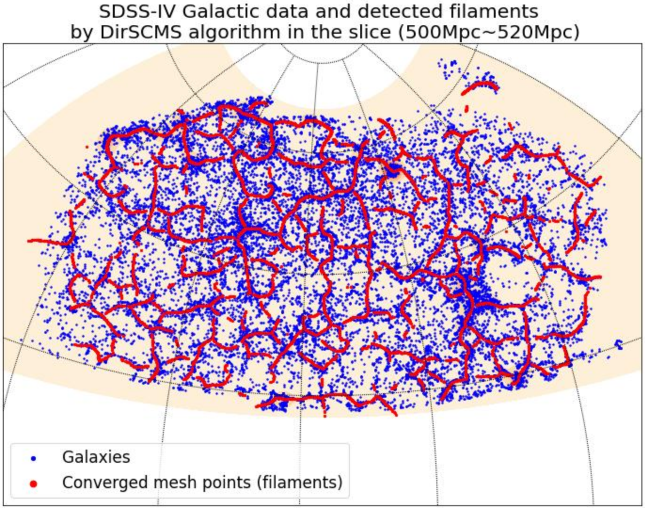


Figure: DirSCMS Iterations (Final).

Step 7 (Mode and Knot Estimation): We seek out the local modes and knots on the filaments as cosmic nodes.

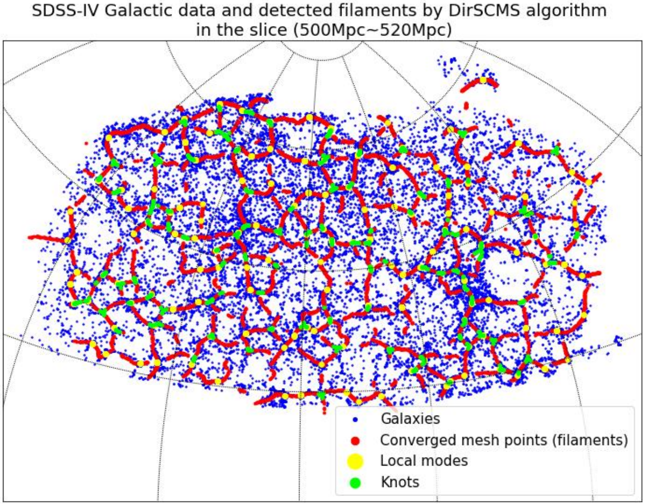
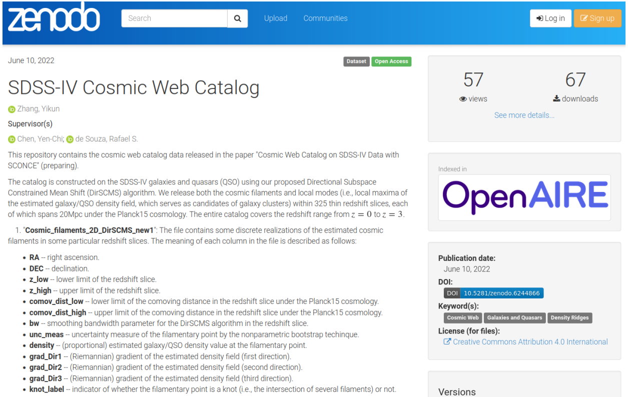


Figure: Nodes on the detected filaments.

Final Cosmic Web Catalog on SDSS-IV Data

- The input data incorporate not only galaxy but also quasar (QSO) observations so as to dive deeper into the Universe.
- We compute the uncertainty measure and other features for each detected filamentary point.
- The final catalog is available at https://doi.org/10.5281/zenodo.6244866.



Conclusion and Future Works

Conclusions

In this talk, we discuss our method for estimating cosmic filament structures from observed galactic data and its statistical theory.

In this talk, we discuss our method for estimating cosmic filament structures from observed galactic data and its statistical theory.

1. The cosmic filaments are modeled by directional density ridges, which can be consistently estimated by directional KDE.

In this talk, we discuss our method for estimating cosmic filament structures from observed galactic data and its statistical theory.

1. The cosmic filaments are modeled by directional density ridges, which can be consistently estimated by directional KDE.

2. We design an efficient algorithm (DirSCMS) to find the directional density ridges in practical applications.

Conclusions

In this talk, we discuss our method for estimating cosmic filament structures from observed galactic data and its statistical theory.

1. The cosmic filaments are modeled by directional density ridges, which can be consistently estimated by directional KDE.
2. We design an efficient algorithm (DirSCMS) to find the directional density ridges in practical applications.
3. The cosmic web catalog based on our proposed method is publicly available.

Conclusions

In this talk, we discuss our method for estimating cosmic filament structures from observed galactic data and its statistical theory.

1. The cosmic filaments are modeled by directional density ridges, which can be consistently estimated by directional KDE.

2. We design an efficient algorithm (DirSCMS) to find the directional density ridges in practical applications.

3. The cosmic web catalog based on our proposed method is publicly available.

Along this line of research, we are planning to

- Leverage our cosmic filament catalog to identify cosmic voids and infer the precise cosmology (Sánchez et al., 2016).

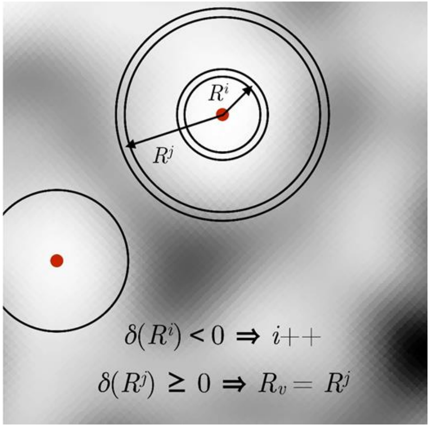
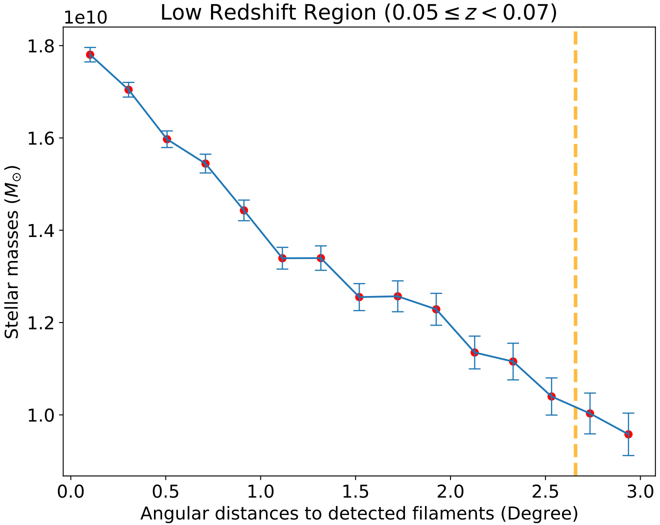


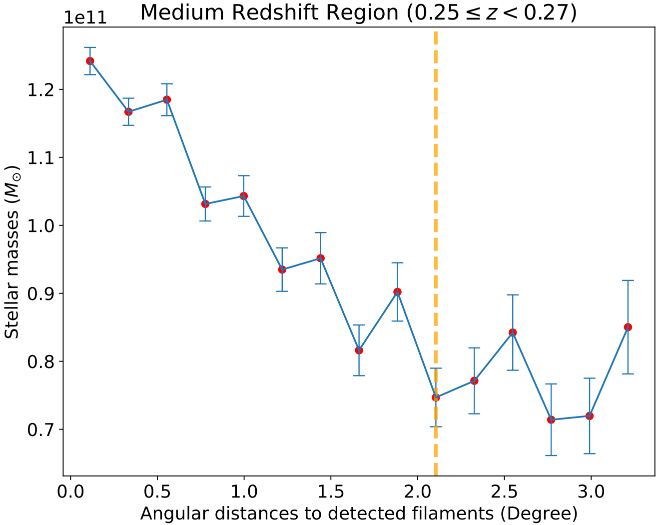
Figure: Simple void-finding algorithm (Sánchez et al., 2016).

Future Work: Filament Effects of Galaxy Properties

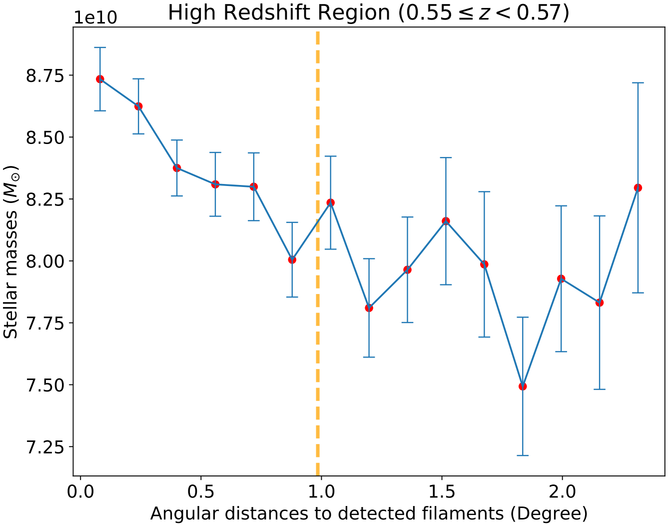
- Analyze if galaxy properties, such as stellar mass, color, and star formation rate, are correlated with our detected cosmic web structures (Chen et al., 2017; Kotecha, 2020)...



- Analyze if galaxy properties, such as stellar mass, color, and star formation rate, are correlated with our detected cosmic web structures (Chen et al., 2017; Kotecha, 2020)...



- Analyze if galaxy properties, such as stellar mass, color, and star formation rate, are correlated with our detected cosmic web structures (Chen et al., 2017; Kotecha, 2020)...



Thank you!

More details can be found in

[1] Y. Zhang and Y.-C. Chen. Kernel Smoothing, Mean Shift, and Their Learning Theory with Directional Data. *Journal of Machine Learning Research*, 22(154):1–92, 2021. https://arxiv.org/abs/2010.13523

[2] Y. Zhang and Y.-C. Chen. The EM Perspective of Directional Mean Shift Algorithm. arXiv preprint, 2021. https://arxiv.org/abs/2101.10058

[3] Y. Zhang and Y.-C. Chen. Linear Convergence of the Subspace Constrained Mean Shift Algorithm: From Euclidean to Directional Data. *Information and Inference: A Journal of the IMA*, 12(1): 210-311, 2022. https://arxiv.org/abs/2104.14977

[4] Y. Zhang and Y.-C. Chen. Mode and Ridge Estimation in Euclidean and Directional Product Spaces: A Mean Shift Approach. *Journal of Computational and Graphical Statistics*, (just-accepted): 1-20, 2025. https://arxiv.org/abs/2110.08505

[5] Y. Zhang, R. S. de Souza, and Y.-C. Chen. SCONCE: A Cosmic Web Finder for Spherical and Conic Geometries. *Monthly Notices of the Royal Astronomical Society*, 517(1): 1197-1217, 2022. https://arxiv.org/abs/2207.07001

Reference

R. Ahumada, C. A. Prieto, A. Almeida, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, R. Arcodia, E. Armengaud, M. Aubert, et al. The 16th data release of the sloan digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020.

Z. Bai, C. Rao, and L. Zhao. Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27(1):24 – 39, 1988.

J. R. Bond, L. Kofman, and D. Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380 (6575):603–606, 1996.

M. Cautun, R. van de Weygaert, and B. J. Jones. Nexus: tracing the cosmic web connection. *Monthly Notices of the Royal Astronomical Society*, 429(2):1286–1308, 2013.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5): 1896–1928, 2015a.

Y.-C. Chen, S. Ho, P. E. Freeman, C. R. Genovese, and L. Wasserman. Cosmic web reconstruction through density ridges: method and algorithm. *Monthly Notices of the Royal Astronomical Society*, 454(1):1140–1156, 2015b.

Y.-C. Chen, S. Ho, R. Mandelbaum, N. A. Bahcall, J. R. Brownstein, P. E. Freeman, C. R. Genovese, D. P. Schneider, and L. Wasserman. Detecting effects of filaments on galaxy properties in the sloan digital sky survey iii. *Monthly Notices of the Royal Astronomical Society*, 466(2):1880–1893, 2017.

J. C. Duque, M. Migliaccio, D. Marinucci, and N. Vittorio. A novel cosmic filament catalogue from sdss data. *Astronomy & Astrophysics*, 659:A166, 2022.

E. García-Portugués. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7:1655–1685, 2013.

Reference

E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga. Central limit theorems for directional and linear random variables with applications. *Statistica Sinica*, pages 1207–1229, 2015.

P. Hall, G. S. Watson, and J. Cabrara. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 12 1987. ISSN 0006-3444. URL https://doi.org/10.1093/biomet/74.4.751.

S. He, S. Alam, S. Ferraro, Y.-C. Chen, and S. Ho. The detection of the imprint of filaments on cosmic microwave background lensing. *Nature Astronomy*, 2(5):401–406, 2018.

N. Kaiser. Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society*, 227(1):1–21, 1987.

S. Kotecha. *The Effect of Cosmic Web Filaments on Quenching in Galaxy Clusters*. PhD thesis, 2020.

U. Kuchner, A. Aragón-Salamanca, A. Rost, F. R. Pearce, M. E. Gray, W. Cui, A. Knebe, E. Rasia, and G. Yepes. Cosmic filaments in galaxy cluster outskirts: quantifying finding filaments in redshift space. *Monthly Notices of the Royal Astronomical Society*, 503(2):2065–2076, 2021.

C. Laigle, C. Pichon, S. Arnouts, H. J. Mccracken, Y. Dubois, J. Devriendt, A. Slyz, D. Le Borgne, A. Benoit-Levy, H. S. Hwang, et al. Cosmos2015 photometric redshifts probe the impact of filaments on galaxy properties. *Monthly Notices of the Royal Astronomical Society*, 474(4):5437–5458, 2018.

N. I. Libeskind, R. Van De Weygaert, M. Cautun, B. Falck, E. Tempel, T. Abel, M. Alpaslan, M. A. Aragón-Calvo, J. E. Forero-Romero, R. Gonzalez, et al. Tracing the cosmic web. *Monthly Notices of the Royal Astronomical Society*, 473 (1):1195–1217, 2018.

Reference

D. Lynden-Bell, S. Faber, D. Burstein, R. L. Davies, A. Dressler, R. Terlevich, and G. Wegner. Spectroscopy and photometry of elliptical galaxies. v-galaxy streaming toward the new supergalactic center. *The Astrophysical Journal*, 326:19–49, 1988.

U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12 (34):1249–1286, 2011.

C. Sánchez, J. Clampitt, A. Kovacs, B. Jain, J. García-Bellido, S. Nadathur, D. Gruen, N. Hamaus, D. Huterer, P. Vielzeuf, et al. Cosmic voids and void lensing in the dark energy survey science verification data. *Monthly Notices of the Royal Astronomical Society*, 465(1):746–759, 10 2016.

W. Sargent and E. Turner. A statistical method for determining the cosmological density parameter from the redshifts of a complete sample of galaxies. *The Astrophysical Journal*, 212:L3–L7, 1977.

V. Springel, S. D. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, et al. Simulations of the formation, evolution and clustering of galaxies and quasars. *nature*, 435(7042):629–636, 2005.

E. Tempel, R. Stoica, V. J. Martinez, L. Liivamägi, G. Castellan, and E. Saar. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the sdss. *Monthly Notices of the Royal Astronomical Society*, 438(4): 3465–3482, 2014.

M. Vogelsberger, S. Genel, V. Springel, P. Torrey, D. Sijacki, D. Xu, G. Snyder, D. Nelson, and L. Hernquist. Introducing the illustris project: simulating the coevolution of dark and visible matter in the universe. *Monthly Notices of the Royal Astronomical Society*, 444(2):1518–1547, 2014.

Y. B. Zel'Dovich. Gravitational instability: An approximate theory for large density perturbations. *Astronomy and astrophysics*, 5:84–89, 1970.

Y. Zhang and Y.-C. Chen. Kernel smoothing, mean shift, and their learning theory with directional data. *Journal of Machine Learning Research*, 22(154):1–92, 2021.

Y. Zhang and Y.-C. Chen. Linear convergence of the subspace constrained mean shift algorithm: From euclidean to directional data. *Information and Inference: A Journal of the IMA*, 2022. URL https://doi.org/10.1093/imaiai/iaac005.

Y. Zhang and Y.-C. Chen. Mode and ridge estimation in euclidean and directional product spaces: A mean shift approach. *Journal of Computational and Graphical Statistics*, (just-accepted):1–20, 2025.

Y. Zhang, X. Yang, A. Faltenbacher, V. Springel, W. Lin, and H. Wang. The spin and orientation of dark matter halos within cosmic filaments. *The Astrophysical Journal*, 706(1):747, 2009.

Y. Zhang, R. S. de Souza, and Y.-C. Chen. Sconce: A cosmic web finder for spherical and conic geometries. *arXiv preprint arXiv:2207.07001*, 2022. URL https://arxiv.org/abs/2207.07001.

Riemannian Gradient and Hessian on Ω_q

Assume tentatively that the directional function f is well-defined and smooth in $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ (or at least in an open neighborhood $U \supset \Omega_q$).

- *Riemannian gradient* $\mathtt{grad}\, f(\boldsymbol{x})$ on Ω_q:

$$\mathtt{grad}\, f(\boldsymbol{x}) = \left(\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}),$$

  where $\boldsymbol{I}_{q+1}$ is the identity matrix in $\mathbb{R}^{(q+1)\times(q+1)}$.

- *Riemannian Hessian* $\mathcal{H}f(\boldsymbol{x})$ on Ω_q (Zhang and Chen, 2021):

$$\mathcal{H}f(\boldsymbol{x}) = (\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} \cdot \boldsymbol{I}_{q+1}\right] (\boldsymbol{I}_{q+1} - \boldsymbol{x}\boldsymbol{x}^T).$$

Here, $\boldsymbol{I}_{q+1}$ is the identity matrix in $\mathbb{R}^{(q+1)\times(q+1)}$, while $\nabla f(\boldsymbol{x})$ and $\nabla\nabla f(\boldsymbol{x})$ are total gradient and Hessian in $\mathbb{R}^{q+1}$.

Input:

- A directional data sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim f(\boldsymbol{x})$ on Ω_q
- The order d of the directional ridge, smoothing bandwidth $h > 0$, and tolerance level $\epsilon > 0$.
- A suitable mesh $\mathcal{M}_D \subset \Omega_q$ of initial points.

Step 1: Compute the directional KDE $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{L,q}(h)}{n} \sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$ on the mesh $\mathcal{M}_D$.

Step 2: For each $\widehat{\boldsymbol{x}}^{(0)} \in \mathcal{M}_D$, iterate the following DirSCMS update until convergence:

while $\left\| \sum_{i=1}^{n} \widehat{V}_d(\widehat{\boldsymbol{x}}^{(0)}) \widehat{V}_d(\widehat{\boldsymbol{x}}^{(0)})^T \boldsymbol{X}_i \cdot L'\left(\frac{1-\boldsymbol{X}_i^T \widehat{\boldsymbol{x}}^{(0)}}{h^2}\right) \right\|_2 > \epsilon$ **do**:

Detailed Procedures of DirSCMS Algorithm

- **Step 2-1**: Compute the scaled version of the estimated Hessian matrix as:

$$\frac{nh^2}{c_{L,q}(h)}\mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{x}}^{(t)}) = \left[\boldsymbol{I}_{q+1} - \widehat{\boldsymbol{x}}^{(t)}\left(\widehat{\boldsymbol{x}}^{(t)}\right)^T\right]\left[\frac{1}{h^2}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \cdot L''\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{x}}^{(t)}}{h^2}\right)\right.$$
$$\left. + \sum_{i=1}^{n}\boldsymbol{X}_i^T\widehat{\boldsymbol{x}}^{(t)}\boldsymbol{I}_{q+1} \cdot L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{x}}^{(t)}}{h^2}\right)\right]\left[\boldsymbol{I}_{q+1} - \widehat{\boldsymbol{x}}^{(t)}\left(\widehat{\boldsymbol{x}}^{(t)}\right)^T\right].$$

- **Step 2-2**: Perform the spectral decomposition on $\frac{nh^2}{c_{L,q}(h)}\mathcal{H}\widehat{f}_h\left(\widehat{\boldsymbol{x}}^{(t)}\right)$ and compute $\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)}) = \left[\boldsymbol{v}_{d+1}(\widehat{\boldsymbol{x}}^{(t)}), ..., \boldsymbol{v}_q(\widehat{\boldsymbol{x}}^{(t)})\right]$, whose columns are orthonormal eigenvectors corresponding to the smallest $q-d$ eigenvalues inside the tangent space $T_{\widehat{\boldsymbol{x}}^{(t)}}$.
- **Step 2-3**: Update

$$\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \widehat{\boldsymbol{x}}^{(t)} - \widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})\widehat{V}_d(\widehat{\boldsymbol{x}}^{(t)})^T\left[\frac{\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{x}}^{(t)}}{h^2}\right)}{\sum_{i=1}^{n}\boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{X}_i^T\widehat{\boldsymbol{x}}^{(t)}}{h^2}\right)}\right].$$

- **Step 2-4**: Standardize $\widehat{\boldsymbol{x}}^{(t+1)}$ as $\widehat{\boldsymbol{x}}^{(t+1)} \leftarrow \frac{\widehat{\boldsymbol{x}}^{(t+1)}}{\left|\left|\widehat{\boldsymbol{x}}^{(t+1)}\right|\right|_2}$.

Output: An estimated directional d-ridge $\widehat{\mathcal{R}}_d$ represented by the collection of resulting points.

Under some regularity conditions, we prove the following (Theorem 4.6 in Zhang and Chen 2022):

1. **R-Linear convergence of** $d(\boldsymbol{x}^{(k)}, \mathcal{R}_d)$ **with** f. When the step size $\underline{\eta} > 0$ is sufficiently small and the initial point $\boldsymbol{x}^{(0)}$ lies within a small neighborhood of its limiting point $\boldsymbol{x}^*$ in $\mathcal{R}_d$,

$$d\left(\boldsymbol{x}^{(k)}, \mathcal{R}_d\right) \leq \underline{\Upsilon}^k \cdot d\left(\boldsymbol{x}^{(0)}, \boldsymbol{x}^*\right) \quad \text{with} \quad \underline{\Upsilon} = \sqrt{1 - \frac{\underline{\Upsilon}\beta_0}{4}},$$

where $\beta_0 > 0$ is the eigengap between the d-th and $(d+1)$-th eigenvalues of $\mathcal{H}f(\boldsymbol{x})$.

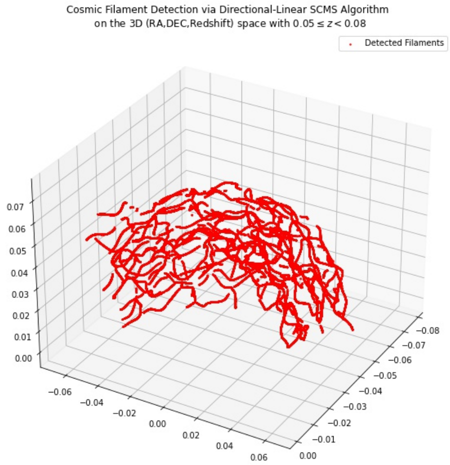
2. **R-Linear convergence of** $d(\widehat{\boldsymbol{x}}^{(k)}, \mathcal{R}_d)$ **with** $\widehat{f}_h$. When the step size $\underline{\eta} > 0$ is sufficiently small and the initial point $\widehat{\boldsymbol{x}}^{(0)}$ lies within a small neighborhood of $\boldsymbol{x}^*$ in $\mathcal{R}_d$,

$$d\left(\boldsymbol{x}^{(k)}, \mathcal{R}_d\right) \leq \underline{\Upsilon}^k \cdot d\left(\boldsymbol{x}^{(0)}, \boldsymbol{x}^*\right) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right)$$

with probability tending to 1, as $h \to 0$ and $\frac{nh^{q+4}}{|\log h|} \to 0$.

- The linear convergence results can also be proved for the subspace constrained gradient ascent method but under some stricter conditions (Zhang and Chen, 2022).
- The (directional) mean shift and SCMS algorithms can be viewed as variants of the (subspace constrained) gradient ascent methods (on Ω_q) but with adaptive step sizes.
- The step sizes can be made sufficiently small as the bandwidth h is small and the sample size n is large, but also universally bounded away from 0 with respect to the iteration number t.

Application of DirLinSCMS to SDSS-IV Galaxy Data



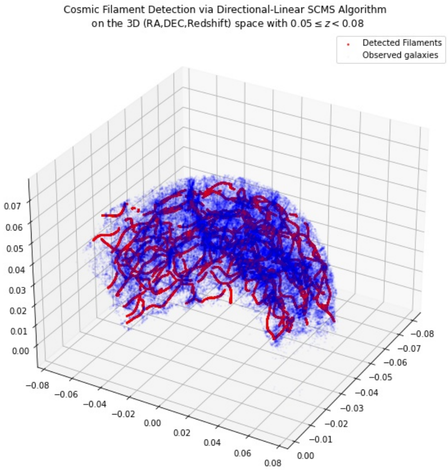


Figure: Cosmic filament detection in the 3D (RA,DEC,Redshift) space with our directional-linear SCMS algorithm.

There are some potential drawbacks of detecting filaments with survey data in the 3D space:

- The determination of $d(\cdot)$ relies on complex cosmological models.
- The galaxy distribution is distorted along the line of sight due to the peculiar velocities of galaxies (i.e., the so-called *finger-of-god* (Sargent and Turner, 1977) and *Kaiser* (Kaiser, 1987) effects).

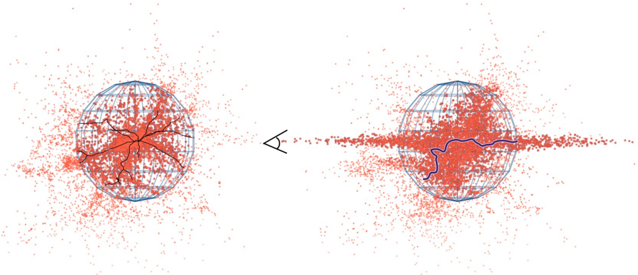


Figure: Redshift distortions along the line of sight (Kuchner et al., 2021).

- The number of galaxies varies across different redshift values, so applying 3D approaches will be computationally intensive.