Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **454,** 1140–1156 (2015) doi:10.1093/mnras/stv1996

Cosmic web reconstruction through density ridges: method and algorithm

Yen-Chi Chen,[1,2★] Shirley Ho,[2,3★] Peter E. Freeman,[1,2] Christopher R. Genovese[1,2] and Larry Wasserman[1,2]

[1] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2] *McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[3] *Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

Accepted 2015 August 24. Received 2015 July 18; in original form 2015 April 1

ABSTRACT

The detection and characterization of filamentary structures in the cosmic web allows cosmologists to constrain parameters that dictate the evolution of the Universe. While many filament estimators have been proposed, they generally lack estimates of uncertainty, reducing their inferential power. In this paper, we demonstrate how one may apply the subspace constrained mean shift (SCMS) algorithm (Ozertem & Erdogmus 2011; Genovese et al. 2014) to uncover filamentary structure in galaxy data. The SCMS algorithm is a gradient ascent method that models filaments as density ridges, one-dimensional smooth curves that trace high-density regions within the point cloud. We also demonstrate how augmenting the SCMS algorithm with bootstrap-based methods of uncertainty estimation allows one to place uncertainty bands around putative filaments. We apply the SCMS first to the data set generated from the Voronoi model. The density ridges show strong agreement with the filaments from Voronoi method. We then apply the SCMS method data sets sampled from a P3M *N*-body simulation, with galaxy number densities consistent with SDSS and *WFIRST-AFTA*, and to LOWZ and CMASS data from the Baryon Oscillation Spectroscopic Survey (BOSS). To further assess the efficacy of SCMS, we compare the relative locations of BOSS filaments with galaxy clusters in the redMaPPer catalogue, and find that redMaPPer clusters are significantly closer (with p-values $<10^{-9}$) to SCMS-detected filaments than to randomly selected galaxies.

Key words: methods: data analysis – methods: statistical – cosmology: observations – large-scale structure of Universe.

1 INTRODUCTION

Observations of the local universe made over the last four decades show that on megaparsec scales, matter is distributed in web-like structures – clusters, filaments, sheets, and voids – that arise naturally from the non-linear evolution of initially small density fluctuations (Peebles 1980; Bond, Kofman & Pogosyan 1996; Jenkins et al. 1998; Colberg, Krughoff & Connolly 2005; Springel et al. 2005; Dolag et al. 2006). Of particular interest to us are the filaments, one-dimensional structures that connect galaxy clusters and form at the boundaries of empty voids. Filaments are of interest for several reasons. The detection and characterization of filaments at a range of redshifts provides a means by which cosmologists can constrain theories of the universe's evolution (Bond et al. 1996; Zhang et al. 2009, 2013). Filaments also influence the shape, angular momentum, and peculiar velocities of dark matter haloes (Hahn et al. 2007b,a; Paz, Stasyszyn & Padilla 2008; Hahn et al. 2009; Zhang et al. 2009; Jones, van de Weygaert & Aragón-Calvo 2010; Zhang et al. 2013; Forero-Romero, Contreras & Padilla 2014), as well as the intrinsic alignments and luminosities of nearby galaxies (Clampitt et al. 2014; Codis et al. 2014; Guo, Tempel & Libeskind 2015).

As the review of Cautun et al. (2014) amply demonstrates, the detection of filamentary structure is a non-trivial problem for which many solutions have been proposed. These solutions include methods that examine the Hessian matrix of the galaxy density field, such as the multiscale morphology filter (MMF; Aragón-Calvo et al. 2007; Aragón-Calvo, van de Weygaert & Jones 2010a) and NEXUS and NEXUS+ (Cautun, van de Weygaert & Jones 2013), as well as segmentation-based methods, such as the Candy model (Stoica et al. 2005; Stoica, Martinez & Saar 2007), the skeleton (Novikov, Colombi & Doré 2006), the Spine method (Aragón-Calvo et al. 2010b), and DisPerSE models (Sousbie 2011), and the path density method (Genovese et al.

★E-mail: yenchic@andrew.cmu.edu (Y-CC); shirleyh@andrew.cmu.edu (SH)

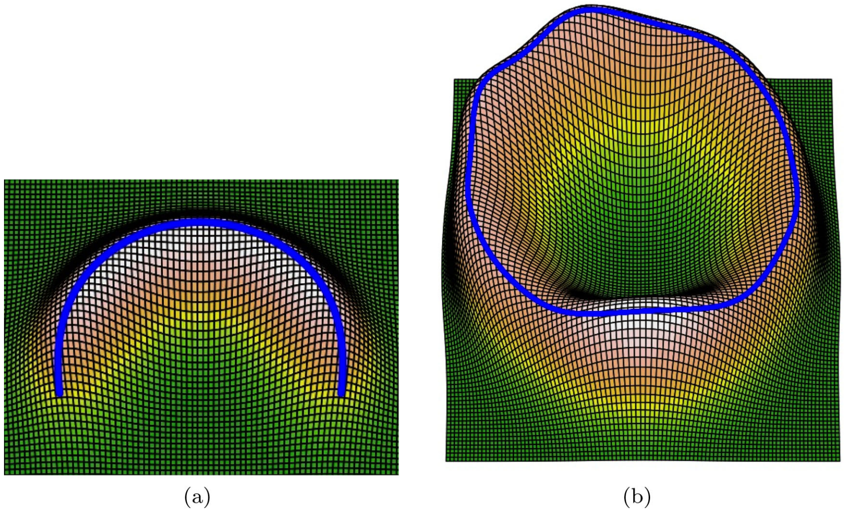


Figure 1. Examples of ridges (blue curves) in a smooth function.

2009). While all of these methods provide estimates of filamentary structure, none provide an assessment of estimator uncertainty. The fact that filament estimates are random sets presents a significant challenge to the construction of valid uncertainty measures (Molchanov 2005).

In this paper, we introduce a new method for filament detection based on the Subspace Constrained Mean Shift (SCMS) algorithm of Ozertem & Erdogmus (2011). The statistical properties of SCMS were studied in Genovese et al. (2014). The mathematical properties of density ridges and the statistical consistency of SCMS are discussed in Eberly (1996); Genovese et al. (2014), and Chen, Genovese & Wasserman (2015a), respectively, while Chen et al. (2015a) introduce an uncertainty measure to the ridge formalism that allows one to quantitatively assess, in the context of the current paper, putative cosmic filaments.

In Section 2, we describe the SCMS algorithm and the methods we use to assess the uncertainty of its filament estimates. In Section 3, we apply SCMS, first to a P3M *N*-body simulation output (Trac, Cen & Mansfield 2015), and then to low-redshift ($0.235 \le z \le 0.240$) and high-redshift ($0.530 \le z \le 0.535$) data collected by the Baryon Oscillation Spectroscopic Survey (BOSS), which was released as part of SDSS Data Release 11. We also demonstrate the consistency between filaments detected by SCMS and galaxy clusters listed in the redMaPPer catalogue. In Section 4, we summarize our results and offer possible avenues for future methodological development. In Appendix A, we provide further detail on how to optimally select values for the tuning parameters of the SCMS algorithm, while in Appendix B, we apply the algorithm to labelled simulated data generated via the Voronoi model of van de Weygaert (1994) to show that it preferentially detects structures labelled as filaments. In a second paper, we will provide a full catalogue of filaments detected in SDSS data.

2 SCMS: ALGORITHM

2.1 Density ridge formalism

Assume that we observe n galaxies with locations $X_1, \ldots, X_n$ that are d-dimension points; for data from typical astronomical surveys, $d = 2$ (if the galaxies are constrained to a redshift shell) or $d = 3$. We model $X_1, \ldots, X_n$ as random variables sampled from an unknown density function p.

Formally, a *density ridge* (Eberly 1996; Ozertem & Erdogmus 2011; Chen, Genovese & Wasserman 2014; Chen et al. 2015a; Genovese et al. 2014) of p is defined as follows. Let $g(x) = \nabla p(x)$ and $H(x)$ be the gradient and Hessian, respectively, of $p(x)$ and let $v_1(x), \ldots, v_d(x)$ be the eigenvectors of the Hessian matrix, with associated eigenvalues $\lambda_1(x) \ge \lambda_2(x) \ge \ldots \ge \lambda_d(x)$. We define $V(x)$ to be the matrix of all eigenvectors orthogonal to the first, $[v_2(x), \ldots, v_d(x)]$, and the ridge set R as

$$R \equiv \mathrm{Ridge}(p) = \{x : G(x) = 0,\ \lambda_2(x) < 0\}, \tag{1}$$

where

$$G(x) = V(x)V(x)^T g(x) \tag{2}$$

is the projected gradient. The fact that ridges have projected a gradient of 0 (and second eigenvalues being negative) means that ridges are local maximums in the subspace spanned by eigenvectors $v_2(x), \ldots, v_d(x)$. When p is smooth and the *eigengap*

$$\beta(x) = \lambda_1(x) - \lambda_2(x) \tag{3}$$

is positive, the ridges have the properties of filaments, i.e. smooth curve-like structures with high density (see Fig. 1). Note that R will include modes of the density p which, in the context of cosmic filament detection, means that R contains both filaments and galaxy clusters. Also note that density ridges are more general objects than the skeleton models proposed in Novikov et al. (2006); Sousbie et al. (2008) and the Spine method (Aragón-Calvo et al. 2010b). Essentially, when $d = 2, 3$, density ridges are the same as skeletons.

Compared with other models, density ridges adapt information from both gradient and Hessian matrix of density. In contrast, MMF (Aragón-Calvo et al. 2007, 2010a), NEXUS, and NEXUS+ (Cautun et al. 2013) only use the information of second derivatives (they define filaments as the regions with $\lambda_2(x) < 0$ and $\lambda_1(x) \approx \lambda_2(x) > \lambda_3(x)$). DisPerSE models (Sousbie 2011) define filaments as those gradient flows that start from saddle points and end up at local maximums, which utilize only the first derivatives.

An attractive feature for the density ridge model is that the statistical theory for consistently estimating the density ridge has been well established (Chen et al. 2014, 2015a; Genovese et al. 2014).

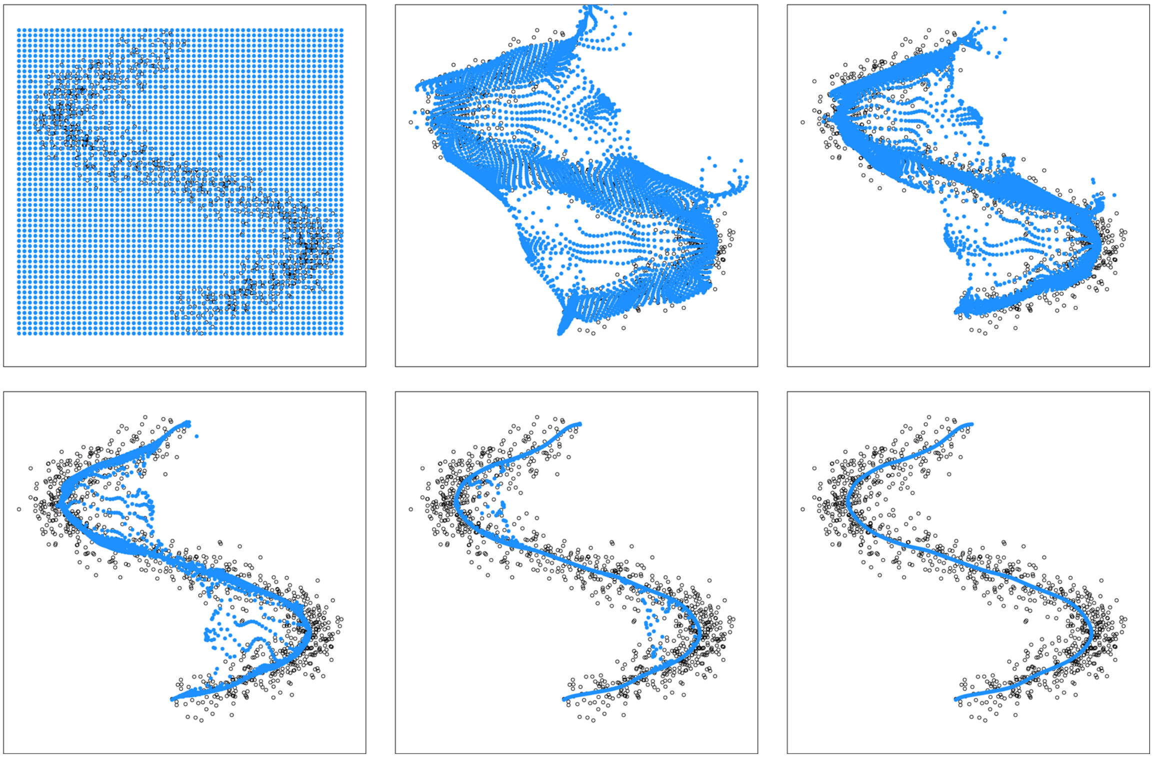


Figure 2. Pictorial overview of the SCMS algorithm (Step 4 in Algorithm 1). Each point in an initially uniform mesh (the blue dots in the top-left panel) is moved to the closest density ridge (bottom right). The top-middle, top-right, bottom-left, bottom-middle, and bottom-right panels indicate the locations of the mesh points after 1, 2, 4, 8, and 16 iterations of the algorithm, respectively.

We also use N-body simulation to verify the convergence of density ridges when we subsample different number of galaxies (see Section 3.2).

2.2 SCMS: filament detection

The algorithm consists of three steps described below and listed in Algorithm 1. The first is to estimate the underlying density function $p(x)$ given $X_1, \ldots, X_n$, the observed locations of galaxies. We use the standard kernel density estimator (see e.g. Wasserman 2006):

$$\widehat{p}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{h}\right), \tag{4}$$

where $K(\cdot)$ is the smoothing kernel (e.g. a Gaussian), $\|x - X_i\|$ is the Euclidean distance between the point x, and the ith galaxy location X_i, and h is the smoothing bandwidth (the selection of which is discussed in Appendix A).

In the second step, we denoise by applying a threshold to the estimated density function $\widehat{p}(x)$ to eliminate the effect that galaxies in low-probability density regions, i.e. where $\widehat{p}(x) < \tau$, would have on filament estimation. How one selects τ is also discussed in Appendix A. The denoising step is not part of the original SCMS algorithm but is important to increase its statistical power in low-density regions (see Fig. 3). We note that a thresholding step is included in several filament-detection algorithms, including those of e.g. Novikov et al. (2006) and Sousbie (2011).

For the final step, given a set of galaxies in high-density regions, we apply the original version of the SCMS (Ozertem & Erdogmus 2011) to detect filamentary structures. Given a point x on a defined, uniform mesh, SCMS moves it according to an 'estimated projected gradient' given by

$$\widehat{G}(x) = \widehat{V}(x)\widehat{V}(x)^{\mathrm{T}}\widehat{g}(x), \tag{5}$$

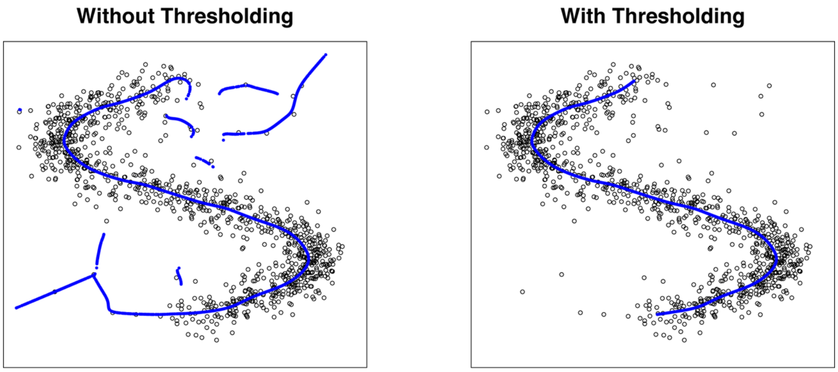
where $\widehat{V}(x)$, $\widehat{g}(x)$ are estimates of the quantities $V(x)$, $g(x)$ that we define above in Section 2.1. One may view this procedure as estimating a ridge set R by applying the Ridge operator to $\widehat{p}$:

$$\widehat{R} = \mathrm{Ridge}(\widehat{p}). \tag{6}$$

Essentially, $\widehat{R}$ is very similar to the filaments defined in Sousbie et al. (2008), Bond, Strauss & Cen (2010), and Choi et al. (2010). Note that a putative filament is, in the context of this algorithm, a set of points and not a one-dimensional curve. In step 4 of Algorithm 1, We further describe how we apply SCMS. In Fig. 2, we illustrate the application of SCMS to uniform mesh of points, and in Fig. 3 we demonstrate the importance of the thresholding step: the left and right panels show putative filaments detected without and with thresholding, respectively. We observe that thresholding greatly decreases the rate of false filament detection. Fig. 4 presents an example for applying algorithm 1.

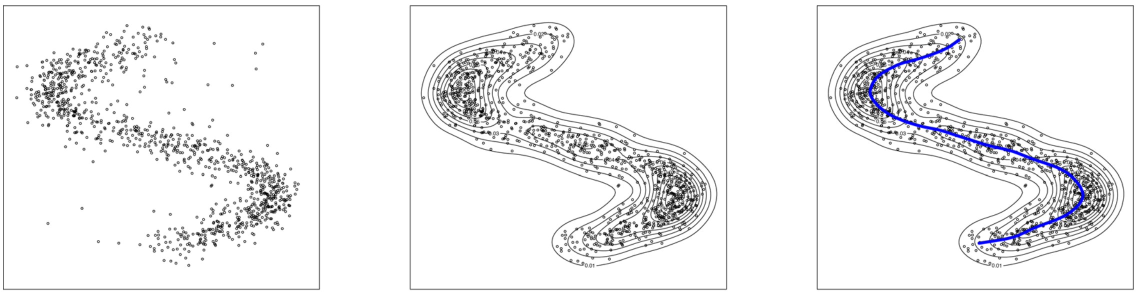
2.3 SCMS: filament uncertainty estimation

We quantify the uncertainty in the filament estimates produced by SCMS using the concept of *local uncertainty* (Chen et al. 2015a). The local uncertainty in an estimated filament $\widehat{R}$ at a point x on the



(a) Without thresholding (b) Thresholding

Figure 3. An example of the comparison of SCMS with and without noise removal. This is a simple simulated data set with clutter noise. As can be seen easily, thresholding the density removes problems of clutter noise.



(a) (b) (c)

Figure 4. An example of the application of SCMS. (a) The original data. (b) Contour plot showing the kernel density estimate of the density p. (c) The ridge estimate (blue curve). Note that in (b), we remove points where the estimated density is less than a threshold τ.

Algorithm 1 SCMS (Subspace Constrained Mean Shift)

Input: Data $\{X_1, \ldots, X_n\}$. Smoothing bandwidth h. Threshold τ.

Step 1. Compute the density estimator $\hat{p}(x)$ via equation (4).

Step 2. Select a mesh $\mathcal{M}$ of points. By default, we can take $\mathcal{M} = \{X_1, \ldots, X_n\}$.

Step 3. Thresholding: remove $m \in \mathcal{M}$ if $\hat{p}(m) < \tau$. Let the remaining mesh points be denoted $\mathcal{M}'$.

Step 4. For each $x \in \mathcal{M}'$, perform the following subspace constrained mean shift until convergence:

Step 4-1. For $i = 1, \ldots, n$, compute

$$\mu_i = \frac{x - X_i}{h^2}, \qquad c_i = K\left(\frac{x - X_i}{h}\right)$$

Step 4-2. Compute the Hessian matrix

$$H(x) = \frac{1}{n}\sum_{i=1}^{n} c_i \left(\mu_i \mu_i^T - \frac{1}{h^2}\mathbf{I}\right). \tag{7}$$

Step 4-3. Perform spectral decomposition on $H(x)$ and compute $V(x) = (v_2(x), \ldots, v_d(x))$, the eigenvectors corresponding to the smallest $d - 1$ eigenvalues.

Step 4-4. Update $x \longleftarrow V(x)V(x)^T m(x) + x$ until convergence, where

$$m(x) = \frac{\sum_{i=1}^{n} c_i X_i}{\sum_{i=1}^{n} c_i} - x \tag{8}$$

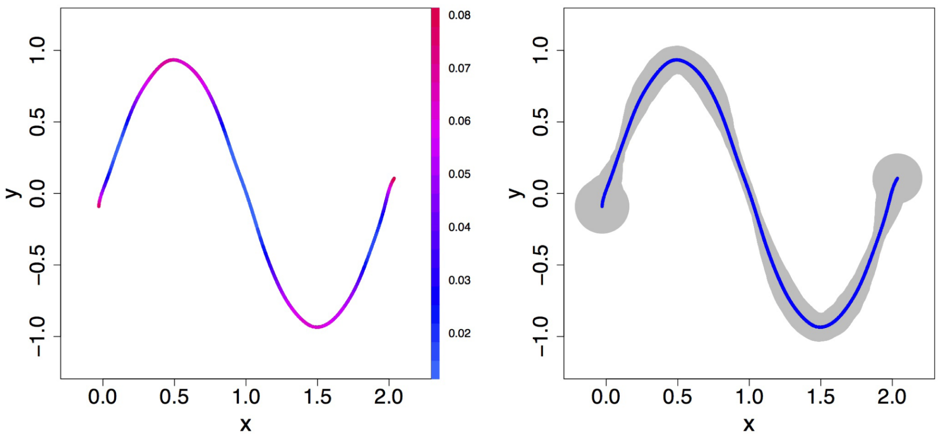
is called the mean shift vector.

Output: The collection of all remaining points.

true filament R is the expected distance between x and the closest point to x on $\widehat{R}$. This is denoted by $\rho(x)$ and is given by

$$\rho(x) = \begin{cases} \sqrt{\mathbb{E}\left[d^2_{\widehat{R}}(x)\right]} & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $d_{\widehat{R}}(x) = \min\{\|x - y\| : y \in \widehat{R}\}$ and the notation $\mathbb{E}[\cdot]$ denotes the expected value operator. $\rho(x)$ is the radius of a local *confidence ball* that surrounds the point x: the more uncertain the true location of the estimated filament, the larger the value of $\rho(x)$. We estimate $\rho(x)$, which is defined as a function of the unknown density field p and the unknown filament set R, by utilizing



(a) Local uncertainty. (b) Uncertainty band.

Figure 5. An illustration of the uncertainty measure for SCMS. In (a), we display the uncertainty measures with different colour (red: highly uncertain). The unit to the colour is the same as x and Y axis. In (b), we show the uncertainty measures by a grey region around the filament (blue). Note that this shows that the SCMS has more uncertainty measures around the highly curved regions and the end points.

Algorithm 2 Uncertainty Measure for SCMS

Input: Data $\{X_1, \ldots, X_n\}$. Smoothing bandwidth h. Threshold τ.

Step 1. Perform SCMS on $\{X_1, \ldots, X_n\}$ to detect filaments; denote the estimated filaments by $\hat{R}$...

Step 2. Generate B bootstrap samples: $X_1^{*(b)}, \ldots, X_n^{*(b)}$ for $b = 1, \ldots, B$.

Step 3. For each bootstrap sample, apply SCMS which yields $\hat{R}^{*(b)}$ for $b = 1, \ldots, B$.

Step 4. For each $x \in \hat{R}$, calculate $\rho^2_{(b)}(x) = d^2(x, \hat{R}^{*(b)})$, $b = 1, \ldots, B$.

Step 5. Compute $\hat{\rho}(x) = \left[\text{mean}\{\rho_1^2(x), \ldots, \rho_B^2(x)\}\right]^{1/2}$.

Output: $\hat{\rho}(x)$.

bootstrap resampling. Fig 5 presents an example for uncertainty measures.

In this paper, we consider both the original version of bootstrap (Efron 1979) and the smooth bootstrap (SB). The *SB* (see e.g. Silverman & Young 1987) is a variant of the bootstrap that is useful in functional estimation problems in which the bootstrap sample is drawn from the estimated density $\widehat{p}$ instead of the original data. When the smoothing kernel is a bivariate Gaussian, we generate the SB sample via the following two steps.

1. Generate the bootstrap sample.
2. Add independent and identically distributed Gaussian noise with variance h^2.

Unlike the bootstrap, the SB takes into account both the variance and the bias of filament estimation, but with less precision in variance estimation with respect to the bootstrap.

Assume we generate B bootstrap samples, and each of them is denoted as $\{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$, $b = 1, \ldots, B$. For each bootstrap sample, say $X_1^{*(b)}, \ldots, X_n^{*(b)}$, we compute the density estimate $\widehat{p}^{*(b)}$, the ridge estimate $\widehat{R}^{*(b)} = \text{Ridge}(\widehat{p}^{*(b)})$, and the confidence ball radii $\rho_{(b)}(x)$ for all $x \in \widehat{R}$. We estimate $\rho(x)$ by adding the B radius estimates in quadrature:

$$\widehat{\rho}(x) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\rho^2_{(b)}(x)}, \tag{10}$$

In Algorithm 2, we outline the computational steps that one must follow to derive $\widehat{\rho}(x)$.

Note that calculating the uncertainty measure is not part to the SCMS algorithm – we can detect filaments without using the uncertainty measure. However, this uncertainty measure is a feature that SCMS filaments have. This measure has a geometric interpretation and can be consistently estimated. See Chen et al. (2015a) for more involved discussion. Note that other filament finders do have such a statistically consistent error measurement.

2.4 SCMS: boundary bias

When computed with a kernel density estimator as in equation (4), SCMS filament estimates suffer from *boundary bias* within approximately two bandwidths of the edge of the observation region. This is a systematic deviation from the true filament caused by the density estimator averaging over a region where no data can be observed, and it can degrade the confidence band coverage probabilities near the boundary. One remedy for boundary bias is to include additional data immediately outside of the region of interest. Including galaxies within $2h$ of the boundaries eliminates most of the boundary bias, since very little of the volume under a bivariate Gaussian kernel lies beyond that point. If one cannot include additional data points outside the boundaries (for instance, due to overall survey limits), then one must be careful when interpreting filaments detected near the boundaries.

2.5 Filament coverage

Here, we introduce some useful geometric concepts about coverage. Given two sets A and B. The *coverage* of B by A is defined as

$$\mathsf{Cov}_B(A) = \frac{\text{Number of points in } (A \cap B)}{\text{Number of points in } B}. \tag{11}$$

Note that when A and B are curves, they will contain infinite number of points. In this case, we will replace 'number of points in' by 'the length of'. Similarly, we can define the coverage of A by B as $\mathsf{Cov}_A(B)$.

Given two collections of filaments R_1 and R_2, since R_1 and R_2 are curves so that they may not intersect each other in general so that the coverage is 0. Thus, instead of directly compute their coverage, we consider a flatten version of R_1 (and R_2, respectively). We define

$$R_1 \oplus r = \{x : d(x, R_1) \leq r\} \tag{12}$$

as the r-flatten set of R_1. Then we define the coverage of R_2 by R_1 as a function of r as

$$\mathsf{Cov}_{R_2}(r; R_1) = \frac{\text{Number of points in } (R_2 \cap (R_1 \oplus r))}{\text{Number of points in } R_2}. \tag{13}$$

Similarly, we can define $\mathsf{Cov}_{R_1}(r; R_2)$. The two functions $\mathsf{Cov}_{R_1}(r; R_2)$ and $\mathsf{Cov}_{R_2}(r; R_1)$ contain information about the similarity between R_1 and R_2.

In simulation, we are able to define true filaments, say R_{true}, and we will have an estimate filament, denoted as $\widehat{R}_n$. Then we call the quantity $\mathsf{Cov}_{R_{\text{true}}}(r; \widehat{R}_n)$ the *true positive coverage* (ratio of true filaments being covered by estimated filaments) and we call $1 - \mathsf{Cov}_{\widehat{R}_n}(r; R_{\text{true}})$ the *false positive coverage* ($\mathsf{Cov}_{\widehat{R}_n}(r; R_{\text{true}})$ is the ratio of estimated filament being covered by truth so that 1 minus this ratio is the ratio of false positive).

Combining the uncertainty measures and the coverage, we can study the properties of the *uncertainty band.*. An uncertainty band for a detected filament is simply the union of the confidence balls computed for each point on the filament, i.e.

$$\widehat{U}(k) = \widehat{R} \oplus k\widehat{\rho} \equiv \bigcup_{x \in \widehat{R}} B(x, k\widehat{\rho}(x)), \tag{14}$$

where $B(x, r) = \{y: \|x - y\| \leq r\}$ represents the set of points within a ball centred at x and with radius r. Denote the region within the uncertainty band as A. The coverage for A is then

$$\begin{aligned} \mathsf{FCov}(A) &= \mathsf{Cov}_{R_{\text{true}}}(A) \\ &= \frac{\text{Number of points in } (\mathrm{A} \cap \mathrm{R}_{\text{true}})}{\text{Number of points in } \mathrm{R}_{\text{true}}}. \end{aligned} \tag{15}$$

One can think of $\mathsf{FCov}(A)$ as the true positive coverage using a set A. For instance, if $\mathsf{FCov}(A) = 0.8$, then on average, 80 per cent of the points on any given true filament lie within its associated uncertainty band, and 20 per cent lie outside the band. This interpretation of coverage differs from the standard interpretation of confidence band coverage, thus motivating our use of the term 'uncertainty band' instead of 'confidence band'.

3 SCMS: APPLICATIONS

3.1 Voronoi data set

To show the effectiveness of capturing filaments, we compare the SCMS filaments (density ridges) to the filaments in the Voronoi model. The Voronoi model (van de Weygaert 1994) applies Voronoi tessellation to compute a density estimate for galaxies as well as the curvature of that estimate. Given a curvature estimate, the Voronoi method assigns a class label to each galaxy, indicating the type of large-scale structure to which to associate the galaxy. There are four possible classes: `cluster`, `filament`, `wall`, and `void`.

We use the SCMS algorithm to analyse a simulated data set (256^3 galaxies, each with a class label, that span a $100 \times 100 \times 100$ Mpc^3 box) generated with the Voronoi model (M. A. Aragón-Calvo, private communication). Fig. 6 shows a comparison between our density ridges (blue curves) and galaxies with different class labels (brown: `cluster`; red: `filament`; green: `wall`; pink: `void`). The two methods generate remarkably similar results: Voronoi clusters (i.e. galaxies labelled `cluster`) occur at the intersection points of density ridges; Voronoi filaments surround the density ridges; and Voronoi walls span surfaces on which the density ridges lay.

To further quantify the association between density ridges and each Voronoi model class, we study their projection distances on to each other. Note that the distribution of projection distances is related to filament coverage; further discussion of this may be found in Chen et al. (2015b). Fig. 7 displays the distributions of projection distances. In both the panels, we see that the distribution for ridges versus the Voronoi filaments peaks at distances $\lesssim 1\ h^{-1}$ Mpc. This indicates that the density ridges and the Voronoi filaments are very similar. On the other hand, the projection distances from the density ridges increases as we consider clusters, walls, and voids; the distributions exhibit increasing positive skewness.

3.2 P3M *N*-body simulation

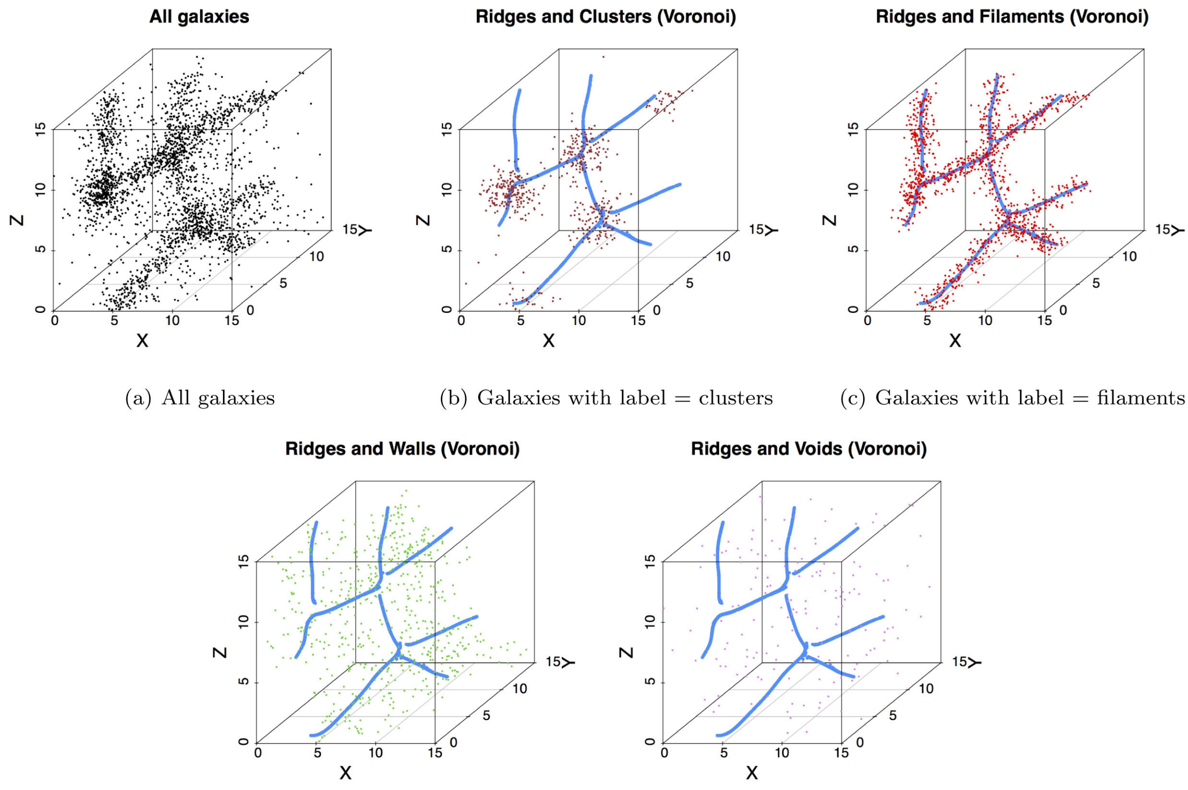
To further demonstrate the efficacy of SCMS, we apply it to P3M *N*-body simulations from Trac et al. (2015), which assume a ΛCDM cosmology with $\Omega_{\mathrm{m}} = 0.3$, $\Omega_l = 0.7$, $\Omega_b = 0.045$, $h = 0.7$, $\sigma_8 = 0.8$ and $n_s = 0.96$. Each side of the simulation box is of length 1 Gpc h^{-1}, and each contains 2048^3 particles.

In Fig. 8, we demonstrate that, as sample size increases, SCMS outputs filament estimates that are closer to the true filaments (defined by the true density function); the uncertainty measures capture SCMS errors due to the sampling variability. We take a slice of the full simulation data ($x, y \in [125, 375]$ Mpc h^{-1} and $z \in [100, 105]$ Mpc h^{-1}) and smooth the data with smoothing bandwidth $h = 5$ (recommended by the selection rule in Appendix A with $A_0 = 0.4$) to get the density function and the filaments (cyan curves). Fig. 8(a) shows a contour plot for the density function. The original sliced data contains 88 406 points (grey dots). We downsample to get three different subsamples; each contains 250 / 2500/10 000 particles. For each subsample (black dots), we apply SCMS to detect filaments (blue curves). Note that the convergence phenomena of Fig. 8 are further quantified by the true positive and false positive coverage plot in Fig. 9.

Note that the sparsest subsample $n = 250$ has a galaxy number density 5.56×10^{-4} Mpc^{-3} which is similar to the number density observed in SDSS CMASS data ($\sim 4 \times 10^{-4}$ Mpc^{-3}). The future survey *Wide-Field Infrared Survey Telescope (WFIRST)*,[1] a NASA mission with science objectives in exoplanet exploration, dark energy research, and galactic and extragalactic surveys, will observe a number density similar to the $n = 2500$ subsample ($\sim 5.56 \times 10^{-3}$ Mpc^{-3}).

We show the uncertainty measures and filament coverage for $n = 2500$ in Fig. 10. We plot filament coverage for confidence regions $\widehat{U}(k)$ for $k \in (0, 3)$ in Fig. 10(a), where $n = 250$ and 2500, and where $\widehat{\rho}$ is estimated by both the bootstrap (BT) and the SB. This range contains sample sizes that are in line with both CMASS ($n \approx 250$) and *WFIRST* ($n \approx 2500$) data. We observe that filament coverage is, as noted above, sensitive to the sample size n and that the SB provides considerably more conservative confidence bands, particularly for $k \lesssim 2$. The grey regions displayed in Fig. 10(b) are the SB confidence regions $\widehat{U}(1)$, which we estimate contain 85 per cent of the true filaments (cyan curves).

[1] http://wfirst.gsfc.nasa.gov/



(a) All galaxies (b) Galaxies with label = clusters (c) Galaxies with label = filaments

(d) Galaxies with label = walls (e) Galaxies with label = voids

Figure 6. A comparison between density ridges and Voronoi model. In each panel, the blue curves are density ridges using all galaxies. Panels (b)–(e) display the comparison of density ridges to the Voronoi clusters, filaments, walls and voids. In the panel (c), we see a remarkable similarity between density ridges and the Voronoi filaments.

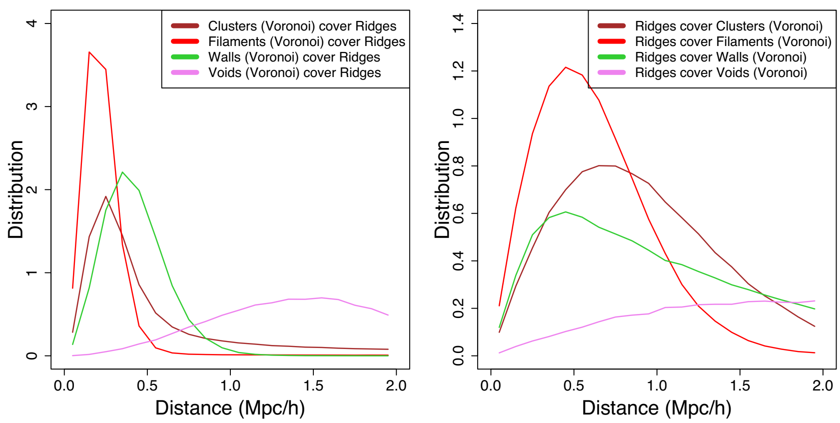
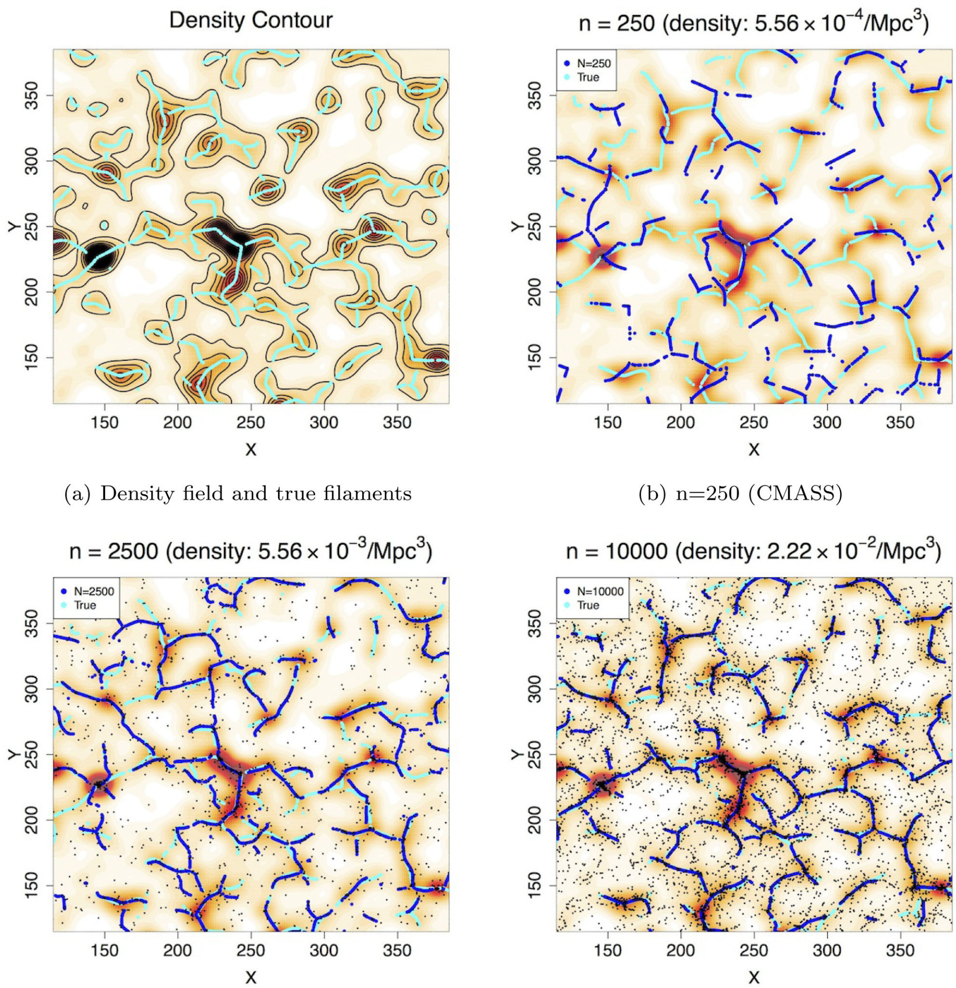


Figure 7. The distributions for projection distances from Voronoi-model-derived structures on to density ridges (left-hand panel) and vice-versa (right-hand panel). Both panels indicate that density ridges trace structures most similar to Voronoi filaments.

Fig. 11 illustrates the effect of boundary bias in the $n = 2500$ subsample by comparing the estimates and uncertainties with padded and unpadded data near the boundary. Panels (a) and (b) show the boundary bias. Note that the red curves are filaments estimated by using only points within the boundary (given by the orange rectangle). The blue curves are filaments detected by SCMS with boundary points (i.e. points outside the orange rectangle). As can be seen, the estimation of filaments without boundary data (red curves) becomes more inaccurate as we approach the boundary. The boundary bias occurs for filaments with distances less than 10 Mpc h^{-1} (2 times smoothing parameter h) to the boundaries. The uncertainty measures also show the influence of boundary bias. Fig. 11(c) and (d) exhibit the uncertainty measures for filaments estimated with and without boundary points. As expected, the uncertainty measures in the panel (d) increase as we move close to the boundary.



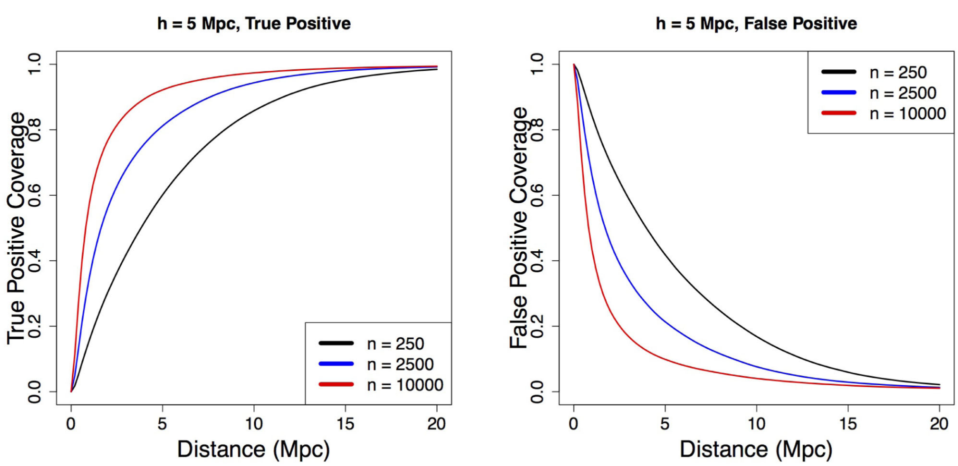
(a) Density field and true filaments

(b) n=250 (CMASS)

(c) n=2500 (*WFIRST*)

(d) n=10000

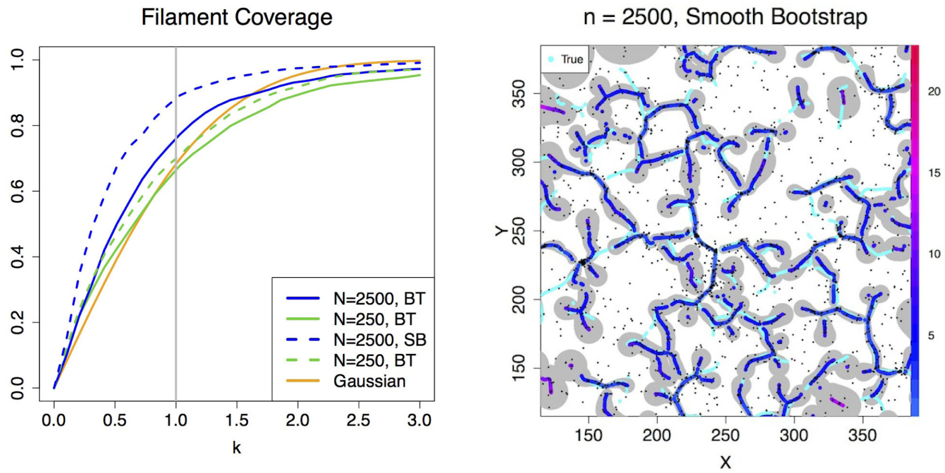
Figure 8. A simulated example to show the consistency of SCMS. This data is a slice of an *N*-body simulation in a box; the unit of *X* and *Y* axes is Mpc h^{-1}. We take a slice with width 5 Mpc h^{-1}. The original sample contains 88 406 particles. The colour contour is the galaxy density field from the original sample with smoothing parameter $h = 5$ and the true filaments (cyan) are density ridges of this density field. We subsample under various sizes. The blue curves are estimated filaments based on the subsample (black dots). One can see a clear pattern; as sample size (for the subsample) increases, the estimated filaments are closer to the true filaments. See Section 3.2 for more details.



(a) True positive coverage.

(b) False positive coverage

Figure 9. True positive and false positive coverage. As can be seen, for all distances, the true positive coverage increases with sample size, whereas the false positive coverage decreases.



(a) Filament coverage. (b) $N = 2500$, smooth bootstrap

Figure 10. Filament coverage based on the uncertainty measure. (a) The filament coverage $\mathsf{FCov}(\widehat{U}(k))$ as a function of k (x-axis). We also provide the coverage for Gaussian distribution (probability being within $k\sigma$ to the centre of Gaussian) as a reference. (b) Visualizing the uncertainty by colour and a confidence set for the subsample with $n = 2500$ with the uncertainty measure estimated via the SB. The cyan curves are the true filaments. Note that the grey regions are $\widehat{U}(k)$, $k = 1$, equivalent to the error bar for 1σ, based on the SB estimate. From panel (a), we know that the grey regions contain about 85 percent true filaments (cyan curves). The unit to the colour in uncertainty band is Mpc, the same as X and Y axes.

3.3 Sloan digital sky survey

3.3.1 Data

We further demonstrate the efficacy of SCMS by applying it to data from Data Release 12 (Alam et al. 2015) of the Sloan Digital Sky Survey (SDSS; York et al. 2000). Together, SDSS I, II (Abazajian et al. 2009), and III (Eisenstein et al. 2011) used a drift-scanning mosaic CCD camera (Gunn et al. 1998) to image over one third of the sky (14 555 deg^2) in five photometric bandpasses (Fukugita et al. 1996; Smith et al. 2002; Doi et al. 2010) to a limiting magnitude of r $\simeq$ 22.5, using the dedicated 2.5-m Sloan Telescope (Gunn et al. 2006) located at Apache Point Observatory in New Mexico. The imaging data were processed through a series of pipelines that perform astrometric calibration (Pier et al. 2003), photometric reduction (Lupton et al. 2001), and photometric calibration (Padmanabhan et al. 2008). All of the imaging was reprocessed as part of SDSS Data Release 8 (Aihara et al. 2011).

The Baryon Oscillation Spectroscopic Survey (BOSS) has obtained spectra and redshifts for 1.35 million galaxies over a footprint covering 10 000 deg^2. These galaxies are selected from the SDSS (Aihara et al. 2011) imaging and are being observed together with 160 000 quasars and approximately 100 000 ancillary targets. The targets are assigned to tiles of diameter 3° using a Blanton et al. (2003) algorithm that is adaptive to the density of targets on the sky (Blanton et al. 2003). Spectra are obtained using the double-armed BOSS spectrographs (Smee et al. 2013). Each observation is performed in a series of 900-s exposures, integrating until a minimum signal-to-noise ratio is achieved for the faint galaxy targets. This ensures a homogeneous data set with a high-redshift completeness of more than 97 per cent over the full survey footprint. Redshifts are extracted from the spectra using the methods described in Bolton et al. (2012). A summary of the survey design appears in Eisenstein et al. (2011), and a full description is provided in Dawson et al. (2013).

BOSS selects two classes of galaxies to be targeted for spectroscopy using SDSS (Aihara et al. 2011) imaging: 'LOWZ' and 'CMASS' (we refer the reader to Anderson et al. 2014 for further description of these classes). For the LOWZ sample, the effective redshift is $z_{\rm eff} = 0.32$, slightly lower than that of the SDSS-II luminous red galaxies (LRGs) as we place a redshift cut $z < 0.43$. The CMASS selection yields a sample with a median redshift $z = 0.57$ and a stellar mass that peaks at $\log_{10}(M/\mathrm{M}_{\odot}) = 11.3$ (Maraston et al. 2013). Most CMASS targets are central galaxies residing in dark matter haloes of mass $\sim 10^{13}\, h^{-1}\, \mathrm{M}_{\odot}$.

We test SCMS using two slices of data: at low and high redshift. The low-z data set comprises 1158 galaxies in the volume

$$135^\circ \le \mathrm{RA} \le 175^\circ, 5^\circ \le \delta \le 45^\circ, 0.235 \le z \le 0.240$$

while the high-z data set lies in the volume

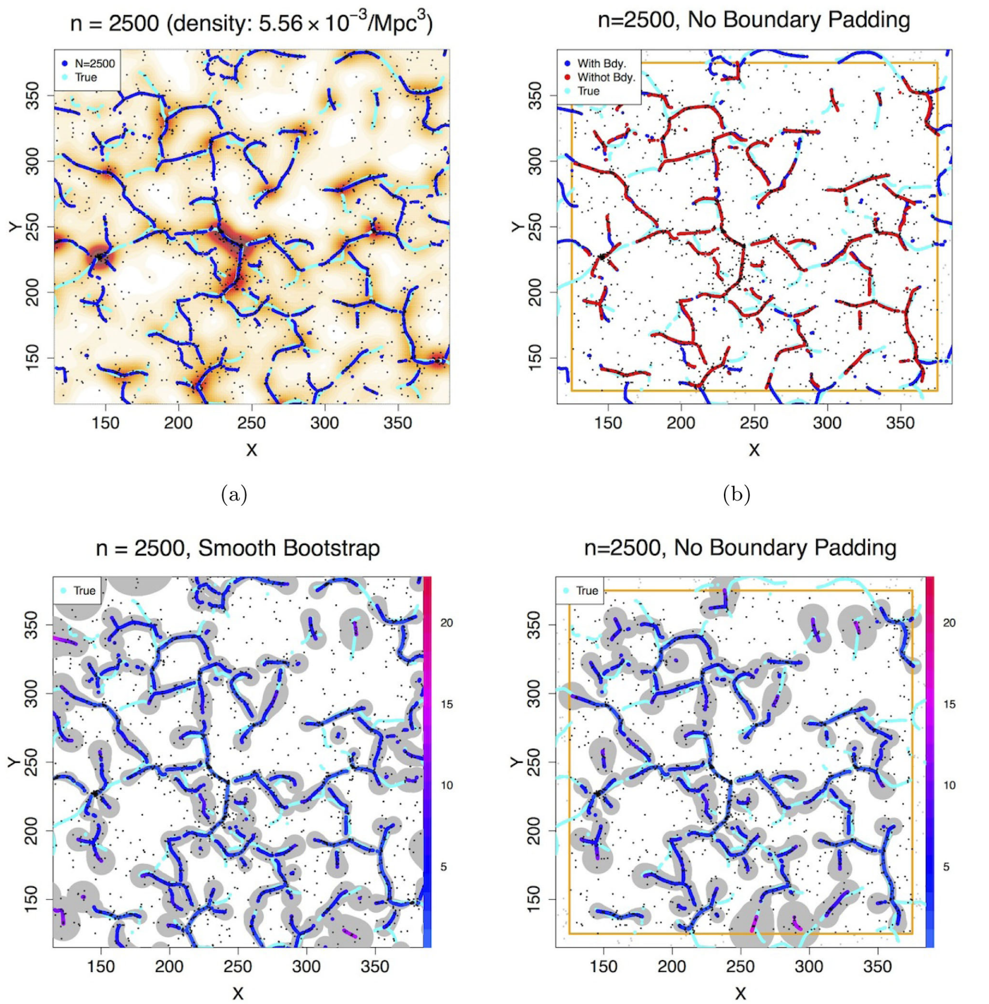
$$135^\circ \le \mathrm{RA} \le 175^\circ, 5^\circ \le \delta \le 45^\circ, 0.530 \le z \le 0.535$$

and contains 4678 galaxies. Both samples have a very thin redshift range $\Delta z = 0.005$ (the corresponding comoving distance is around 14–21 Mpc) so that their constituent galaxies may be viewed as lying on a two-dimensional surface with coordinates (RA,δ).

There are two principal reasons for our choice to perform a two-dimensional analysis of the SDSS data. The first is that there is too large a change in the number density of detected galaxies over the SDSS redshift range. The SCMS algorithm incorporates kernel density estimation to locate density ridges, and KDE requires a fixed smoothing parameter h. However, in low-density regions, h should be large to obtain reliable results, while in high-density regions, h has to be small so as to not oversmooth the point cloud. The second reason is that when $z > 0.2$, the number density is very low, and performing a three-dimensional analysis will produce results with large statistical errors due to the small sample size. Lower dimensional analyses result in decreased statistical error; see e.g. Wasserman 2006.

3.3.2 Results

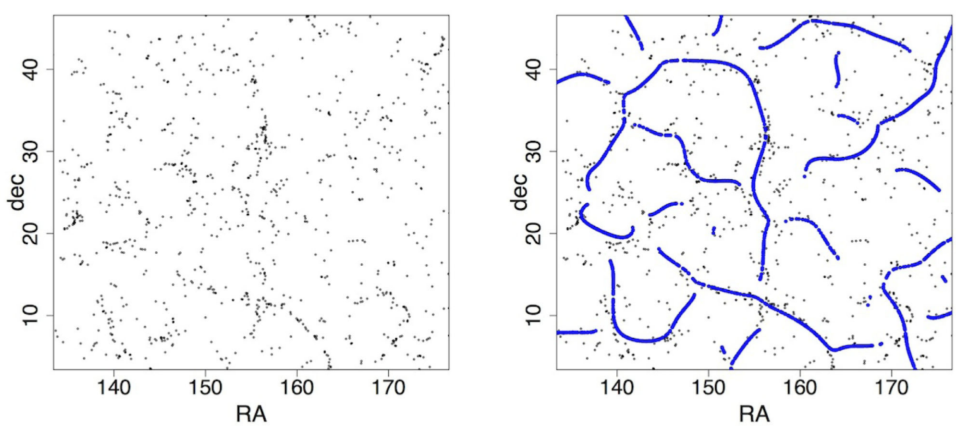
We apply SCMS to the low-z data with smoothing bandwidth $h = 2\overset{\circ}{.}50$ (41.8 Mpc) and threshold level $\tau = 1.02 \times 10^{-3}$; we display our results in Fig. 12. For the high-z data, h and τ are $2\overset{\circ}{.}03$



(a) (b)

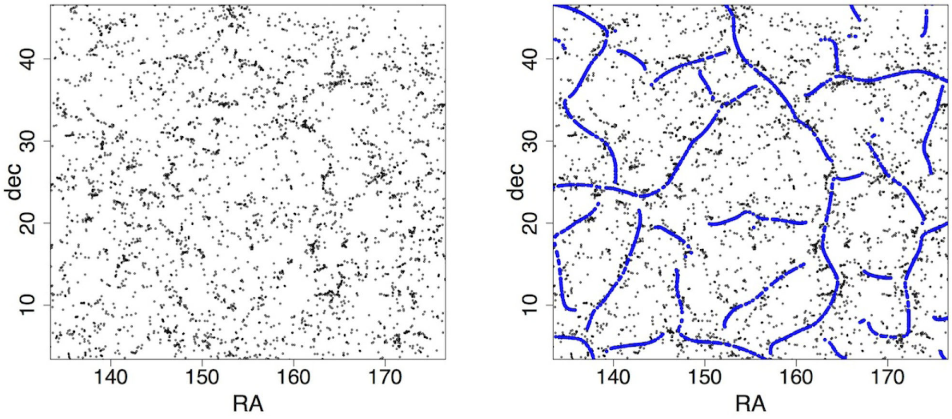
(c) Uncertainty Measure with Boundary Points (d) Uncertainty Measure without Boundary Points

Figure 11. Simulated example with sample size 2500 that demonstrates the boundary bias of SCMS. To demonstrate this bias, we remove points outside the orange rectangle (the so-called boundary points). (a) and (b): comparisons between SCMS with boundary points (blue) and SCMS without boundary points (red). As can be seen, the bias (between red and blue curves) is large for filaments whose distance to the boundary are less than 10 Mpc h^{-1} (2 × smoothing bandwidth h). (c) and (d): uncertainty measure for the filaments with and without boundary points. Notice that in (d), filaments near the boundary tend to have much higher uncertainty. The unit to the colour in uncertainty band is Mpc, the same as X and Y axes.



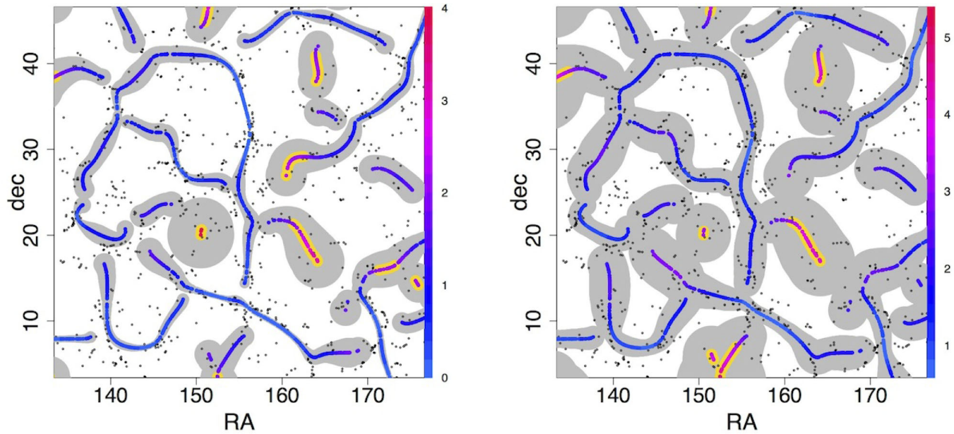
(a) The original data. (b) Filament detection.

Figure 12. Application of SCMS to low-z data ($z = 0.235$–0.240). The blue curves are filaments detected by SCMS.



(a) The original data. (b) Filament detection.

Figure 13. Application of SCMS to high-z data ($z = 0.530$–0.535). The blue curves are filaments detected by SCMS.



(a) The bootstrap estimate. (b) The smooth bootstrap estimate.

Figure 14. Local uncertainty estimates for our low-z SDSS data set ($z = 0.235$–0.240). We display the amount of uncertainty via colour (red: high) and a confidence band in (a), (b) using both ordinary bootstrap and the SB. The filament points surrounded by yellow colours are those with high uncertainty and are declared as 'unstable'. Based on the simulation result in Fig. 10, we expect that the grey regions in plot (a) contain about 50 percent true filaments and in (b) contain 85 percent true filaments. The unit to the colour in uncertainty band is degree.

(71.1 Mpc) and $\tau = 7.52 \times 10^{-4}$, respectively; we display our results in Fig. 13. Note that we have included additional galaxies within 5° of our selected window to mitigate boundary bias.

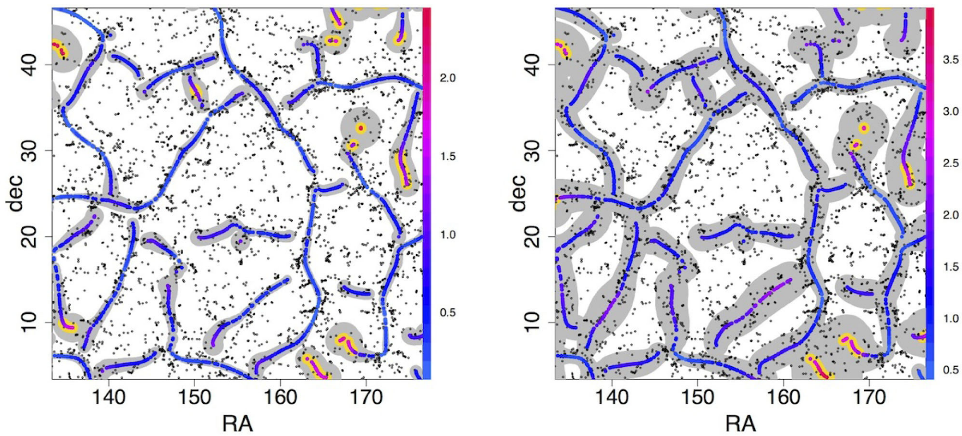
As can be seen in Figs 12 and 13, SCMS filament estimates capture high density regions and they exhibit one-dimensional, nearly connected structures. In addition, SCMS yields smooth filaments; most filament estimators do not output such smooth structures (cf. Stoica et al. 2005, 2007; Sousbie 2011; Aanjaneya et al. 2012; Lecci, Rinaldo & Wasserman 2013). We note that the filaments detected by SCMS will not actually connect with each other; points on merging filaments have eigengap β (equation 3) that asymptote towards 0, making the density ridge ill-defined since the first and second eigenvalues become equal. We note that in both figures there are possibly spurious filaments; for instance, in Fig. 12, at (RA,δ)=(165°, 40°) and (165°, 20°), we see filaments that are associated with a relatively small number of galaxies. As we demonstrate below, these putative filaments have higher estimates of uncertainty.

We derive the uncertainties for the filament estimators as described in Section 2.3 from the two test data sets; the results for low-z and high-z samples are given in Figs 14 and 15, respectively. We visualize local uncertainty using colour, where red indicates locations where the filamentary structure is highly uncertain. We also display uncertainty regions as bands of varying width (shown in grey) centred on the filaments. Our simulation study in Section 3.2 indicates that the filament coverage FCov for the regions in Figs 14(a) and 15(a) is ≈45 percent, while that in Figs 14(b) and 15(b), is ≈60 percent. We find that the overall structure for filaments in the high-z data set is more stable than for the low-z data, due to the significantly larger size of the high-z data set; as shown in Fig. 8, sample size plays a crucial role in determining the size of the uncertainty regions associated with SCMS filament estimates.

As can be inferred from Figs 14 and 15, our measures of local uncertainty provide useful information to determine the quality of filament detections. We declare a point $x \in \widehat{R}$ to be 'unstable' if

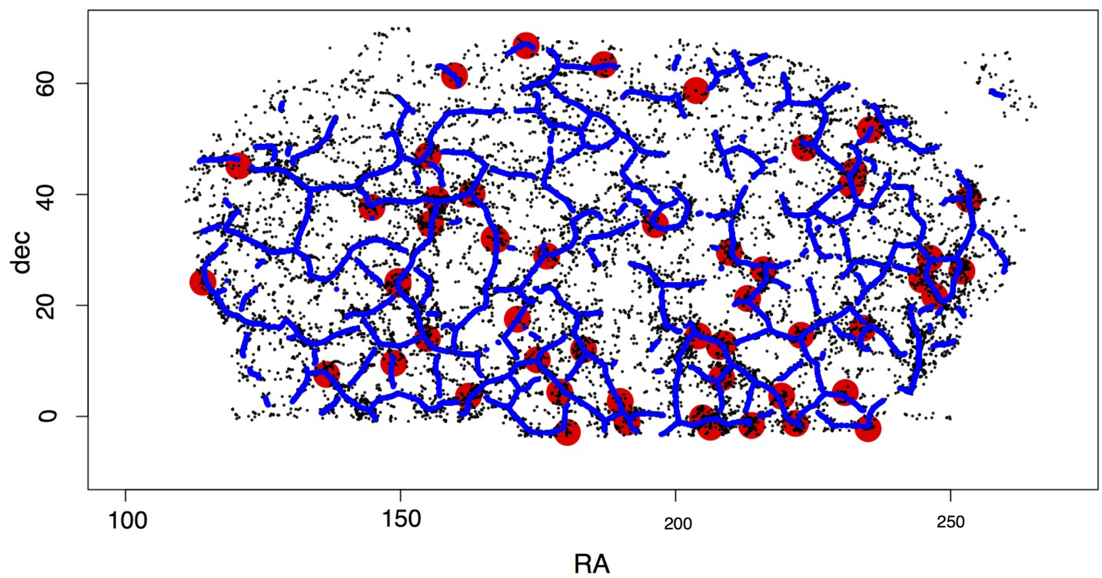
$$\widehat{\rho}(x) \geq \bar{\rho} + 1.69\sigma_{\rho}, \qquad (16)$$

where $\bar{\rho}$ is the mean of uncertainty over all filament points and σ_{ρ} is the root mean square of uncertainty. Namely, if the local uncertainty at x is too large, this point is not stable. The constant 1.69 comes from the width of 90 percent confidence interval for a Gaussian distribution. For instance the two filaments at (RA,δ) = (165°, 40°) and (165°, 20°) in Fig. 12 appear by eye to be spurious, given the



(a) The bootstrap estimate. (b) The smooth bootstrap estimate.

Figure 15. Local uncertainty estimates for our high-z SDSS data set ($z = 0.530$–0.535). We display the amount of uncertainty via colour (red: high) and a confidence band in (a), (b) using both ordinary bootstrap and the SB. The filament points surrounded by yellow colours are those with high uncertainty and are declared as 'unstable'. Based on the simulation result in Fig. 10, we expect that the grey regions in plot (a) contain about 50 percent true filaments and in (b) contain 85 percent true filaments. The unit to the colour in uncertainty band is degree.



(a)

Figure 16. Comparison of SCMS filaments to redMaPPer galaxy clusters, for $z = 0.145$–0.150. Shown are SDSS galaxies (black), putative filaments (blue), and redMaPPer galaxy clusters (red). As shown in Table 1, the redMaPPer clusters lie significantly closer to filaments than randomly selected points in the analysis window.

relative lack of galaxies in their vicinity. Based on the uncertainty measures and our stability test (equation 16), these filaments are declared as unstable (yellow colour in Fig. 14).

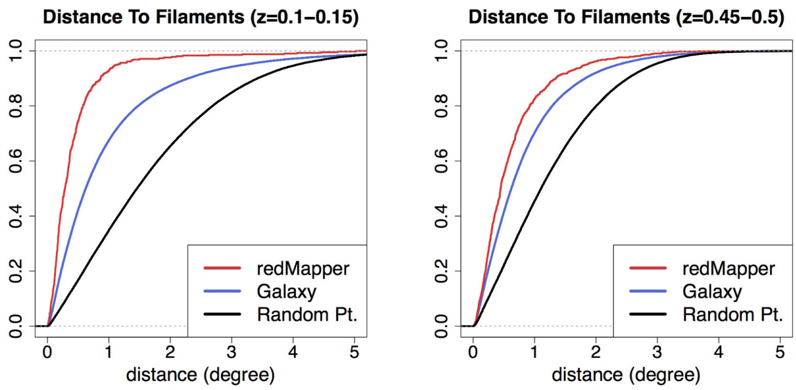
3.3.3 Test Data: comparison to redMaPPer clusters

As one last demonstration of the efficacy of SCMS, we examine the consistency between our filament maps and the galaxy clusters in the redMaPPer catalogue (Rozo & Rykoff 2014; Rykoff et al. 2014; Rozo et al. 2015). We make this comparison within the window

$$100^\circ \le \mathrm{RA} \le 270^\circ, -10^\circ \le \delta \le 70^\circ$$

and within annuli of width $\Delta z = 0.005$ from $z_{\mathrm{lo}} = 0.100$ to $z_{\mathrm{hi}} = 0.500$ (a range that includes 10 602 galaxy clusters with spectroscopically determined redshifts, or 93.1 percent of the redMaPPer sample). Note that we also include SDSS DR7 main sample galaxy from NYU VAGC (Blanton et al. 2005; Adelman-McCarthy et al. 2008; Padmanabhan et al. 2008) to detect filaments for low-redshift regions ($z < 0.25$). We slice the data primarily for computational efficiency, since SCMS is an $O(n^2)$ algorithm, but slicing has the ancillary benefit of simplifying visualization. In total, we examine 80 slices, each of which contains ≈100 galaxy clusters. Within each slice, we determine optimal values of h and τ using the criteria described in Appendix A.

In Fig. 16, we display SCMS-detected filaments along with redMaPPer clusters (in red). As can be seen, nearly all galaxy clusters are associated with detected filaments. Qualitatively similar results hold for all other slices. To quantify the association of



(a) $z = 0.100 - 0.150$ (b) $z = 0.450 - 0.500$

Figure 17. The cumulative distribution of the distance statistics from galaxies to filaments (blue) versus galaxy clusters to filaments (red) at different redshifts. We also display the distribution for random points (black) as a reference. The galaxy clusters are from redMaPPer catalogue. The unit of distance is 'degree'. We only display the first ($z = 0.100$–0.150) and the last sub-region ($z = 0.450$–0.500) since other regions have a similar result. The p-value for each region is given in Table 1.

Table 1. Significances generated from a one-sided, two-sample KS test, for the null hypothesis that galaxy clusters lie at the same average distance from filaments as field galaxies. p-value is a statistical quantity to measure the significance. Typically, the usual rejection rule requires $p < 0.05$. The p-values show strong evidence that clusters lie much closer to filaments than field galaxies.

Redshift	p-value	Redshift	p-value
0.100–0.150	4.38×10^{-40}	0.300–0.350	1.73×10^{-18}
0.150–0.200	1.01×10^{-31}	0.350–0.400	1.53×10^{-10}
0.200–0.250	1.66×10^{-26}	0.400–0.450	7.56×10^{-14}
0.250–0.300	2.26×10^{-19}	0.450–0.500	1.95×10^{-19}

galaxy clusters and filaments, we compare the distance to filaments for three types of objects: galaxy clusters, galaxies and randomly generated points within the regions where galaxies are observed. We divide the whole redshift range $z = 0.100$–0.500 evenly into eight sub-regions (each sub-region contains 10 slices); within each sub-region we compute distance statistics. Ideally, galaxy clusters should be systematically closer to filaments than galaxies are, and both galaxies and galaxy clusters should be far closer to filaments than randomly generated points. Fig. 17 and Table 1 confirm this hypothesis. Fig. 17 shows the cumulative distribution for these distance statistics. For a collection of values $x_1, \ldots, x_n$, the cumulative distribution function (CDF) is a non-decreasing function ranging from 0 to 1 defined as

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x). \tag{17}$$

Both galaxies (blue curves) and galaxy clusters (red curves) tend to be much closer to the filaments than random points; this suggests that galaxies and galaxy clusters are indeed concentrated around the detected filaments. When we compare galaxies and clusters, we observe that galaxy clusters are much more right skewed in the CDF plot for every redshift sub-region. That is, galaxy clusters tend to distribute around low-distance-to-filament regions compared to a random galaxy. We conduct the two-sample, one-sided KS test (Stephens 1974), which compares the distributions of distance statistics for galaxy clusters and randomly generated points, for all eight sub-regions. Table 1 shows the p-values, a statistical quantity measuring the significance of observations, for the eight KS tests that we carry out. A smaller p-value indicates stronger evidence for clusters being closer to a filament than galaxies. Typically, we declare significance as p-value being less than 0.05. We observe an increasing trend in p-value as the redshift increases, due to the decrease in the number density of galaxies along the line of sight. The sharp reversal in this trend at the last sub-region ($z = 0.450$–0.500) is due to the large size of the CMASS sample at $z > 0.430$: the number density of galaxies in our sample actually increases from $z = 0.430$–0.500.

Note that in Fig. 16, many clusters appear to be located near the intersections of filaments. However, we do not construct a statistics to summarize this phenomenon since defining the intersections of filaments detected by SCMS is a non-trivial problem. The main difficulty is due to the 'gap' between filaments; the SCMS filaments will not intersect each other but with a small gap. This gap can be explained by the model of density ridges. In the density ridges model, we require the eigengap $\beta > 0$ (recall equation 3) to ensure the properties of filaments. Therefore, when one ridge merges with another, the eigengap vanishes at some point (i.e. $\beta = 0$). This leaves a small gap between one ridge and another.

4 SUMMARY AND DISCUSSION

In this paper, we demonstrate how one may apply the SCMS algorithm of Ozertem & Erdogmus (2011) to uncover filamentary structure in galaxy point cloud data. The density ridge model behind the SCMS algorithm ensures that galaxies will concentrate around detected filaments. In addition, we introduce an uncertainty measure for detected filaments that is based on the bootstrap, allowing us to study the significance of these filaments.

In Section 3, we first show that the SCMS filaments are very similar to the Voronoi filaments. Then, we demonstrate the efficacy of our SCMS-based filament-finding algorithm by applying it both to P3M N-body simulation output and to SDSS DR 12 data (including the NYU main sample galaxy, LOWZ and CMASS data sets). By applying SCMS to simulated data, we are able to estimate the coverage of our bootstrap-generated uncertainty bands, i.e. the fraction of any one true filament that lies within its associated band (see Fig. 10). We find that the coverage depends sensitively on the number of galaxies in an analysed sample, with the SB algorithm generating more conservative uncertainty bands with 1σ coverage $\approx$0.6–0.8 (cf. 0.683 for a 1σ confidence band) for galaxy number densities $\approx 5 \times 10^{-4}$–5×10^{-3} (densities observed/to be observed by SDSS CMASS and *WFIRST*, respectively).

In Figs 12–15, we show the results of applying the SCMS algorithm to SDSS spectroscopically observed galaxies in the redshift slices $0.235 \le z \le 0.240$ and $0.530 \le z \le 0.535$, respectively. To test the hypothesis that our estimated filaments are associated with real filamentary structures, we compare the distances between filaments and redMaPPer galaxy clusters, random field galaxies, and random points in the galaxy field. By using the one-sided, two-sample KS test, we find that we can safely reject the null hypothesis that galaxy clusters and field galaxies reside at similar distances from filaments; the p-values are $\lesssim 10^{-9}$ (cf. the usual rejection criterion that $p < 0.05$; see Table 1).

The SCMS algorithm models filaments as one-dimensional ridges that trace high-density regions within the point cloud; as such, SCMS may be grouped with other filament-modelling algorithms that use the eigenvalues and eigenvectors of the Hessian matrix associated with the point cloud density function, such as MMF (Aragón-Calvo et al. 2007, 2010a) and NEXUS/NEXUS+ (Cautun et al. 2013). However, in contrast to these methods, which output filament estimates as two-dimensional regions, SCMS filament estimates are smooth, one-dimensional curves; the filament orientations are well-defined. Also in contrast to these methods, we offer measures of uncertainty by augmenting the SCMS algorithm with bootstrap-based uncertainty estimation algorithms that allow one to e.g. place bands around putative filaments, whose relative sizes indicate uncertainty in filament location (as in e.g. Fig. 5). We note that the segmentation-based DisPerSE algorithm of Sousbie (2011) uses the persistence ratio, a metric encapsulating the evolution of topological structure in the galaxy field, to define the *significance* of putative filaments, but not their spatial uncertainty. Finally, we compare SCMS filaments to those generated by the Spine (Aragón-Calvo et al. 2010b; Aragon-Calvo & Yang 2014) and Skeleton (Novikov et al. 2006) algorithms. Both the Skeleton and Spine models look for ridges within a density field. However, the Skeleton model does not provide a means by which to compute density ridges. In contrast, the SCMS algorithm allows us to efficiently compute ridges of the field's kernel density estimate. The Spine method outputs ridges as points on grids, so that resolution is an issue. On the other hand, the SCMS algorithm yields points that are on continuous curves (ridges), so there is no resolution issue to address.

We conclude by stating that one may extend the use of the SCMS algorithm beyond the analysis of galaxy point cloud data. For instance, Chen et al. (2014) discuss how to apply the algorithm to pixelized image data; in particular, they modify the algorithm (calling it the weighted SCMS algorithm) to find intensity ridges caused by e.g. tidal tails. In addition, the authors also discuss how one would incorporate the mass of a galaxy to achieve a better estimate of the local density as well as of corresponding ridges.

ACKNOWLEDGEMENTS

We thank Hy Trac for providing the *N*-body simulations and Rachel Mandelbaum for useful discussions. We also thank the referee Miguel A. Aragón-Calvo for useful comments and for providing the Voronoi data set that we analyse in Section 3.1. This work is supported in part by the Department of Energy under grant DESC0011114; YC is supported by DOE; SH is supported in part by DOE-ASC, NASA and NSF; CG is supported in part by DOE and NSF; LW is supported by NSF.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

REFERENCES

Aanjaneya M., Chazal F., Chen D., Glisse M., Guibas L., Morozov D., 2012, Int. J. Comput. Geom. Appl., 22, 305
Abazajian K. N. et al., 2009, ApJS, 182, 543
Adelman-McCarthy J. K. et al., 2008, ApJS, 175, 297
Aihara H. et al., 2011, ApJS, 193, 29
Alam S. et al., 2015, ApJS, 219, 12
Anderson L. et al., 2014, MNRAS, 439, 83
Aragon-Calvo M. A., Yang L. F., 2014, MNRAS, 440, L46
Aragón-Calvo M. A., Jones B. J. T., van de Weygaert R., van der Hulst J. M., 2007, A&A, 474, 315
Aragón-Calvo M. A., van de Weygaert R., Jones B. J. T., 2010a, MNRAS, 408, 2163
Aragón-Calvo M. A., Platen E., van de Weygaert R., Szalay A. S., 2010b, ApJ, 723, 364
Blanton M. R., Lin H., Lupton R. H., Maley F. M., Young N., Zehavi I., Loveday J., 2003, AJ, 125, 2276
Blanton M. R. et al., 2005, AJ, 129, 2562
Bolton A. S. et al., 2012, AJ, 144, 144
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Bond N. A., Strauss M. A., Cen R., 2010, MNRAS, 409, 156
Cautun M., van de Weygaert R., Jones B. J. T., 2013, MNRAS, 429, 1286
Cautun M., van de Weygaert R., Jones B. J. T., Frenk C. S., 2014, MNRAS, 441, 2923
Chacón J. E., Duong T., Wand M., 2011, Stat. Sin., 21
Chacón J. E., Duong T., 2013, Electron. J. Stat., 7, 499
Chen Y.-C., Genovese C. R., Wasserman L., 2014, preprint (arXiv:1406.1803)
Chen Y.-C., Genovese C. R., Wasserman L., 2015a, Ann. Stat., 43, 1896
Chen Y.-C., Genovese C. R., Ho S., Wasserman L., 2015b, preprint (arXiv:1506.02278)
Choi E., Bond N. A., Strauss M. A., Coil A. L., Davis M., Willmer C. N. A., 2010, MNRAS, 406, 320
Clampitt J., Miyatake H., Jain B., Takada M., 2014, preprint (arXiv:1402.3302)
Codis S., Dubois Y., Pichon C., Devriendt J., Slyz A., 2014, preprint (arXiv:1409.2838)
Colberg J. M., Krughoff K. S., Connolly A. J., 2005, MNRAS, 359, 272
Dawson K. S. et al., 2013, AJ, 145, 10
Doi M. et al., 2010, AJ, 139, 1628
Dolag K., Meneghetti M., Moscardini L., Rasia E., Bonaldi A., 2006, MNRAS, 370, 656

Eberly D., 1996, Ridges in Image and Data Analysis. Springer-Verlag, Berlin
Efron B., 1979, Ann. Stat., 7, 1
Eisenstein D. J. et al., 2011, AJ, 142, 72
Forero-Romero J. E., Contreras S., Padilla N., 2014, MNRAS, 443, 1090
Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748
Genovese C. R., Verdinelli I., Perone-Pacifico M., Wasserman L., 2009, Ann. Stat., 37, 3236
Genovese C. R., Perone-Pacifico M., Verdinelli I., Wasserman L., 2014, Ann. Stat., 42, 1511
Gunn J. E. et al., 1998, AJ, 116, 3040
Gunn J. E. et al., 2006, AJ, 131, 2332
Guo Q., Tempel E., Libeskind N. I., 2015, ApJ, 800, 112
Hahn O., Porciani C., Carollo C. M., Dekel A., 2007a, MNRAS, 375, 489
Hahn O., Carollo C. M., Porciani C., Dekel A., 2007b, MNRAS, 381, 41
Hahn O., Porciani C., Dekel A., Carollo C. M., 2009, MNRAS, 398, 1742
Jenkins A. et al., 1998, ApJ, 499, 20
Jones B. J. T., van de Weygaert R., Aragón-Calvo M. A., 2010, MNRAS, 408, 897
Lecci F., Rinaldo A., Wasserman L., 2013, preprint (arXiv:1305.1212)
Lupton R., Gunn J. E., Ivezić Z., Knapp G. R., Kent S., 2001, in Harnden F. R., Jr, Primini F. A., Payne H. E., eds, ASP Conf. Ser. Vol. 238, Astronomical Data Analysis Software and Systems X. Soc. Pac., San Francisco, p. 269
Maraston C. et al., 2013, MNRAS, 435, 2764
Molchanov I., 2005, Theory of Random Sets. Springer-Verlag, Berlin
Novikov D., Colombi S., Doré O., 2006, MNRAS, 366, 1201
Ozertem U., Erdogmus D., 2011, J. Mach. Learn. Res., 12, 1249
Padmanabhan N. et al., 2008, ApJ, 674, 1217
Paz D. J., Stasyszyn F., Padilla N. D., 2008, MNRAS, 389, 1127
Peebles P. J. E., 1980, The Large Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Pier J. R., Munn J. A., Hindsley R. B., Hennessy G. S., Kent S. M., Lupton R. H., Ivezić Ž., 2003, AJ, 125, 1559
Rozo E., Rykoff E. S., 2014, ApJ, 783, 80
Rozo E., Rykoff E. S., Bartlett J. G., Melin J.-B., 2015, MNRAS, 450, 592
Rykoff E. S. et al., 2014, ApJ, 785, 104
Sheather S. J., 2004, Stat. Sci., 19, 588
Silverman B. W., 1986, Density Estimation for Statistics and Data Analysis. Chapman & Hall, UK
Silverman B. W., Young G. A., 1987, Biometrika, 74, 469
Smee S. A. et al., 2013, AJ, 146, 32
Smith J. A. et al., 2002, AJ, 123, 2121
Sousbie T., 2011, MNRAS, 414, 350
Sousbie T., Pichon C., Colombi S., Novikov D., Pogosyan D., 2008, MNRAS, 383, 1655
Springel V. et al., 2005, Nature, 435, 629
Stephens M. A., 1974, J. Am. Stat. Assoc., 69, 730
Stoica R. S., Martínez V. J., Mateu J., Saar E., 2005, A&A, 434, 423
Stoica R. S., Martinez V. J., Saar E., 2007, J. R. Stat. Soc. C, 56, 459
Trac H., Cen R., Mansfield P., 2015, preprint (arXiv:1507.02685)
van de Weygaert R., 1994, A&A, 283, 361
Wasserman L., 2006, All of Nonparametric Statistics. Springer-Verlag, Berlin
York D. G. et al., 2000, AJ, 120, 1579
Zhang Y., Yang X., Faltenbacher A., Springel V., Lin W., Wang H., 2009, ApJ, 706, 747
Zhang Y., Yang X., Wang H., Wang L., Mo H. J., van den Bosch F. C., 2013, ApJ, 779, 160

APPENDIX A: PARAMETER SELECTION

Our version of SCMS has two key parameters, the smoothing bandwidth h and the threshold level τ. In this section, we show how we select optimal values for each.

The smoothing bandwidth h represents the amount by which we smooth the observed point cloud of galaxies when estimating p. One can choose h by applying prior knowledge or by letting h adapt to the sample. There is a large body of literature on the choice of bandwidth, e.g. Sheather et al. (2004) and Chacón, Duong & Wand (2011), Chacón et al. (2013). Among all methods, we recommend choosing h via

$$h = A_0 \times \left(\frac{1}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \sigma_{\min}, \tag{A1}$$

where A_0 is a constant that we discuss below, n is the sample size, d is the dimension (in our case $d = 2$), and $\sigma_{\min}$ is the minimal value for the standard deviation along each coordinate. Note that the reference rule (equation A1) will choose smaller h values as the sample size increases.

If A_0 is 1, (equation A1) corresponds to Silverman's rule (Silverman 1986). Silverman's rule selects h via minimizing the mean integrated error

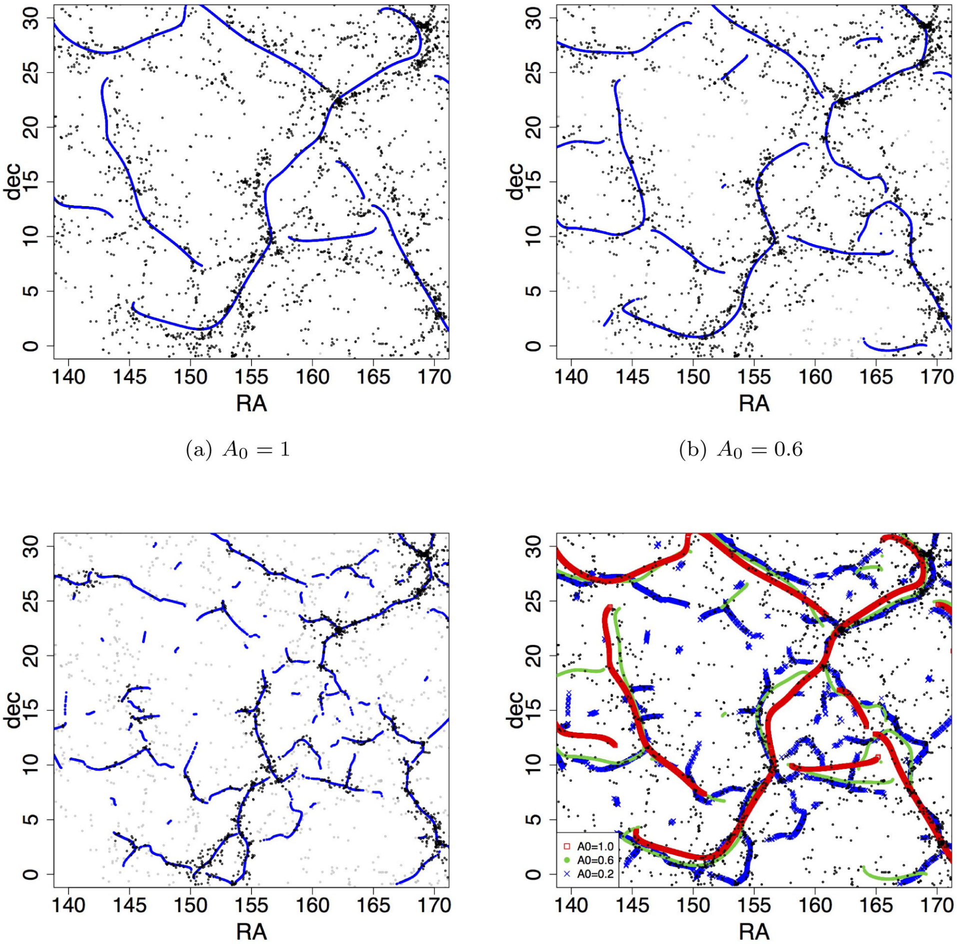
$$\mathbb{E}\left(\int |\widehat{p}(x) - p(x)|^2 dx\right) \tag{A2}$$

when p is a Gaussian. When the data include filaments, p is no longer Gaussian and A_0 must be optimized as a free parameter. A smaller A_0 yields more filaments in a given data set but more spurious filaments as well. There is no general rule for selecting A_0 since the optimality criterion involves the unknown density p. Fig. A1 shows how varying A_0 affects the estimation of filamentary structures. Our results indicate that the optimal A_0 lies in the range [0.4, 0.8]. This is further confirmed by true positive and false positive coverage of N-body simulation (described in Section 3.2) as shown in Fig. A2. In N-body simulation, $A_0 = 0.4$ corresponds to $h = 5$ Mpc (actual value is 4.82) while $A_0 = 0.8$ corresponds to $h = 10$ Mpc (actual value is 9.65). Both values are better than h being too large or too small (compared with $h = 2$ Mpc and $h = 15$ Mpc cases). In our analyses of SDSS data, we adopt the value $A_0 = 0.4$.

Thresholding stabilizes the ridge-finding process since random noise may cause small bumps in the estimated density field. However, if the threshold is set too high, we will remove useful information about the field. We recommend selecting the thresholding level according to the root mean square (RMS) in the density fluctuation:

$$\tau = \sigma(\widehat{p}) \equiv \left(\int_{\mathbb{K}} (\widehat{p}(x) - \bar{p}(\mathbb{K}))^2 dx\right)^{1/2} \sim \widehat{p} - \bar{p}, \tag{A3}$$

where $\mathbb{K}$ is the region we are interested in and $\bar{p}(\mathbb{K})$ is the average density in $\mathbb{K}$. Note that thresholding is also utilized by the MMF (Aragón-Calvo et al. 2010a) and NEXUS (Cautun et al. 2013) filament- (and galaxy cluster-) detection algorithms.



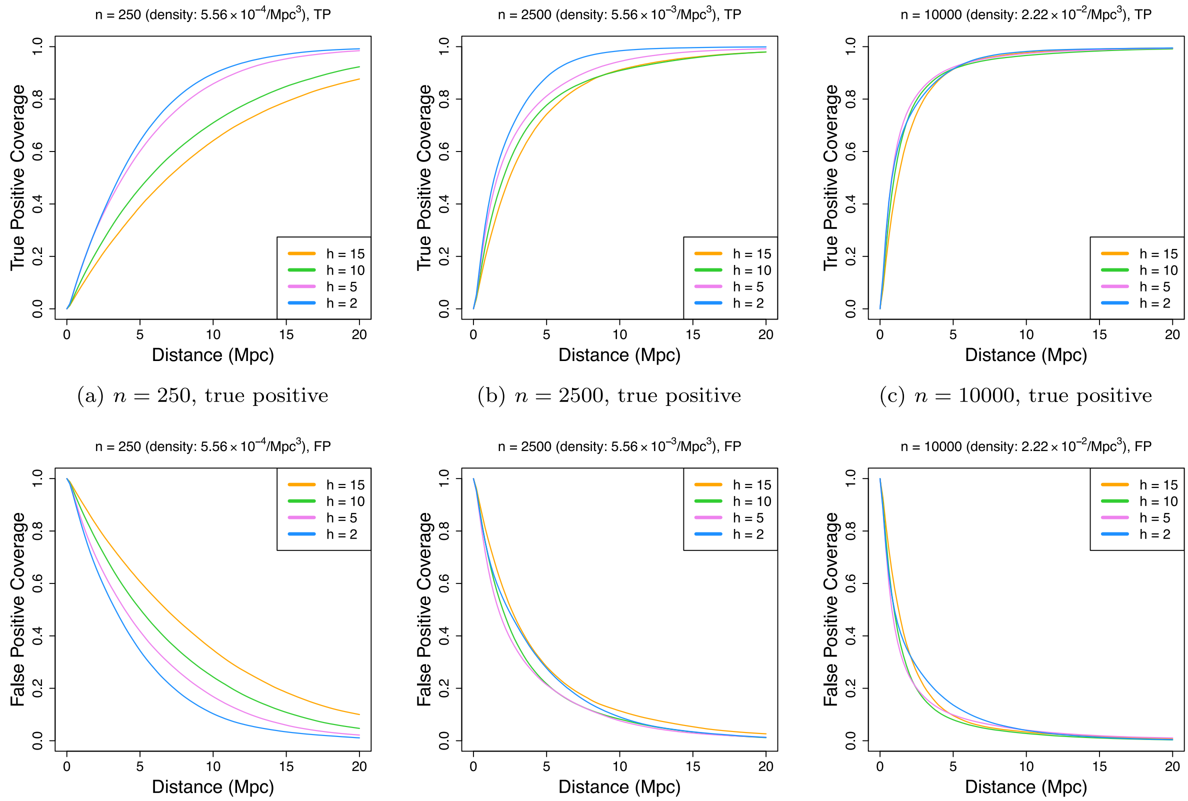
(a) $A_0 = 1$

(b) $A_0 = 0.6$

(c) $A_0 = 0.2$

(d) The comparison.

Figure A1. The bandwidth selection for equation (A1) for various A_0. The data are galaxies observed spectroscopically by SDSS in the redshift range $z = 0.045$–0.050. The grey dots are galaxies with density under $\tau = \sigma(\widehat{p})$. The black dots are galaxies with density above τ. In panels (a)–(c), blue curves are filaments detected by SCMS. In the panel (d), we compare the filaments from (a)–(c).



(a) $n = 250$, true positive (b) $n = 2500$, true positive (c) $n = 10000$, true positive

(d) $n = 250$, false positive (e) $n = 2500$, false positive (f) $n = 10000$, false positive

Figure A2. The true positive (top row) and false positive (bottom row) coverage for different sample sizes. As can be seen, $h = 5$ or $h = 10$ (corresponds to the reference rule (equation A1) using $A_0 = 0.4$ and 0.8) are good choices for both true and false positive coverage. Note that the reason $h = 2$ has good true positive coverage is because $h = 2$ undersmooths the data, leading to numerous small filaments. Thus, it is more likely that there are some estimated filaments around true filaments, which increases the true positive coverage but also increases the false positive coverage (as true filaments may not appear around some estimated filaments).

This paper has been typeset from a TEX/LATEX file prepared by the author.