Detection of cosmic filaments using galaxy distribution

Bonnaire Tony
June, 11th, 2020
Orsay-Tartu Workshop

Nabila Aghanim
Aurélien Decelle

Institut d’Astrophysique Spatiale, Orsay, France
Some filament finders

Libeskind et al., 2017 (Filaments are in blue)
Historically, this is the first method that has been used to exhibit the filamentary structure (Barrow+85).

It associates to a galaxy distribution a unique graph with no parameter that is minimising the global distance to reach each of them.

Visually traces what one would call « filaments ».
The graph structure acts like a topological prior in a mixture model. Considering $N$ galaxies $\{x_i\}$ and looking for $K$ Gaussian clusters with means $\{f_k\}$, the probability of a galaxy being located at $x_i \in \mathbb{R}^D$ is

$$p(x_i \mid \Theta) = \sum_{k=1}^{K} \pi_k N(x_i \mid f_k, \Sigma_k) + \alpha p(x_i)$$

1 component of uniform background for galaxies outlier of the filamentary pattern

$K$ Gaussians to pave the filamentary structure

With normalized amplitudes $\sum_{k=1}^{K} \pi_k + \alpha = 1$ and $\Theta = \{\alpha, \pi_1, \ldots, \pi_K, f_1, \ldots, f_K, \Sigma_1, \ldots, \Sigma_K\}$ the set of parameters of the mixture model.

To introduce the constraint on the solution (centres positions $f_k$), we set an a priori probability distribution on parameters

$$\log p(\Theta) = -\frac{3}{2} \left( \sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij} \|f_i - f_j\|^2 \right) + \text{cte}$$

with $A_{ij} = \left\{ \begin{array}{ll} 1 \text{ if } i \sim j, \\ 0 \text{ otherwise.} \end{array} \right.$

Fully encodes the graph information

This a priori has a Gaussian form with a $L_2$ constraint like usually done in statistics to constrain the smoothness of an estimate. Here the idea is the same, except that we do not impose the smoothness on the Euclidean space but directly using the graph structure $G$.

In what follows, we will first consider spherical Gaussian clusters, $\forall k \in \{0, \ldots, K\}$, $\Sigma_k = I_D \sigma^2_k$. 

---

T-ReX: Formalism (1)
To solve this problem iteratively and estimate parameters of the model, we use the Expectation-Maximization procedure \((\text{Dempster+77})\) to maximise the posterior distribution

\[
\Theta^{(t)} = \text{argmin}_\Theta \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \frac{\|x_i - f_k\|^2}{\sigma_k^2} + \lambda \sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij} \|f_i - f_j\|^2
\]

EM leads to the following update equations:

\[
\begin{align*}
p_{ik} &= p(z_i = k \mid x_i, \theta_k^{(t)}) = \frac{\pi_k \mathcal{N}(x_i \mid \theta_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_i \mid \theta_k')} + \alpha \rho(x_i) \\
p_{i}^{bkg} &= p(z_i = K + 1 \mid x_i, \theta_k^{(t)}) = \frac{\alpha \rho(x_i)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_i \mid \theta_k')} + \alpha \rho(x_i)
\end{align*}
\]

\[
\begin{align*}
f_k^{(t+1)} &= \frac{\sum_{i=1}^{N} \frac{p_{ik}}{\sigma_k^2} x_i + 2\lambda \sum_{j=1}^{K} A_{kj} f_j^{(t+1)}}{\sum_{i=1}^{N} \frac{p_{ik}}{\sigma_k^2} + 2\lambda \sum_{j=1}^{K} A_{kj}} \\
\sigma_k^{(t+1)} &= \left[ \frac{1}{D \sum_{i=1}^{N} p_{ik}} \sum_{i=1}^{N} p_{ik} (f_k - x_i)^T (f_k - x_i) \right]^{1/2} \\
\pi_k^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} p_{ik} \\
\alpha^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} p_{i}^{bkg}
\end{align*}
\]
T-ReX: Formalism (3)

- To solve this problem iteratively and estimate parameters of the model, we use the Expectation-Maximization procedure (Dempster+77) to maximise the posterior distribution.

\[ \Theta(t) = \arg\min_{\Theta} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \frac{\|x_i - f_k\|_2^2}{\sigma_k^2} + \lambda \sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij} \|f_i - f_j\|_2^2 \]

The MST allows the natural connectivity of clusters.
The Mixture Model allows the probabilistic modelling of the data set.

- The MST allows the natural connectivity of clusters.
- The Mixture Model allows the probabilistic modelling of the data set.
The algorithm learns a regularised graph structure given by the MST having better geometric properties while keeping the same definition as usually used (branches of the graph).
To assess the reliability of the path obtained using the MST and get rid of the tree topology (not allowing cycles), we can use a bootstrap method to get $B$ tree estimates.

Those trees allow to compute a probability (frequency) to cross a given region of the space during all the realisations.
Features: Heteroscedastic learning

- Note that in the update equations, we can optimize the value of $\sigma_k^2$ leading to an estimate of the local extension of the filamentary pattern

$$\sigma_k^2 = \frac{1}{D \sum_{i=1}^{N} p_{ik}} \sum_{i=1}^{N} p_{ik} (f_k - x_i)^T (f_k - x_i)$$

Weighted covariance matrix for component $k$

- 25% background noise
- A decade of difference in std in the two extremes branches

Work in progress, Bonnaire+, in prep
The probabilistic formalism of mixture model allows the assignment of each data point (galaxy) to the component, among the $K + 1$, having generate it the most probably (background or node of the graph).

In practice, we denote $z_i \in \{0, \ldots, K + 1\}$ the assignment of the galaxy $i$ at position $x_i$. During the E-step, we compute

$$p_{ik} = p(z_i = k \mid x_i, \theta_k)$$

Hence we can estimate $z_i = \arg\max_k p_{ik}$ and get the most probable component.

_Bonnaire+, in prep._
Investigating other topologies

- Even if it has a lot of convenient features, the MST topology cannot represent cycles.
- In the model, the graph intervenes only through its algebraic representation (like the adjacency or Laplacian matrices).
- We are hence not restricted to the MST topology and can obtain a regularised version of any kind of graph as long as we can compute those matrices.