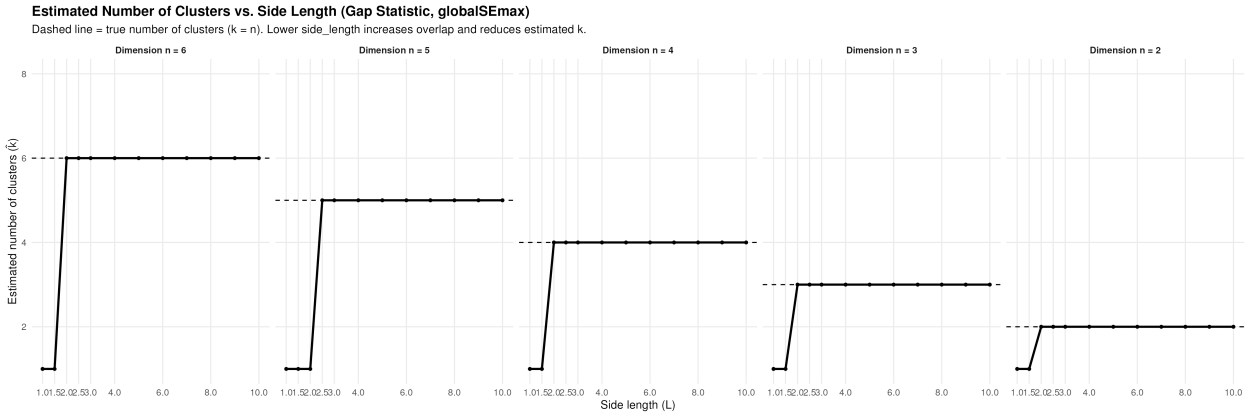
Clustering Assignment

Yuanyuan Yan

October 2025

GitHub link: https://github.com/yuanyuanyan669/BIOS611.git

Task1. Clusters and the Gap Statistic



Interpretation: When clusters are well-separated (larger side length bigger than 2), the Gap Statistic correctly identifies the true number of clusters in all dimensions. As L decreases and clusters start to overlap, the estimated number of clusters collapses sharply to 1(when side length equal to the sd of noise), indicating that the method can no longer distinguish the groups. This pattern reflects the resolution limit of the Gap Statistic, the minimum separation needed to reliably detect distinct clusters.

Task2. Spectral Clustering on Concentric Shells

The 3D shell example dataset was generated using the `generate_shell_clusters` function with four concentric shells ($n_{\text{shells}} = 4$), 600 points per shell, a maximum radius of 3, and Gaussian noise with a standard deviation of 0.1, and was visualized interactively using the `plotly` and `htmlwidgets` packages.

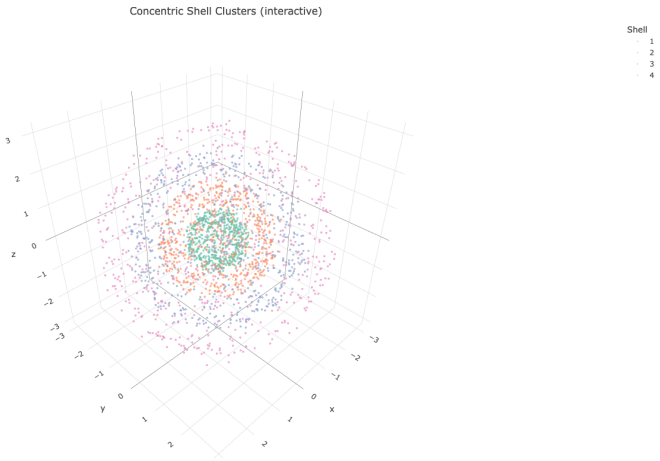


Figure 1: Interactive 3D generated shell

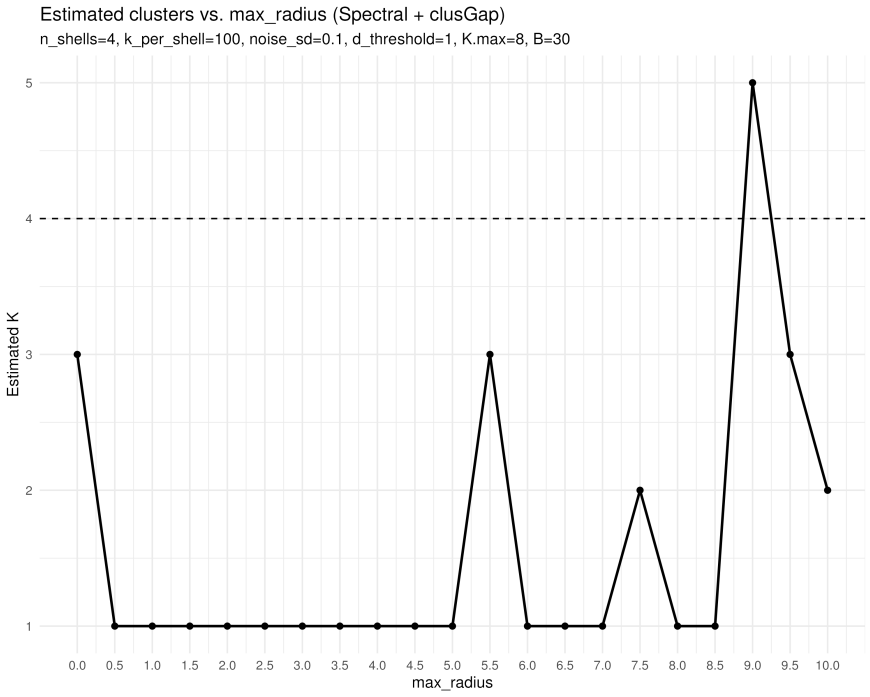


Figure 2: spectral clustering result with D Threshold 1.0

Interpretation: When the distance threshold is set to $d_{\text{threshold}} = 1.0$, spectral clustering performs reasonably well only when the shells are well separated, recovering close to the true four clusters at larger radii ($R_{\max} \geq 8$). As the maximum radius decreases, however, the shells begin to overlap and the similarity graph becomes overly connected, causing the eigenvectors of the Laplacian to lose discriminative power. This leads the estimated number of clusters to collapse toward one or two, marking the point where the algorithm can no longer distinguish between adjacent shells.

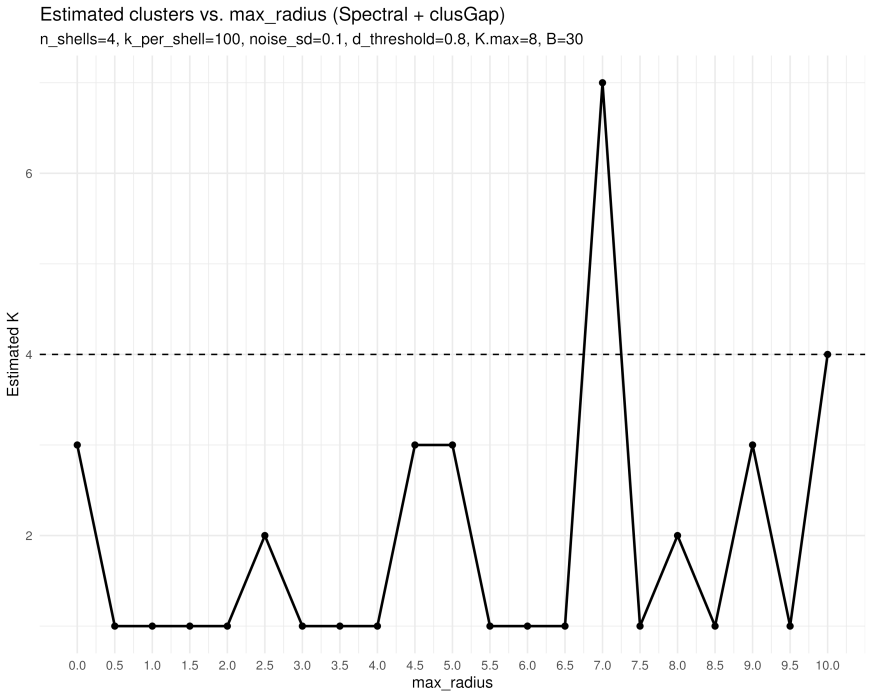


Figure 3: spectral clustering result with D Threshold 0.8

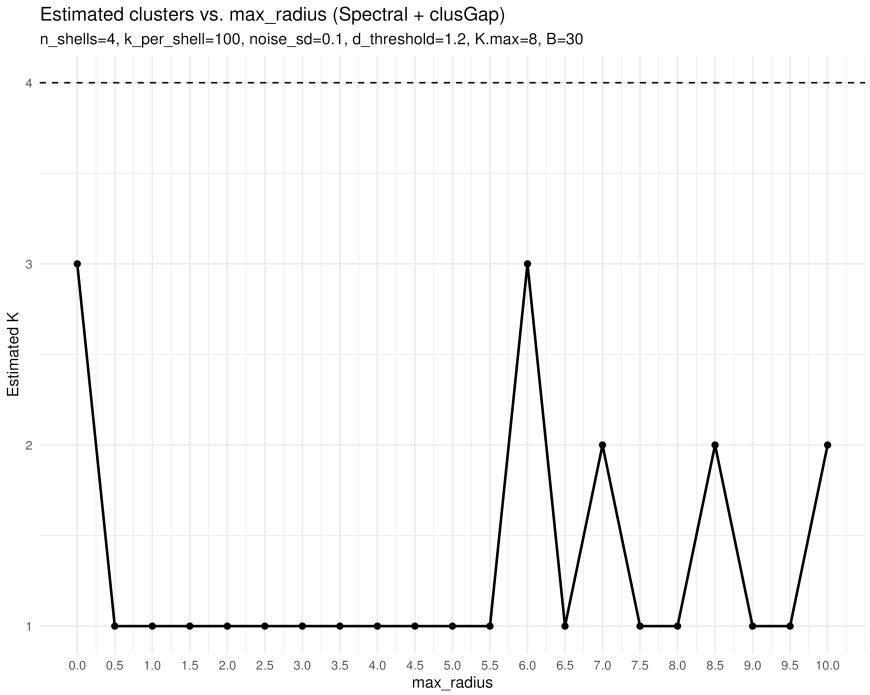


Figure 4: spectral clustering result with D Threshold 1.2

Interpretation: As the maximum radius decreases, the concentric shells become increasingly close, and the similarity graph formed by spectral clustering grows denser, eventually connecting points across adjacent shells. When this happens, the Laplacian eigenvectors fail to preserve the true shell boundaries, and the algorithm collapses to a single dominant cluster—this marks the failure point, observed around $R_{\max} \approx 3$–4 in the plot. Changing the distance threshold $d_{\text{threshold}}$ shifts this boundary: a smaller value (e.g., 0.8) yields a sparser graph that tends to fragment clusters and produce noisy, unstable k estimates, while a larger value (e.g., 1.2) over-connects the graph, smoothing out distinctions between shells and causing early merging. The intermediate threshold ($d_{\text{threshold}} = 1.0$) provides the most balanced trade-off between graph sparsity and connectivity, resulting in the most coherent clustering performance across radii.