k-means clustering in galaxy feature data

Seb Turner

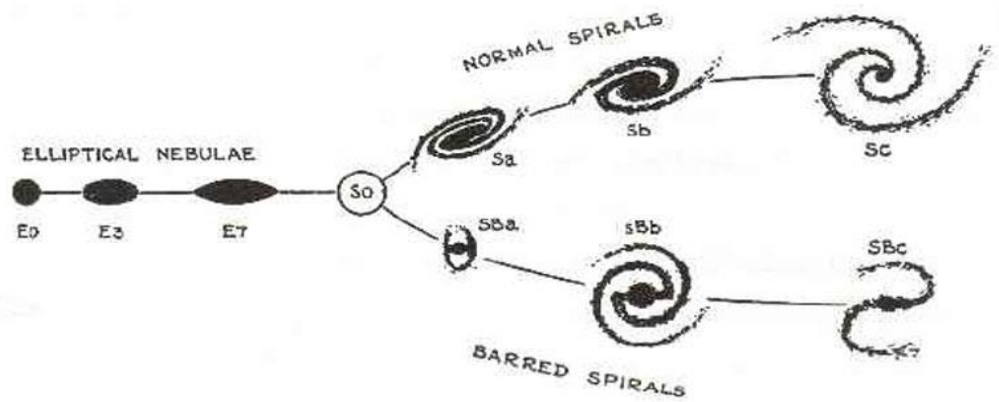
Lee Kelvin, Ivan Baldry, Paulo Lisboa, Steve Longmore, Chris Collins

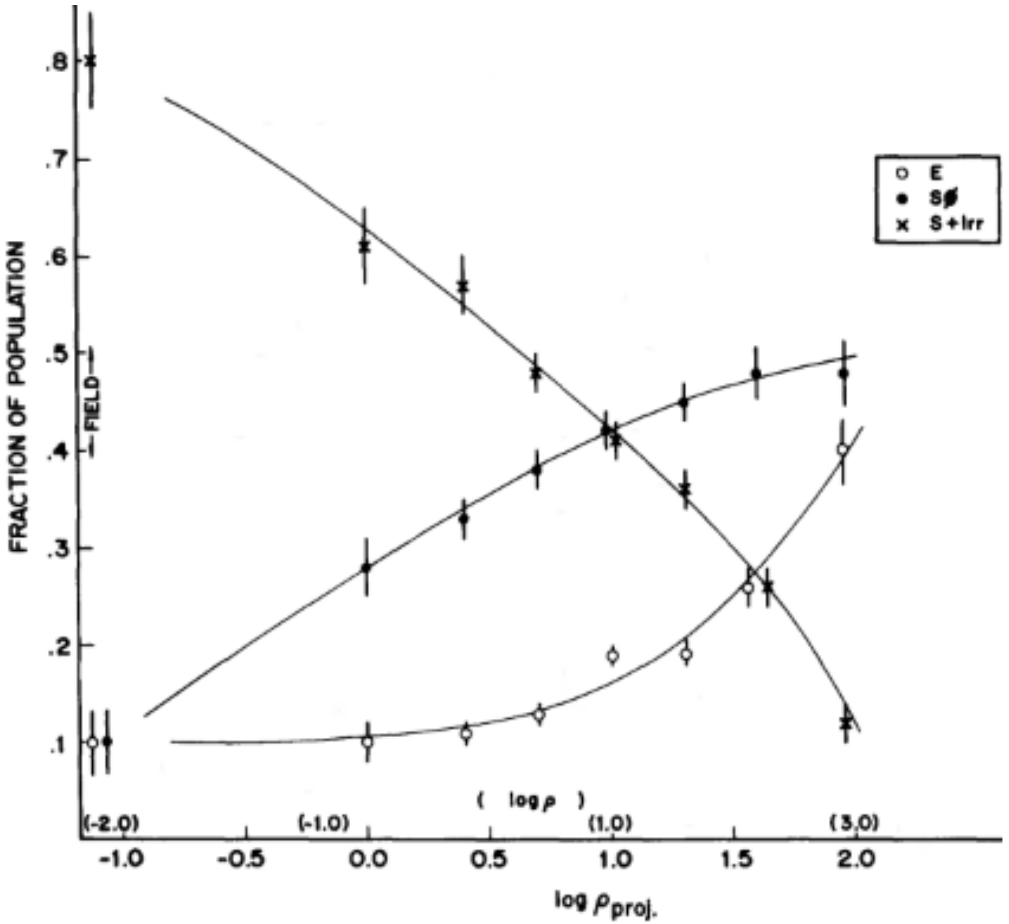
LIVERPOOL JOHN MOORES UNIVERSITY

s.turner1@2012.ljmu.ac.uk

@sebturne

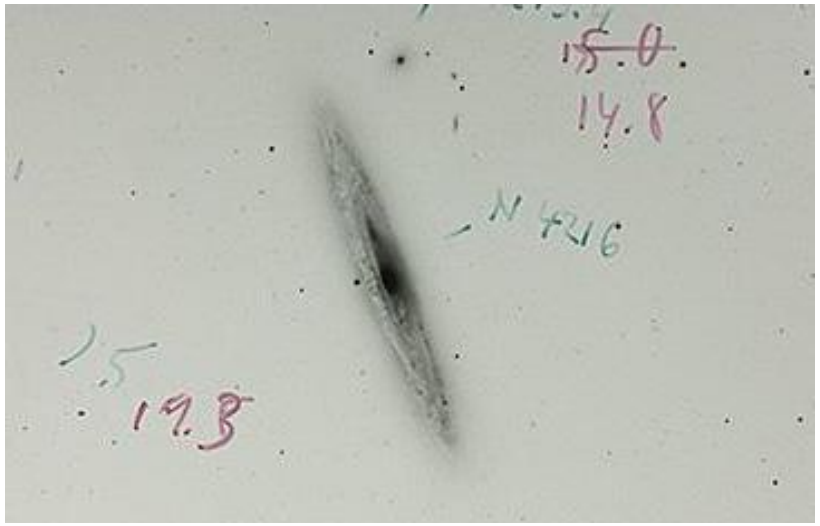
Galaxy classifications: background



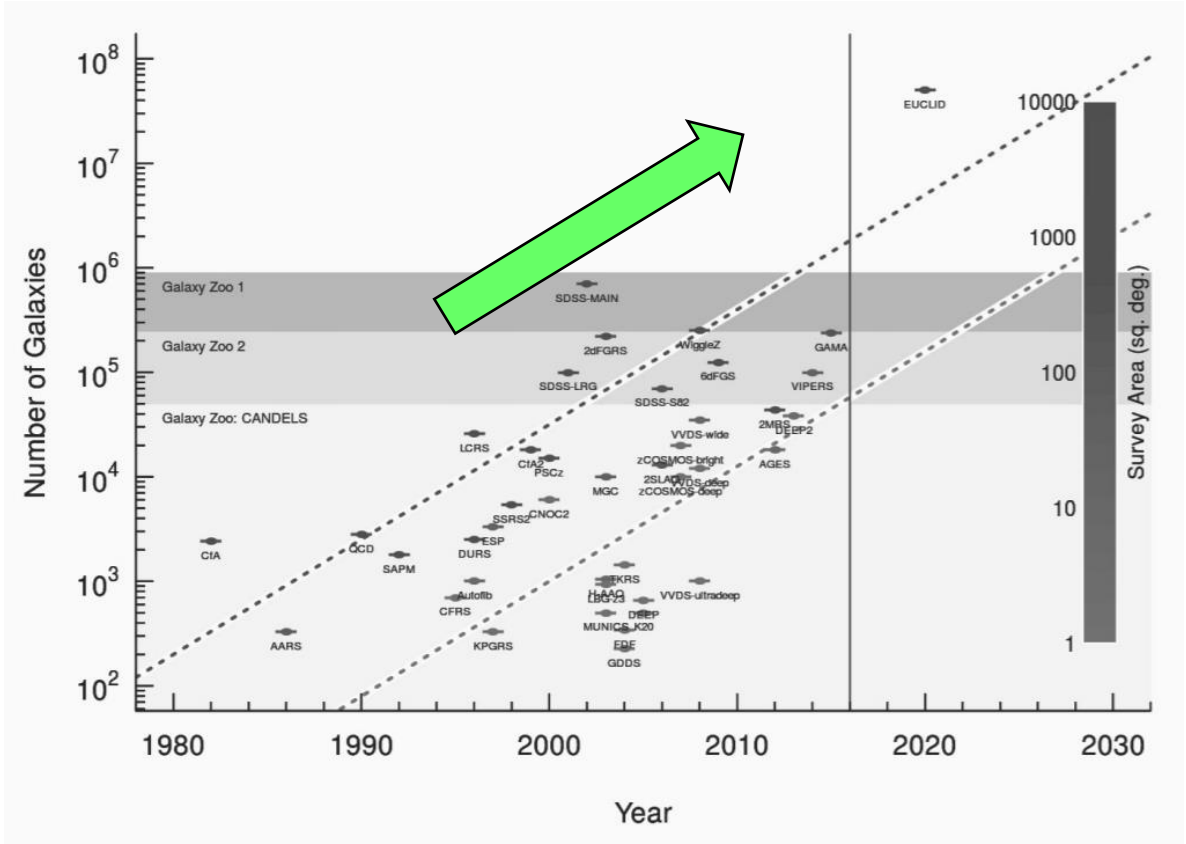


Hubble 1936; Dressler 1980

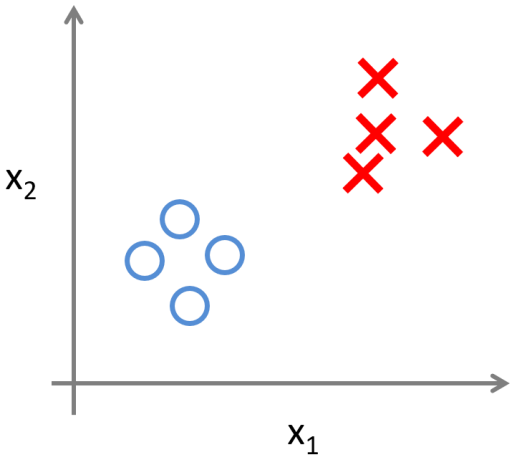
Galaxy classifications: moving forward

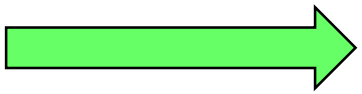


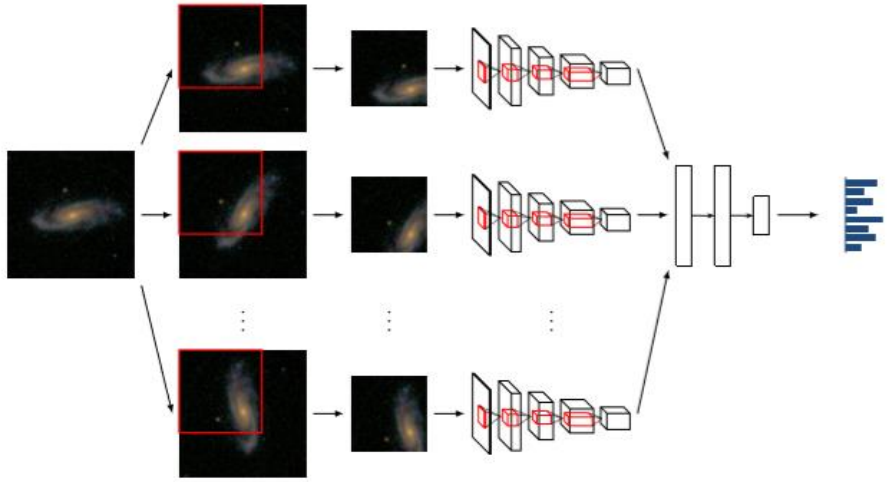
- Traditional: ~$10^{2\text{-}4}$
- Crowdsourced: ~$10^{5\text{-}6}$
- Automated (future): >10^{7}!!

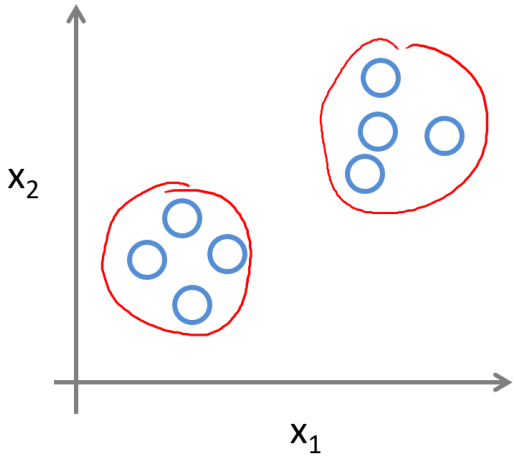


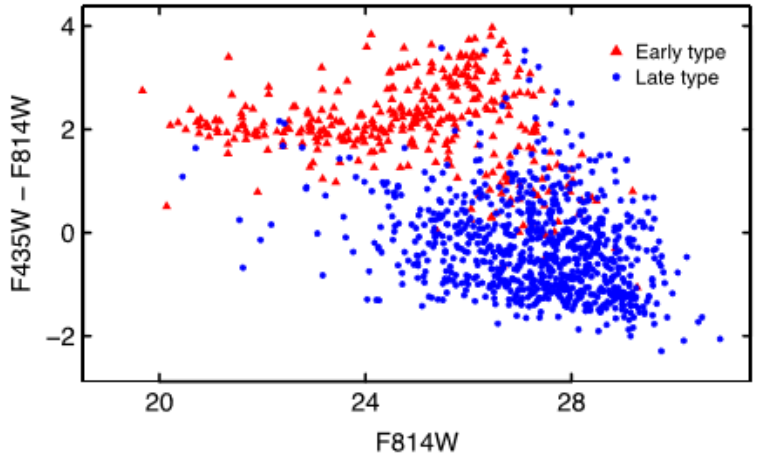
Supervised vs. unsupervised methods











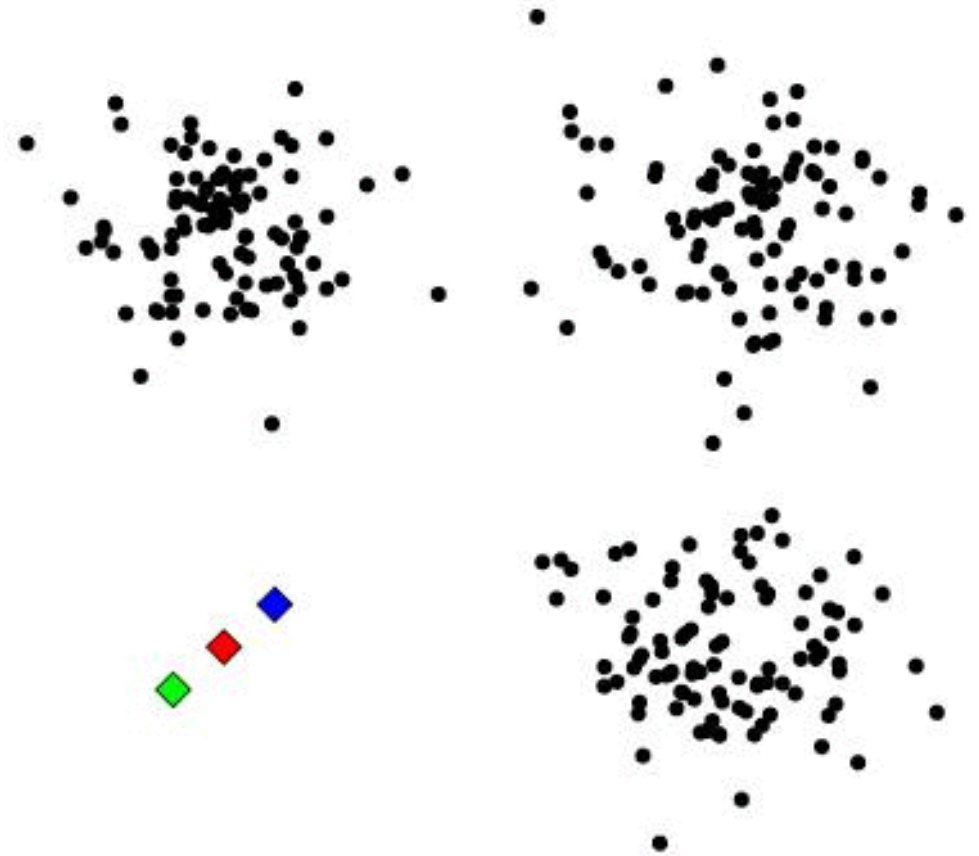
dosSantos 2017, gitbook.com; Dieleman+2015; Hocking+2018

k-means

- Simple, quick, popular.

- Step 0: initialise centres.
- Step 1: assign data points.
- Step 2: update centres.

- Repeat 1 & 2 until convergence.
- Compact, spherical clusters.

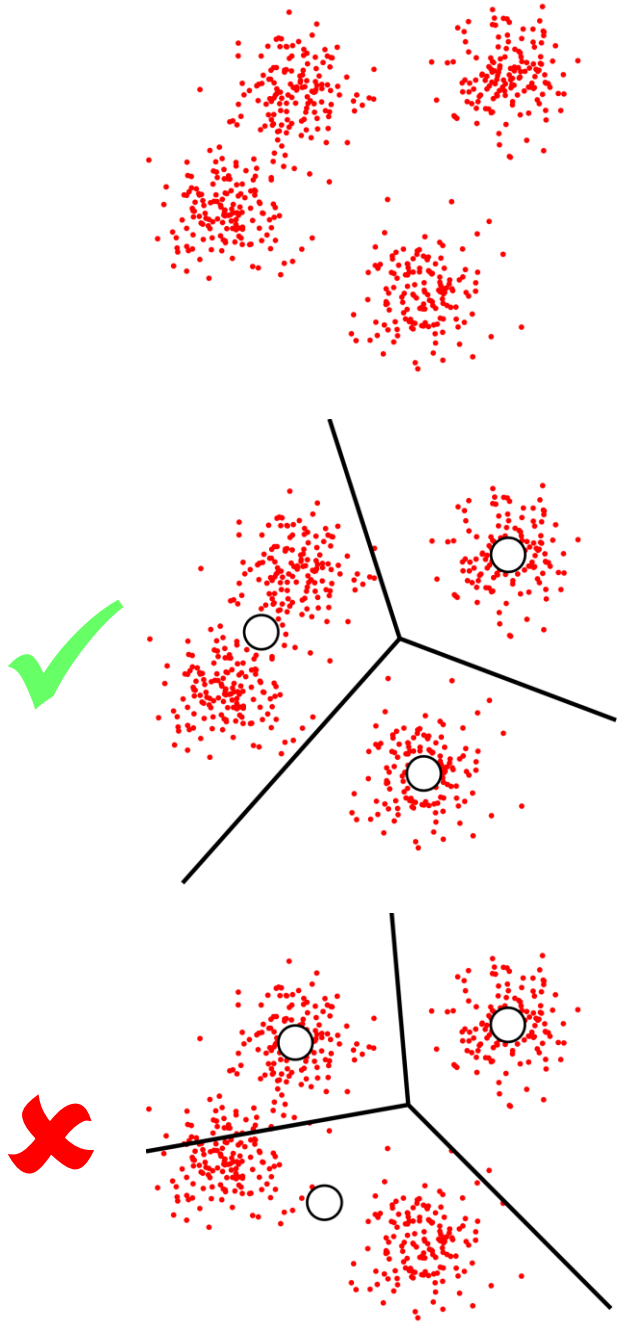


MacQueen 1967; Lloyd 1982; Arthur+2006; https://picoledelimao.github.io/blog/2016/03/12/multithreaded-k-means-in-java/

Cluster evaluation

- Local search heuristic.
- Outcome depends on initialisation!
- Best solution: most compact.

- $SSQW_j = \sum_{\boldsymbol{c} \epsilon C_j} \left|\left| \boldsymbol{c} - \bar{\boldsymbol{c}}_j \right|\right|^2$
- $\phi = \sum_{j=1}^{k} SSQW_j = \sum_{j=1}^{k} \sum_{\boldsymbol{c} \epsilon C_j} \left|\left| \boldsymbol{c} - \bar{\boldsymbol{c}}_j \right|\right|^2$

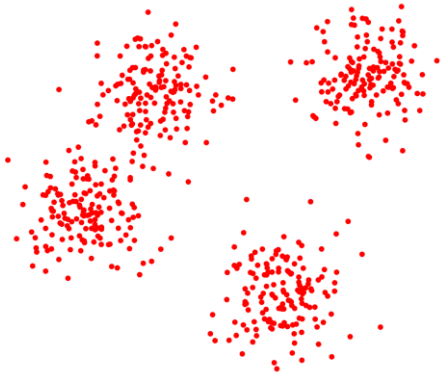


Cluster evaluation

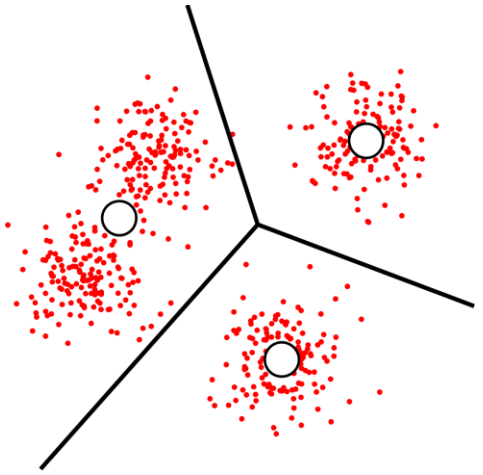
- What value of *k*?
- Try several!
- Best *k*: most stable.

- Cramer's V index of association:

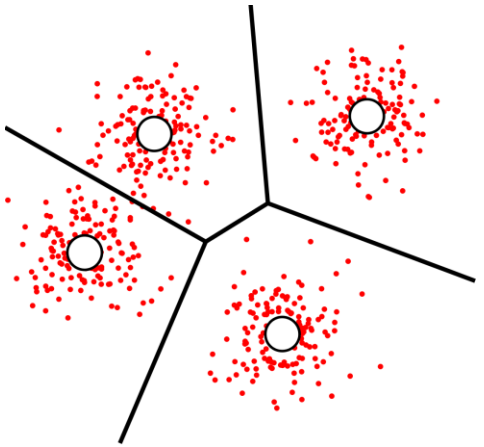
$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$



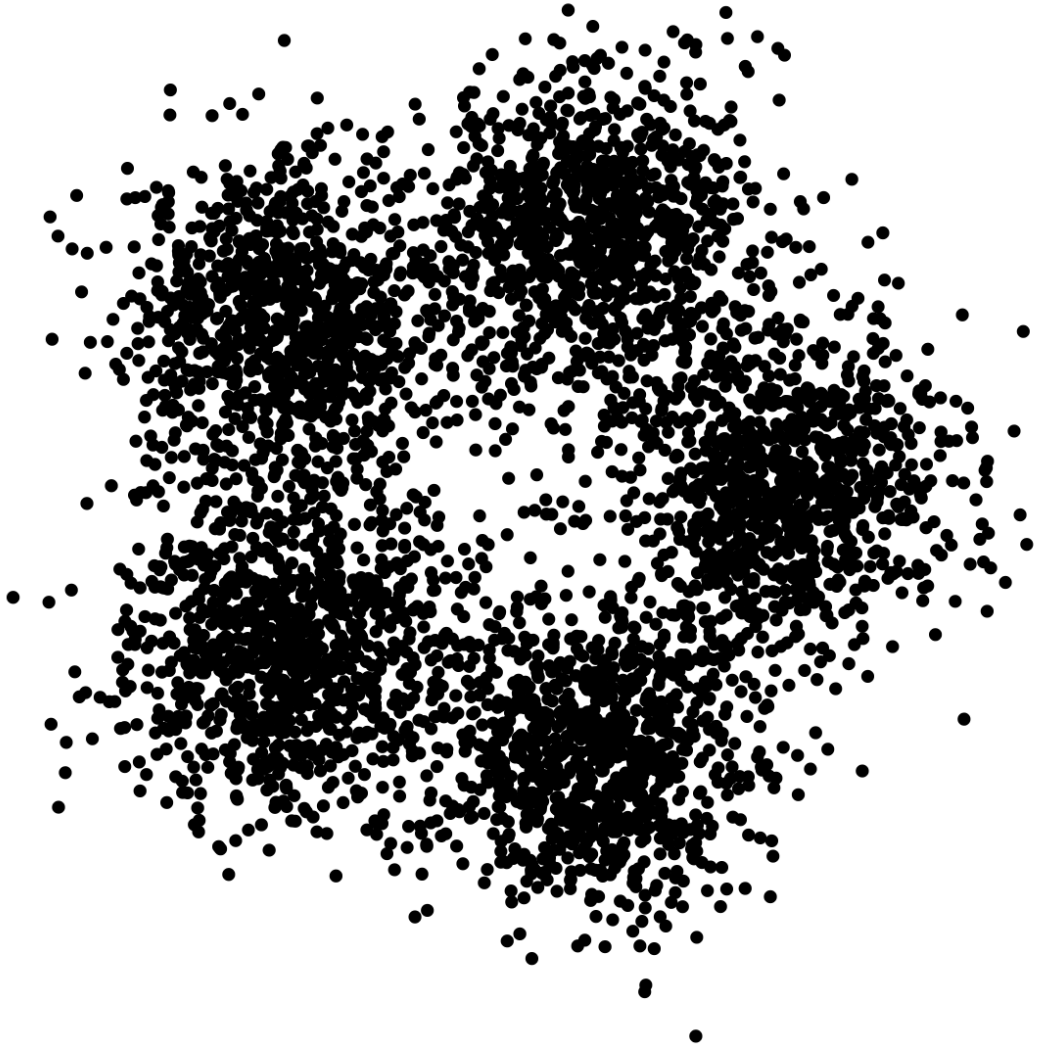
$k = 3$

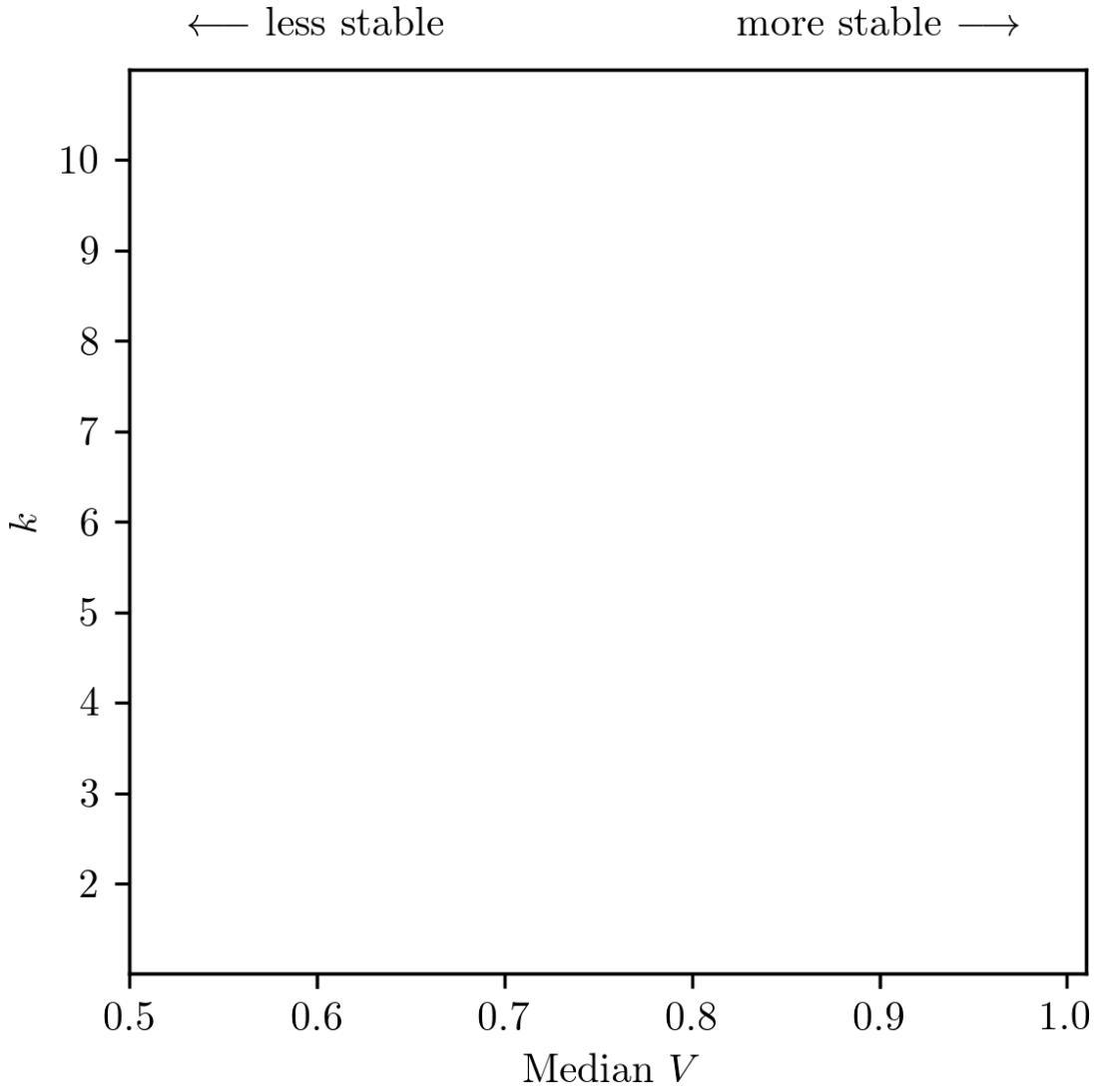


$k = 4$

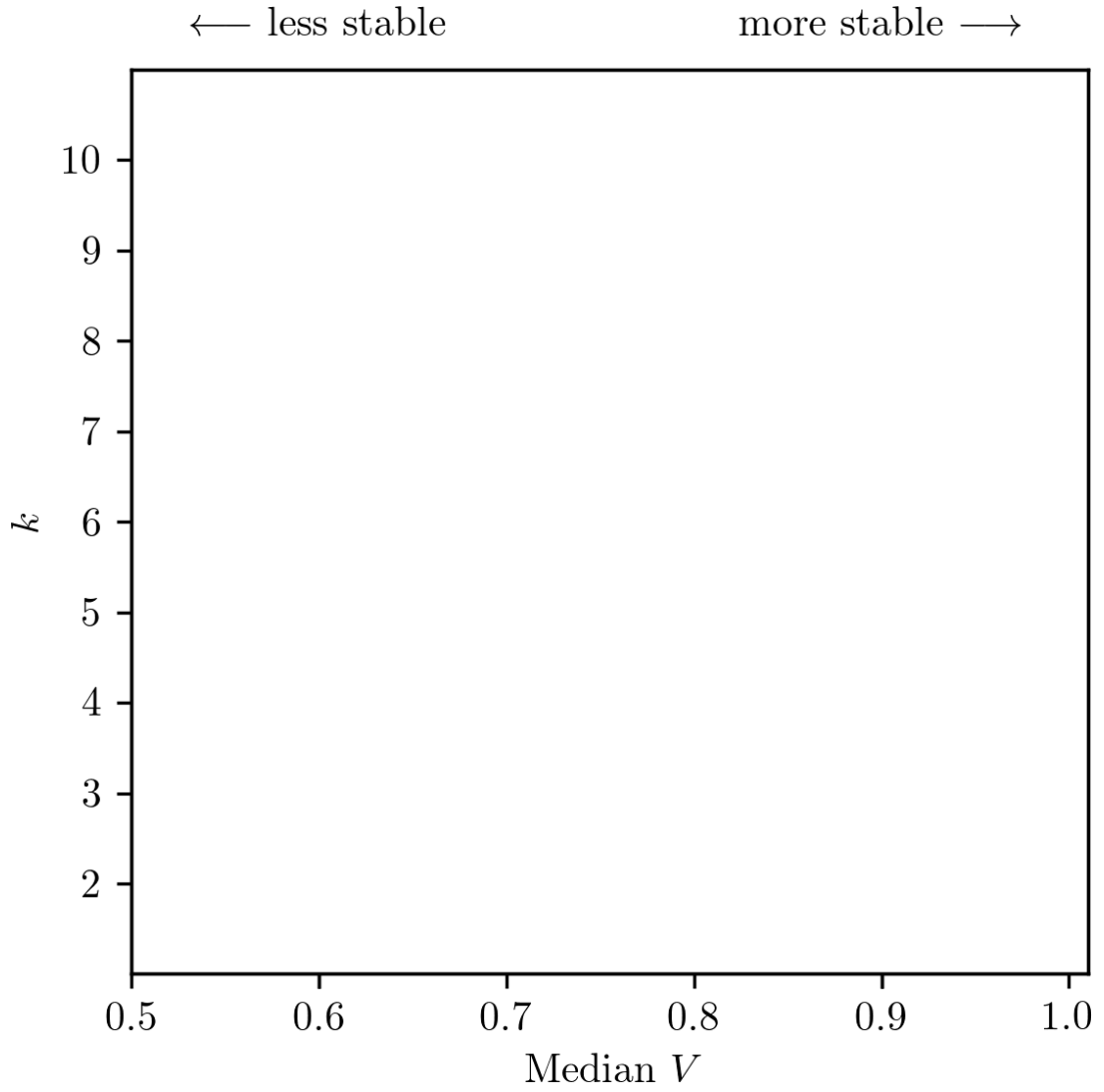


Simulated 2D example

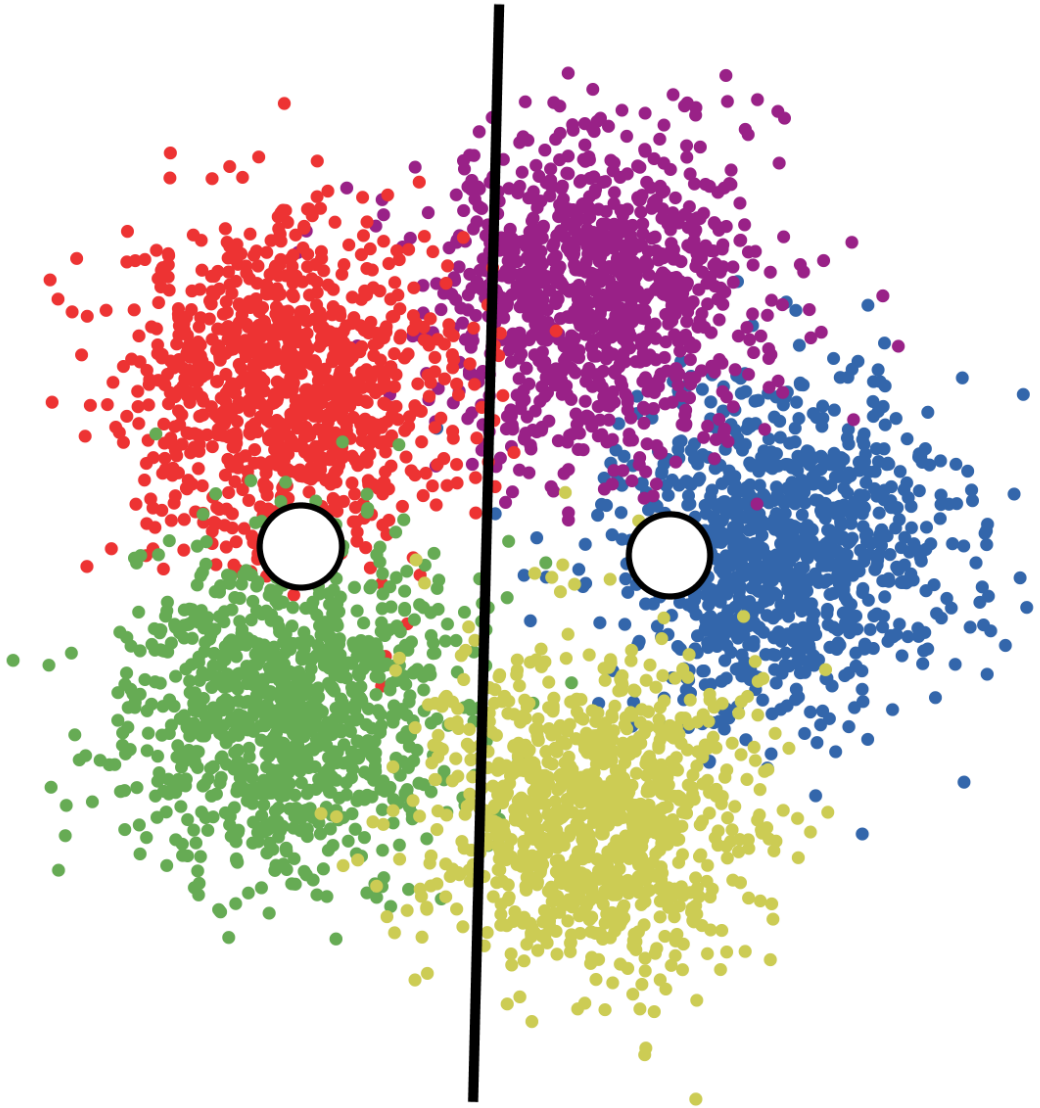


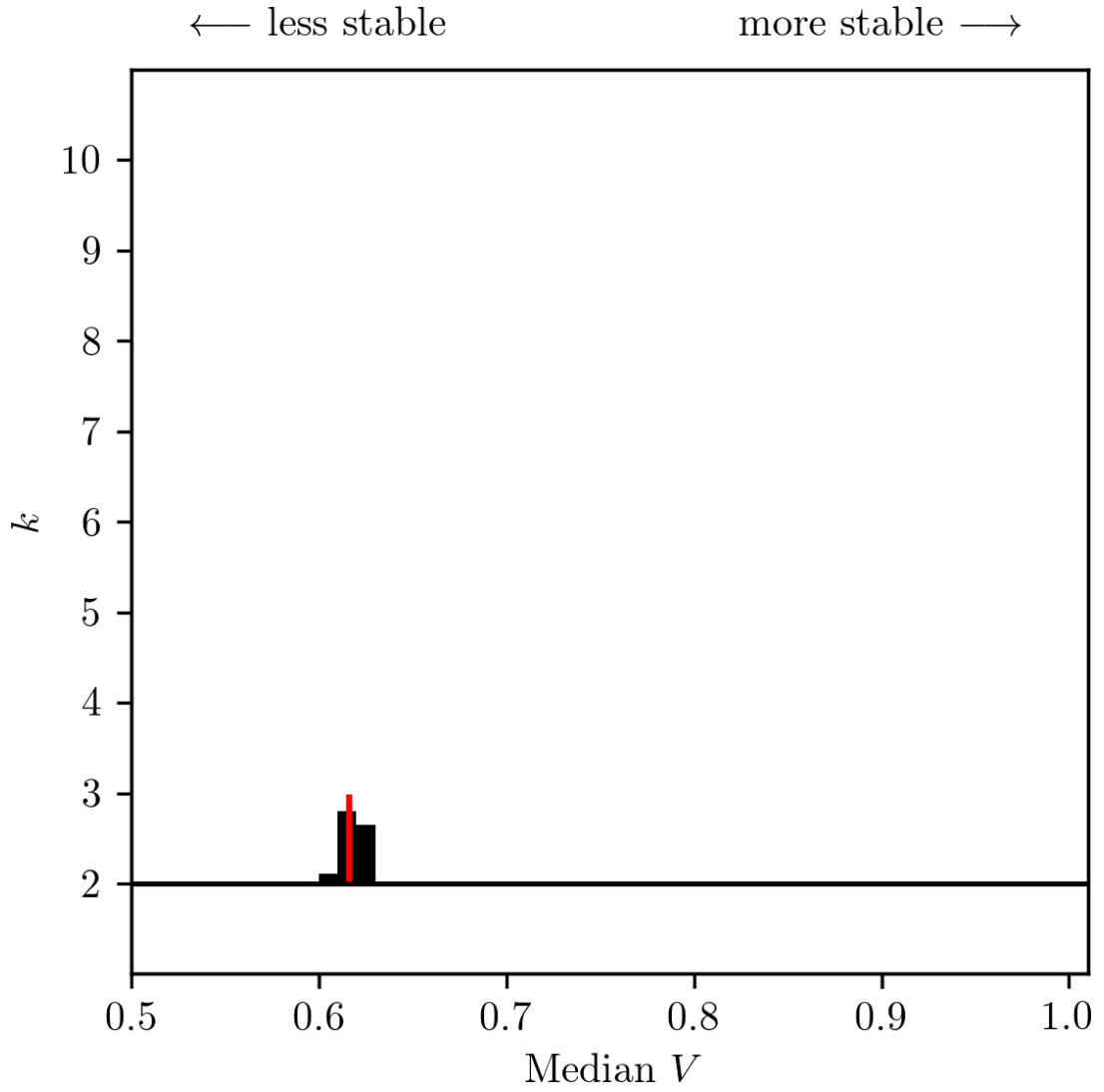


Simulated 2D example

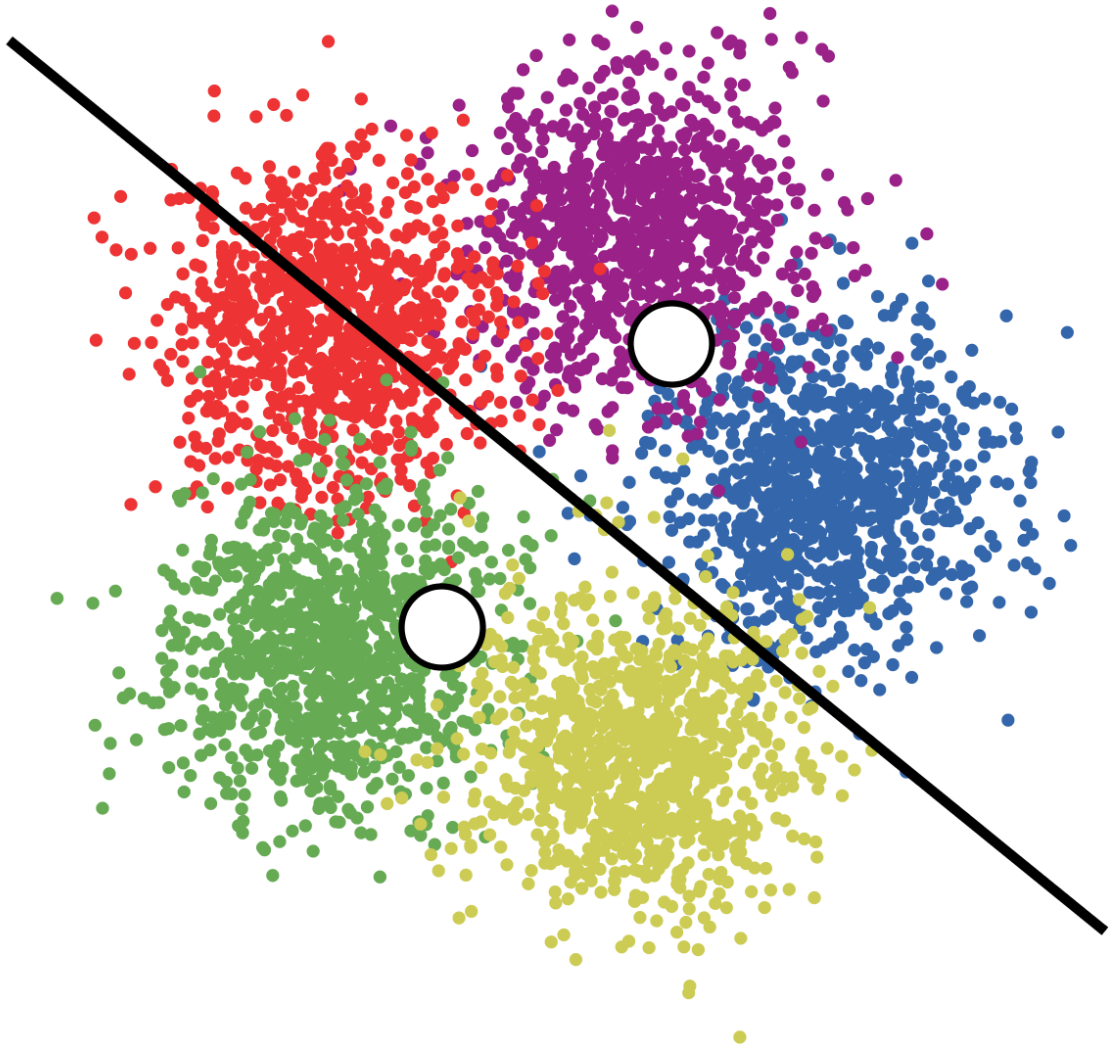


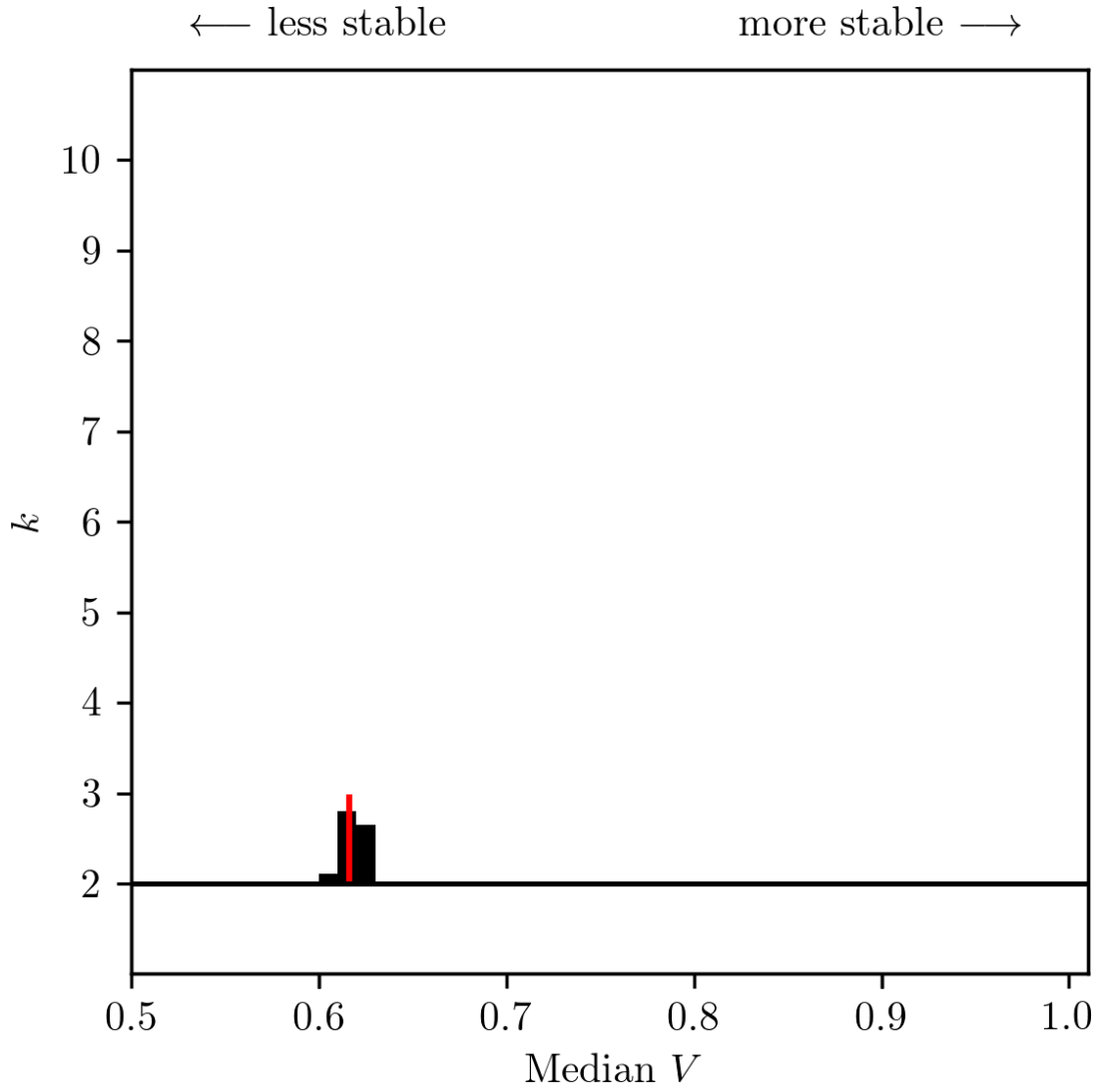
Simulated 2D example: $k = 2$



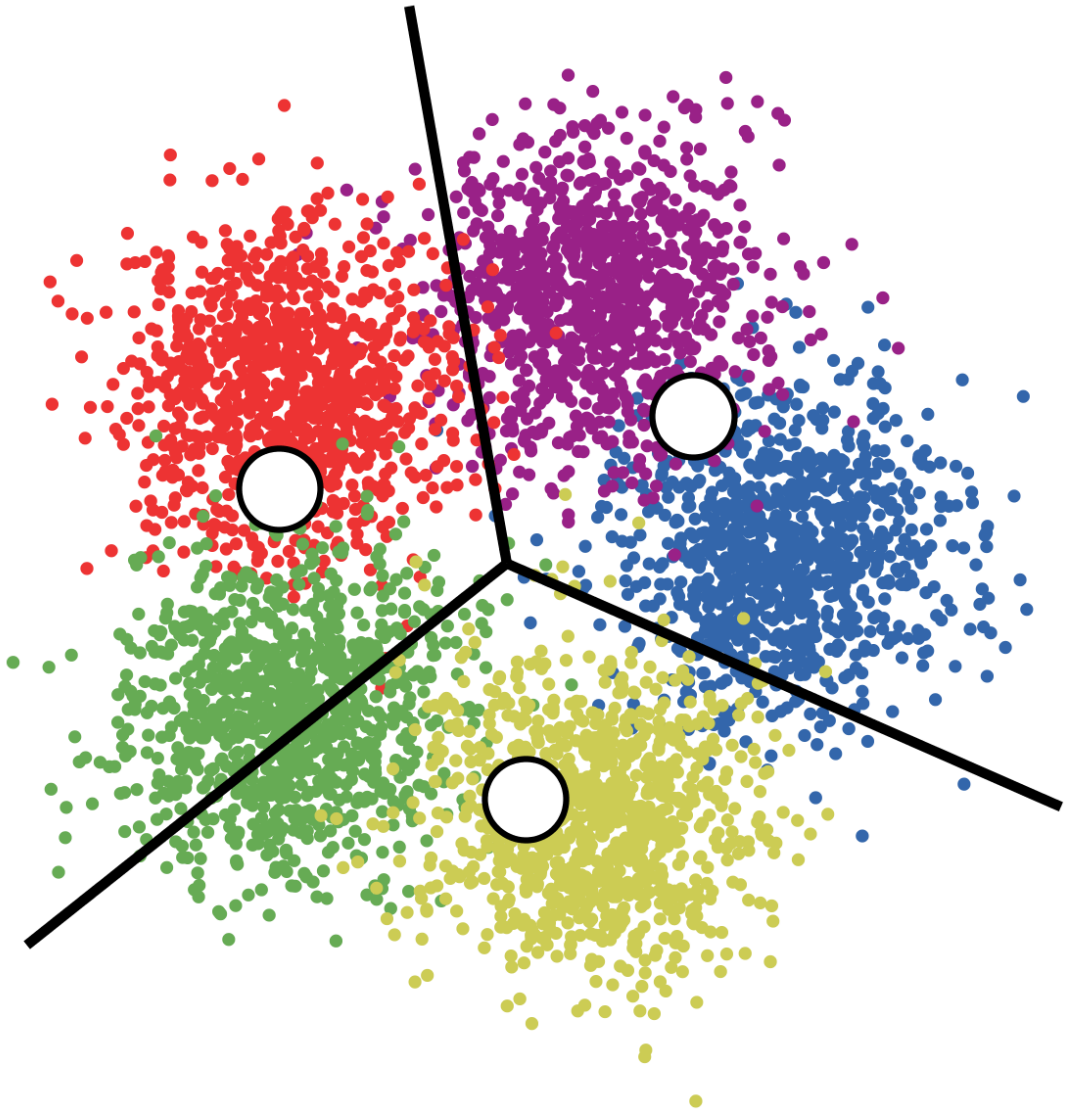


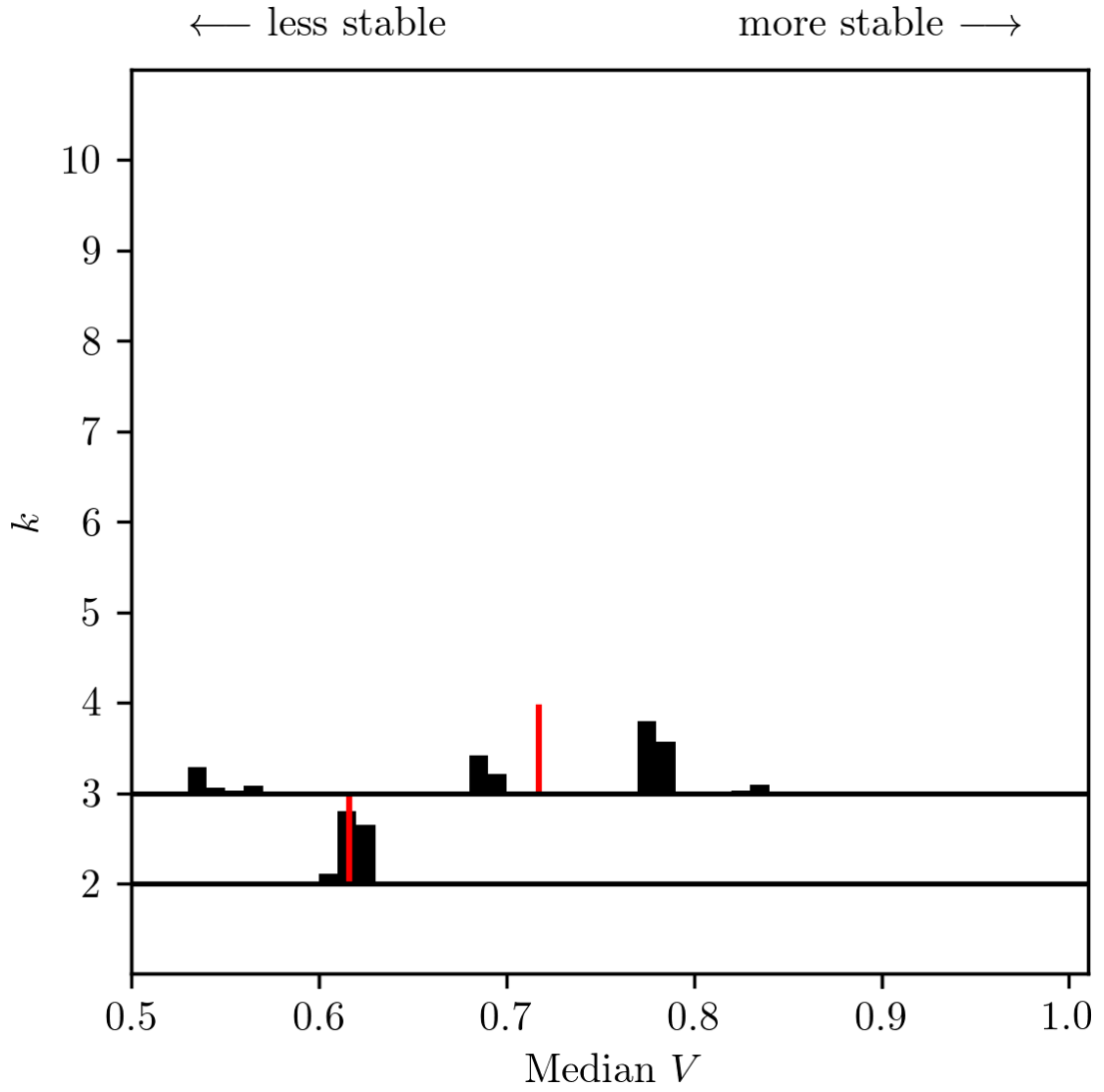
Simulated 2D example: $k = 2$



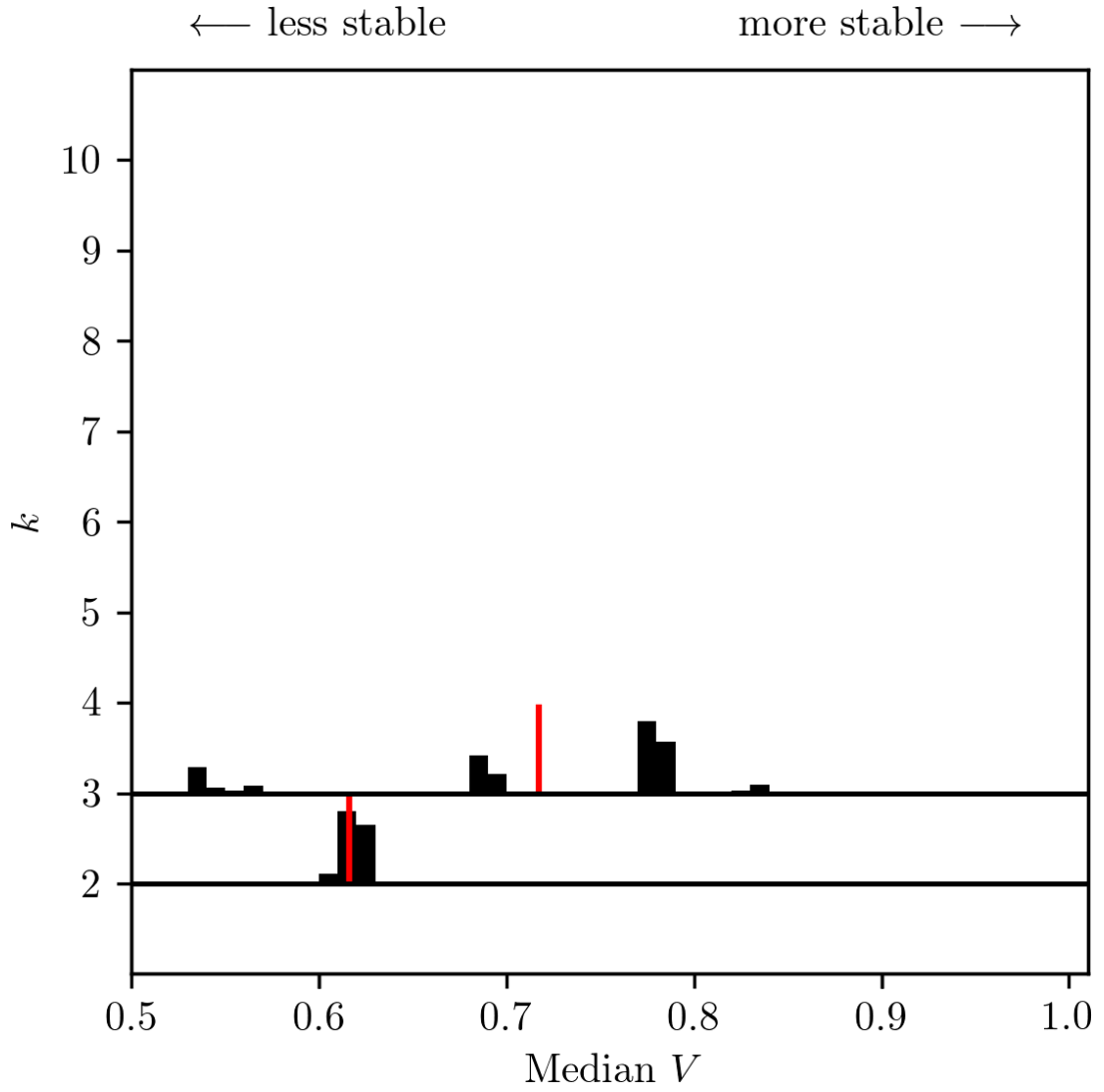


Simulated 2D example: $k = 3$

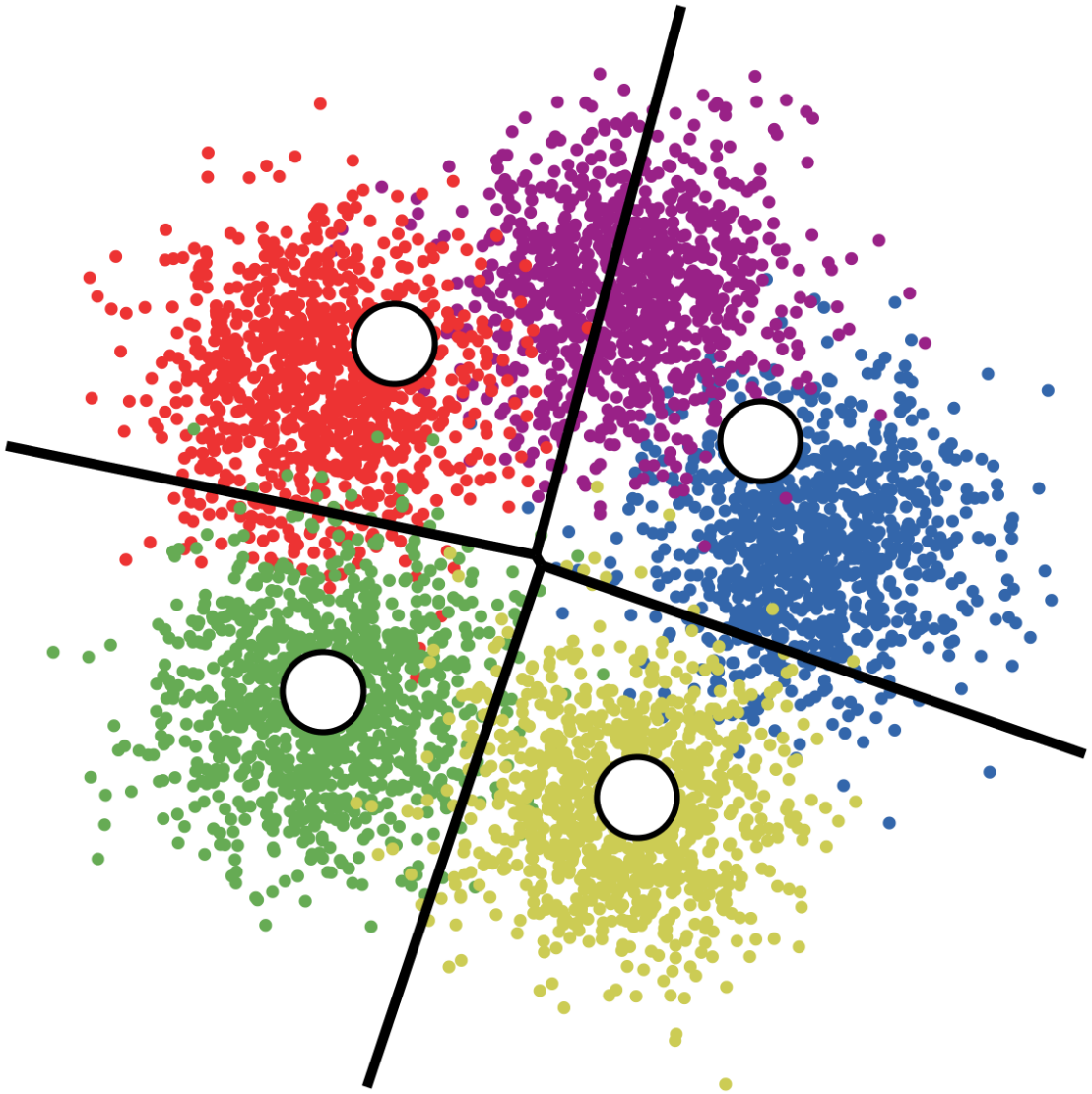


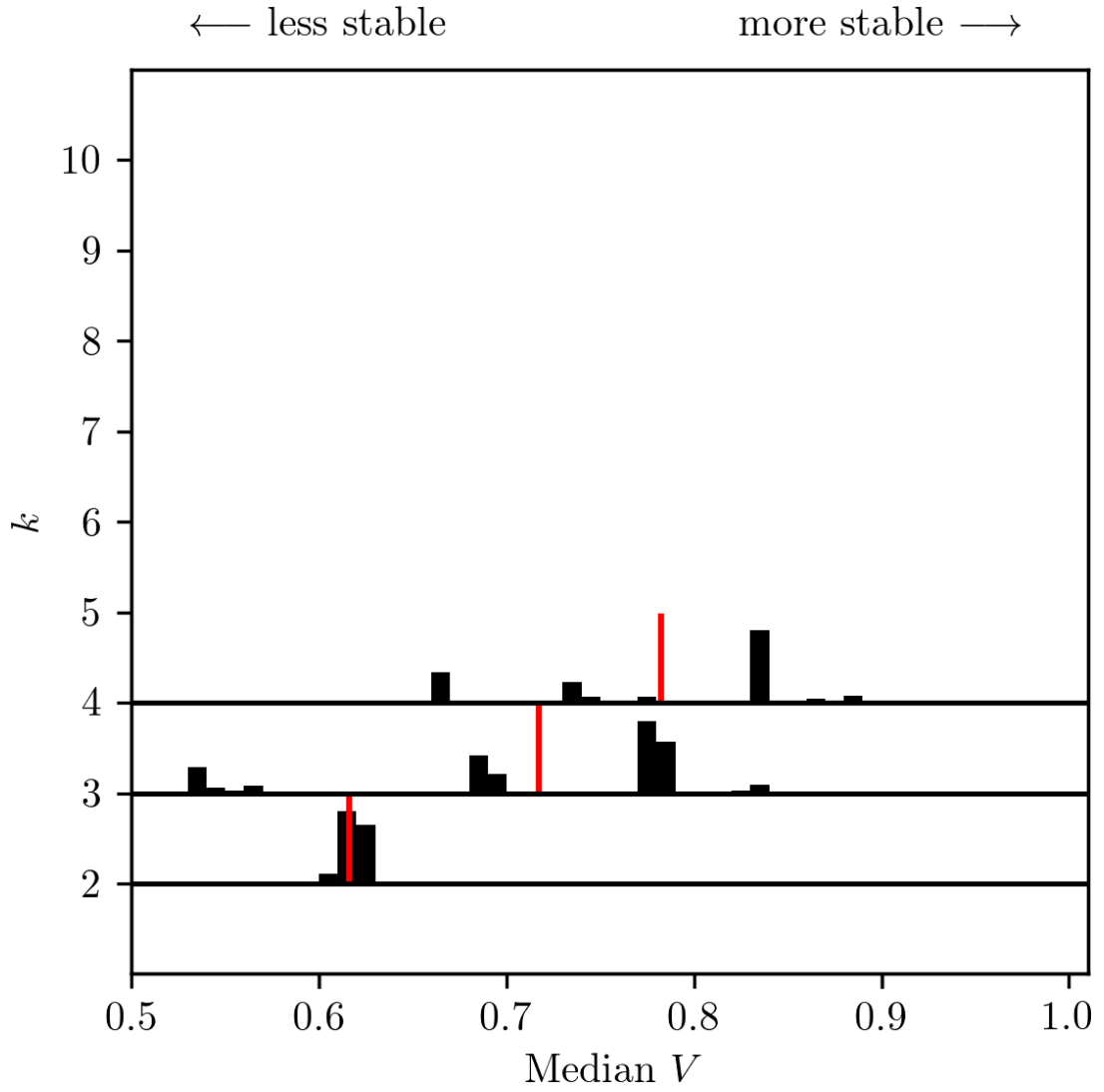


Simulated 2D example: $k = 3$

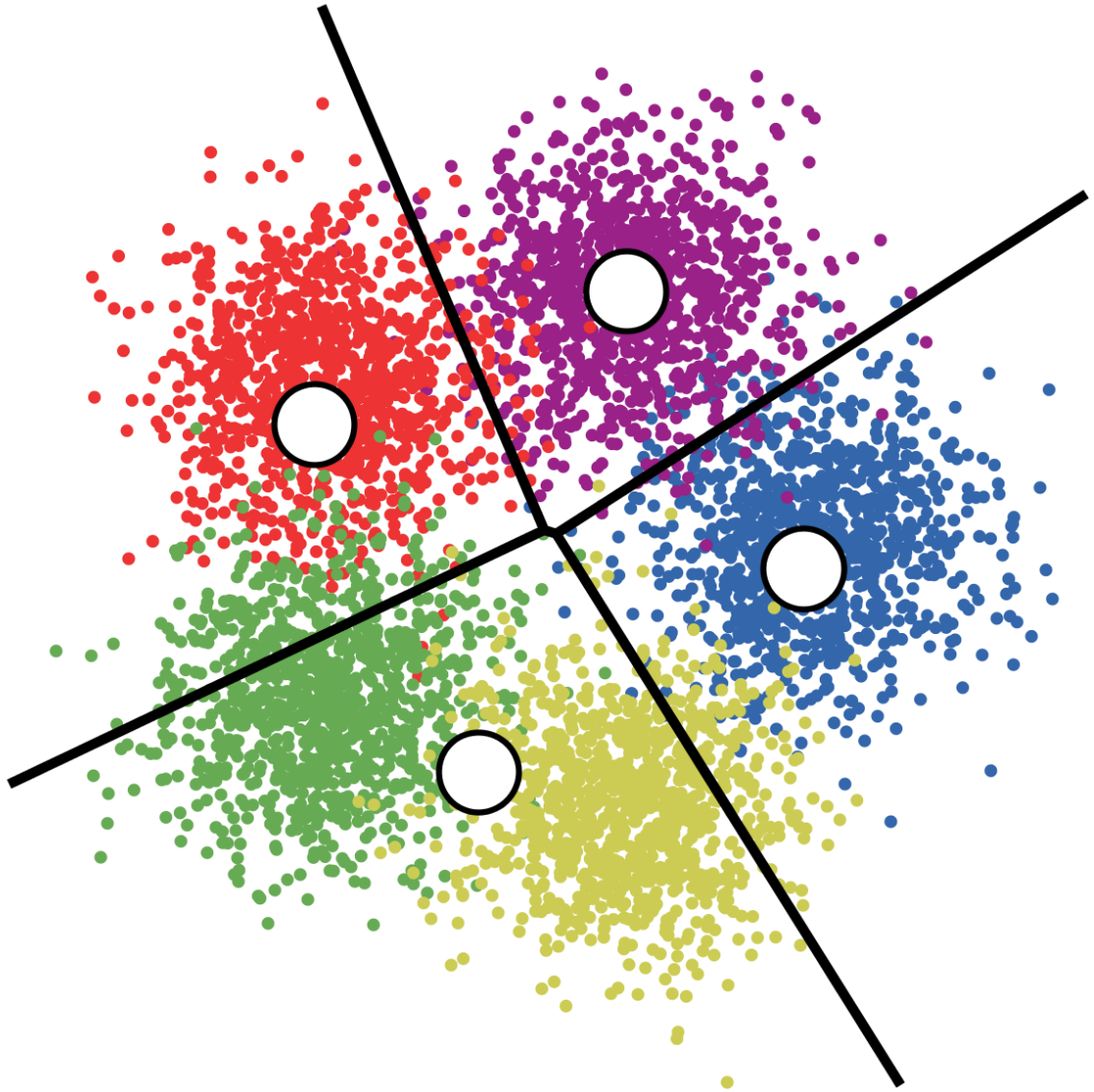


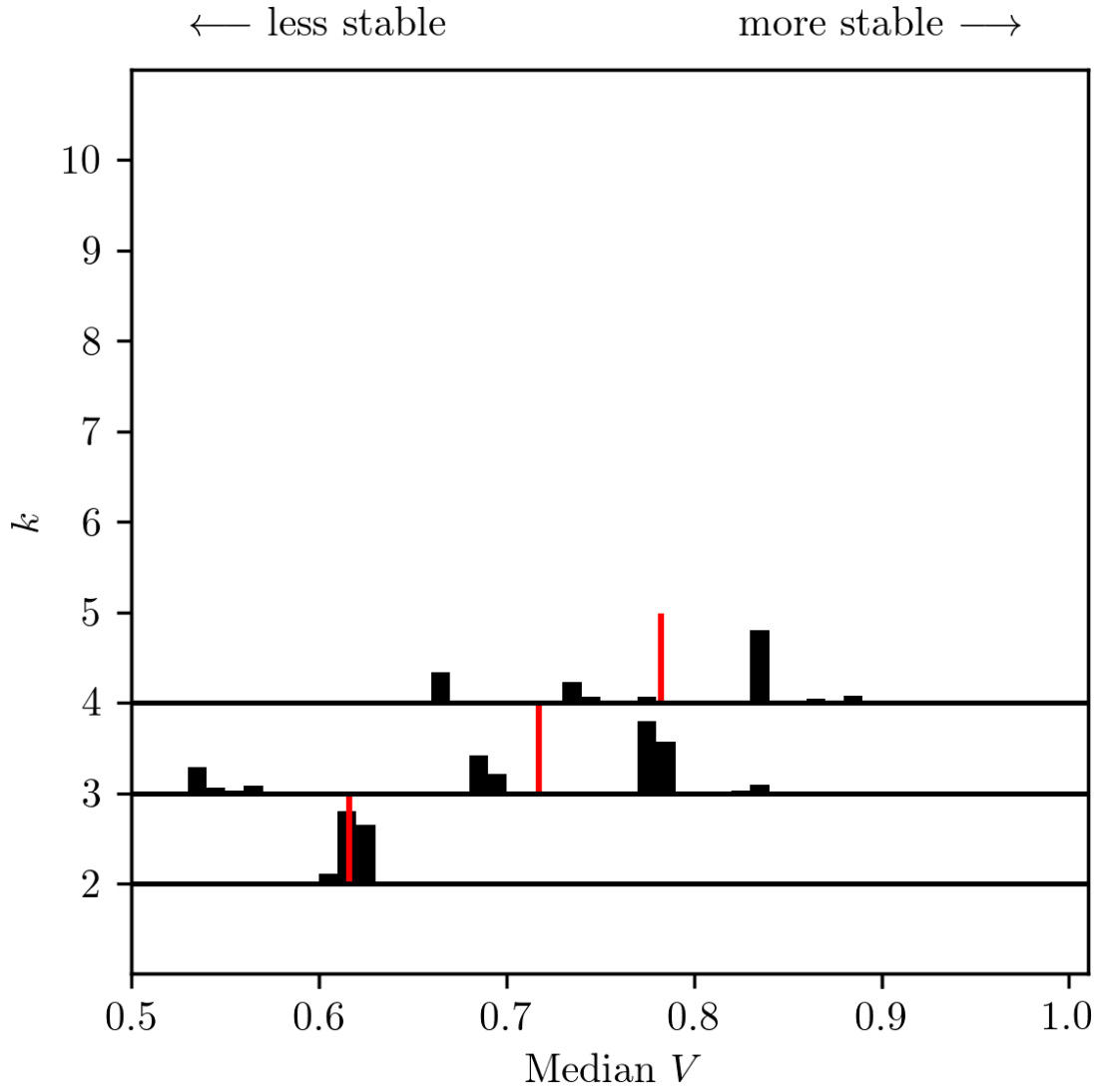
Simulated 2D example: $k = 4$



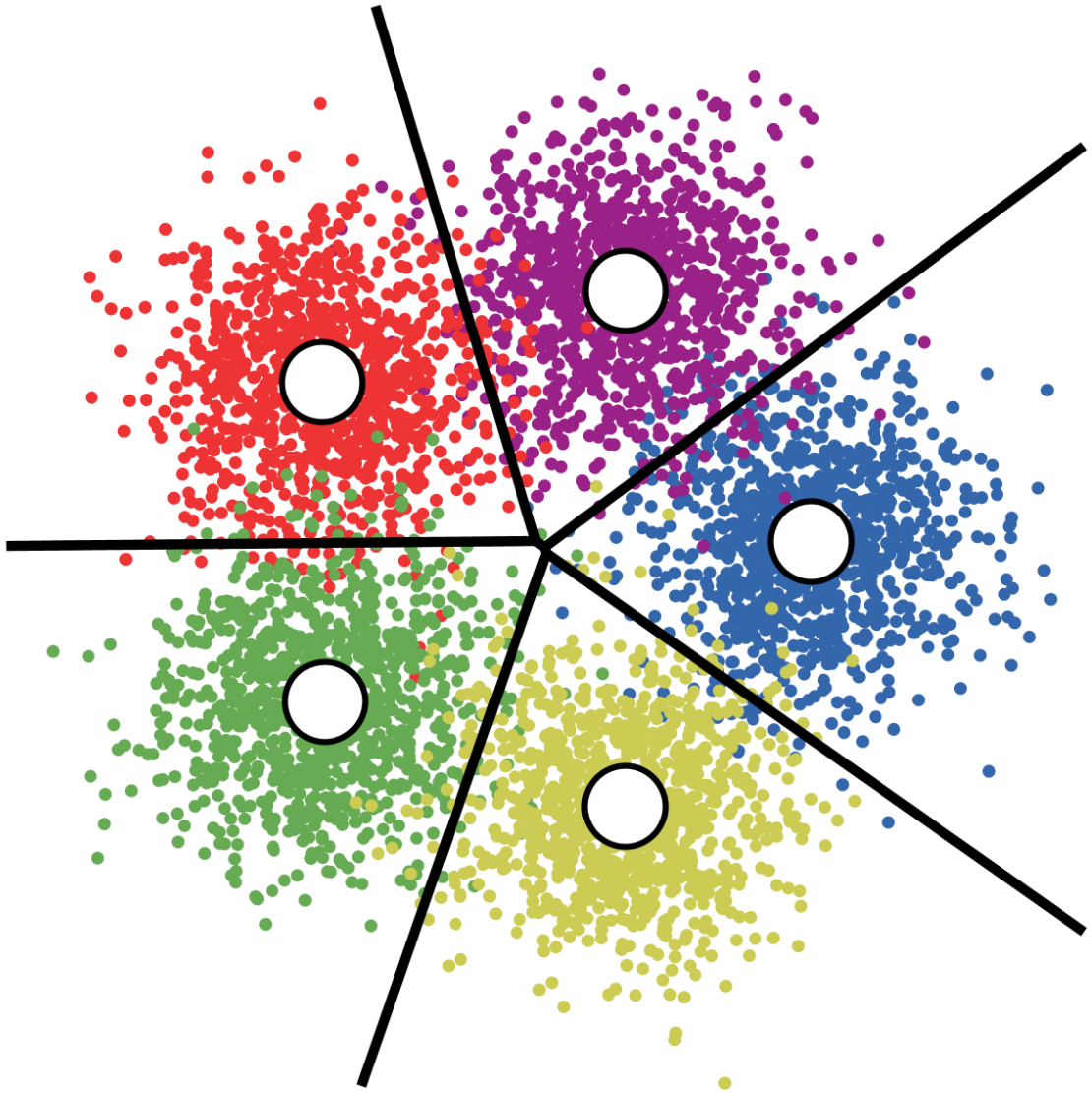


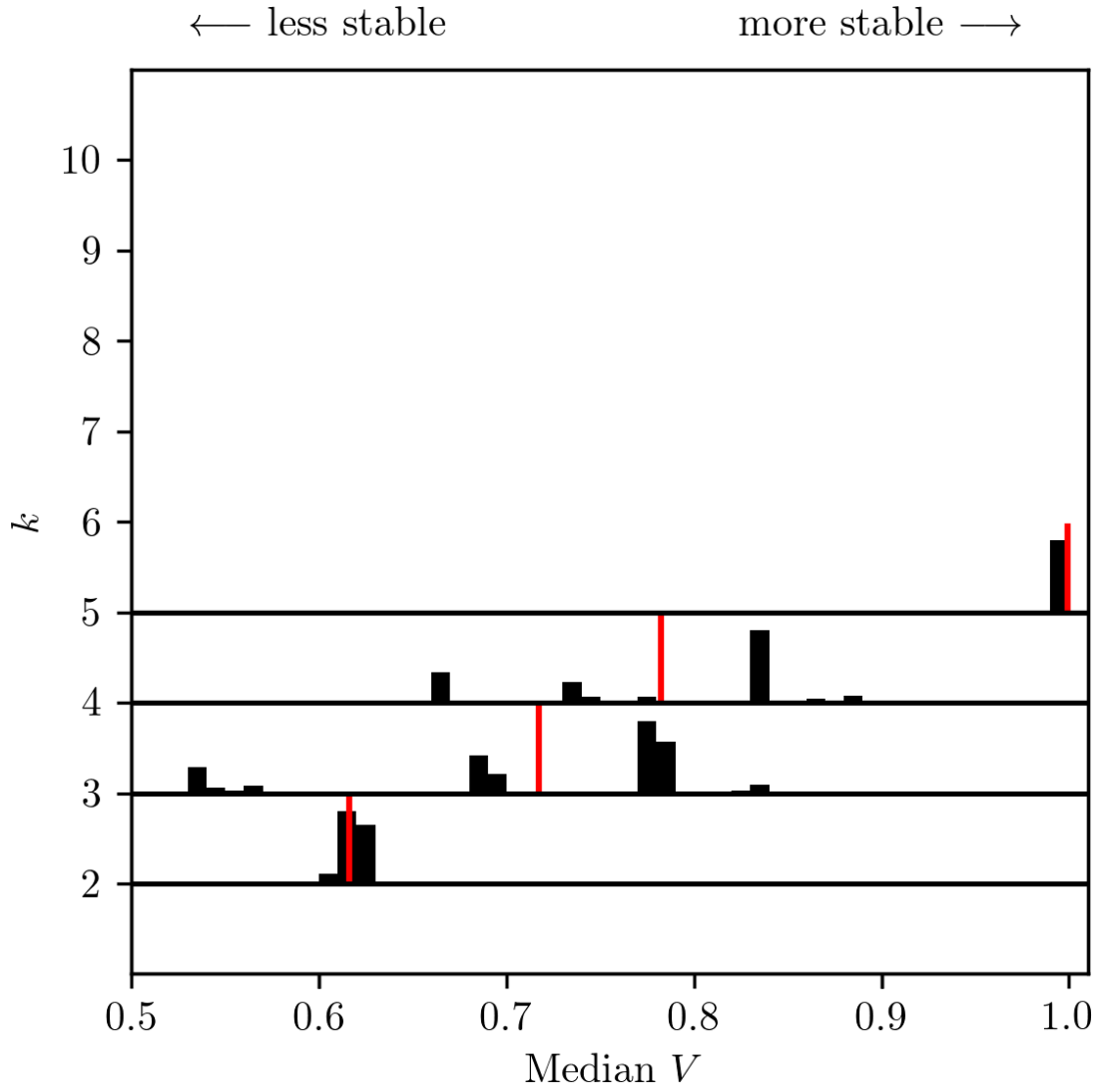
Simulated 2D example: $k = 4$



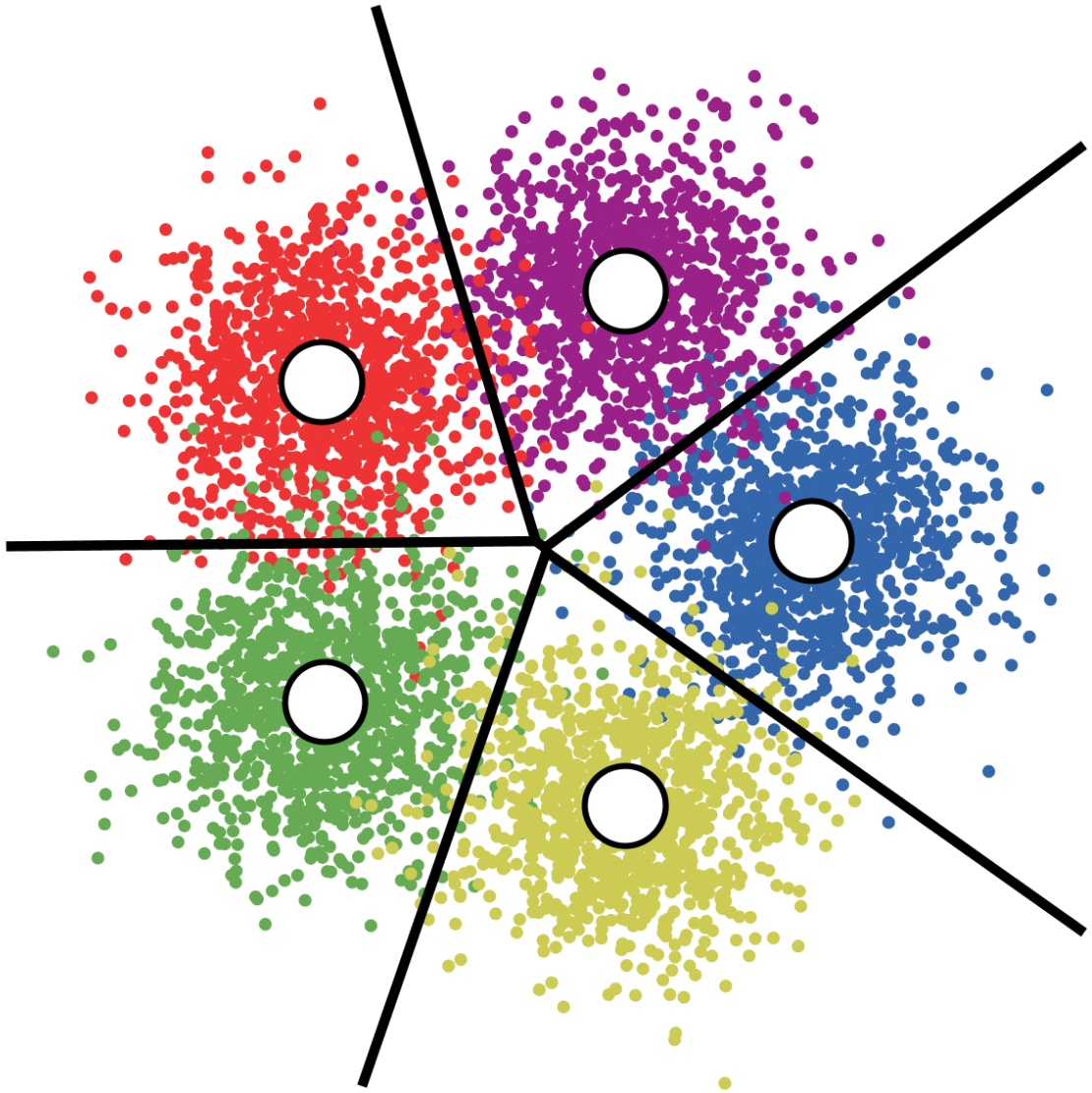


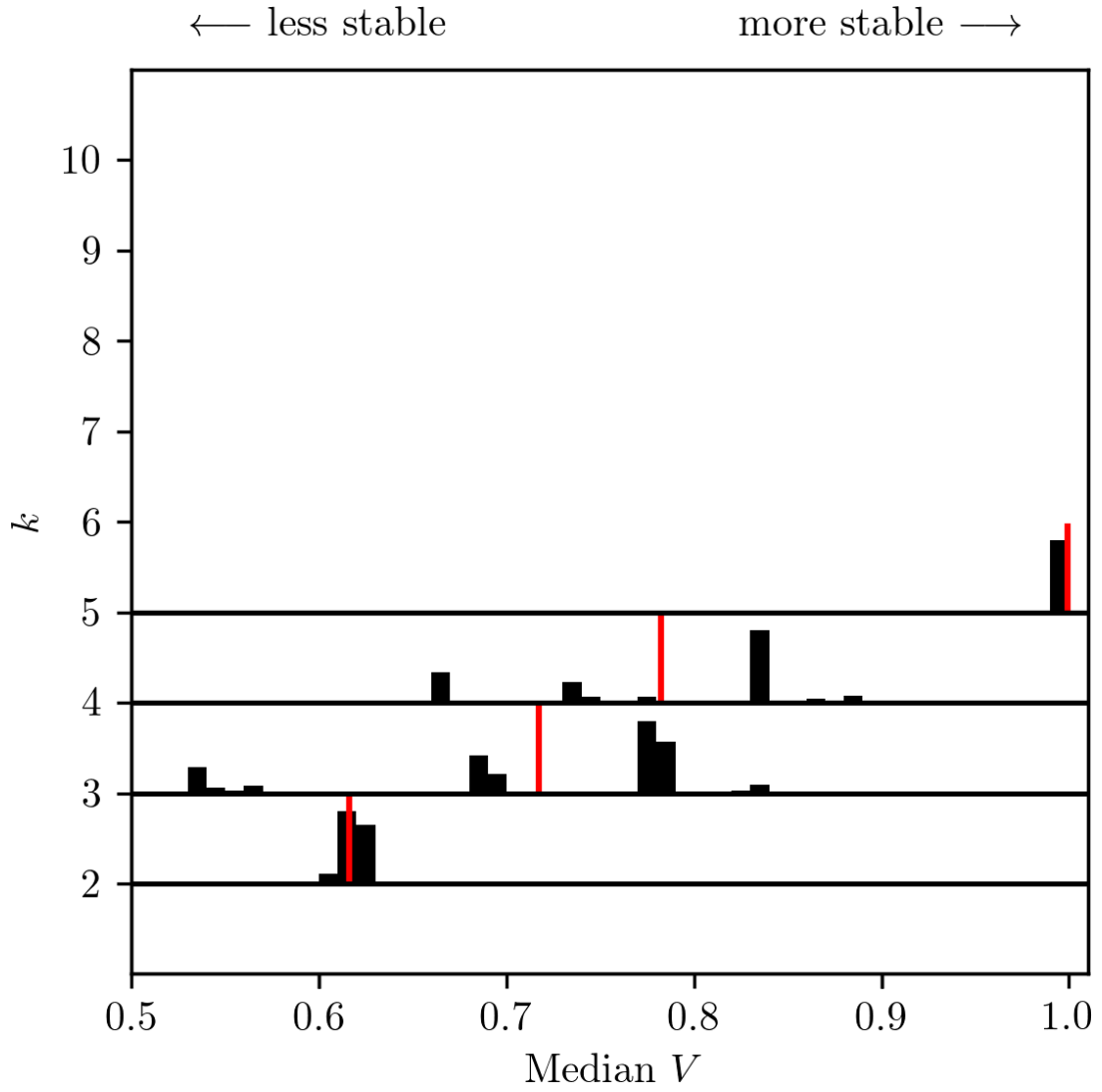
Simulated 2D example: $k = 5$



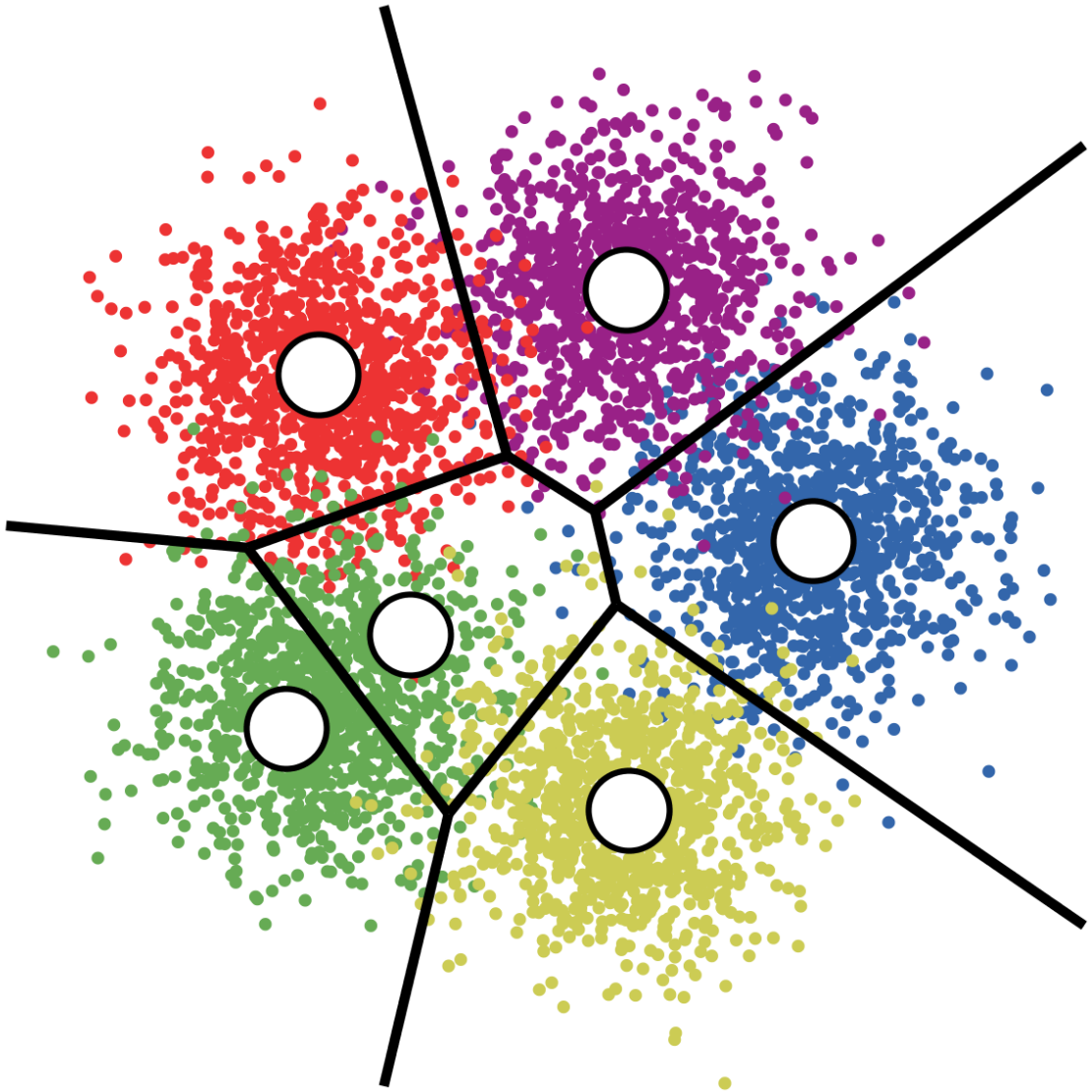


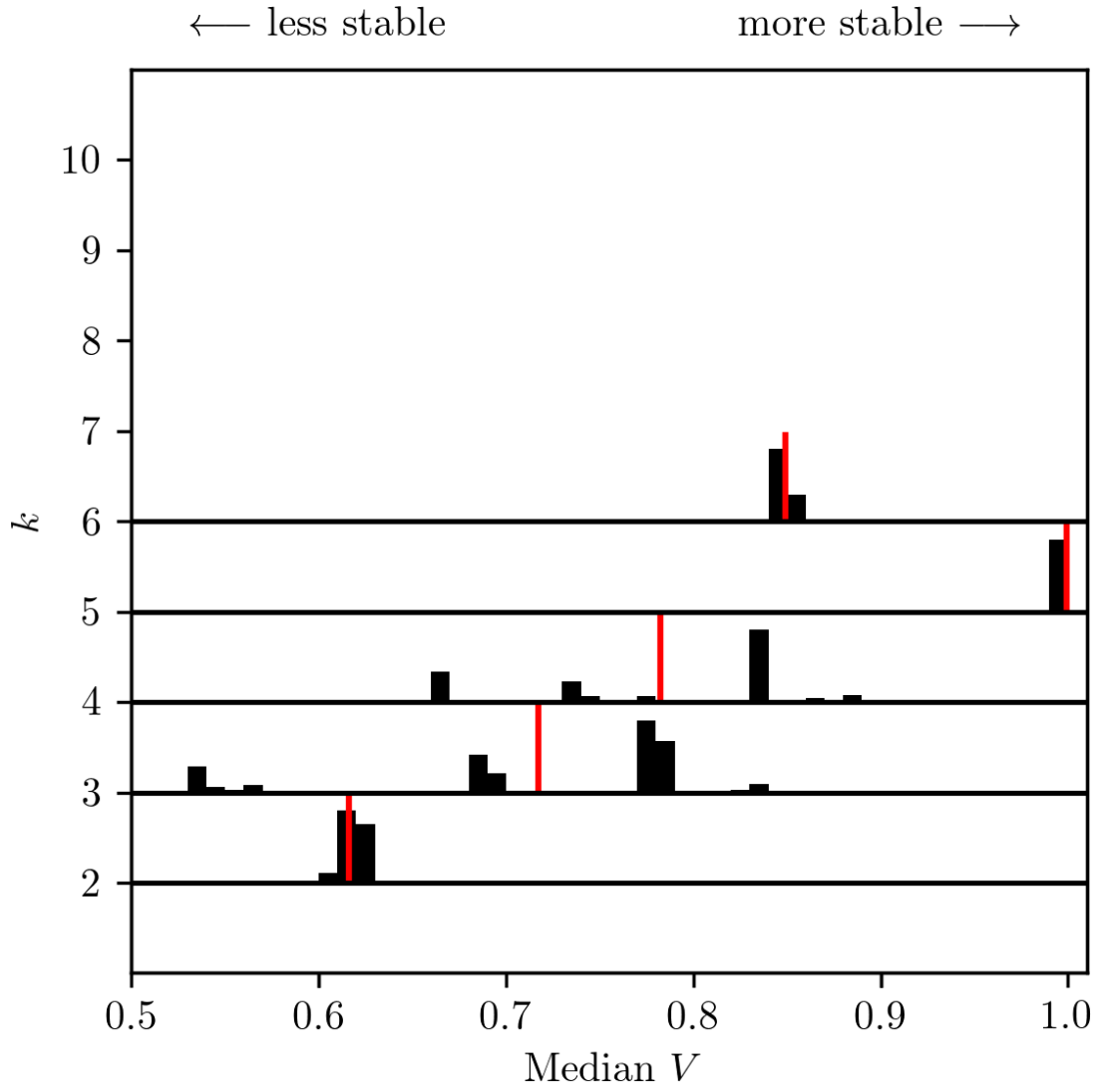
Simulated 2D example: $k = 5$



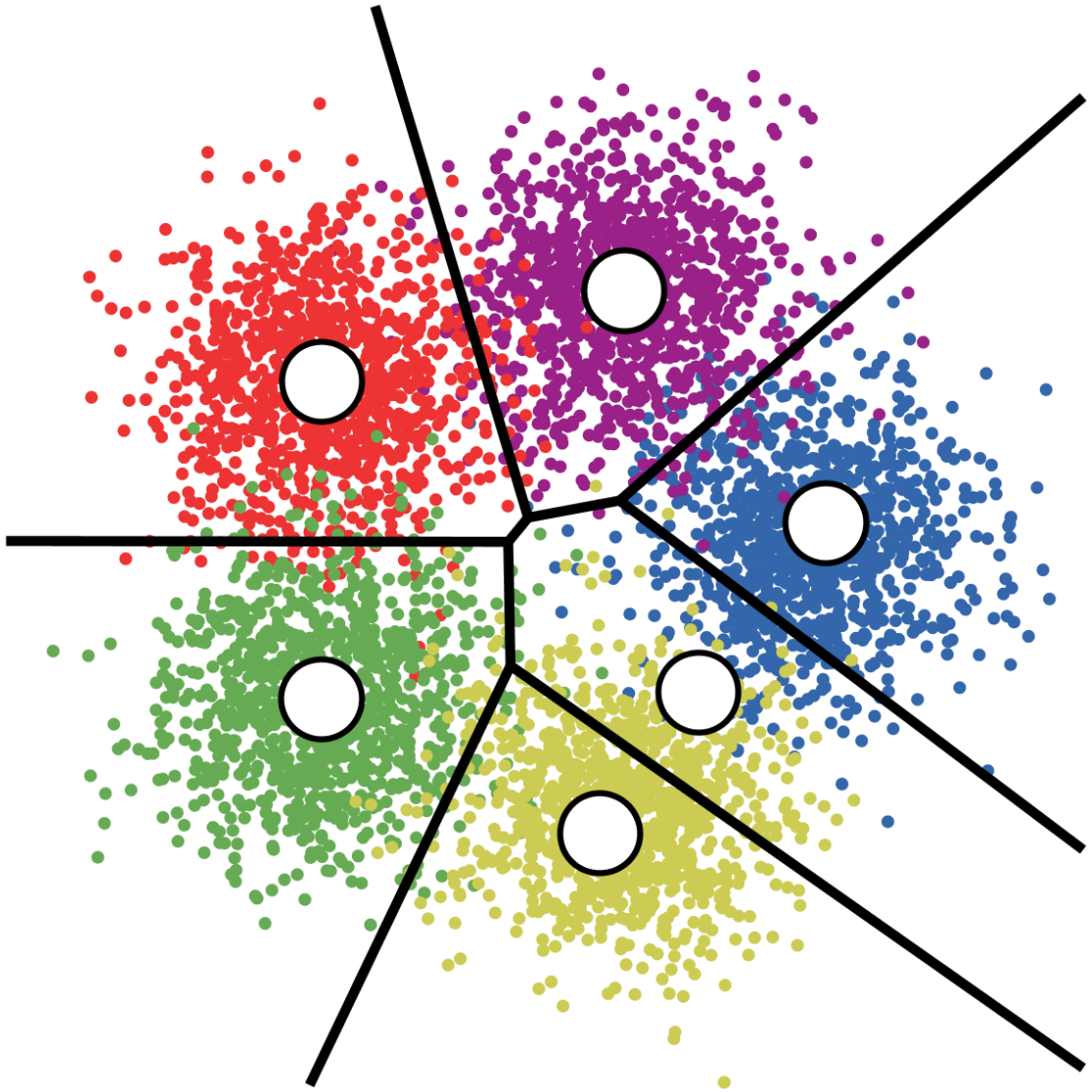


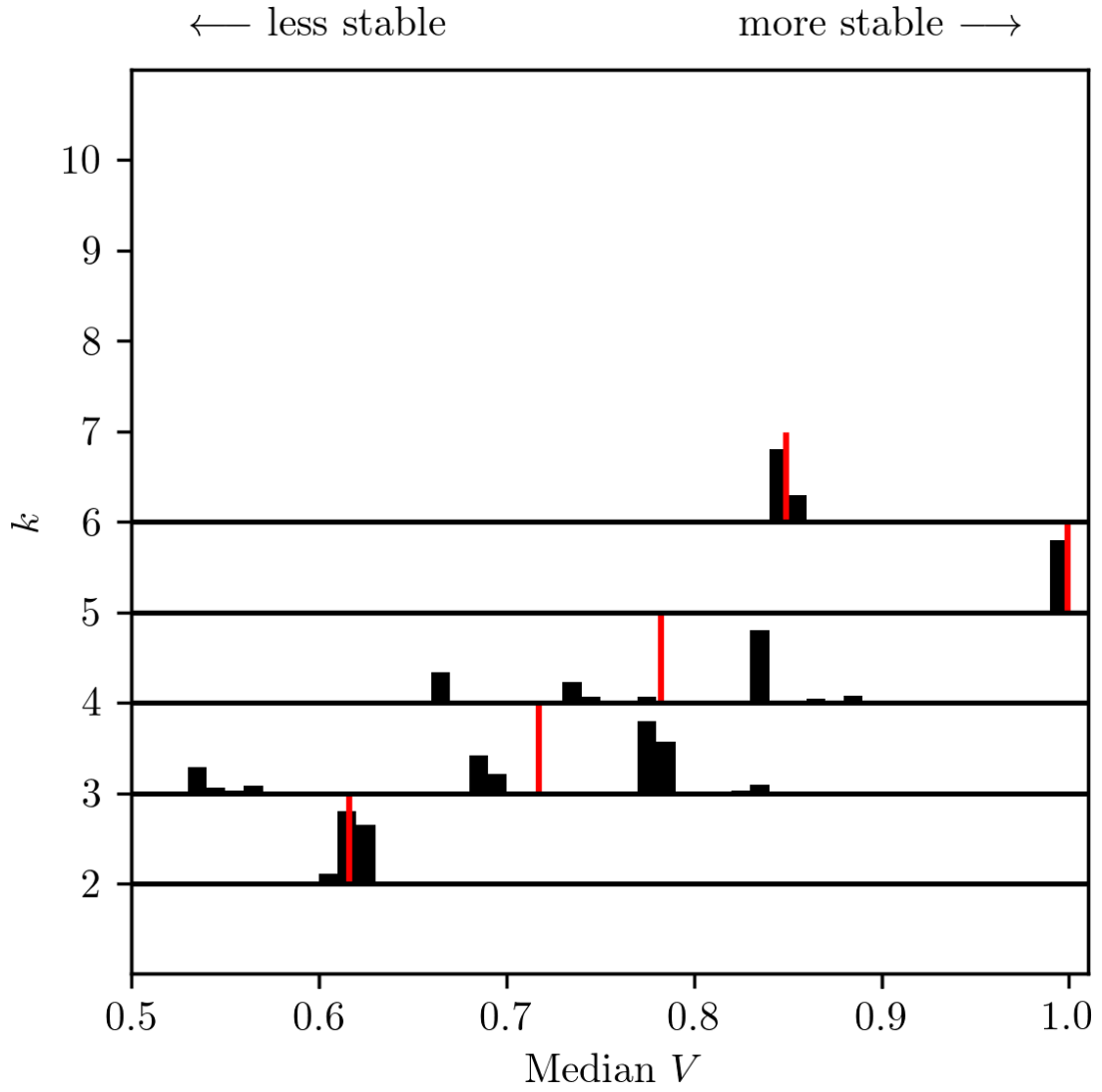
Simulated 2D example: $k = 6$



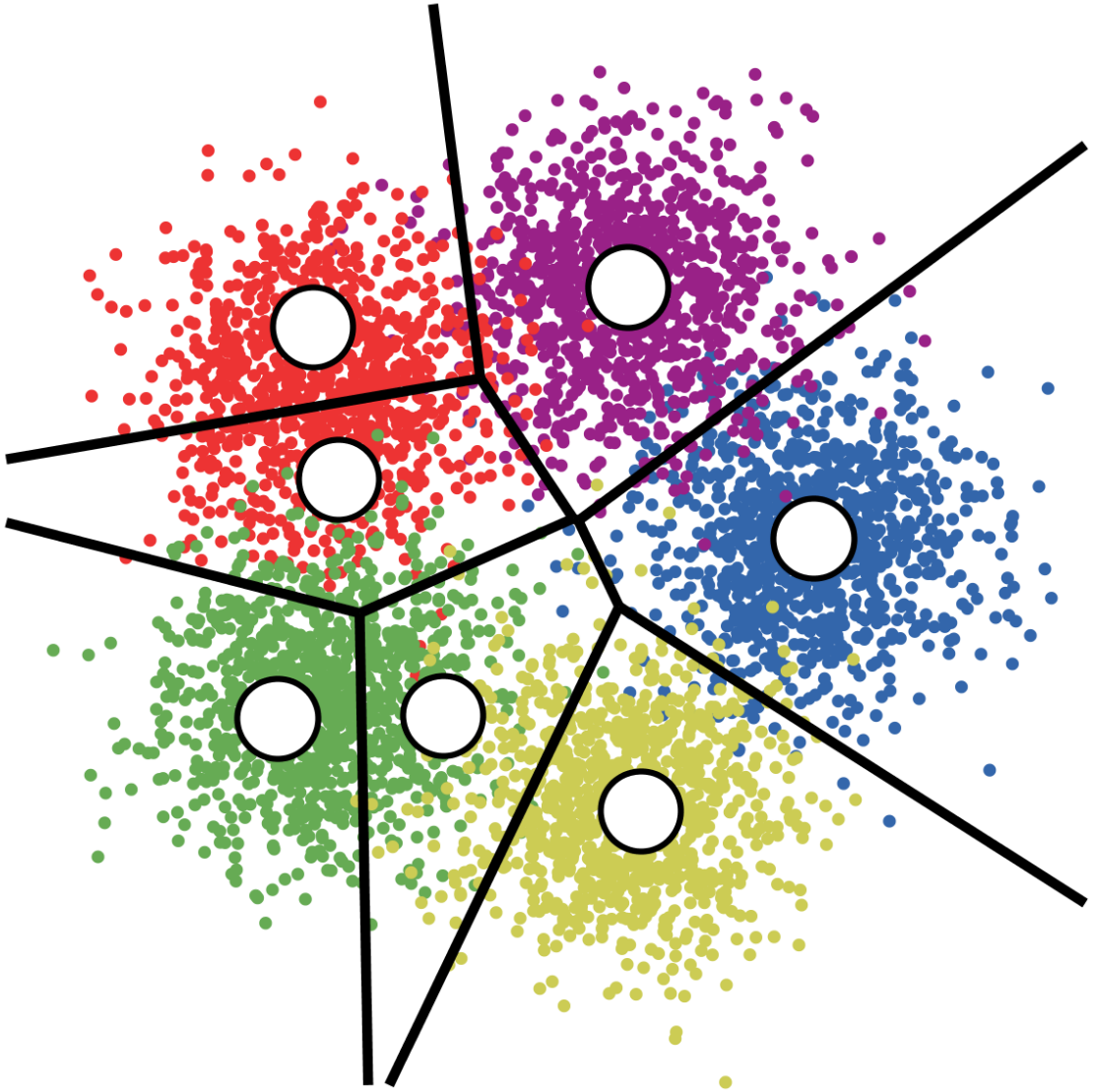


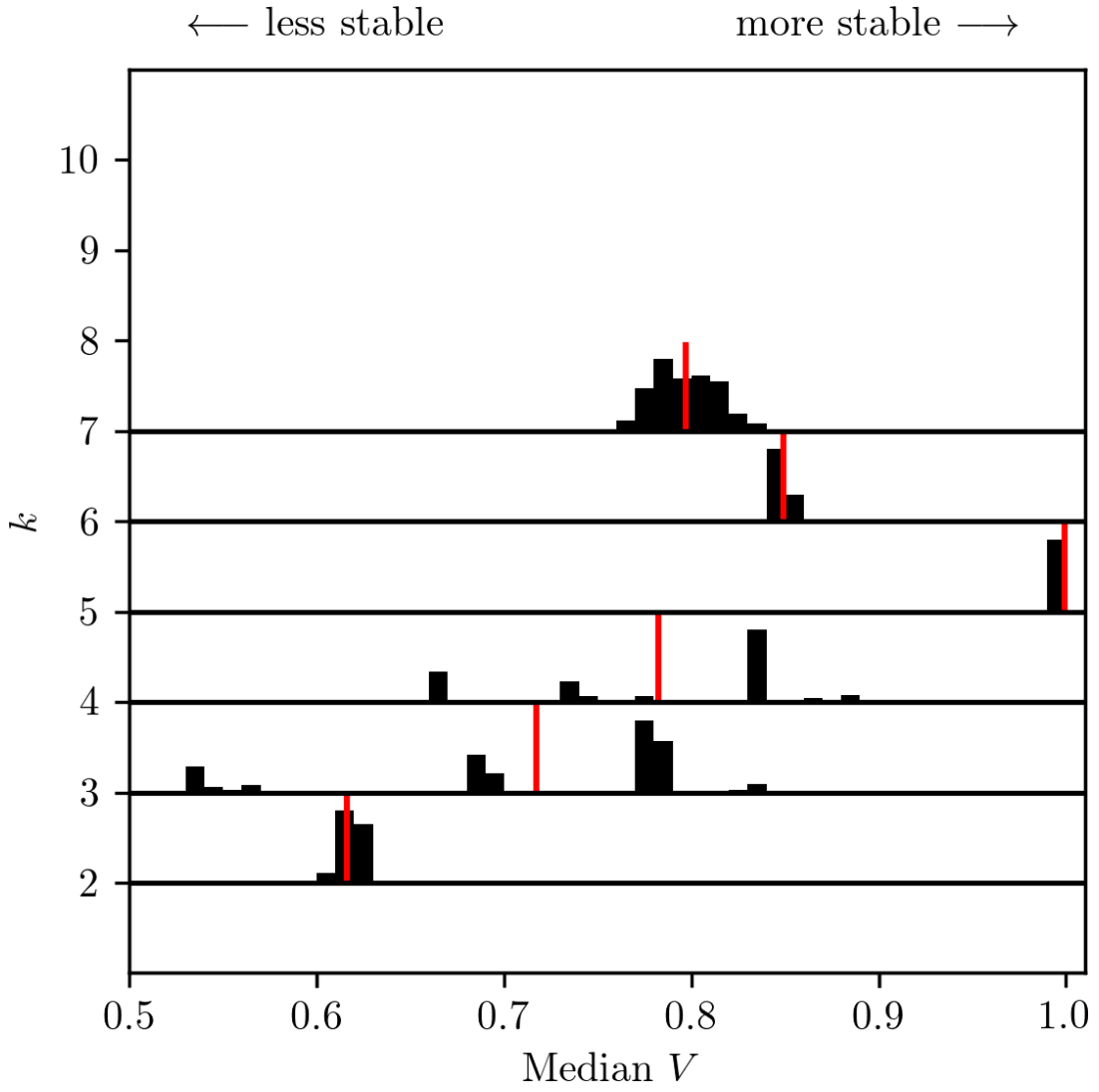
Simulated 2D example: $k = 6$



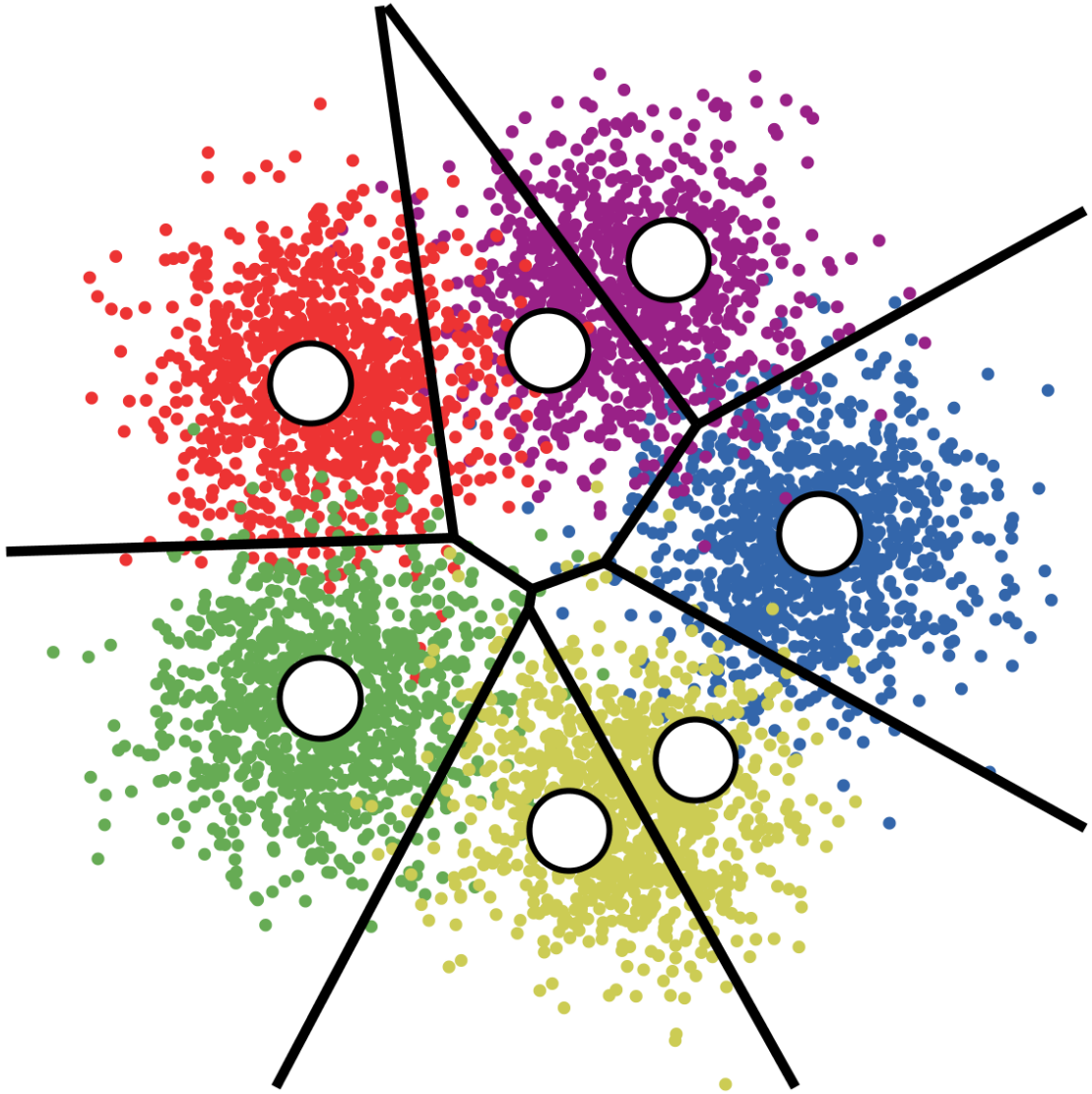


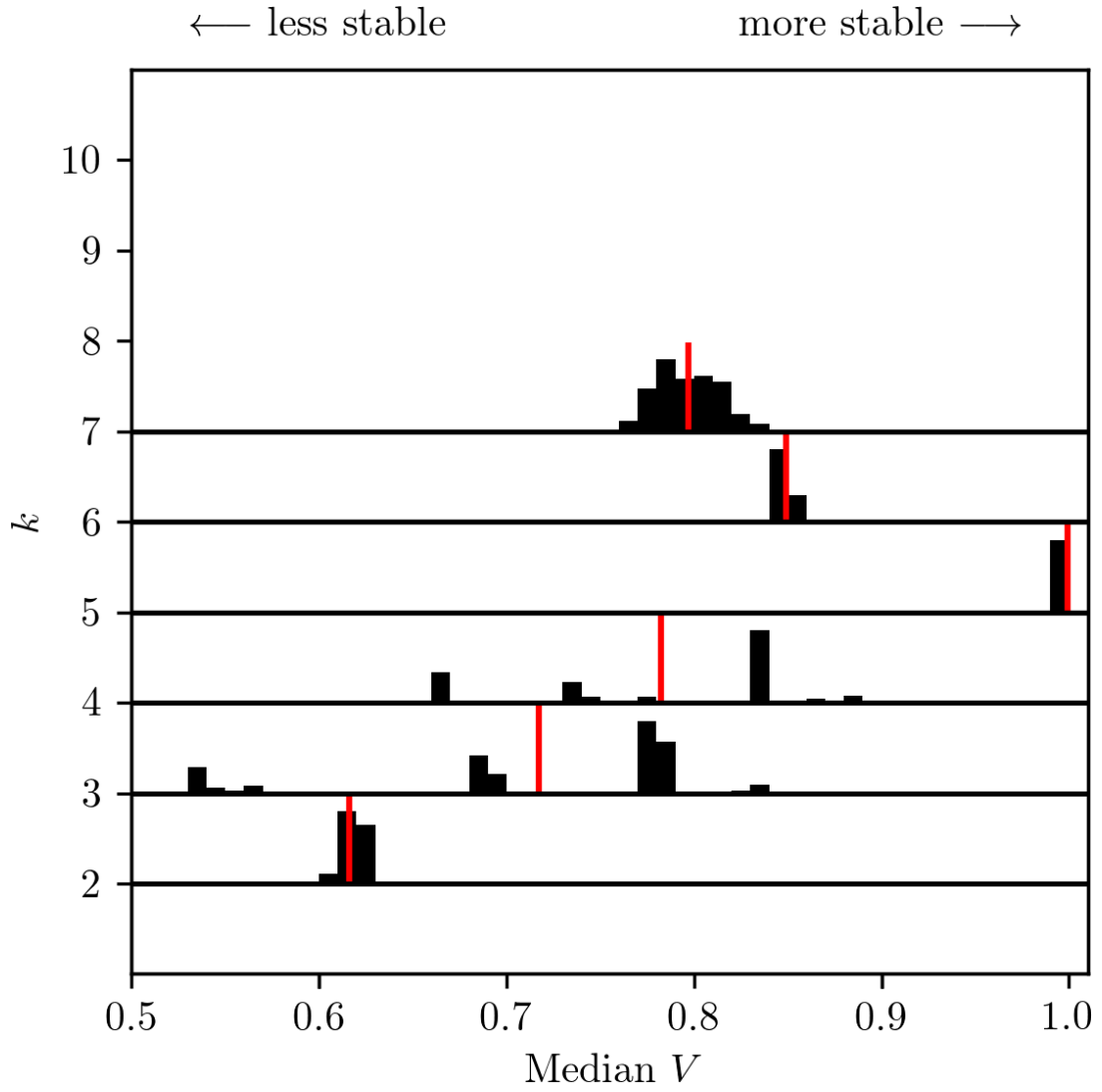
Simulated 2D example: $k = 7$



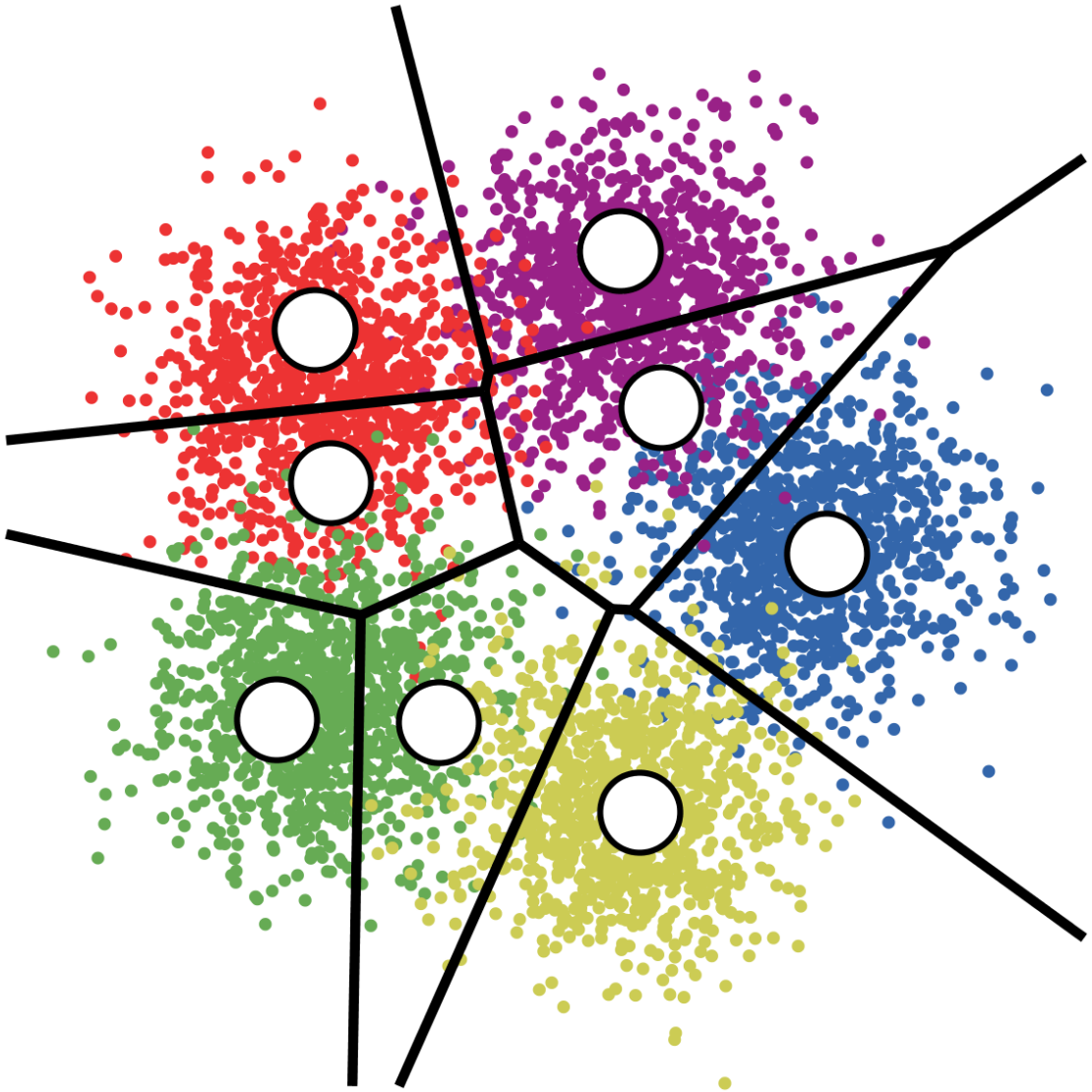


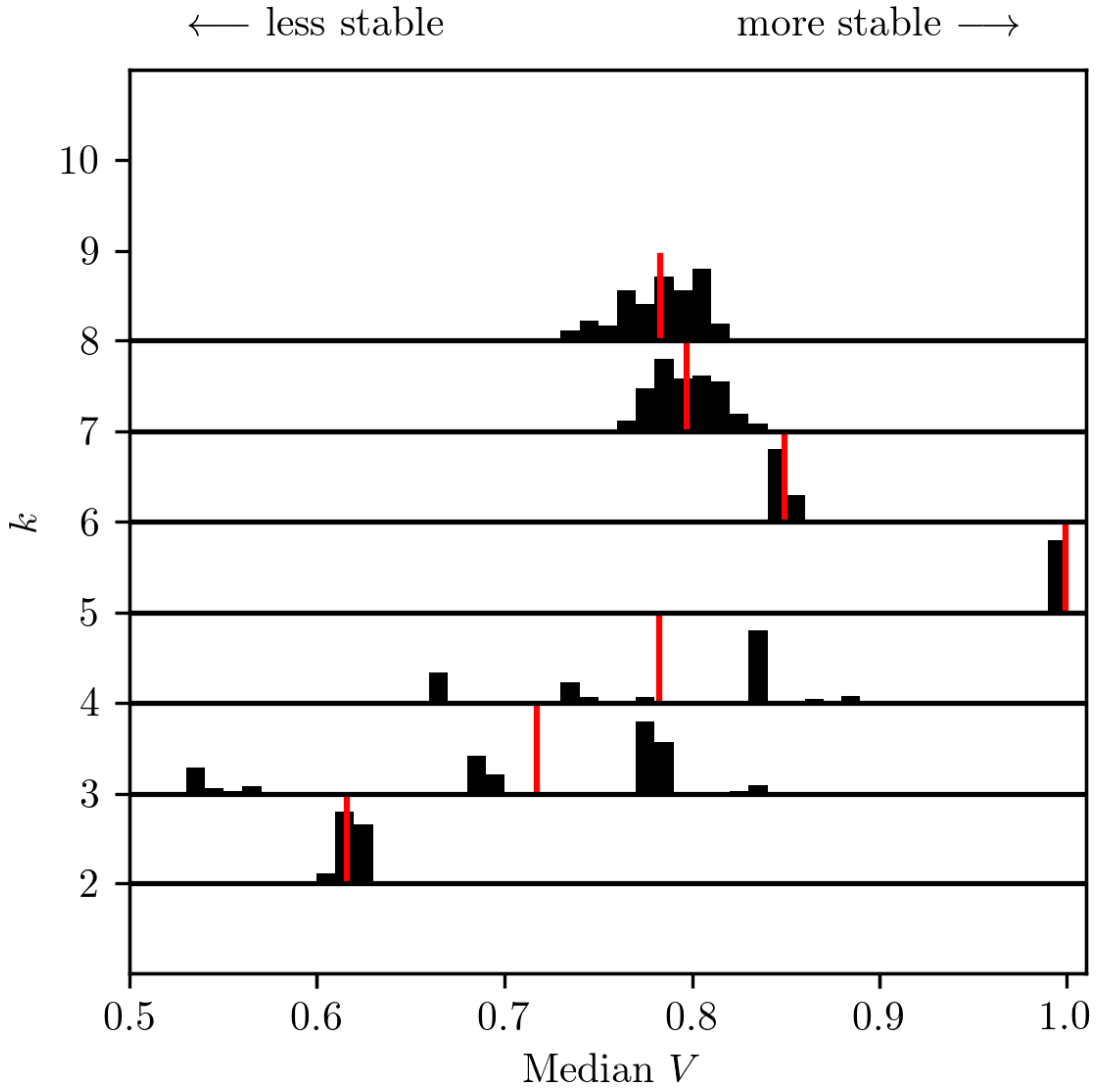
Simulated 2D example: $k = 7$



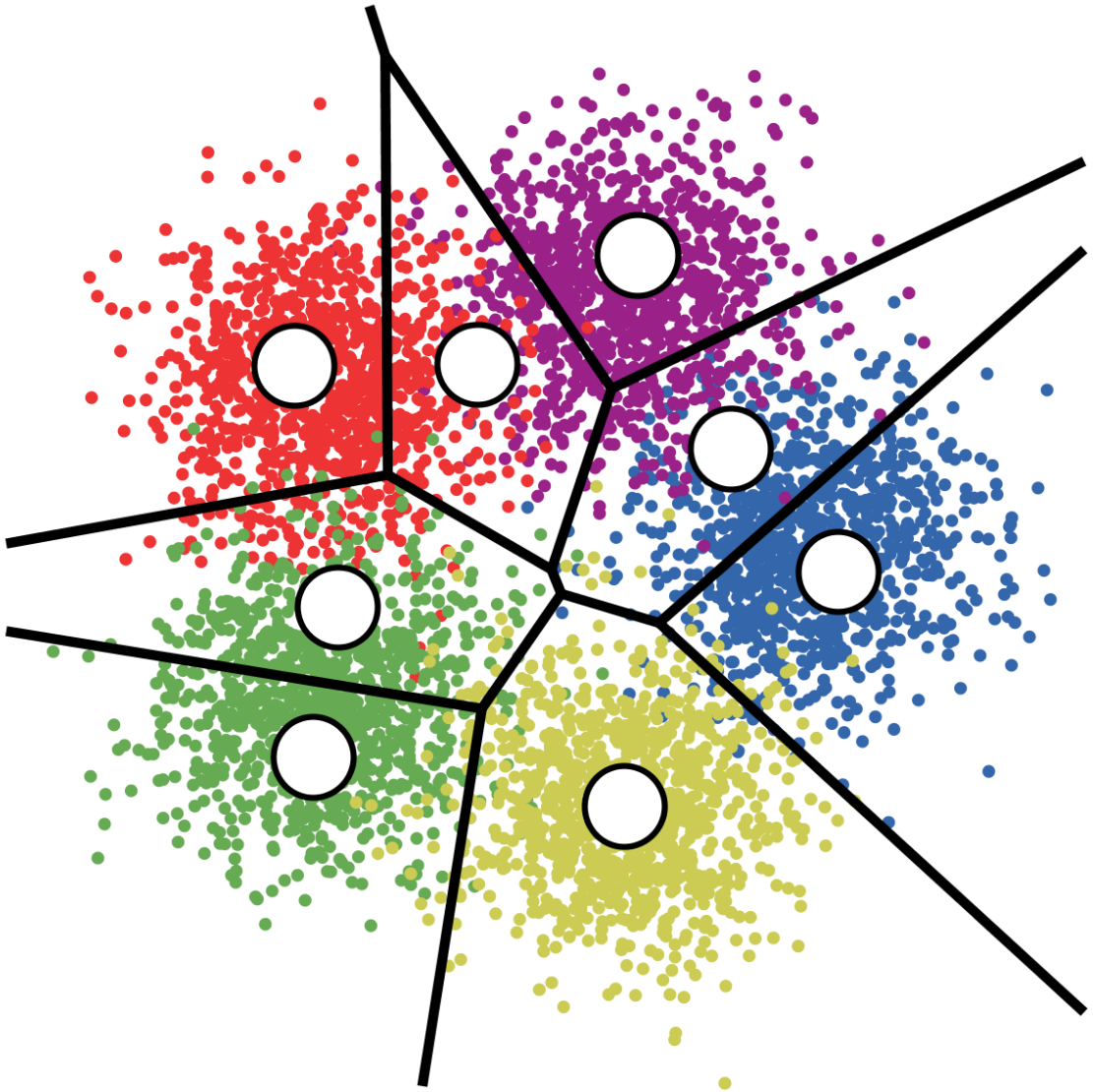


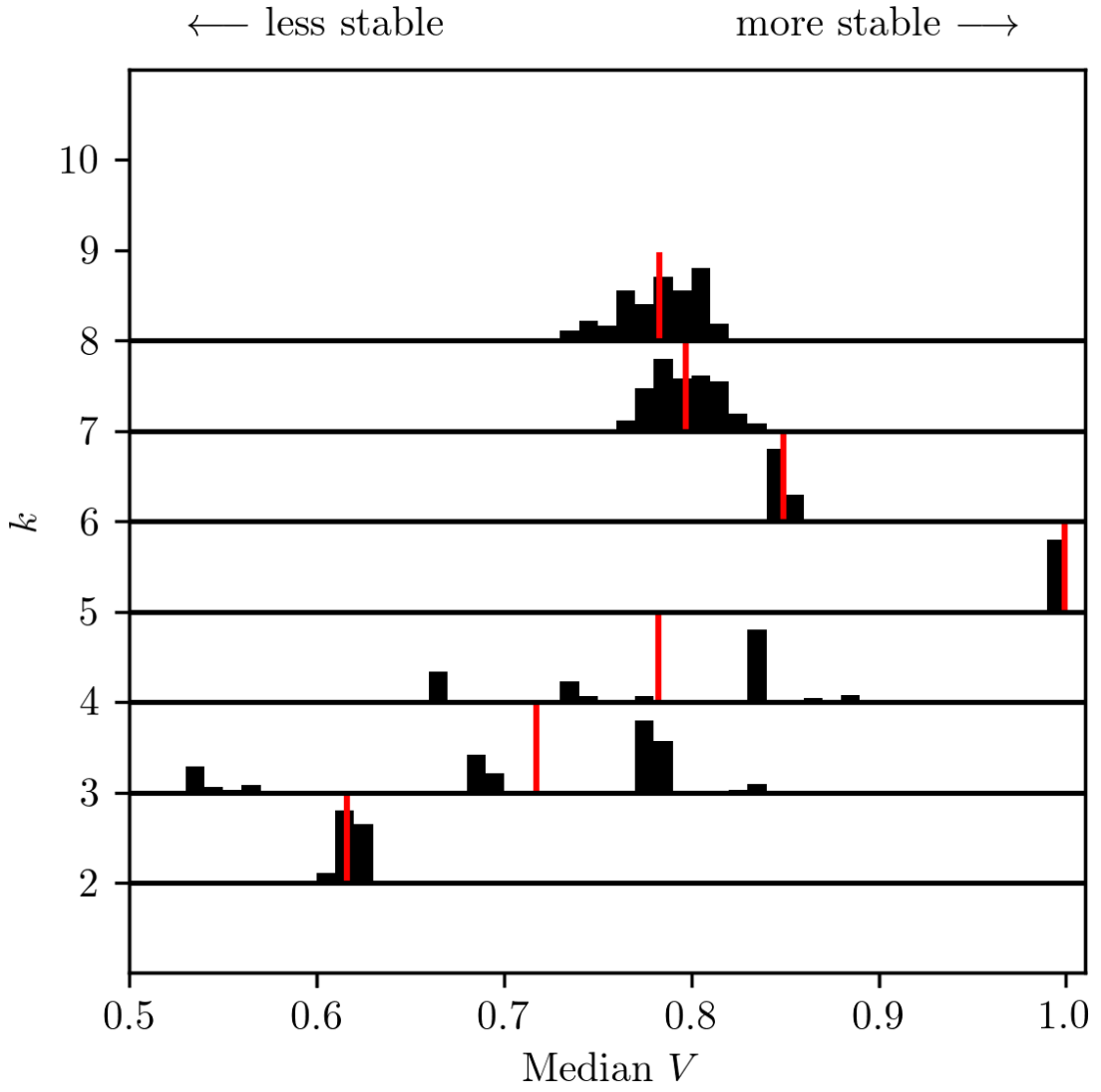
Simulated 2D example: $k = 8$



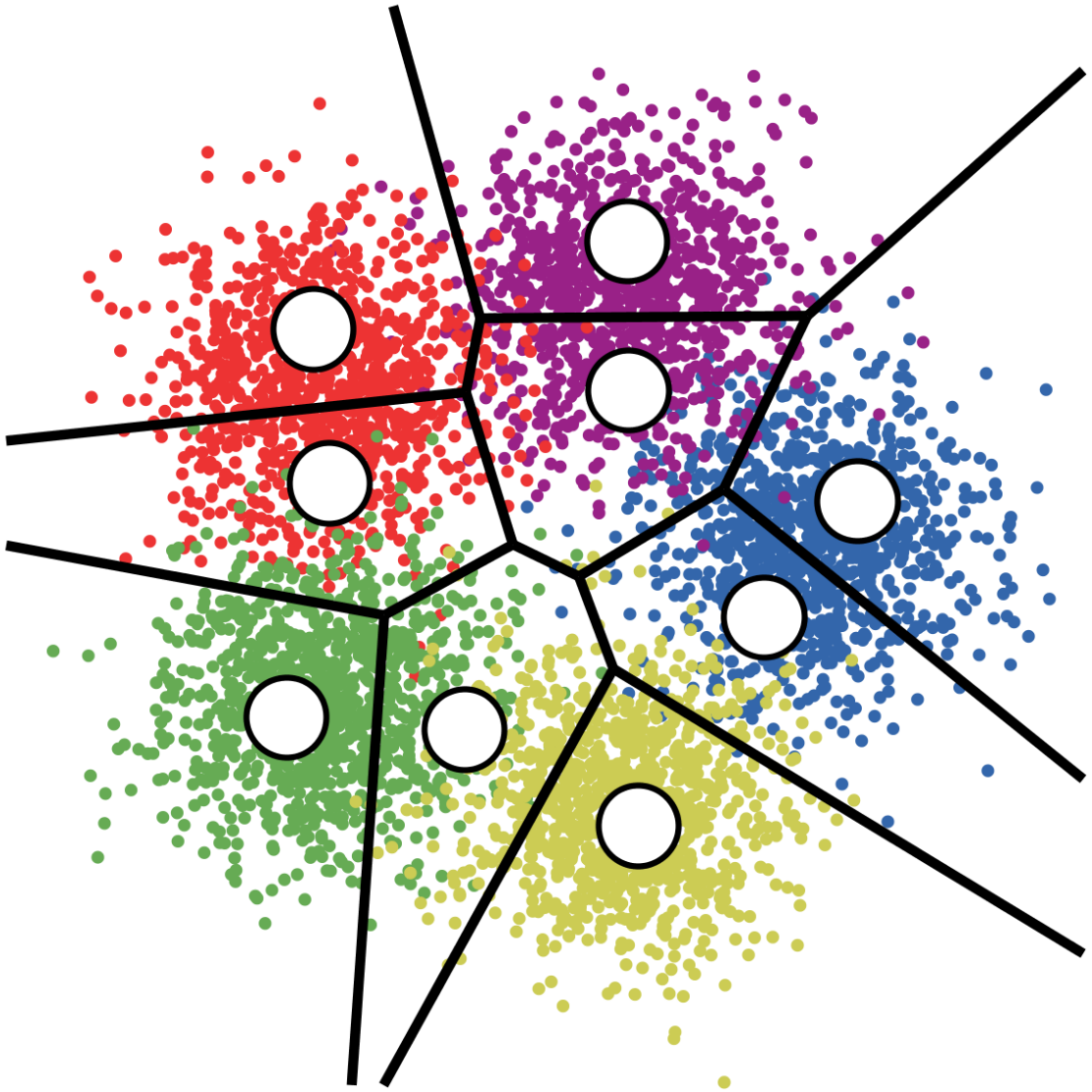


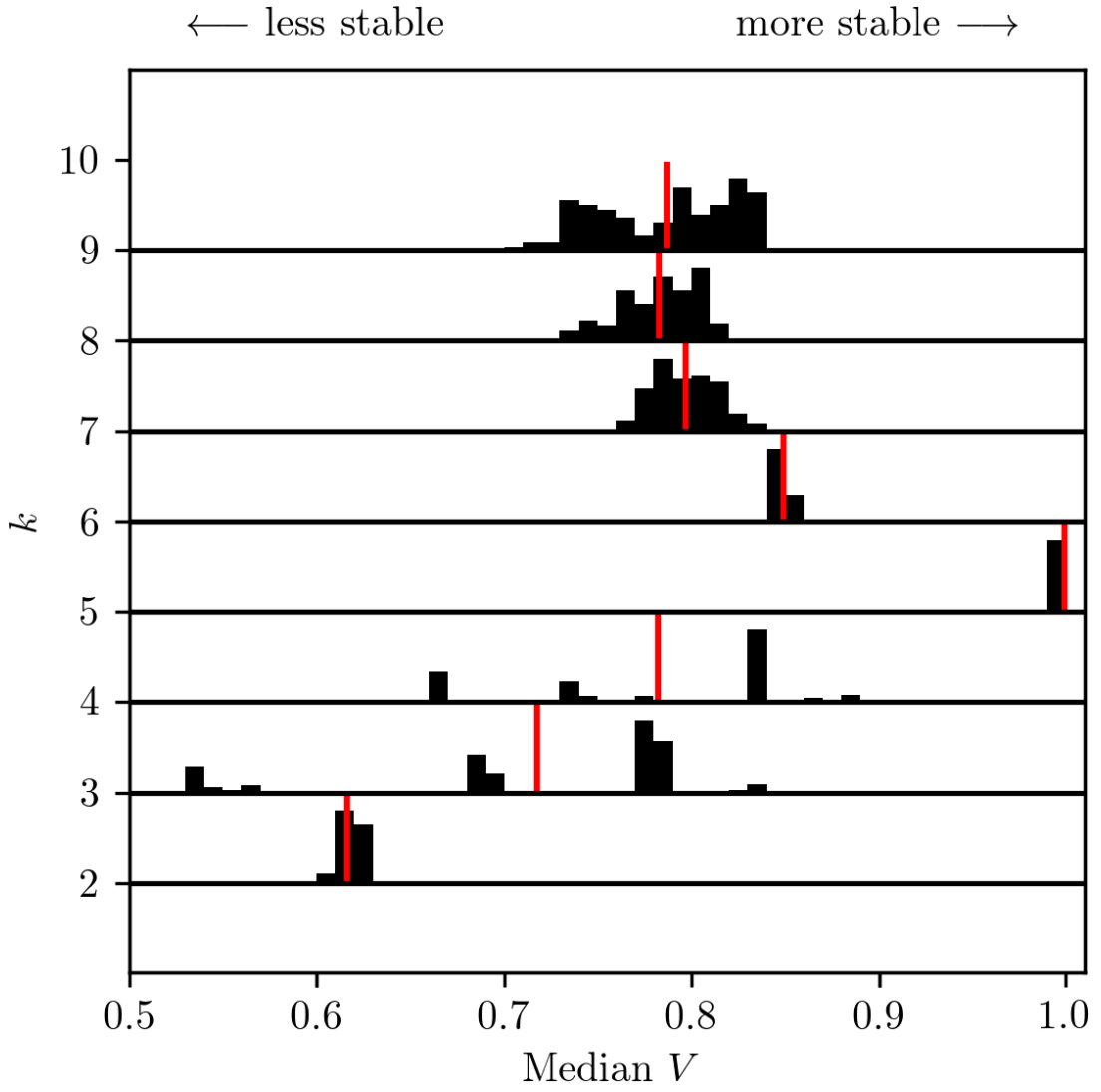
Simulated 2D example: $k = 8$



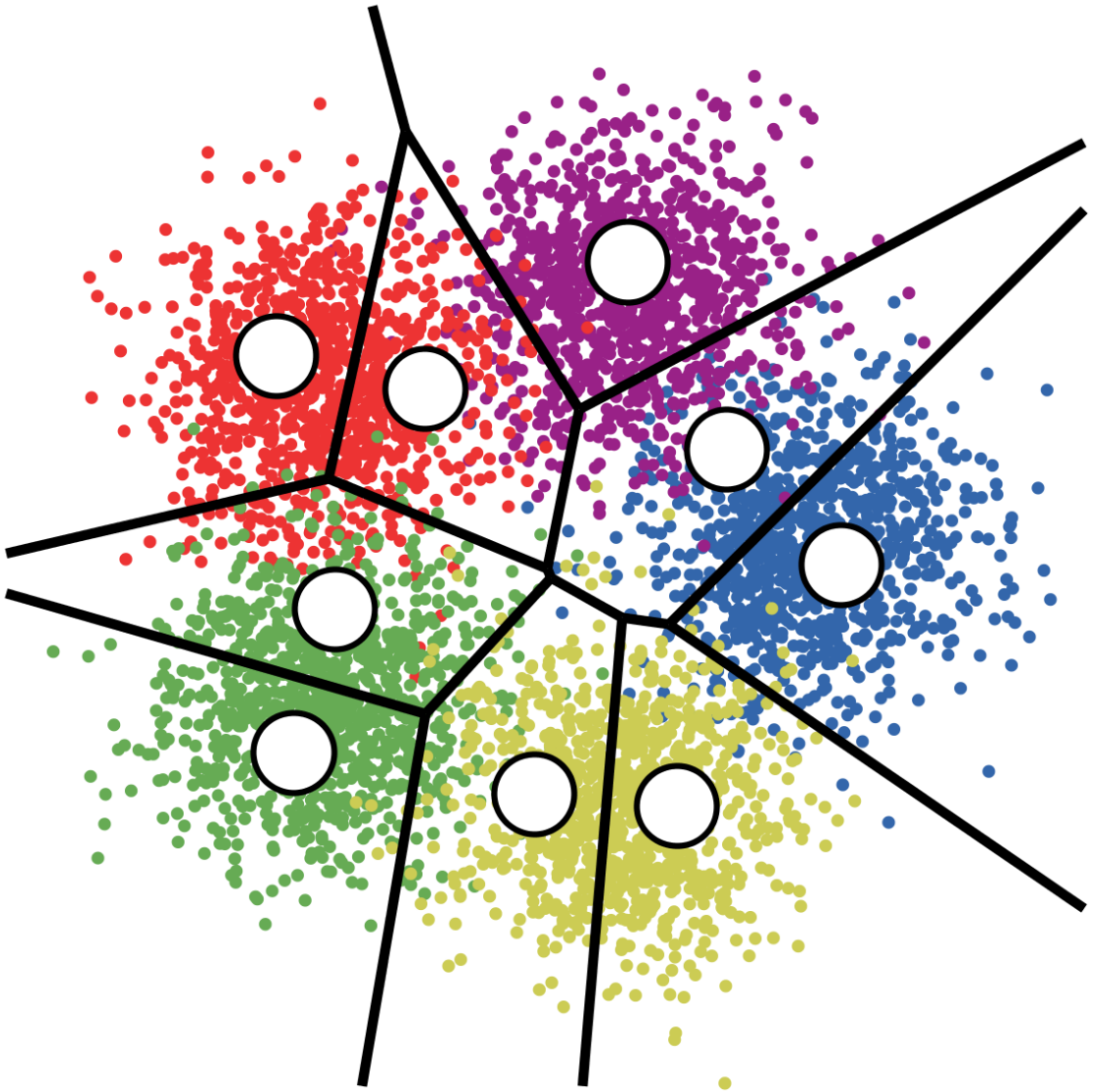


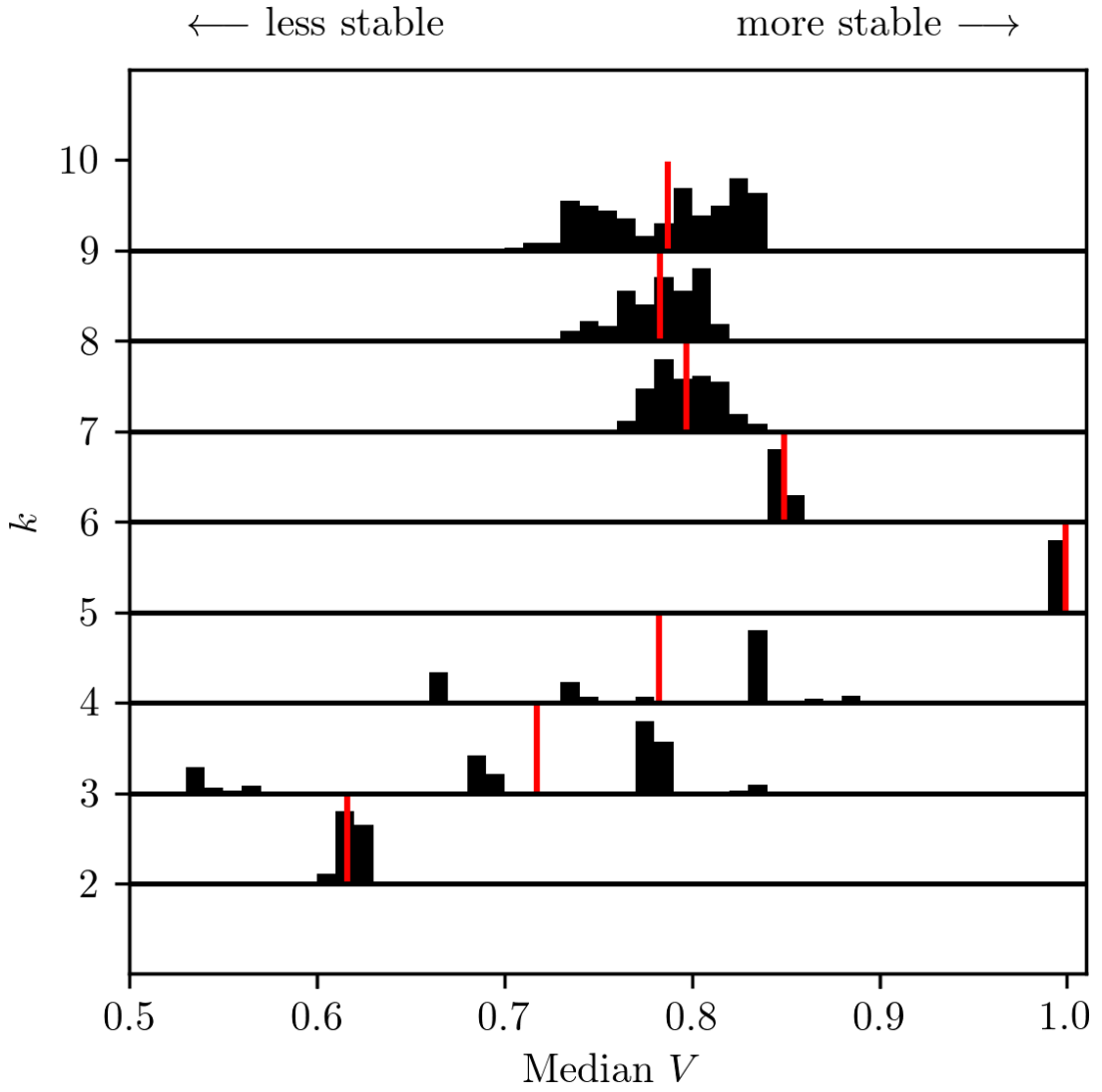
Simulated 2D example: $k = 9$



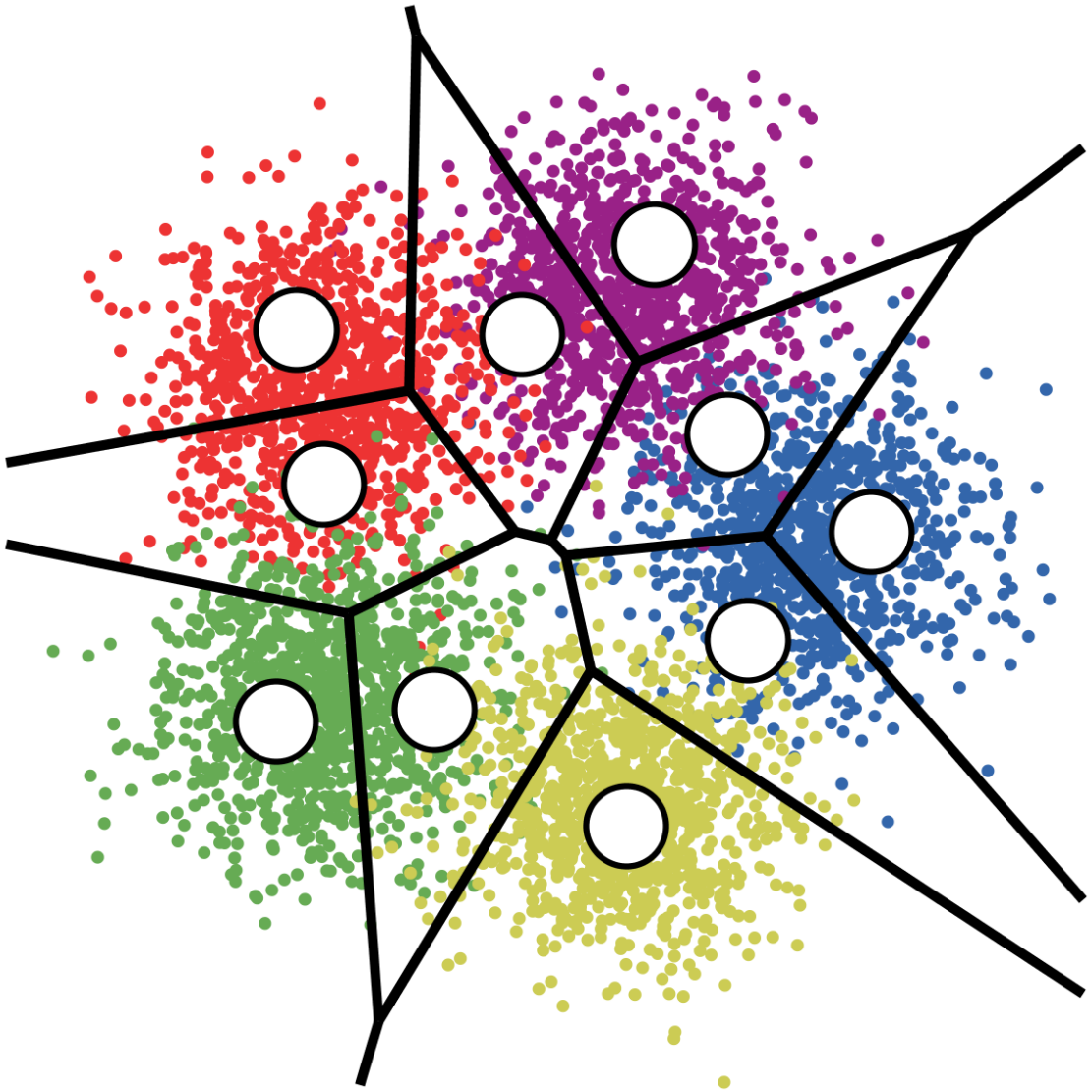


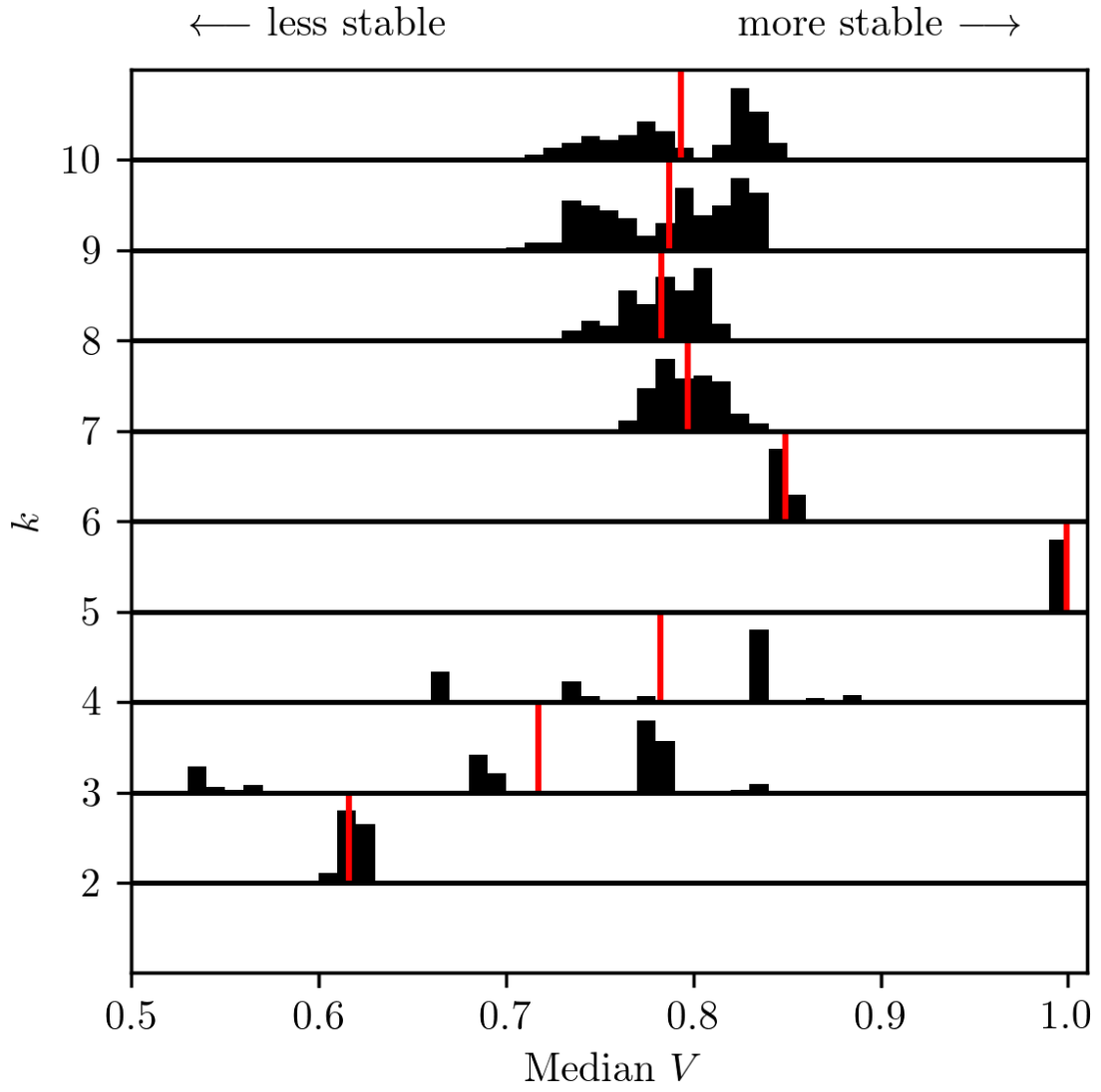
Simulated 2D example: $k = 9$



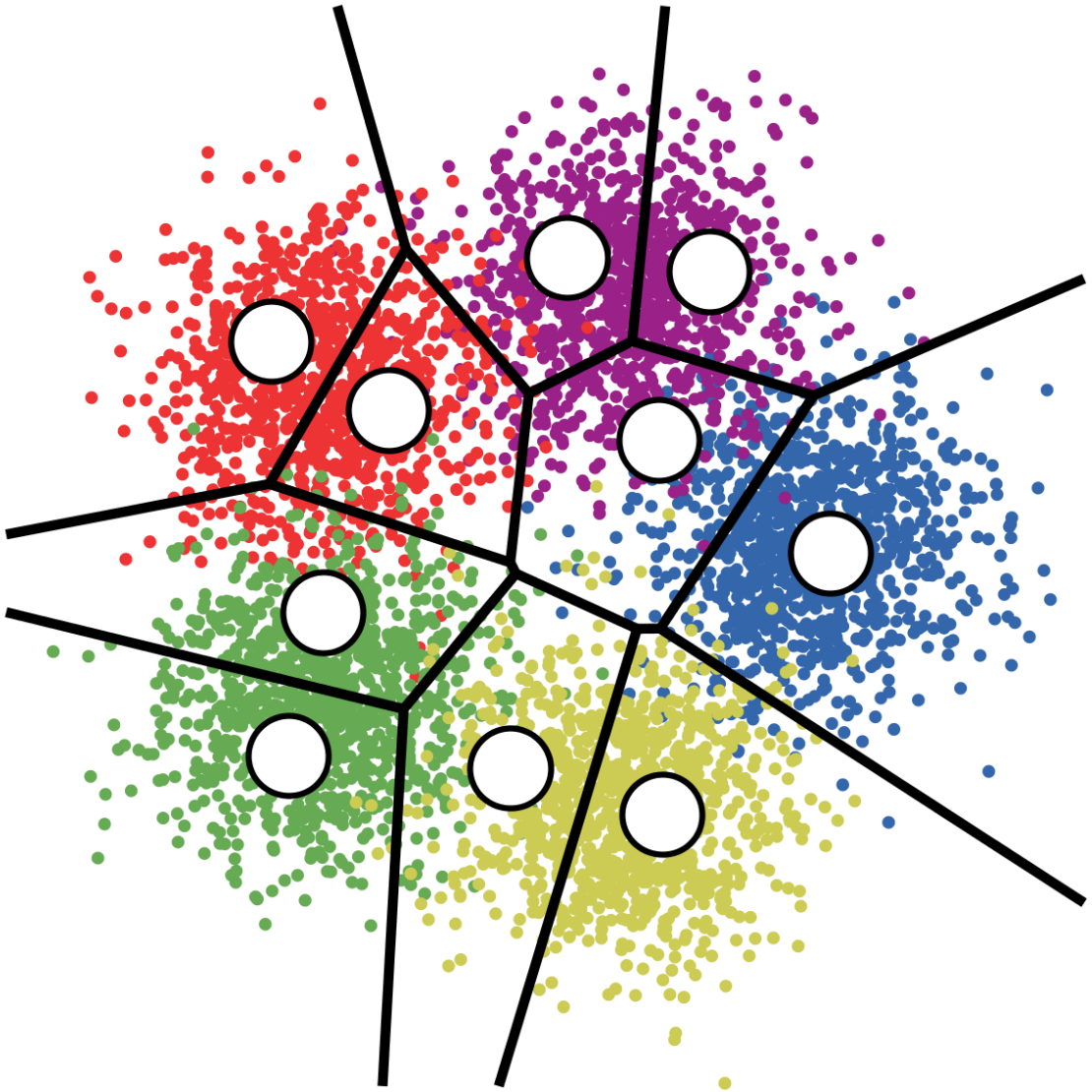


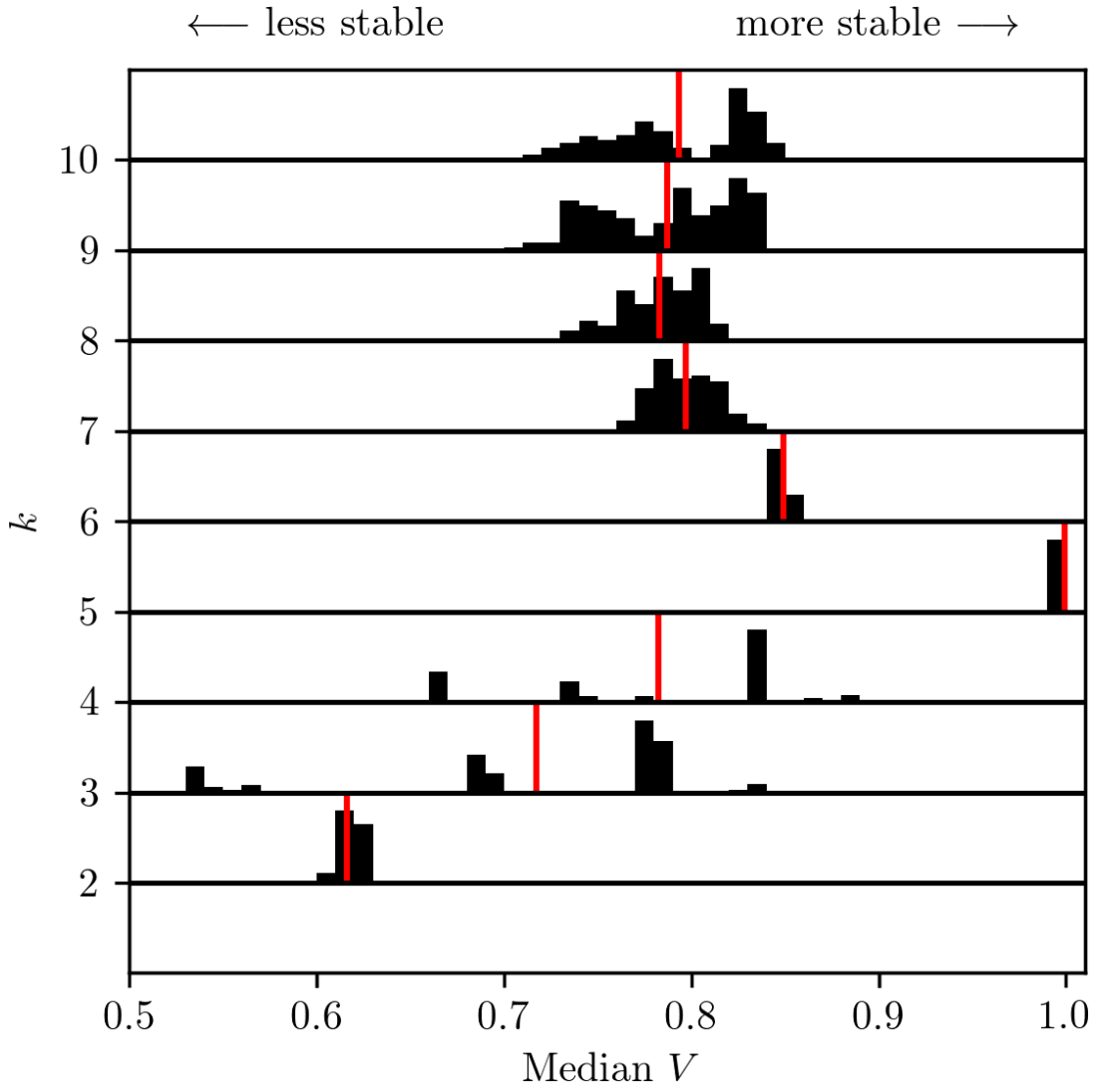
Simulated 2D example: $k = 10$



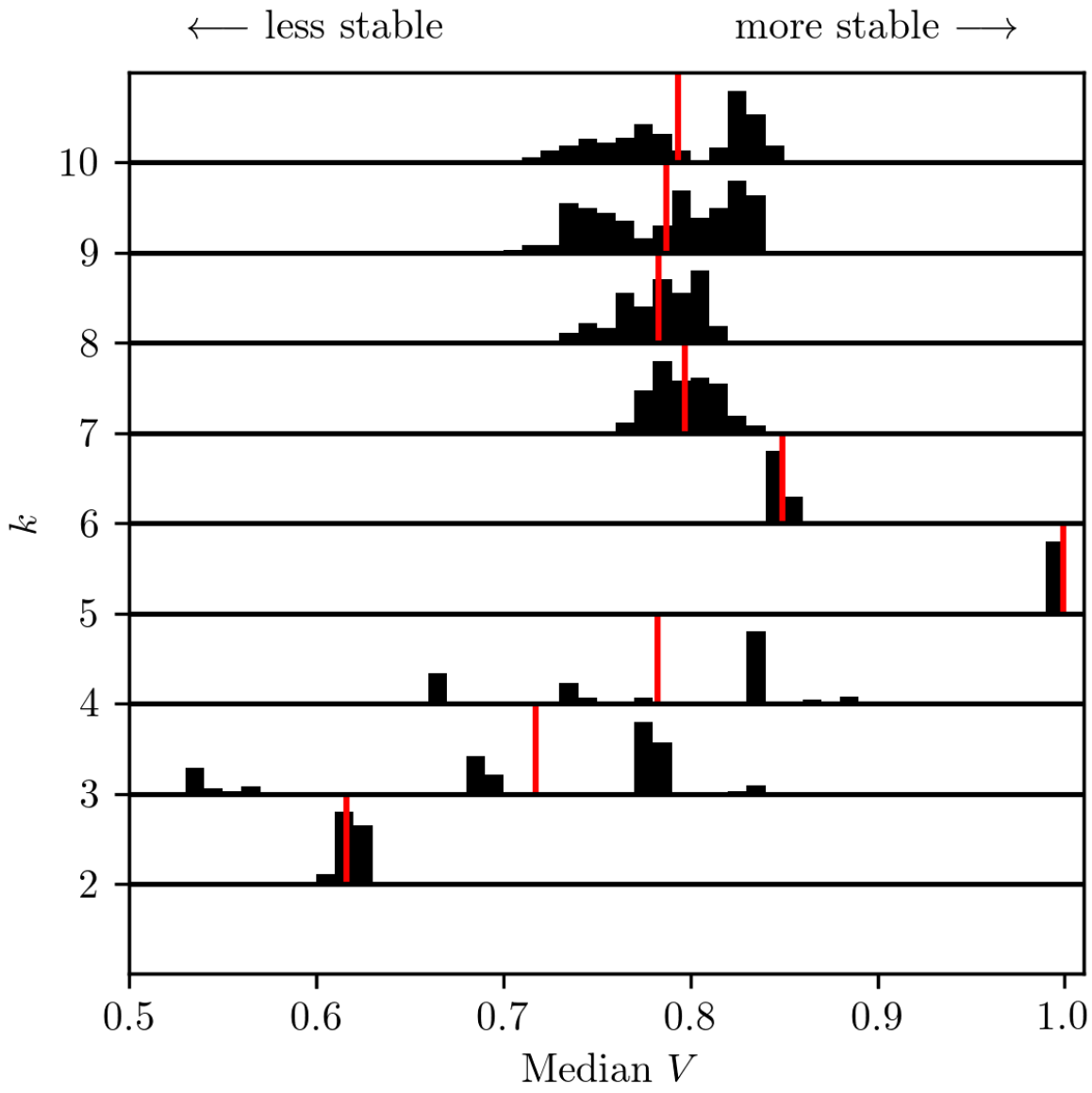


Simulated 2D example: $k = 10$





Simulated 2D example: stability map



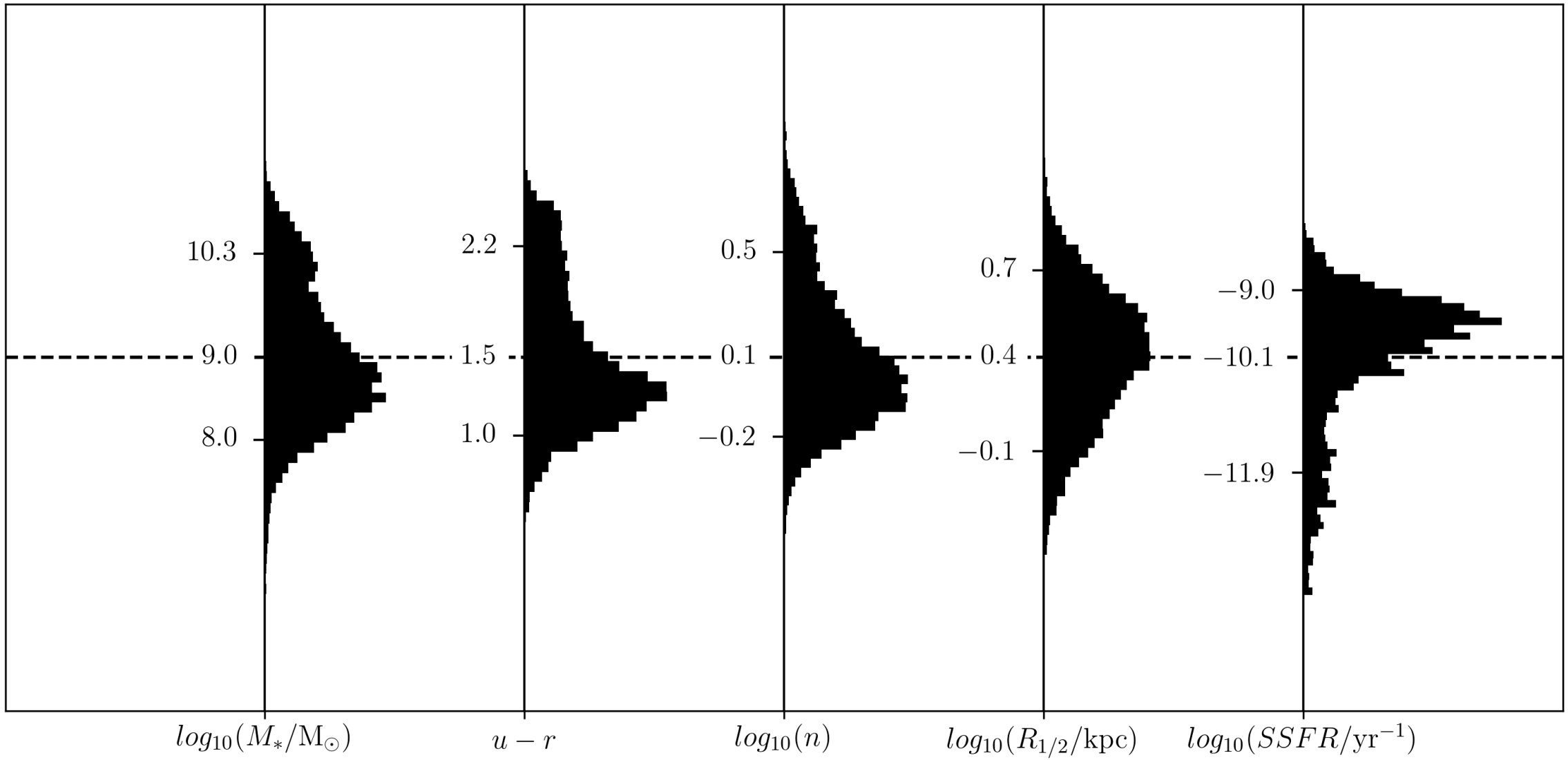
Galaxy sample: basics

- 7338 galaxies from the GAMA survey.
- Redshift- (< 0.06), magnitude- (r_{PETRO} < 19.8) limited sample.
- Features:

• Stellar mass	dex(M⊙)	MagPhysv06
• u-r colour	dex	StellarMassesv20
• Half-light radius	dex(kpc)	SersicCatSDSSv09
• Sersic index	dex(*n*)	SersicCatSDSSv09
• Specific star formation rate	dex(yr^{-1})	MagPhysv06

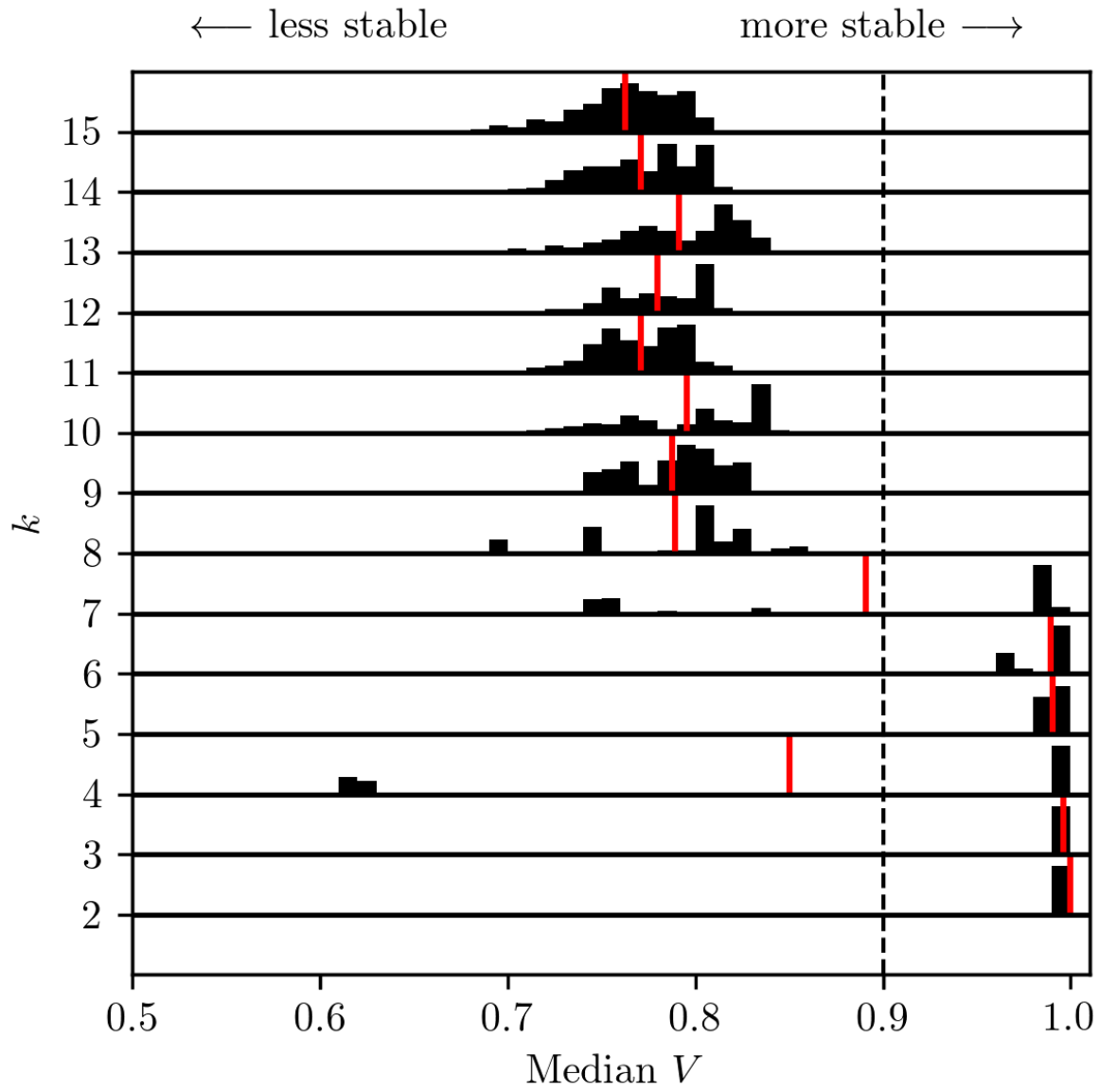
- Data truncated and normalised.

Galaxy sample: features



Seb Turner, LJMU

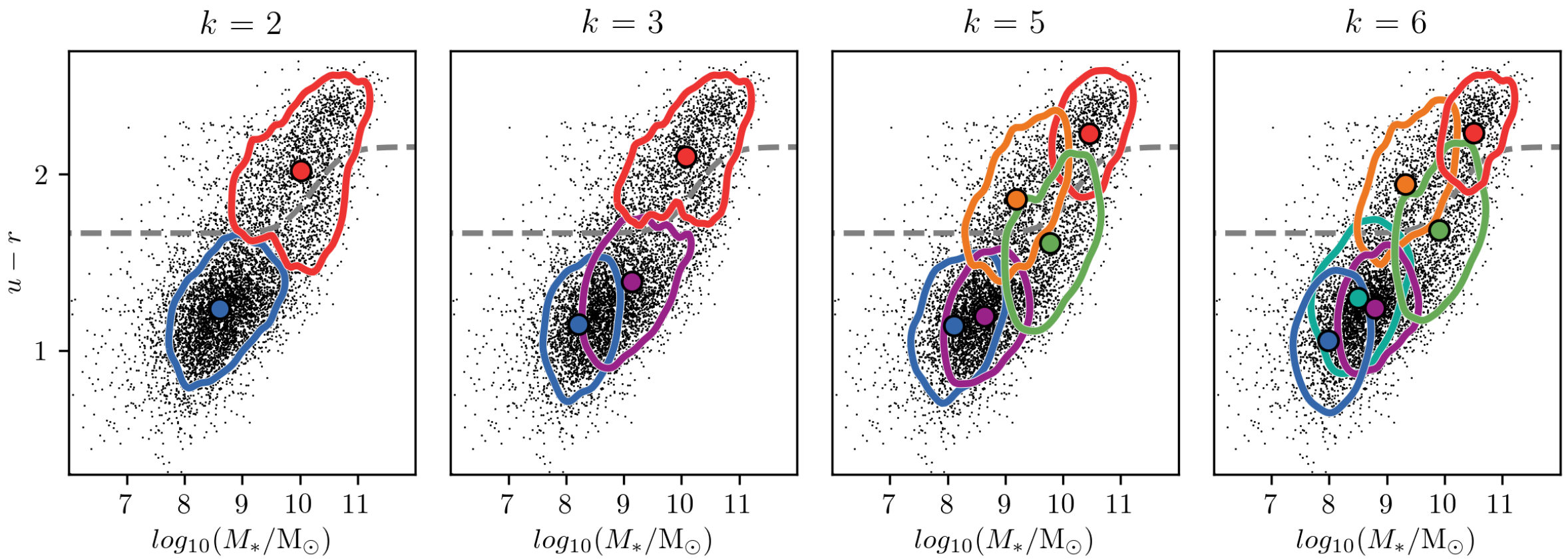
Galaxy clusters: stability map



Turner+ in prep

Seb Turner, LJMU

Galaxy clusters: colour-mass plane



Turner+ in prep

Poster!

k-means clustering in galaxy feature data from the GAMA survey

Sebastian Turner[1], Lee Kelvin[1], Ivan Baldry[1], Paulo Lisboa[2], Steven Longmore[1], Chris Collins[1]
[1]Astrophysics Research Institute, LJMU, 146 Brownlow Hill, Liverpool, L3 5RF, UK
[2]Department of Applied Mathematics, LJMU, Byrom Street, Liverpool, L3 3AF, UK
✉ s.turner1@2012.ljmu.ac.uk @sebturne

Abstract

Using an unsupervised machine learning algorithm, we find that galaxies may be meaningfully divided into five distinct groups. We also explore new perspectives on the established bimodality of galaxies. Our approach will be useful for the analysis of the huge data volumes expected from next generation surveys like Euclid, and for new analysis of existing data sets like Galaxy Zoo.

Introduction

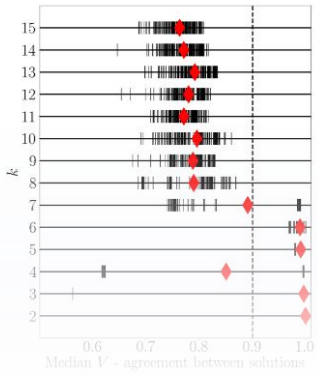
Understanding the diversity of galaxies necessitates a classification scheme that segregates galaxies in a way that reflects their formation and evolution. Galaxies are commonly distinguished as being blue, star-forming, disky, late-type galaxies in low density environments, or red, quiescent, spheroidal, early type galaxies in high density environments. The existence of further, meaningfully distinct subclasses has previously been mostly speculative. We explore this using the *k*-means unsupervised clustering machine learning algorithm.

Data

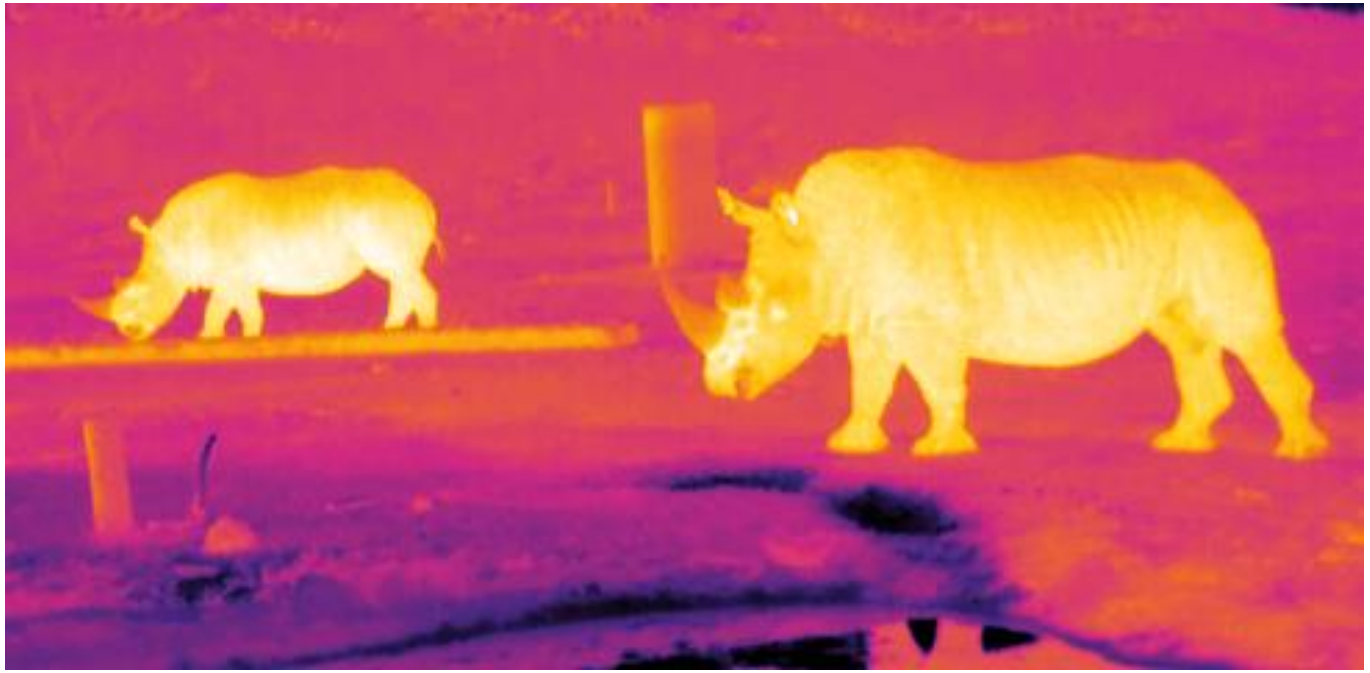
We derive a redshift- ($z < 0.06$) and magnitude- ($r_{PETRO} < 19.8$) limited sample of 7338 morphologically classified galaxies from the GAMA survey. We select a preliminary set of five features to represent our sample: **stellar mass (M_*)**, ***u-r*** colour, Sérsic index (n), half-light radius, and specific star formation rate

k-means

The *k*-means algorithm partitions our sample into k clusters. As a local search heuristic, the outcome of *k*-means is dependent on randomised input initialisations. We sample many varying initialisations to ensure we find globally optimal solutions. Adopting an exploratory approach, we also sample a range of values of k and examine whether clustering at each is stable. We focus below on **stable clustering at k = 2 and k = 5**. Clustering is also stable at k = 3 and k = 6.



Want a job?
Join the astro-ecology group!



https://jobs.ljmu.ac.uk/vacancy/postdoctoral-research-asst-drone-image-analyses-astrophysics-research-inst-345433.html

Summary & Conclusions

- Task of galaxy classification is changing.
- We present approach for identifying stable clustering structure.
- Approach is malleable: any algorithm, any data.
- Our results with galaxies:
  - Up to six clusters, incl. green valley, dwarfs, etc..
  - New perspectives on established bimodality of galaxies and galaxy evolution.
- What's next? New algorithms, new samples, new features.

LIVERPOOL JOHN MOORES UNIVERSITY

s.turner1@2012.ljmu.ac.uk

@sebturne