Clustering with Gap Statistic

by Group01

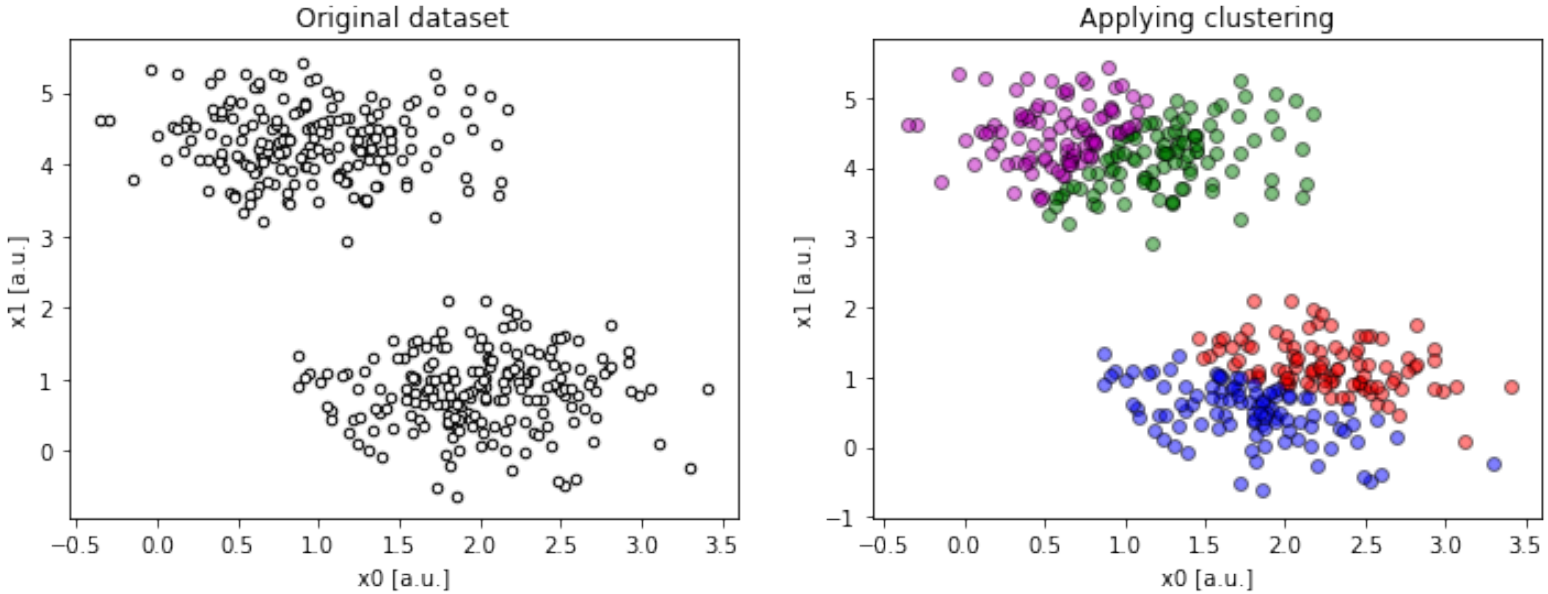
15 May 2020

Content

- What is clustering and cluster analysis?
- Why we want to talk about Gap Statistic?
- Notation
- What is elbow method and how elbow shows the optimal number of clusters?
- The reference distribution
- Gap Statistic step by step
- Some simulations to be clear
- When Gap doesn't work properly? Some alternatives?
- Comments and Conclusions

What is clustering and cluster analysis?

- No rigorous definition
- Subjective
- Scale/Resolution dependent



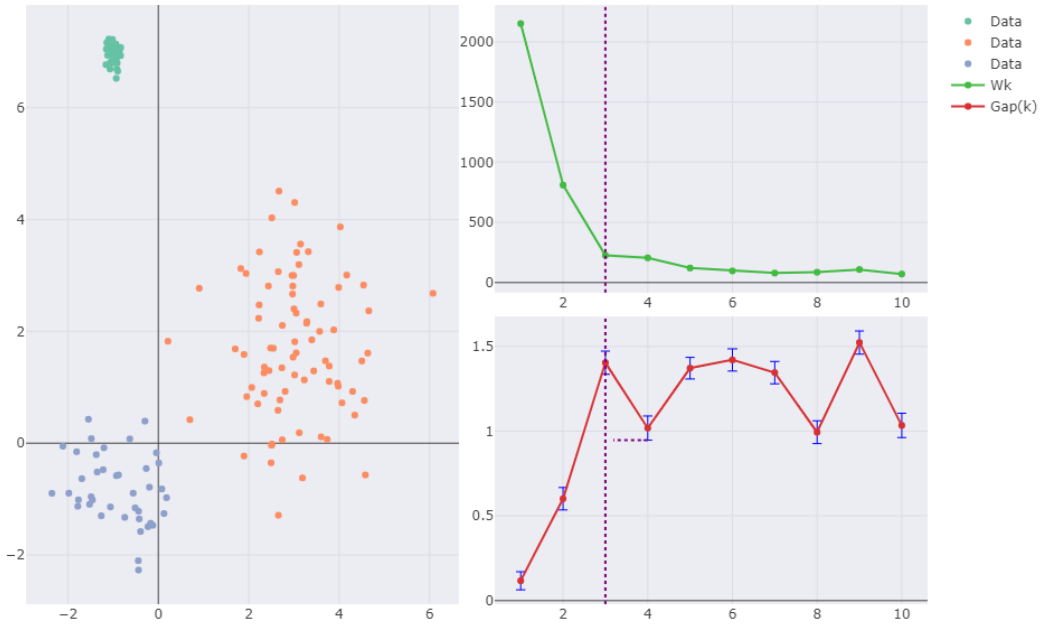
Let's formalize what is clustering

If C(.) is a cluster, partition the observations x_i so that $C(i) = C(j)$ if xi and x_j are 'similar' $C(i) \neq C(j)$ if x_i and x_j are 'dissimilar'

- Being similar is based on a concept of distance that we choose to use!
- But Problem... how many clusters? Input parameters to some clustering algorithms to validate the number of clusters suggested by clustering algorithm.
- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Why we want to talk about Gap Statistic?

- Methods to find the optimal number k of clusters have many problems, we want to find a better procedure.
- It's a relatively easy approach.
- The method is mostly not affected by the data structure.
- It's a global procedure.



Notation

Let's consider our population data $\{x_{ij}\}$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ consists of p features measured on n independent observations.

Then, we can define:

- $d_{ii'}$ as the **distance** between observation i and i'
- Squared Euclidean distance:

$$d_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{1}$$

- we denote clusters as $C_1, C_2, \ldots, C_r$ with C_r denoting the indices of observations in cluster r and $n_r = |C_r|$.

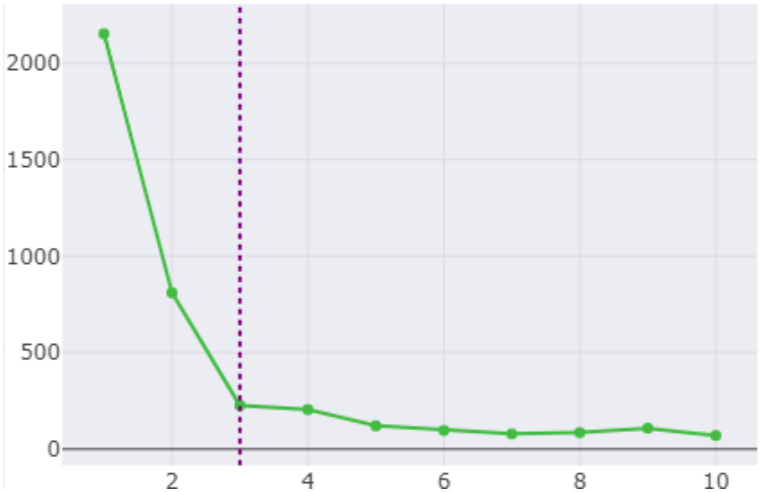
Notation

- Sum of Pairwise Distances for all points in cluster r:

$$D_r = \sum_{i,i' \in C_r} d_{i\,i'} \tag{2}$$

- Pool within cluster sum of squares (around the cluster means)

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \tag{3}$$



What is elbow method and how it shows the optimal number of clusters?

Definition

In clustering analysis, the **elbow method** is an *euristic* used to determine the number of clusters in a dataset

Method: plotting the **explained variation** as a function of the number of clusters k and picking the value where starts the so called *'elbow'*, where diminishing returns are no longer worth the additional cost.

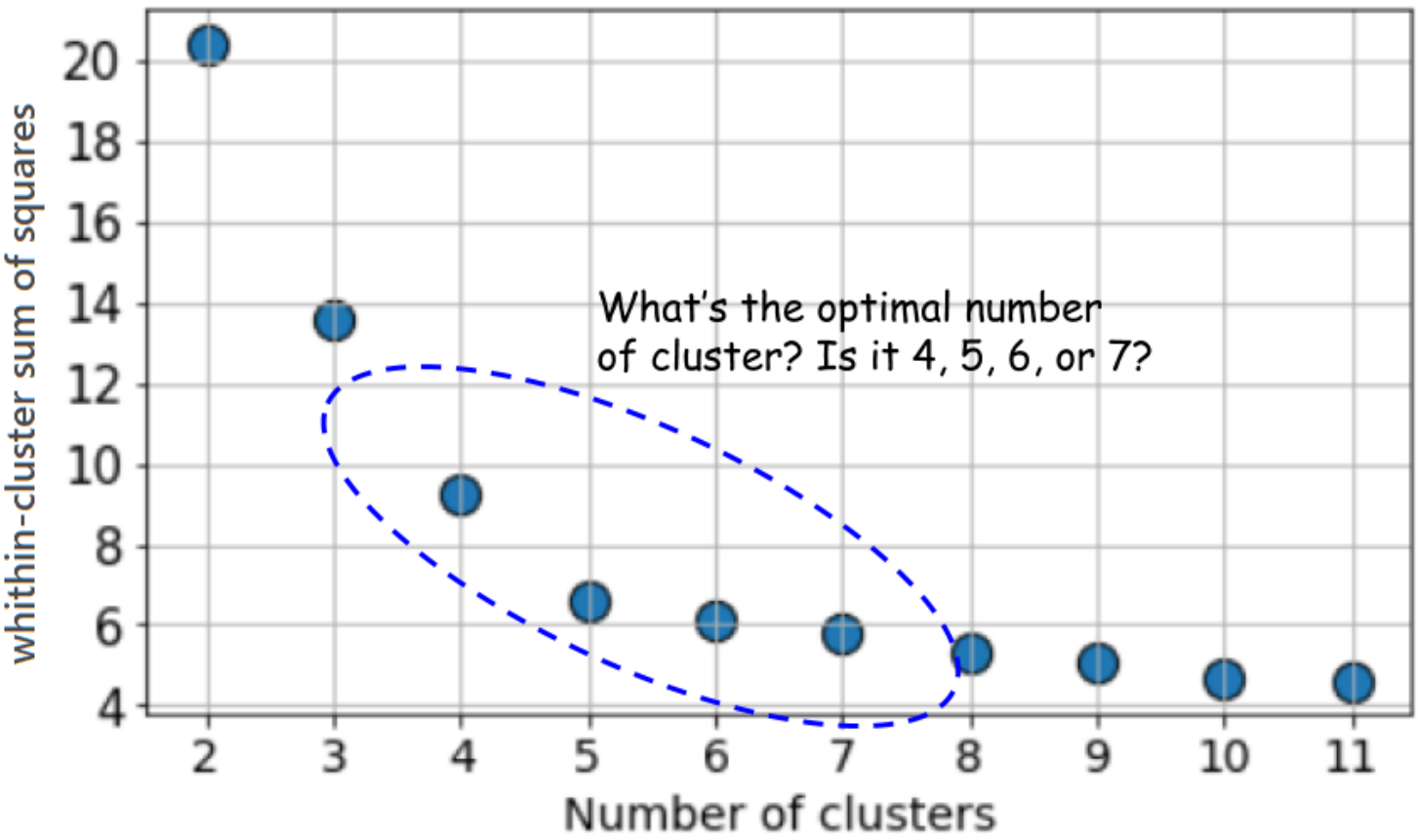
Intuition: increasing the number of clusters will naturally improve the fit, but at some point this is over-fitting, and the elbow reflects this.

- First cluster adds a lot of information
- Information drop, when **number of cluster overcome number of actual groups.**

What is elbow method and how it shows the optimal number of clusters?

But it's not the whole story...

Sometimes It's not easy to find the optimal number of clusters with elbow method



Notation: Formalize the Gap

The **main idea** of our approach is to standardize the $log(W_k)$ graph, comparing it with its **expectation** under an appropriate **null reference distribution** of the data.

Our estimate of the **optimal values** of clusters is the k value that *falls the farthest* below this reference curve.
So we can define the **Gap Function**:

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k) \tag{4}$$

Where:

- E_n^* is the expectation under a sample size of n from the reference distribution.
- Our estimate $\hat{k}$ will be evaluated maximizing the function $Gap_n(k)$.

The reference distribution and log-concave distribution

In our framework, we assume a null model of a single component (1 cluster), and we reject in favour of a k-model component ($k > 1$).

- A single-component is modeled by a **log-concave distribution** (strong unimodality)

$$f(x) = e^{\psi(x)} \tag{5}$$

- ψ is concave
- Counting number of modes in a unimodal distribution doesn't work, it's impossible to set a CI for the number of modes.

Thus we model the components as log-concave densities, instead of unimodal densities. We denote this set as S^p.

K-means Gap Statistic

To see how to find an appropriate reference distribution, let's consider for a moment the population version corresponding to the gap statistic in the case of K-means-clustering:

$$g(k) = \log\left\{\frac{MSE_{X^*}(K)}{MSE_{X^*}(1)}\right\} - \log\left\{\frac{MSE_X(K)}{MSE_X(1)}\right\} \tag{6}$$

where $MSE_X(K) = E(min_{\mu \in A_k}\|X - \mu\|^2)$.

- Note that $g(1) = 0$
- So we are looking for a least favorable single component reference distribution on X^* such that $g(k) \leq 0$ for all $X \in S^p$ and all $k \geq 1$

Univariate and multivariate case in Gap Statistic

The first Theorem shows that in the univariate case, such reference distribution is given by the uniform distribution $U = U[0, 1]$

Theorem 1 (Solution for the 1-D case)

Let $p = 1$, then for all $k \geq 1$

$$inf_{X \in S^p} \left\{ \frac{MSE_X(K)}{MSE_X(1)} \right\} = \frac{MSE_U(K)}{MSE_U(1)} \tag{7}$$

Theorem 2

If $p > 1$ then no distribution $U \in S^p$ can satisfy equation (7) unless its support is degenerate to a subset of line.

Higher dimensional cases

- So, in higher dimensional cases, no log-concave distribution solves (7)

  A possible solution: generate data from the MLE in S^p.

- This estimate, can be shown to exist, and can be computed in 1-dimension...

But we do not know how to compute it in higher dimensions... idea!

Mimic the 1-D case and use a **uniform distribution** $U = U[0, 1]$ as reference in **higher dimensional cases.**

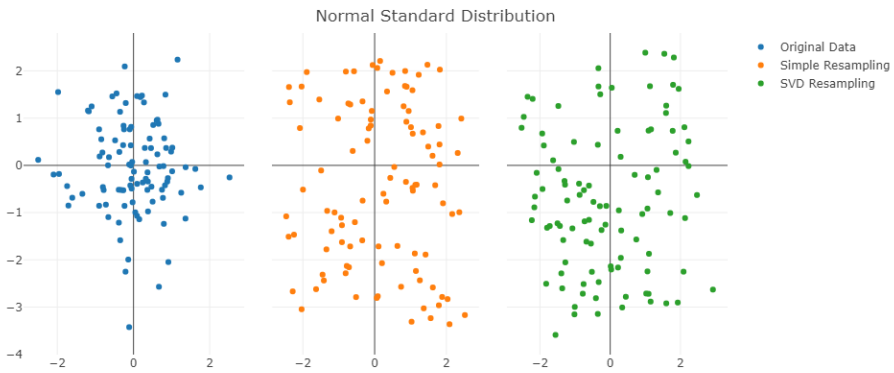
Gap Statistic step by step: Ref generation

- Two option to define a Reference distribution.
  1. **Simple resampling** Generate each reference uniformly over the range of the observed values for that features.

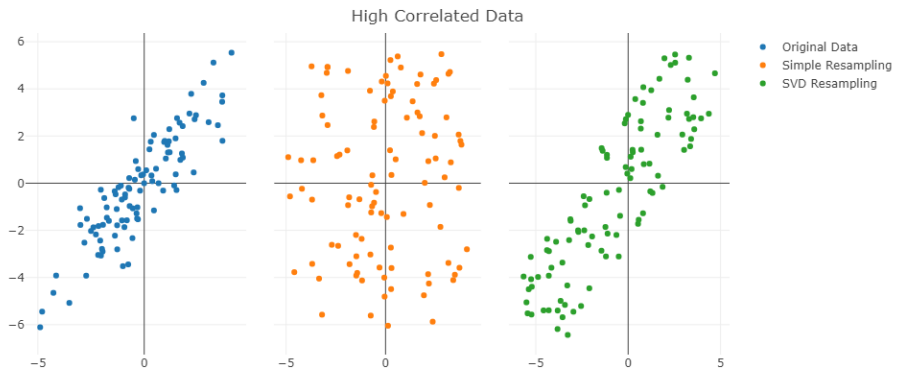
  2. **SVD resampling** Generate the reference feature from a uniform distribution over a box aligned with the principal components of the data. Assuming X is a nxp matrix with mean 0,
     $X = UDV^T \rightarrow X' = XV \rightarrow Z'$
     with uniform features from X',then
     $Z = Z'V^T$ is the final reference.

Let's see the two methods!

These plots show the difference in operating the resampling procedure using the simple method or the SVD method.
As you can see the SVD method retrieves the correlation structure of the data



(a) Normal std distribution



(b) High Correlated Data

Gap Statistic step by step

- In each case we estimate $E_n^* \log(W_k)$ by an average of B copies $\log(W_k)$, each of which is computed by a Monte Carlo Sample $X_1^*, ..., X_n^*$ drawn from our reference distribution
- Let $sd(K)$ denote the standard deviation of the B Monte Carlo replicated, accounting also for the simulation error, we get:

$$s_k = \sqrt{1 - 1/B} sd(k) \tag{8}$$

choosing the k^- as the smallest such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{9}$$

Steps!

- 1) Cluster the observed data, varying the total number of clusters from k = 1, 2,...K, giving within-dispersion measures $W_k, k = 1, 2, ..., K$.
- 2) Generate B reference datasets with method 1 or 2, from the Reference generation slide, and cluster each one giving within dispersion measures $W^*_{kb}, b = 1, 2, ..., B$ and $k = 1, 2, ..., K$. Compute now the estimated Gap statistic:

$$Gap_n(k) = (1/B) \sum_b \log(W^*_{kb}) - \log(W_k) \tag{10}$$

Steps!

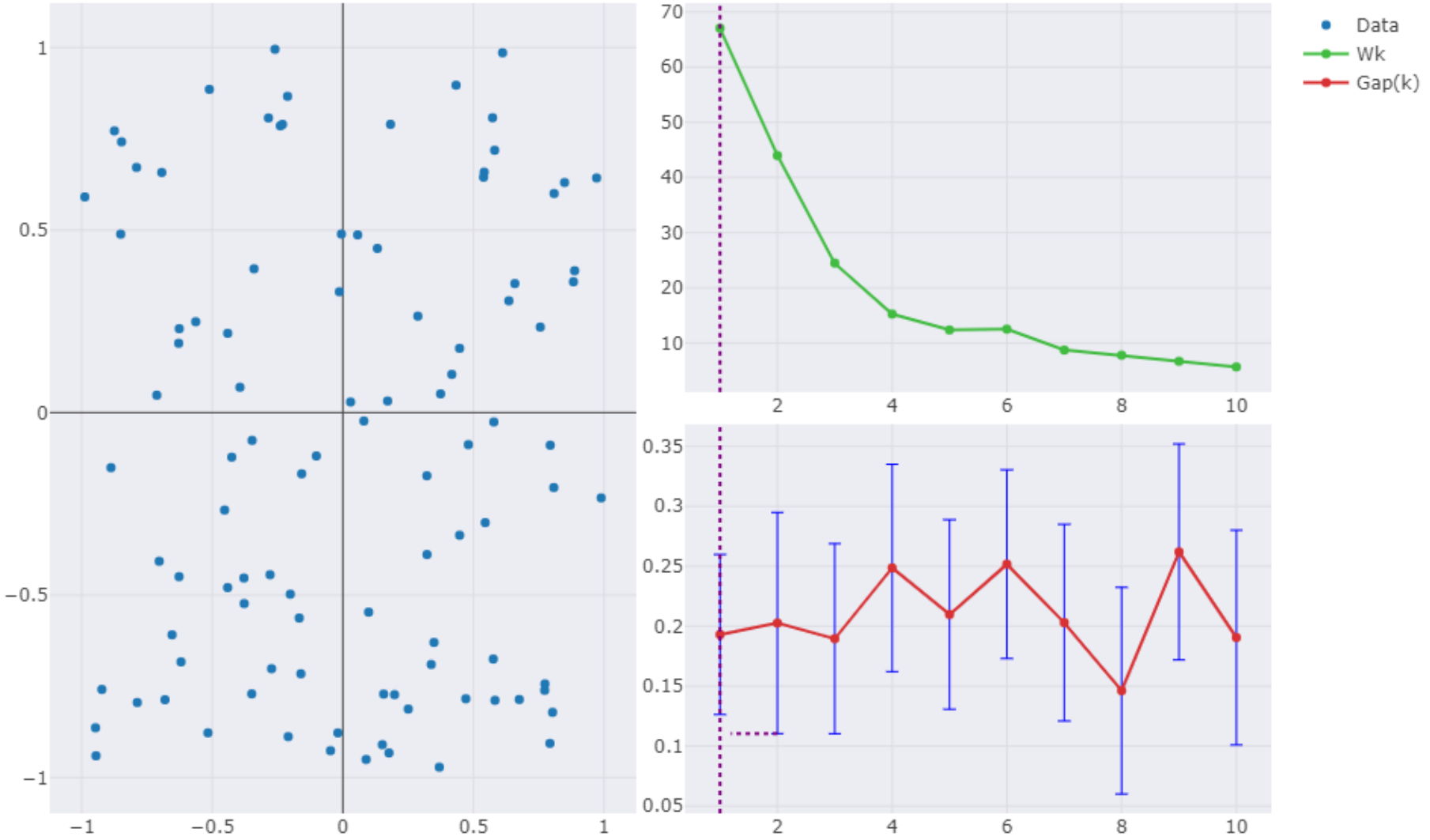
- 3) Let $\bar{l} = (1/B)\sum_b \log(W^*_{kb})$, compute the standard deviation

$$sd_k = [(1/B)\sum_b (\log(W^*_{kb}) - \bar{l})^2]^{1/2} \tag{11}$$

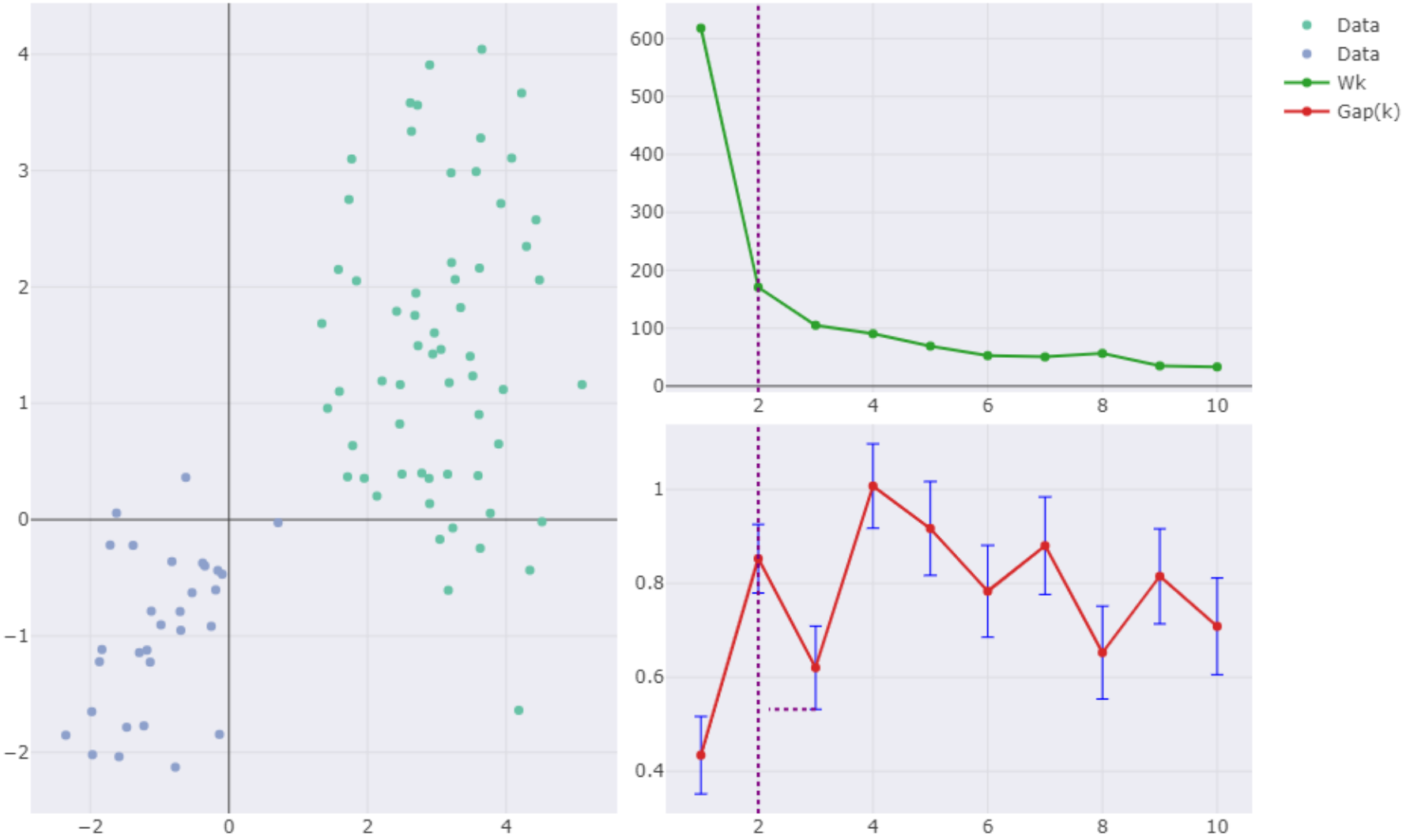
and define $s_k = \sqrt{1 - 1/B}sd(k)$, finally choose that number of clusters via k^- as the smallest such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1} \tag{12}$$

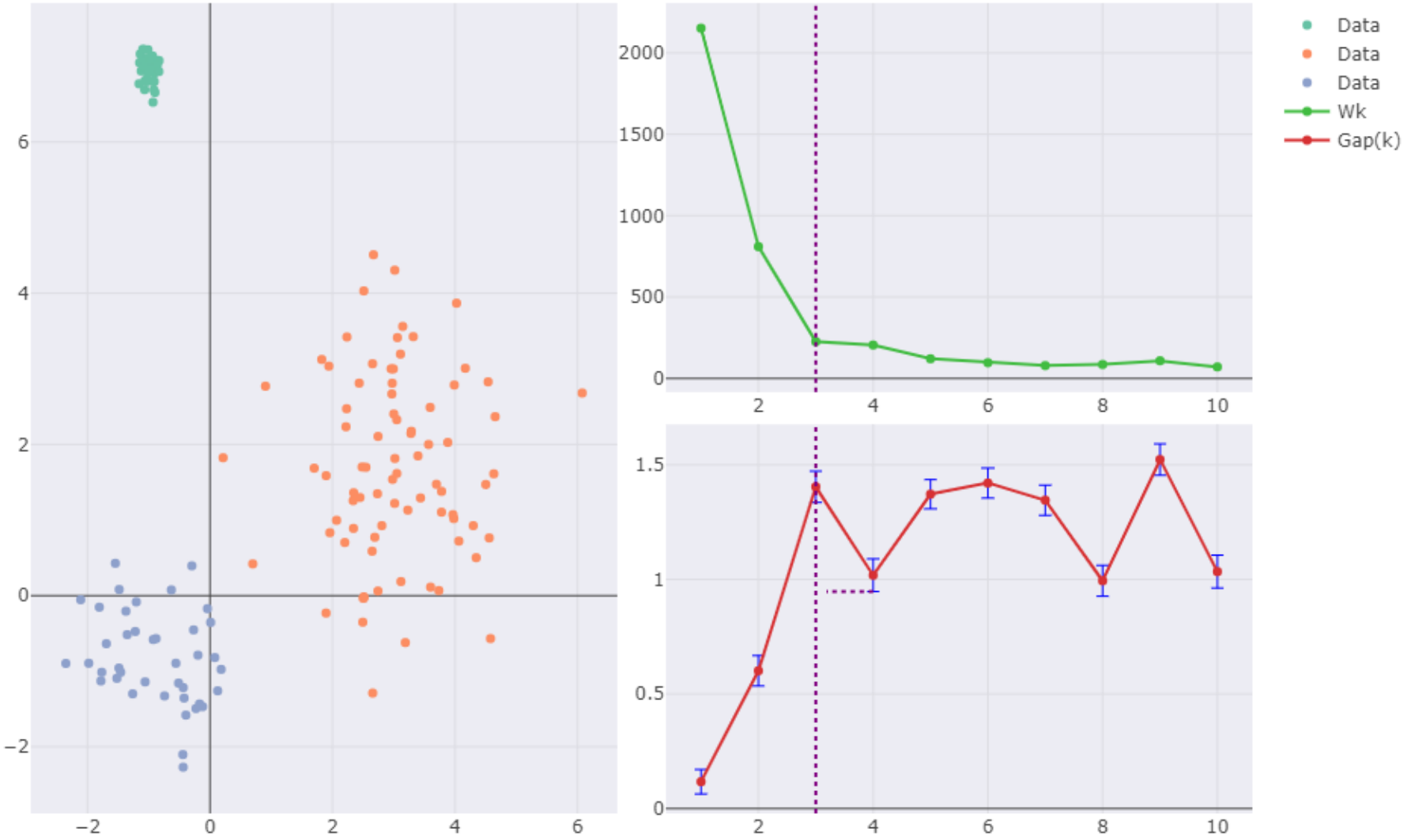
Some simulations: With Normal Data



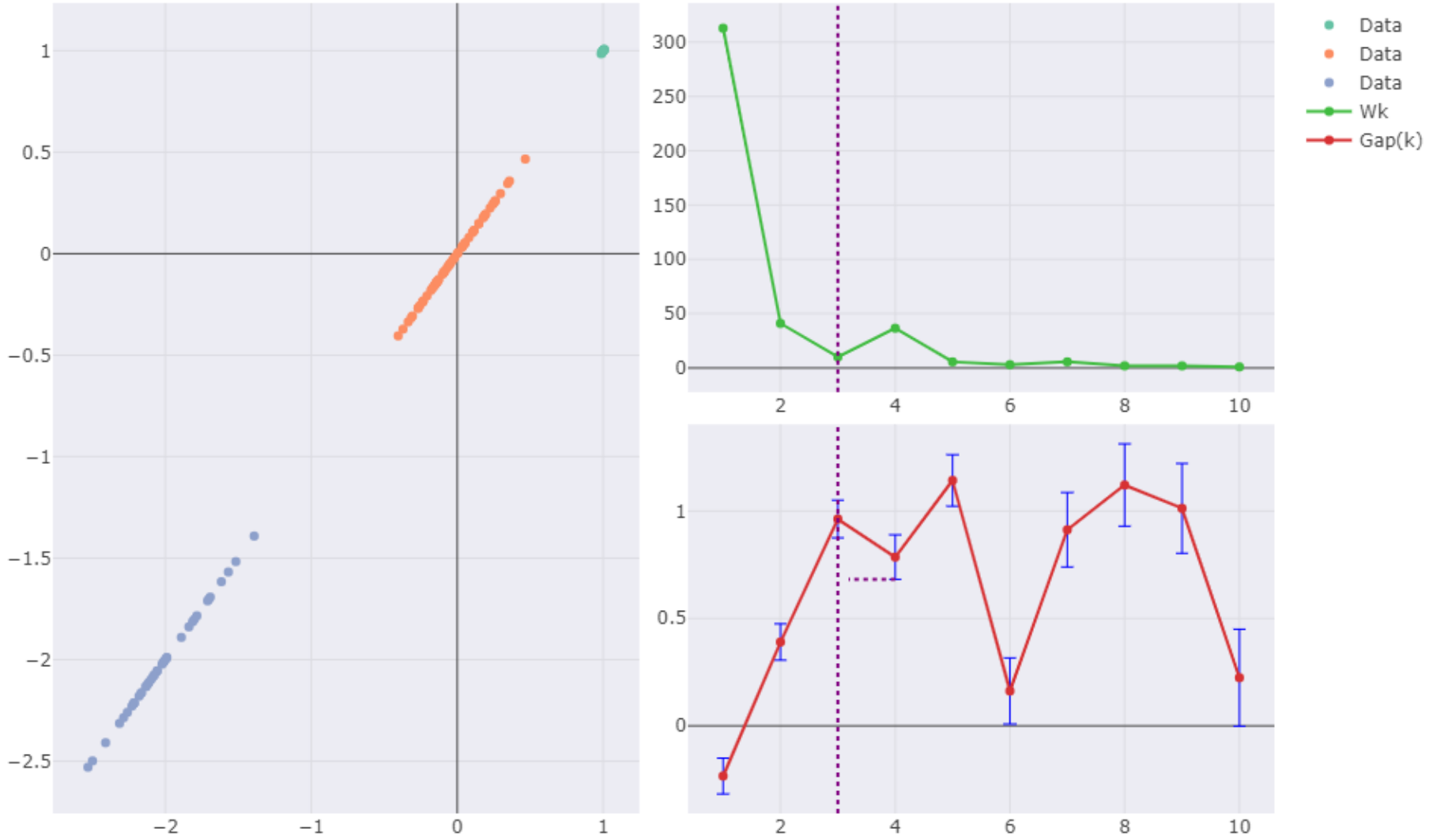
Some simulations: With Normal Data



Some simulations: With Normal Data



Some simulations: With High Correlated Data



When Gap doesn't work properly? Some alternatives?

Downsides of Gap statistic

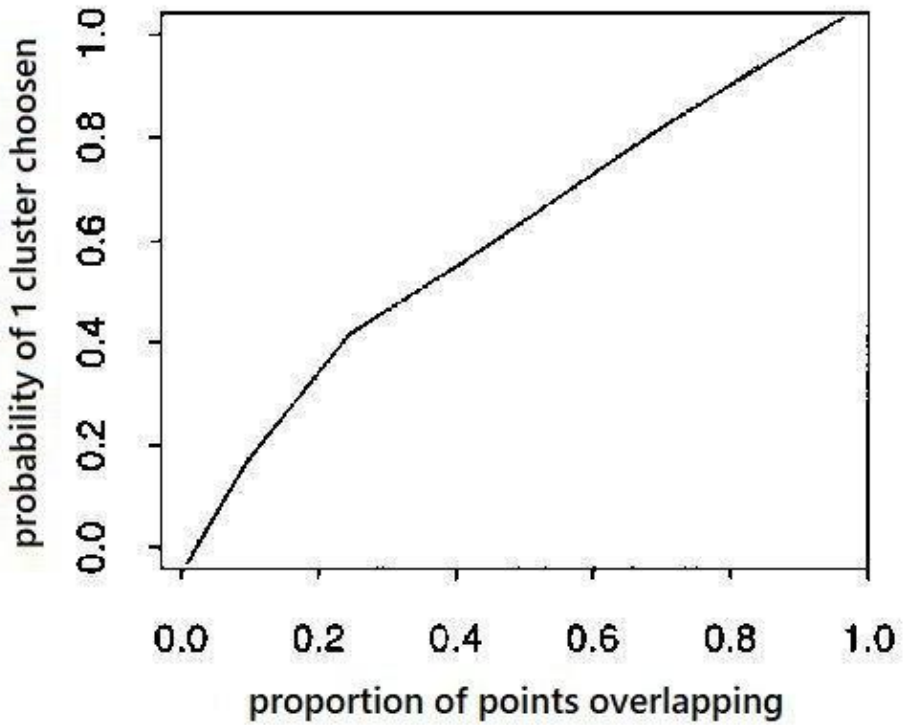
- Sometimes overestimates the number of clusters.
- May not work properly for data driven from exponential distributions $\rightarrow$ use weighted Gap statistic.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r(1-n_r)} D_r.$$

- May fail when a dataset contains clusters of different densities $\rightarrow$ sample reference dataset from Normal distribution instead of uniform.

How does Gap method respond when data is not well-separated?

- Simulating from 2 bivariate normal distribution with means (0,0) and (0, Δ) and identity covariance
- choosing 10 Δ values in [0,5]
- Simulating 10 observations for each Δ value



An alternative approach

A variant of the Gap Statistic does not include a logarithm in the definition of Gap_n :

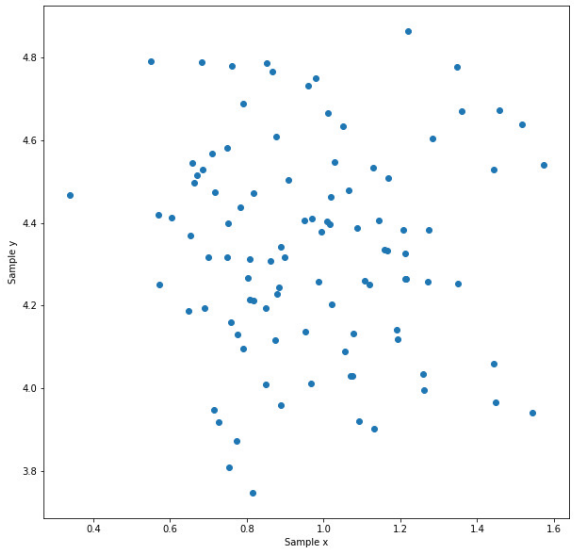
$$Gap_k^* = E_n^* - W_k \tag{13}$$

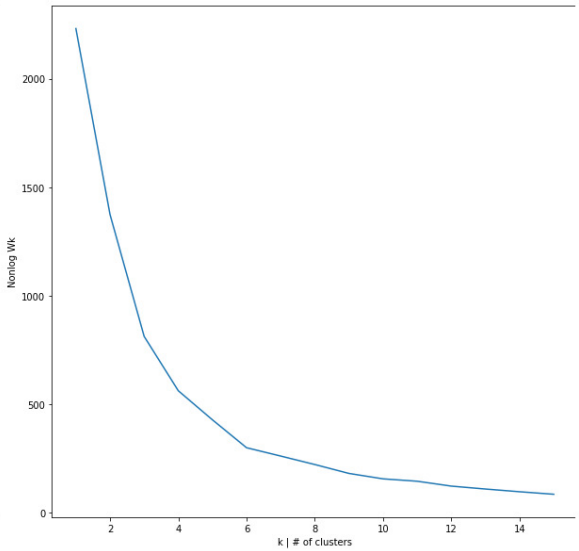
having

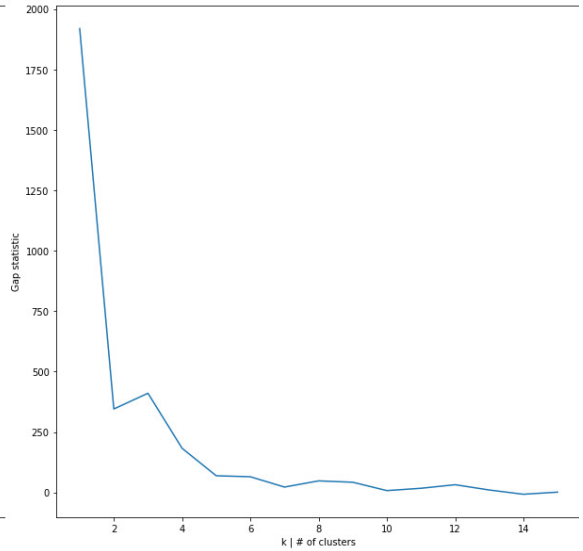
$$E_n^* = \frac{1}{B} \sum_b W_{kb}^* \tag{14}$$

There is no computational advantage when using either of them.
Let us now try to see if there is an advantage via comparison of the two definitions:

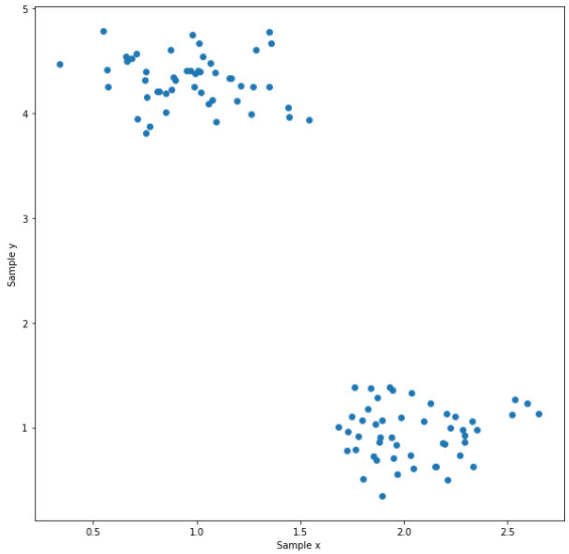
Images: 1 Cluster Case

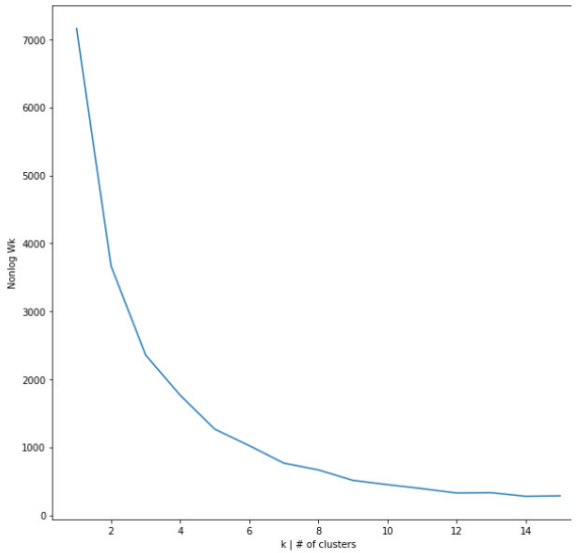


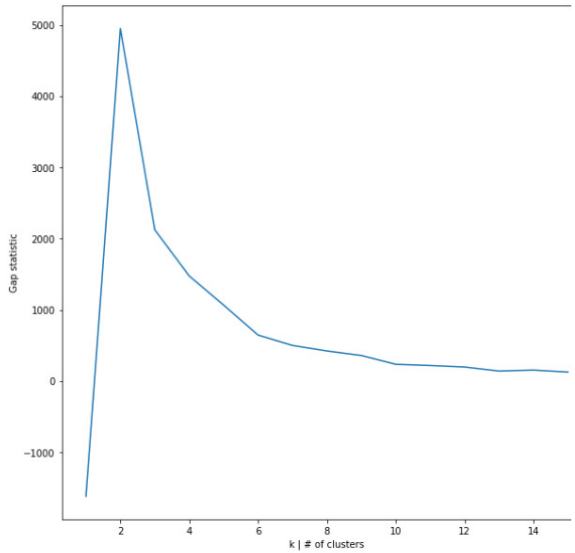




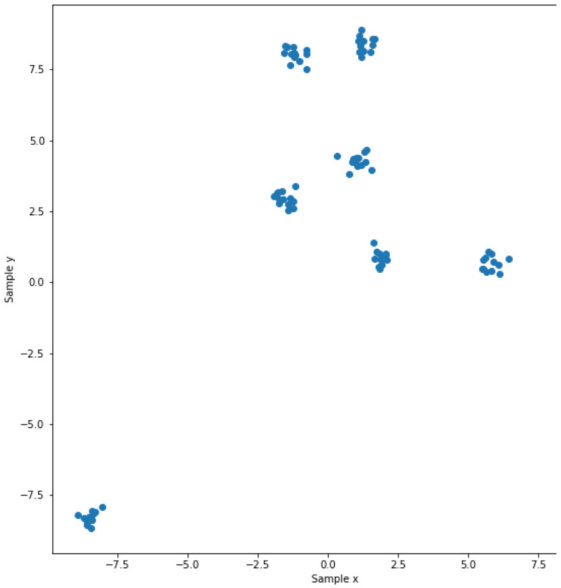
Images: 2 Cluster Case

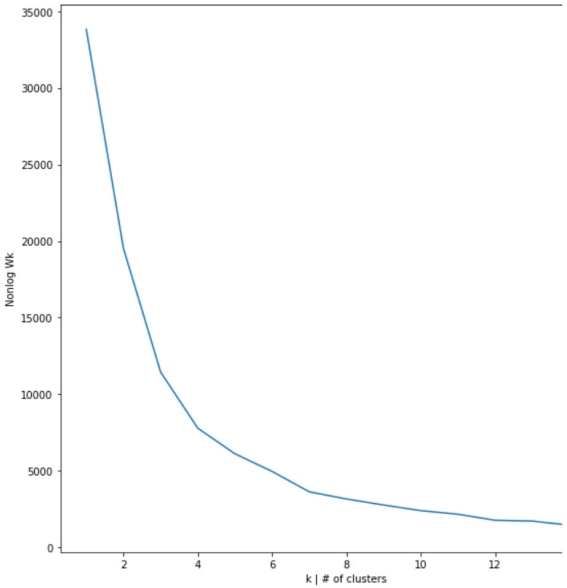


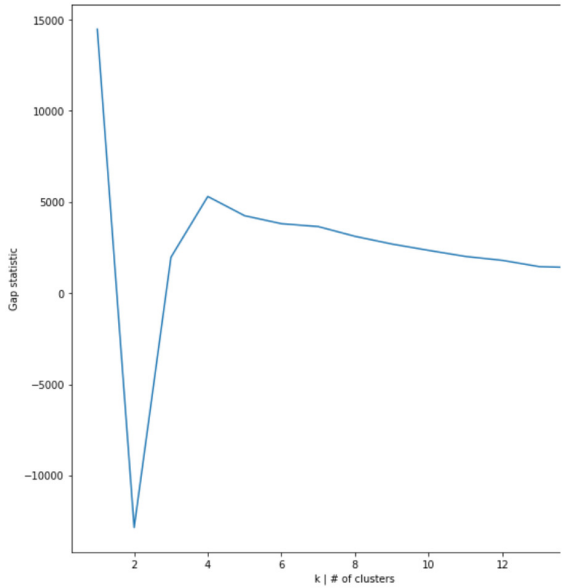




Images: 8 Cluster Case







An alternative: why?

As proved in the paper by Mohajer et al., the non-log definition might be useful when the "classical" definition of the Gap statistic does not allow to define a canditate for k. The Gap_n^* will allow for a k.

- Moreover, getting a k from Gap_n implies that a suitable k will be found applying the Gap_n^*. Vice versa *might* not be true.

More variants: Weighted Gap statistic

This time a new definition comes from W_k. Usually:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{15}$$

but now we can define it as:

$$W'_k = \sum_{r=1}^{k} \frac{2}{n_r(n_r - 1)} D_r \tag{16}$$

This is done to reduce the effect on the metric from within cluster points that result to be further from the cluster center.
It basically allows the sum of squared distances to be averaged out.

Bibliography


'Estimating the number of clusters in a data set via the gap statistic' by Robert Tibshirani, Guenther Walther, Trevor Hastie.


'A comparison of Gap statistic definitions with and without logarithm function' by Mojgan Mohajer, Karl-Hans Englmeier, Volker J. Schmid.

Presentation by

- Negin Amininodoushan
- Leonardo Placidi
- Davide Zingaro
- Stefano Rando
- Marco Muscas

The End