Feature assessment and selection using sparse clustering

Avgoustinos Vouros[1]

[1]PhD student,
Department of Computer Science,
University of Sheffield

Supervised by Prof Eleni Vasilaki

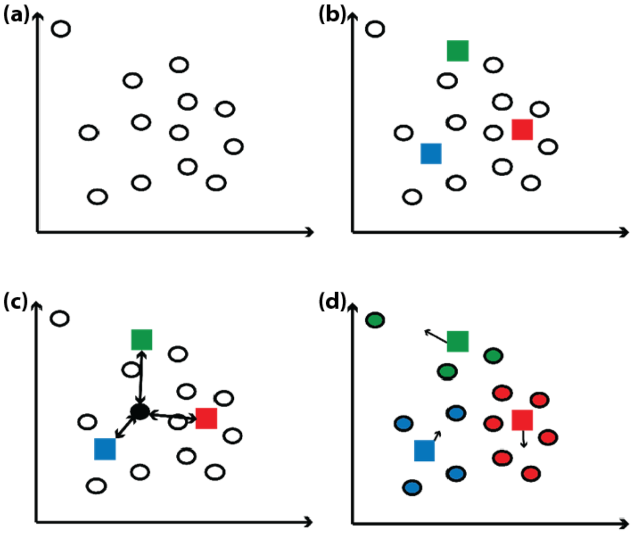
Contents

- The K-Means Algorithm (Lloyd's)
- Sparse K-Means
  - Theory
    - ⋆ Regression (quick)
  - Algorithm
  - Tuning
    - ⋆ Gap Statistic
- Ongoing Research

The University Of Sheffield.

The K-Means Algorithm (Lloyd's)

The K-Means Algorithm (Lloyd's)



[1] Lloyd, Stuart. "Least squares quantization in PCM." IEEE transactions on information theory 28.2 (1982): 129-137.

The K-Means Algorithm (Lloyd's)

Minimize:

$$WCSS = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ (x_{i:} \in c_k)}}^{n} \sum_{j=1}^{p} (x_{ij} - m_{kj})^2$$

Maximize:

$$BCSS = \sum_{j=1}^{p} \left(\sum_{i=1}^{n} (x_{ij} - M_{1j})^2 - \sum_{k=1}^{K} \sum_{\substack{i=1 \\ (x_{i:} \in c_k)}}^{n} (x_{ij} - m_{kj})^2 \right)$$

The K-Means Algorithm (Lloyd's)

Advantages:

- Simple and easy to implement.
- Versatile.
- Guaranteed to converge.
- Invariant to data ordering.

Disadvantages:

- Detects only spherical and well-separated clusters.
- Sensitive to noise and outliers (Euclidean).
- Converges to a local minimum.

[1] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications 40.1 (2013): 200-210.

The K-Means Algorithm (Lloyd's)

Advantages:

- Simple and easy to implement.
- Versatile.
- Guaranteed to converge.
- Invariant to data ordering.

Disadvantages:

- Detects only spherical and well-separated clusters.
- Sensitive to noise and outliers (Euclidean).
- Converges to a local minimum.

In general:

- Sensitive to initial centroids location.

[1] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications 40.1 (2013): 200-210.

The K-Means Algorithm (Lloyd's)

Advantages:

- Simple and easy to implement.
- Versatile.
- Guaranteed to converge.
- Invariant to data ordering.

Disadvantages:

- Detects only spherical and well-separated clusters.
- Sensitive to noise and outliers (Euclidean).
- Converges to a local minimum.

In general:

- Sensitive to initial centroids location.
- Sensitive to features (variables/attributes).

[1] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications 40.1 (2013): 200-210.

Sparse K-Means
- Theory -

Sparse K-Means Theory

$$\underset{c_1,\ldots,c_k,w}{maximize}\left\{\sum_{j=1}^{p} w_{jj}\left(\sum_{i=1}^{n}(x_{ij}-M_{1j})^2-\sum_{k=1}^{K}\sum_{\binom{i=1}{x_{i:}\in c_k}}^{n}(x_{ij}-m_{kj})^2\right)\right\}$$

$$\text{subject to} \qquad \sum_{j=1}^{p} w_{jj}^2 \leq 1, \;\; \sum_{j=1}^{p}|w_{jj}| \leq s, \;\; w_{jj} \geq 0 \;\; \forall j$$

[1] Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726.

Sparse K-Means Theory

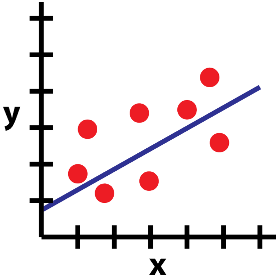
$$\underset{c_1,\ldots,c_k,w}{maximize}\left\{\sum_{j=1}^{p} w_{jj}\left(\sum_{i=1}^{n}(x_{ij}-M_{1j})^2-\sum_{k=1}^{K}\sum_{\binom{i=1}{x_{i:}\in c_k}}^{n}(x_{ij}-m_{kj})^2\right)\right\}$$

$$\text{subject to} \qquad \sum_{j=1}^{p} w_{jj}^2 \leq 1,\ \ \sum_{j=1}^{p}|w_{jj}| \leq s,\ \ w_{jj} \geq 0\ \ \forall j$$

- w is a diagonal square p-by-p matrix.
- $\sum_{j=1}^{p} w_{jj}^2 \leq 1$ is the L_2 penalty or ridge regression ($\|w\|^2 \leq 1$) [2].
- $\sum_{j=1}^{p}|w_{jj}| \leq s$ is the L_1 penalty or lasso regression ($|w| \leq s$) [3].

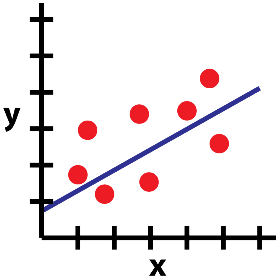
[1] Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726.
[2] Hoerl, Arthur E., and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12.1 (1970): 55-67.
[3] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

Regression



[1] StatQuest with Josh Starmer. "Regularization Part 1: Ridge Regression." YouTube. 2018. Online: https://www.youtube.com/watch?v=Q81RR3yKn30.
[2] StatQuest with Josh Starmer. "Regularization Part 2: Lasso Regression." YouTube. 2018. Online: https://www.youtube.com/watch?v=NGf0voTMlcs.

Regression

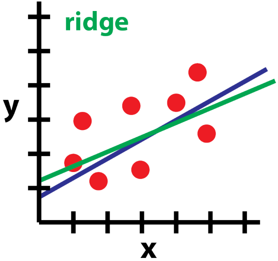


Line fitting using least squares:

$$y = y_{inter} + slope * x$$

Least squares minimizes the sum of the squared residuals.

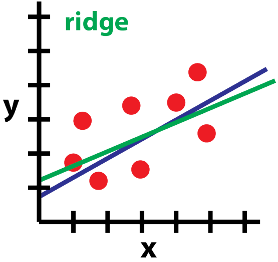
Regression



Line fitting using ridge regression:

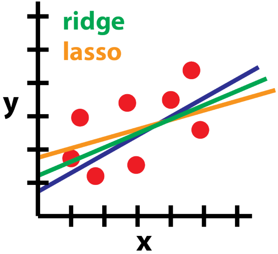
$$y = y_{inter} + slope * x$$
$$+\lambda_2 * slope^2$$

Regression



Line fitting using ridge regression:

$$y = y_{inter} + slope * x$$
$$+\lambda_2 * slope^2$$



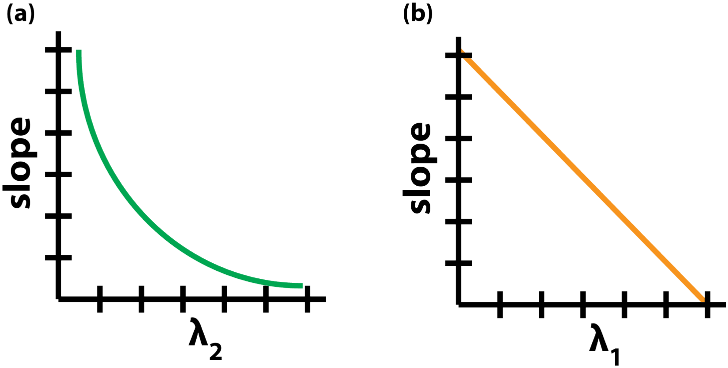
Line fitting using lasso regression:

$$y = y_{inter} + slope * x$$
$$+\lambda_1 * |slope|$$

Regression

$$y = y_{inter} + slope * x + \lambda_2 * slope^2 \quad (a)$$

$$y = y_{inter} + slope * x + \lambda_1 * |slope| \quad (b)$$



Sparse K-Means Theory

$$\underset{c_1,\ldots,c_k,w}{\mathit{maximize}}\left\{\sum_{j=1}^{p} w_{jj}\left(\sum_{i=1}^{n}(x_{ij}-M_{1j})^2-\sum_{k=1}^{K}\sum_{\substack{i=1\\ (x_{i:}\in c_k)}}^{n}(x_{ij}-m_{kj})^2\right)\right\}$$

$$\text{subject to} \qquad \sum_{j=1}^{p} w_{jj}^2 \leq 1, \;\; \sum_{j=1}^{p}|w_{jj}| \leq s, \;\; w_{jj} \geq 0 \;\; \forall j$$

Sparse K-Means Theory

$$\underset{c_1,\ldots,c_k,w}{maximize}\left\{\sum_{j=1}^{p} w_{jj}\left(\sum_{i=1}^{n}(x_{ij}-M_{1j})^2-\sum_{k=1}^{K}\sum_{\substack{i=1\\ \left(x_{i:}\in c_k\right)}}^{n}(x_{ij}-m_{kj})^2\right)\right\}$$

$$\text{subject to} \qquad \sum_{j=1}^{p} w_{jj}^2 \leq 1, \;\; \sum_{j=1}^{p}|w_{jj}| \leq s, \;\; w_{jj} \geq 0 \;\; \forall j$$

$$\underset{w}{maximize}\left\{\sum_{j=1}^{p} w_{jj} a_j + \lambda \sum_{j=1}^{p} w_{jj}^2 + \delta \sum_{j=1}^{p}|w_{jj}|\right\}$$

Sparse K-Means Theory

$$\underset{w}{maximize}\left\{\sum_{j=1}^{p} w_{jj}a_j + \lambda\sum_{j=1}^{p} w_{jj}^2 + \delta\sum_{j=1}^{p}|w_{jj}|\right\}$$

- $\lambda\sum_{j=1}^{p} w_{jj}^2$ and $\delta\sum_{j=1}^{p}|w_{jj}|$ are called Lagrange multipliers.
- When the constraints are having both equalities and inequalities we extend to Karush-Kuhn-Tucker (KKT) conditions.

Sparse K-Means Theory

If $w \neq 0$

$$\frac{\partial}{\partial w}|w| = \frac{\partial}{\partial w}\sqrt{w^2} = \frac{\partial}{\partial w}(w^2)^{\frac{1}{2}} = \frac{1}{2}(w^2)^{\frac{1}{2}} \cdot 2w = \frac{w}{\sqrt{w^2}} = \frac{w}{|w|}$$

else if $w = 0$

$$\frac{\partial}{\partial w}|w| = \lim_{w \to 0} \frac{|w| - 0}{w - 0} = \left\{ \begin{matrix} \lim_{w \to 0^+} \frac{w}{w} & , w > 0 \\ \lim_{w \to 0^-} \frac{-w}{w} & , w < 0 \end{matrix} \right\} = \left\{ \begin{matrix} 1 & , w > 0 \\ -1 & , w < 0 \end{matrix} \right\}$$

[1] proofwiki.org. "Derivative of Absolute Value Function."
wiki. 2018. Online:
https://proofwiki.org/wiki/Derivative_of_Absolute_Value_Function.

Sparse K-Means Theory

If $w \neq 0$

$$\frac{\partial}{\partial w}|w| = \frac{\partial}{\partial w}\sqrt{w^2} = \frac{\partial}{\partial w}(w^2)^{\frac{1}{2}} = \frac{1}{2}(w^2)^{\frac{1}{2}} \cdot 2w = \frac{w}{\sqrt{w^2}} = \frac{w}{|w|}$$

else if $w = 0$

$$\frac{\partial}{\partial w}|w| = \lim_{w\to 0}\frac{|w|-0}{w-0} = \begin{Bmatrix} \lim_{w\to 0^+} \frac{w}{w} & , w > 0 \\ \lim_{w\to 0^-} \frac{-w}{w} & , w < 0 \end{Bmatrix} = \begin{Bmatrix} 1 & , w > 0 \\ -1 & , w < 0 \end{Bmatrix}$$

[1] proofwiki.org. "Derivative of Absolute Value Function." wiki. 2018. Online: https://proofwiki.org/wiki/Derivative_of_Absolute_Value_Function.

Sparse K-Means Theory

Proposition: The solution to this convex problem is,

$$w_{jj} = \frac{(sign(a_j)(|a_j| - \delta))_+}{(sign(a_j)(|a_j| - \delta))_+^2}$$

where the $+$ subscript indicates the positive part of the function, $\delta = 0$ if that results in $\sum_{j=1}^{p} |w_{jj}| < s$ or $\delta > 0$ is chosen so that $\sum_{j=1}^{p} |w_{jj}| = s$ and it is assumed that $1 \leq s \leq \sqrt{p}$.

[1] Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." Biostatistics 10.3 (2009): 515-534.

Sparse K-Means Theory

Proposition: The solution to this convex problem is,

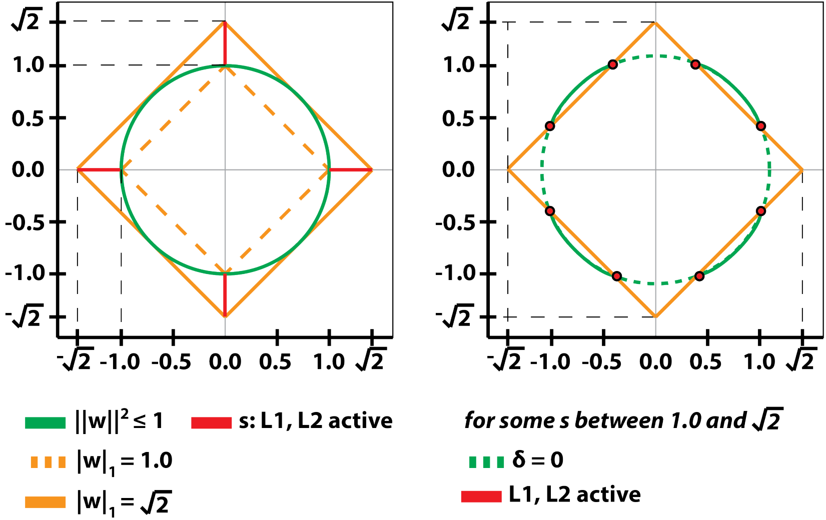
$$w_{jj} = \frac{(sign(a_j)(|a_j| - \delta))_+}{(sign(a_j)(|a_j| - \delta))_+^2}$$

where the $+$ subscript indicates the positive part of the function, $\delta = 0$ if that results in $\sum_{j=1}^{p} |w_{jj}| < s$ or $\delta > 0$ is chosen so that $\sum_{j=1}^{p} |w_{jj}| = s$ and it is assumed that $1 \leq s \leq \sqrt{p}$.

How do we find δ?

[1] Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." Biostatistics 10.3 (2009): 515-534.

Sparse K-Means Theory



[1] Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis." Biostatistics 10.3 (2009): 515-534.

Sparse K-Means Theory

Using the Binary Search algorithm:

- Let δ be between lim_1 and lim_2
- $lim_1 = 0$, $lim_2 = max(a)$
- Iterate...

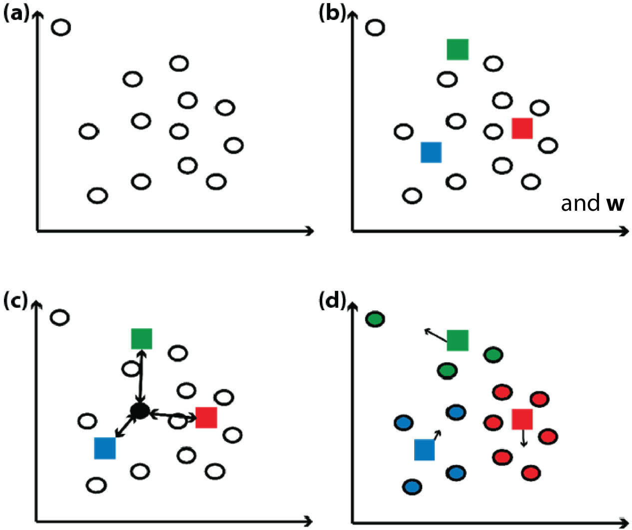
$$u = \frac{\sum_{j=1}^{p}(sign(a_j)(|a_j| - \frac{lim_1+lim_2}{2}))_+}{\sum_{j=1}^{p}(sign(a_j)(|a_j| - \frac{lim_1+lim_2}{2}))^2}$$

$$\begin{Bmatrix} \text{lim}_2 = \frac{lim_1+lim_2}{2} & , u < s \\ \text{lim}_1 = \frac{lim_1+lim_2}{2} & , u \geq s \end{Bmatrix}$$

- $\delta = \frac{lim_1+lim_2}{2}$

Sparse K-Means
- Algorithm -

Phase 1



Sparse K-Means Algorithm

Phase 2

- Execute Binary Search to find δ
  i.e. $\delta = BinarySearch(wBCSS, s)$

- Compute all the new weights using

$$w_{jj} = \frac{(sign(a_j)(|a_j| - \delta))_+}{(sign(a_j)(|a_j| - \delta))_+^2}$$

Phase 3

- Update dataset based on w.

- Repeat from **Phase 1c** until convergence.

Sparse K-Means
- Tuning -

How do we decide k and s?

- Normally for k we use a performance index. But...

[1] Desgraupes, Bernard. "Clustering indices." University of Paris Ouest-Lab Modal'X 1 (2013): 34.

Sparse K-Means Tuning

How do we decide k and s?

- Normally for k we use a performance index. But...

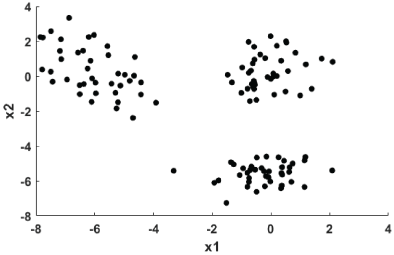
[1] Desgraupes, Bernard. "Clustering indices." University of Paris Ouest-Lab Modal'X 1 (2013): 34.

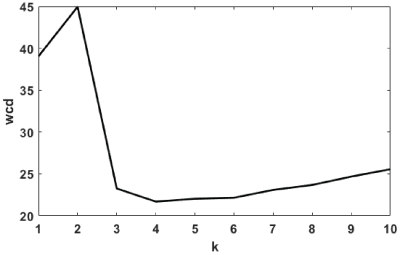
- In the studies of [2] and [3] the gap statistic is proposed. But...

[2] Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726.

[3] Brodinova, Sarka, et al. "Robust and sparse k-means clustering for high-dimensional data." arXiv preprint arXiv:1709.10012 (2017).

The Gap Statistic

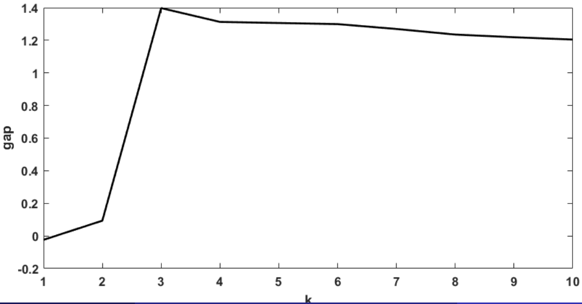




[1]Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.

The Gap Statistic

- Standardize the graph of $\log(WCSS_k)$ by comparing it with its expectation under an appropriate null reference distribution of the dataset.
- The optimal number of clusters is then the k for which $log(WCSS_k)$ falls the farthest below the reference curve.
- $Gap_k = E_b^*\{\log(WCSS_k)\} - \log(WCSS_k)$



The Gap Statistic: Algorithm

- Given dataset D, cluster it with different values for k and keep $\log(J_k)$, where J_k specifies the final value of the objective function of the clustering algorithm for a given k.
- Create B permutations of D and for each repeat the above step, which results to: $\log(J_k^b) = [\log(J_k^1), \ldots, \log(J_k^B)$, for each k.
- Compute the estimated gap statistic,

$$Gap_k = \frac{1}{B} \sum_{b=1}^{B} \log(J_k^b) - \log(J_k)$$

- Given that $E_b^* = \frac{1}{B} \sum_{b=1}^{B} \log(J_k^b)$, compute the simulation error,

$$SE_k = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(\log(J_k^b) - E_b^* \right)^2} + \sqrt{1 + \frac{1}{B}}$$

- Optimal k, $\hat{k}$, is the smallest k for which $Gap_k \geq Gap_{k+1} - SE_{k+1}$

How many permutations, B?

The Gap Statistic: Algorithm

How many permutations, B?

- Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726. **B=25**.

- [3] Brodinova, Sarka, et al. "Robust and sparse k-means clustering for high-dimensional data." arXiv preprint arXiv:1709.10012 (2017). **B=10**.

The Gap Statistic: Algorithm

How many permutations, B?

- Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726. **B=25**.
- [3] Brodinova, Sarka, et al. "Robust and sparse k-means clustering for high-dimensional data." arXiv preprint arXiv:1709.10012 (2017). **B=10**.
- MATLAB, **B=100**.
- R, **B=100**.

The Gap Statistic: Algorithm

How many permutations, B?

- Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726. **B=25**.
- [3] Brodinova, Sarka, et al. "Robust and sparse k-means clustering for high-dimensional data." arXiv preprint arXiv:1709.10012 (2017). **B=10**.
- MATLAB, **B=100**.
- R, **B=100**.
- Let's say $B = 45$. We would like to test 10 different values for k and 5 for s.

The Gap Statistic: Algorithm

How many permutations, B?

- Witten, Daniela M., and Robert Tibshirani. "A framework for feature selection in clustering." Journal of the American Statistical Association 105.490 (2010): 713-726. **B=25**.
- [3] Brodinova, Sarka, et al. "Robust and sparse k-means clustering for high-dimensional data." arXiv preprint arXiv:1709.10012 (2017). **B=10**.
- MATLAB, **B=100**.
- R, **B=100**.
- Let's say $B = 45$. We would like to test 10 different values for k and 5 for s.
- We need to execute our algorithmic framework, $45 + 1(B)\ \ x\ \ 10(k)\ \ x\ \ 5(s) = 2300$ times!

Ongoing Research

Ongoing Research

Ongoing Research

- Find a computationally less expensive criterion than the gap statistic.
- Preliminarily results indicated that:
  - ⋆ Silhouette index for s.
  - ⋆ Correlation for k.