Gaussian Shift (Mean Shift) Clustering and Variance Approximation

Jason Taillon
SUNYIT
31 March 2013

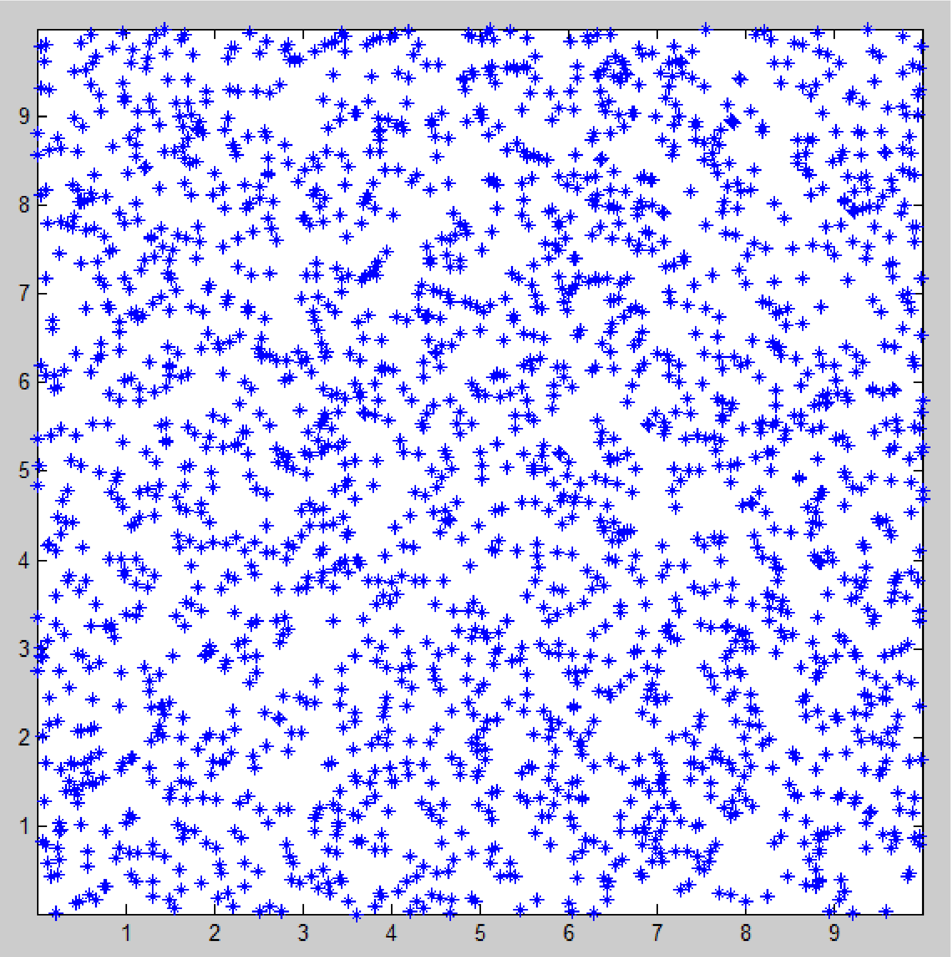
Introduction

- Part 1
  - Clustering
    - Technique & Theory
- Part 2
  - Variance Estimation
    - Sub-grouping
    - Displacement Analysis

Overview & Background

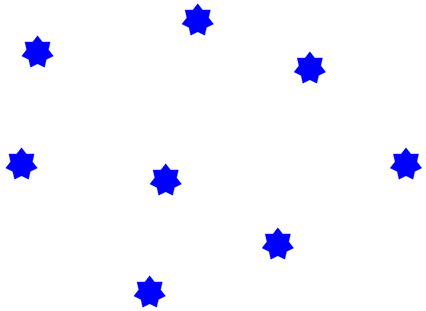
- Mean Shift Clustering is a method of grouping data together, using gradient finding Algorithm
- Applications
  - Genetics
    - Clustering genes with similar expressions or motifs
  - Machine Vision
    - Grouping objects or people together
  - Mathematics
    - Finding groups or underlying distributions in data

Problem?



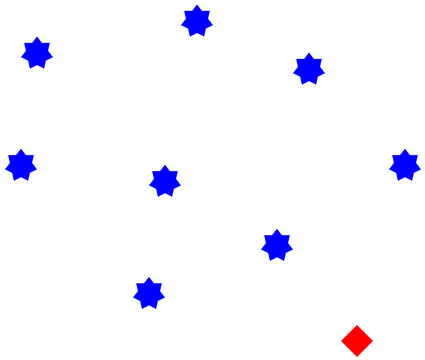
- Given this data:
  - Are there clusters
  - How do we find them

Algorithm 1



Given a set of points in a Cartesian plane

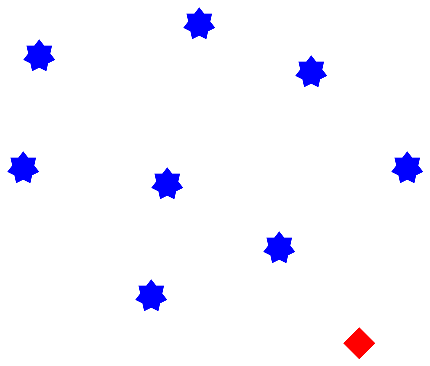
Algorithm 2



Given a set of points in a Cartesian plane

1. Select a Starting Point

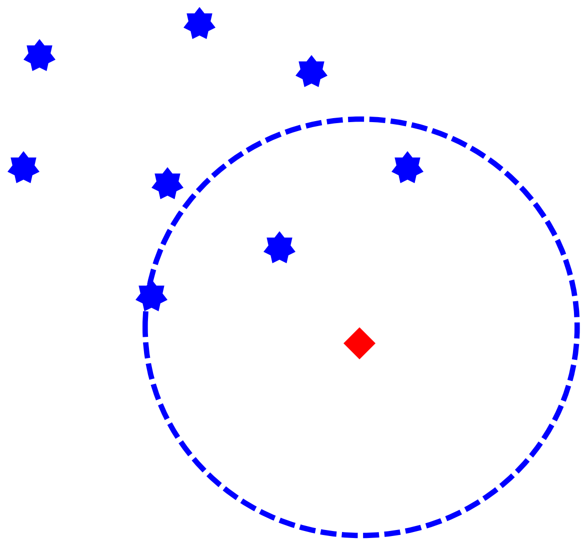
Algorithm 2



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.

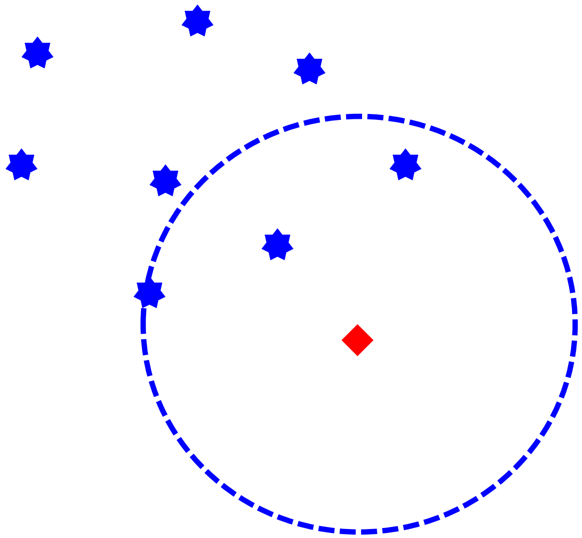
Algorithm 3



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean

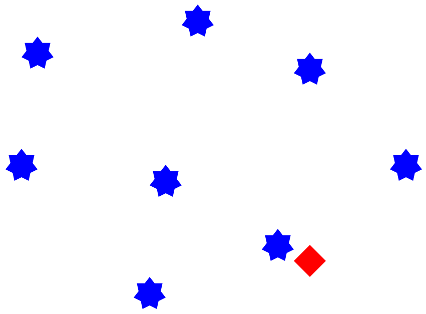
Algorithm 3



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean
4. Take Mean of all points in window, this becomes new local mean

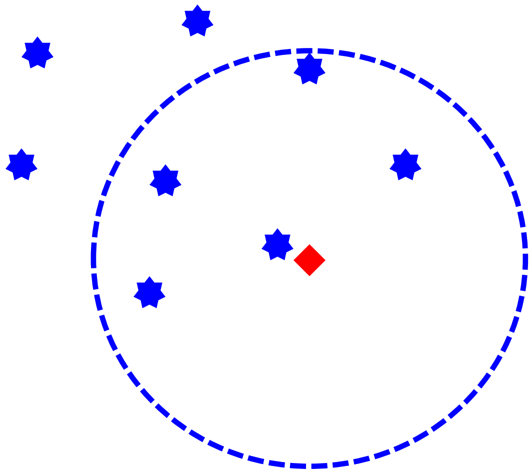
Algorithm 3



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean
4. Take Mean of all points in window, this becomes new local mean
5. Now mean is shifted

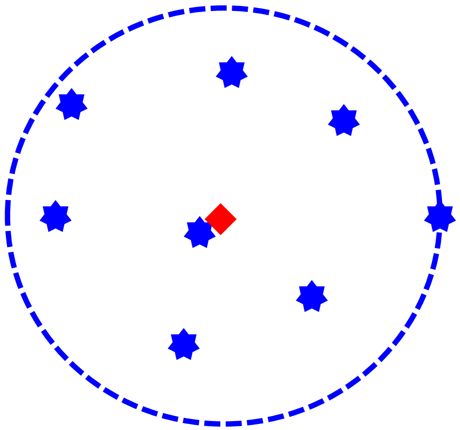
Algorithm 3



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean
4. Take Mean of all points in window, this becomes new local mean
5. Now mean is shifted
6. Recast Window and Repeat Process

Algorithm 3



Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean
4. Take Mean of all points in window, this becomes new local mean
5. Now mean is shifted
6. Recast Window and Repeat Process.
7. Eventually a Convergence is reached and

Formula

- Mean Shift – Is an algorithm for finding the local mode, or modes in a sample population. It is also known as a gradient finding algorithm when used with a Gaussian Kernel

Mean-Shift Formula

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

Kernel: Gaussian

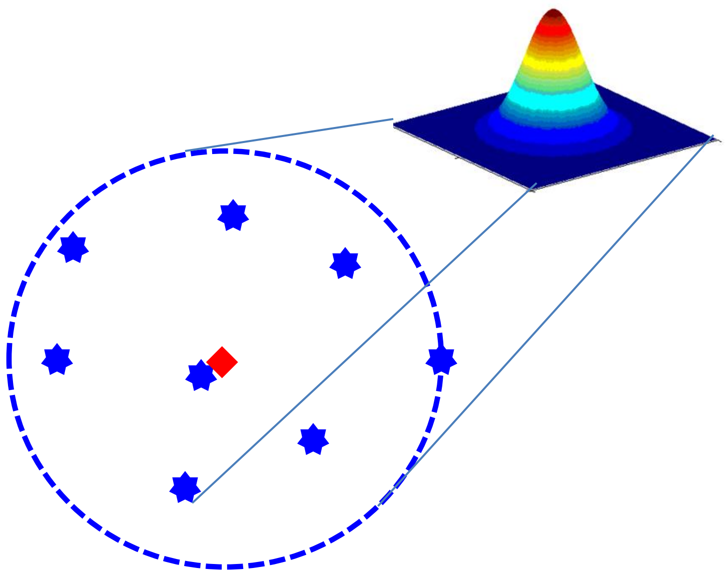
$$K(x_i - x) = e^{-c||x_i - x||^2}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Weighted Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Algorithm



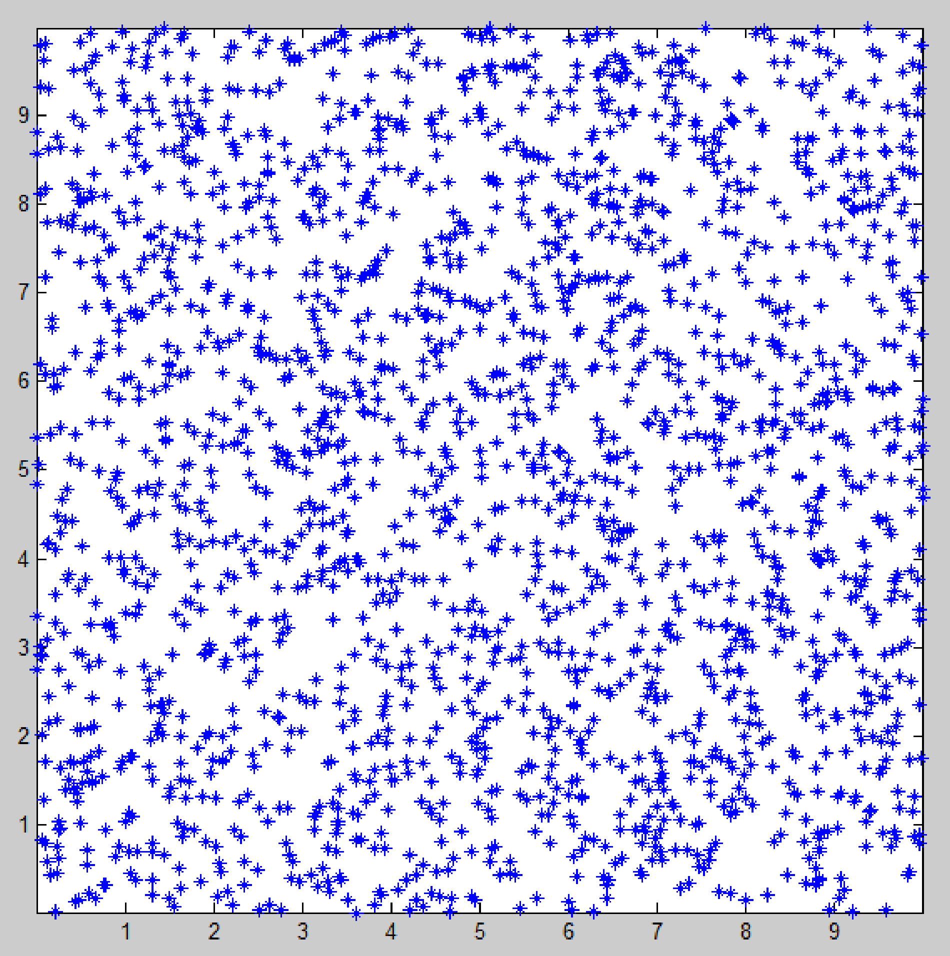
$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

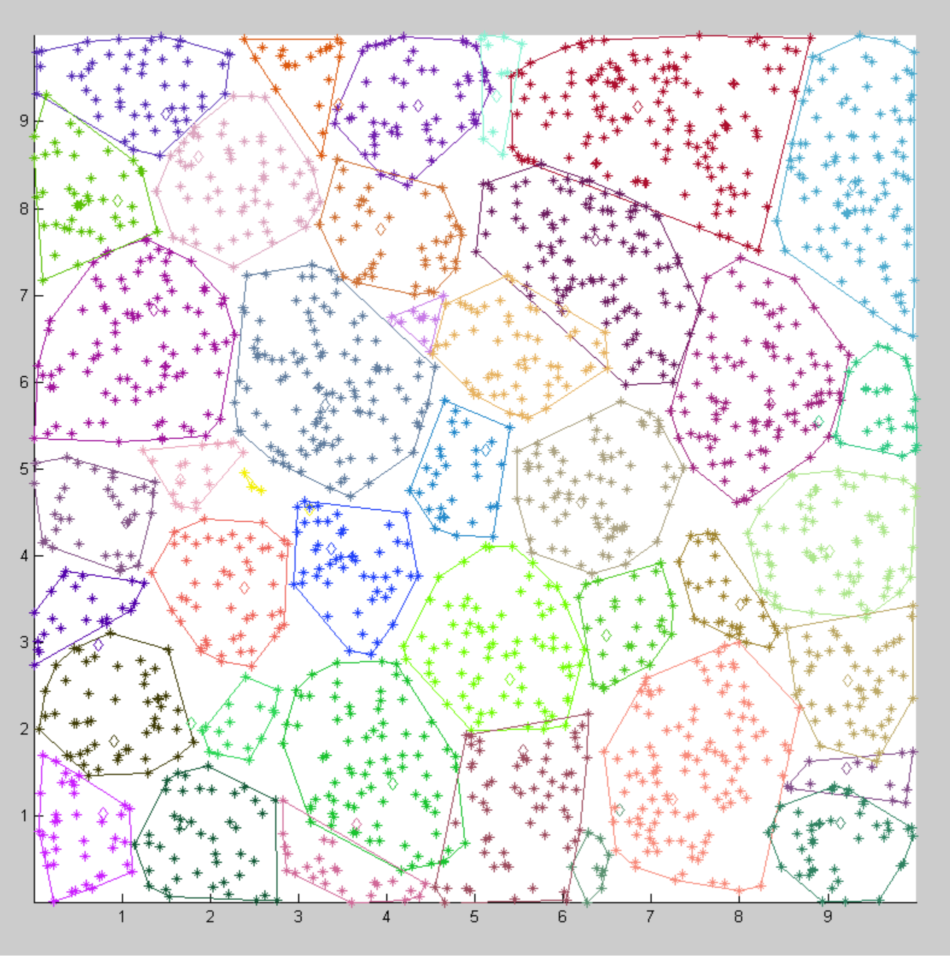
$$K(x_i - x) = e^{-c||x_i - x||^2}$$

Given a set of points in a Cartesian plane

1. Select a Starting Point
2. This point becomes a local mean.
3. Cast a window with radius R, about the local mean
4. Take Mean of all points in window, this becomes new local mean
5. Now mean is shifted
6. Recast Window and Repeat Process.
7. Eventually a Convergence is reached and

Clustering Results





Part II

Variance Estimation

Variance, Normal Distribution

In probability theory, the **normal** (or **Gaussian**) **distribution** is a continuous probability distribution, defined by the formula [1]

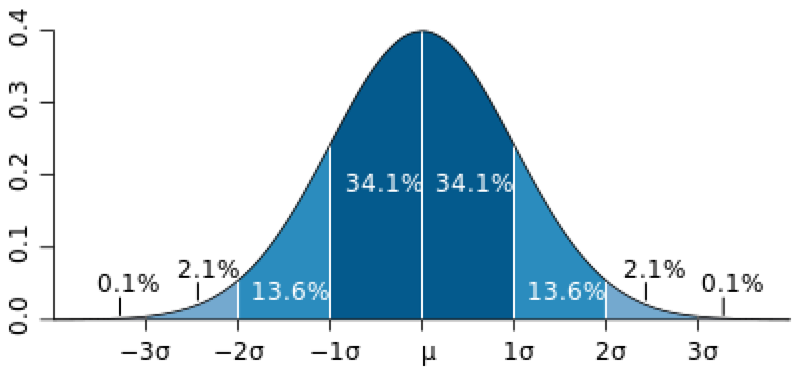
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The parameter σ is its standard deviation; its variance is therefore σ^2.

Discrete random variable

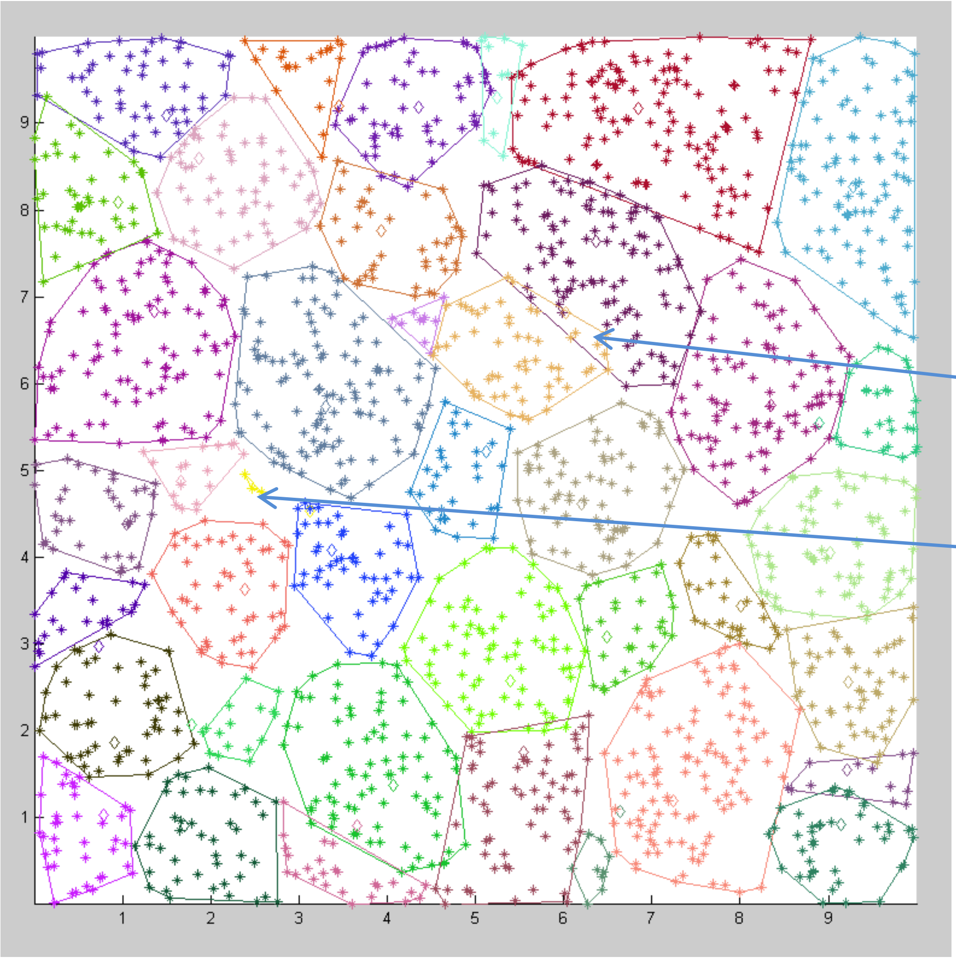
If the random variable X is discrete with probability mass function $x_1 \mapsto p_1, \ldots, x_n \mapsto p_n$, then

$$\operatorname{Var}(X) = \sum_{i=1}^{n} (p_i \cdot (x_i - \mu)^2) = \sum_{i=1}^{n} (p_i \cdot x_i^2) - \mu^2$$



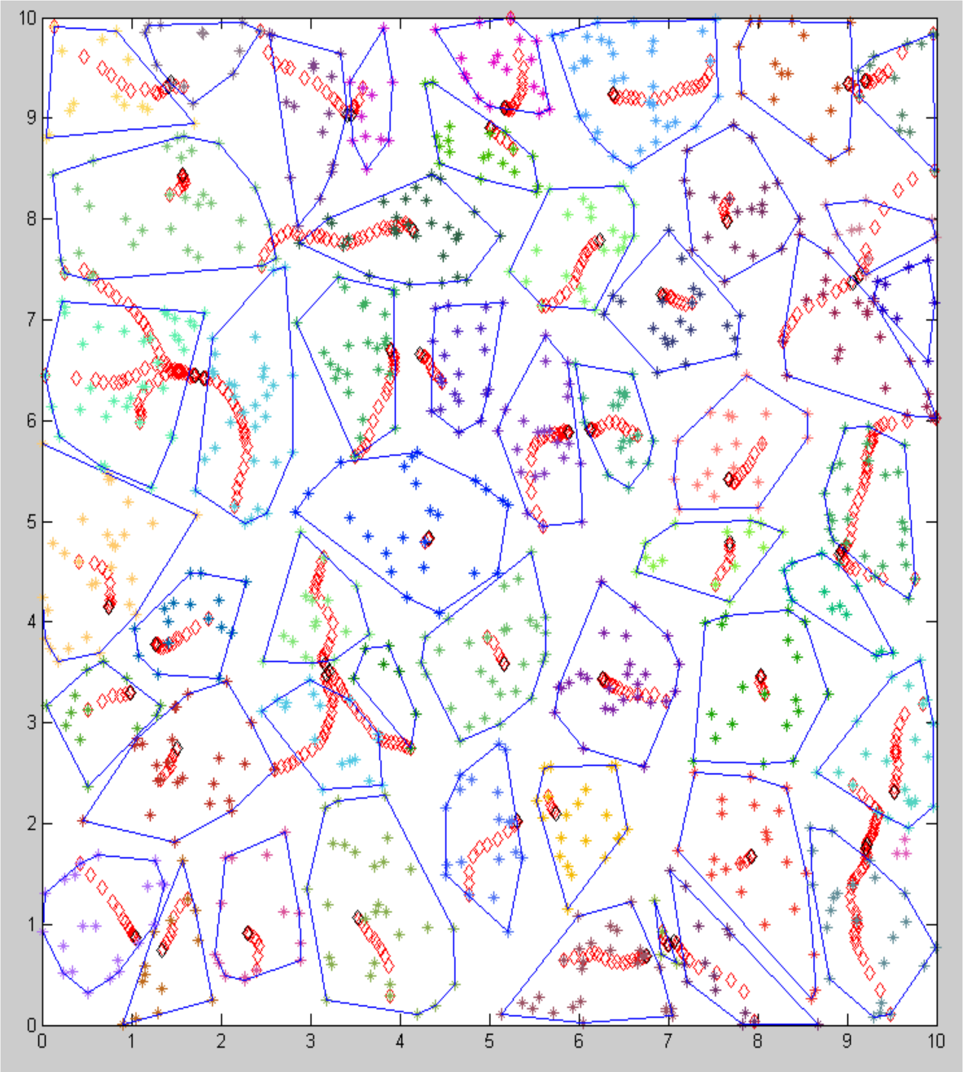
*[1] All data above from wikipedia: http://en.wikipedia.org/wiki/Normal_distribution

Mean Shift Assumes Equal Size Distributions



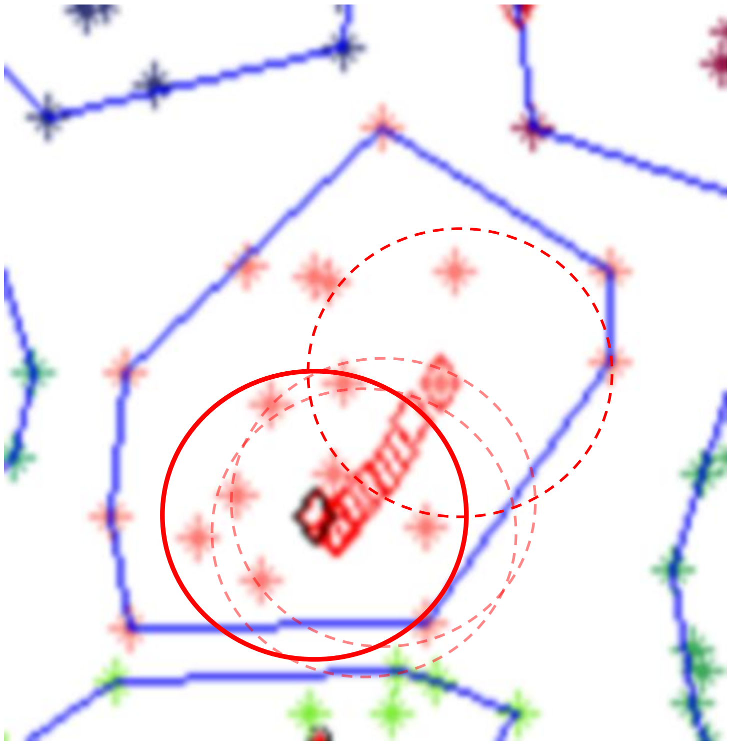
- What Are the true sizes of these clusters
    - We must not assume all our clusters have the same distribution size in terms of variance.

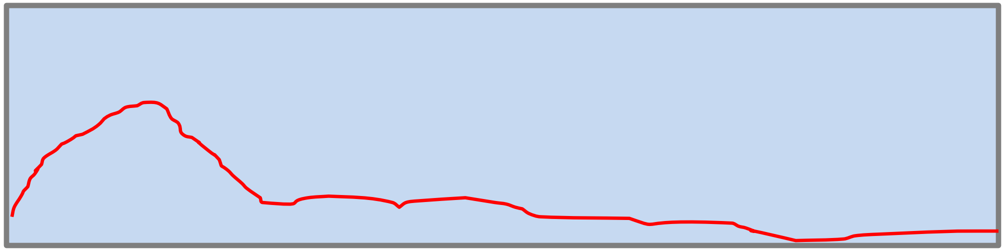
Variance Estimation and Displacement Vector



- The Diamonds Represent the displacement at iterations in the clustering
- The "Trail" is representative of the gradient finding process.
- The displacement trail can be used to determine the variance of an underlying distribution.

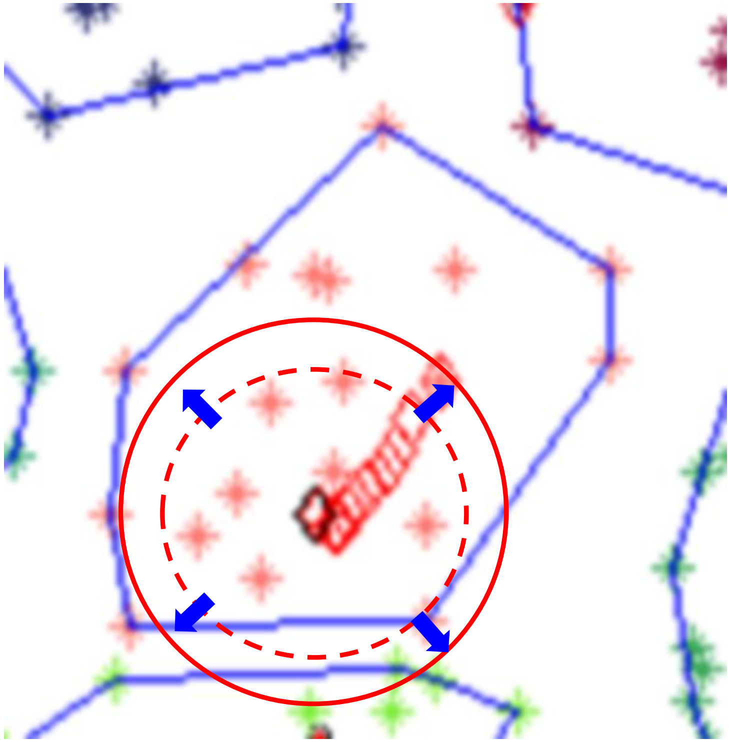
Conceptual Idea

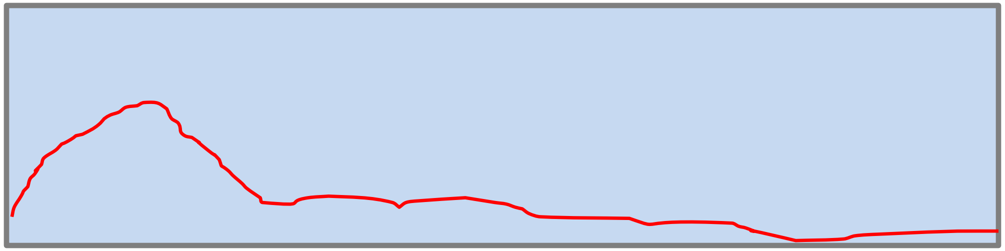




- Displacement Analysis:
  - Is really a change to the process of clustering
  - Normally we have a fixed window size.

Conceptual Idea

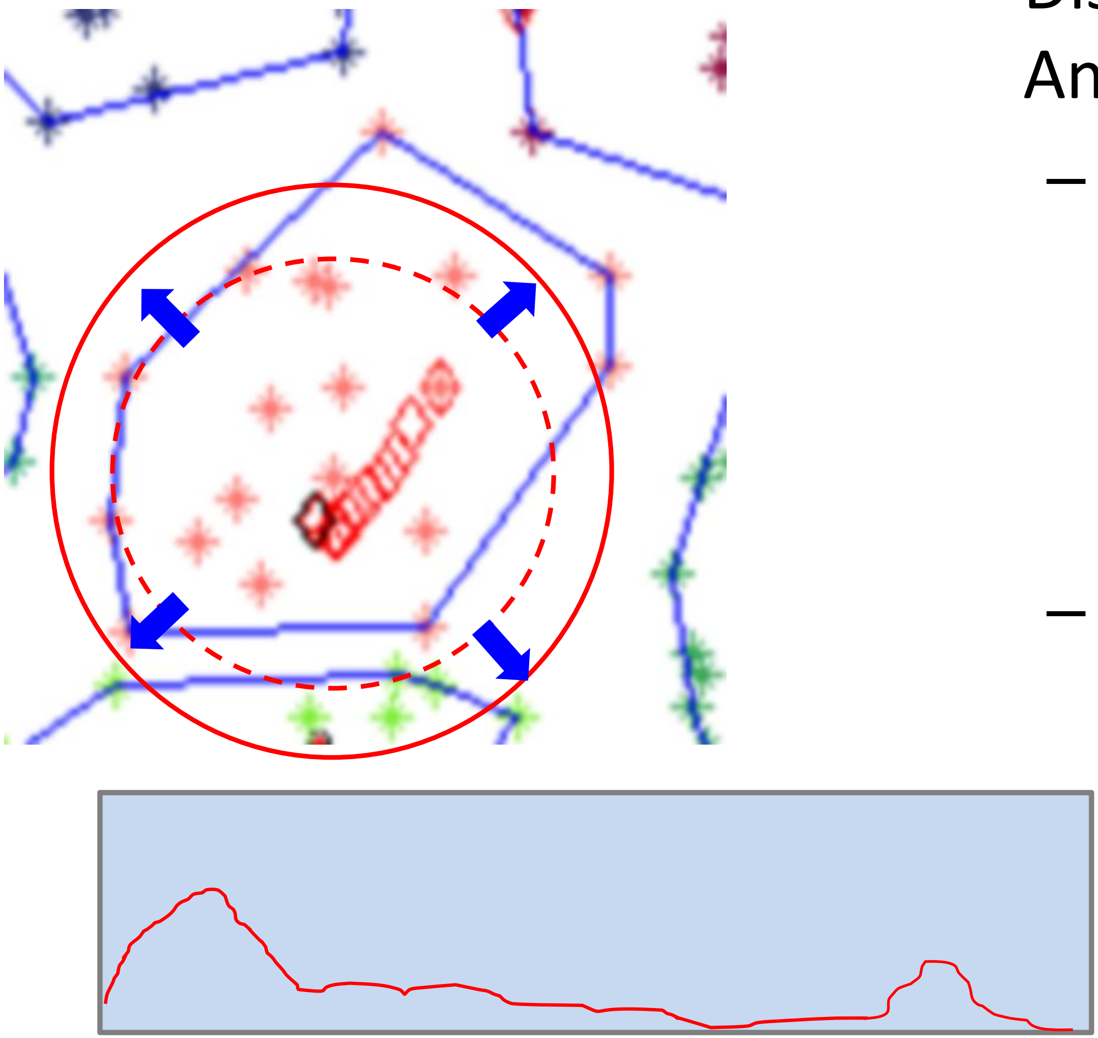




- Displacement Analysis:
  - The difference is when convergence is reached
  - We increase the size of the window incrementally
  - Re-Compute local mean

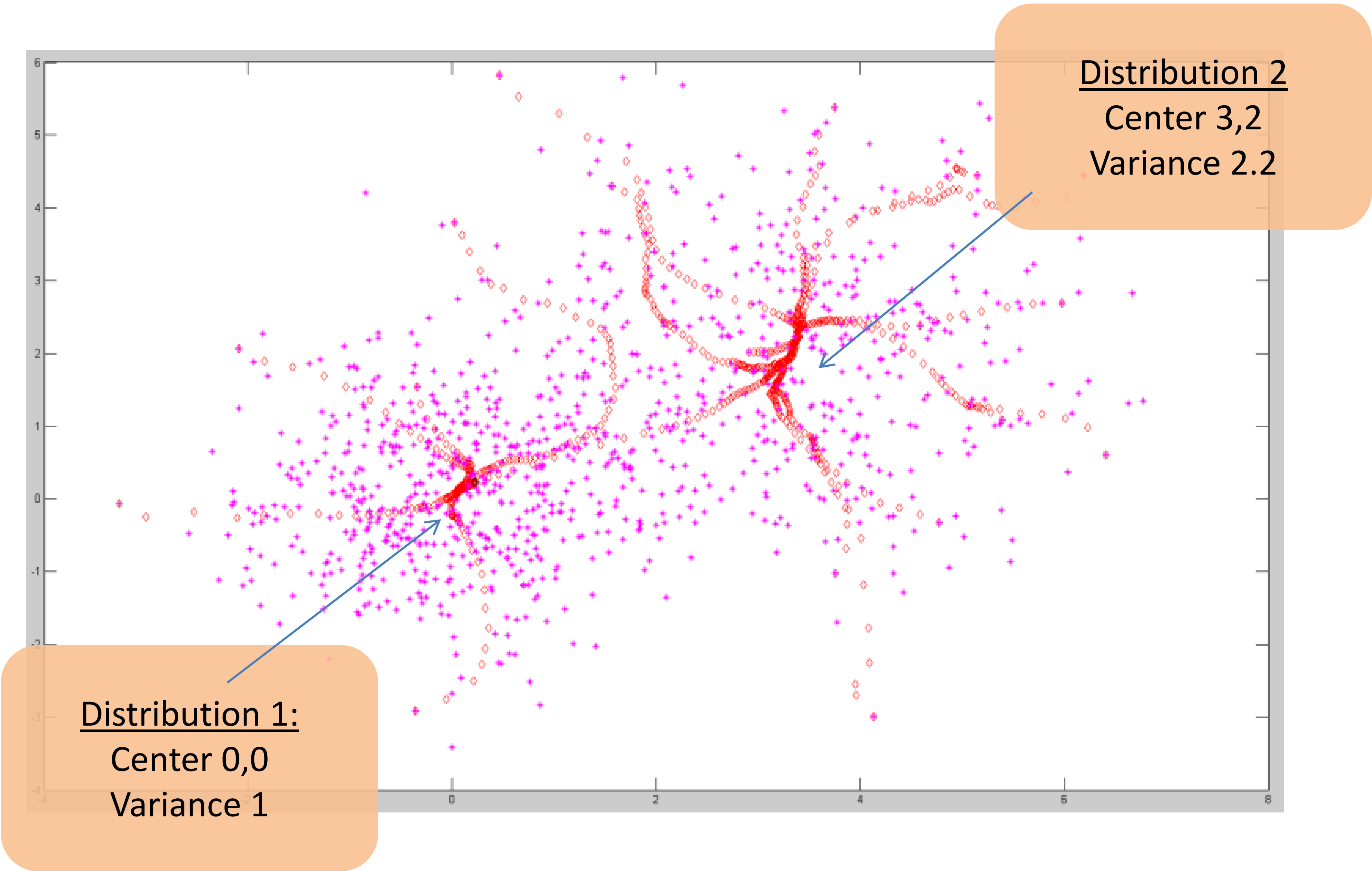
    Let converge

Conceptual Idea

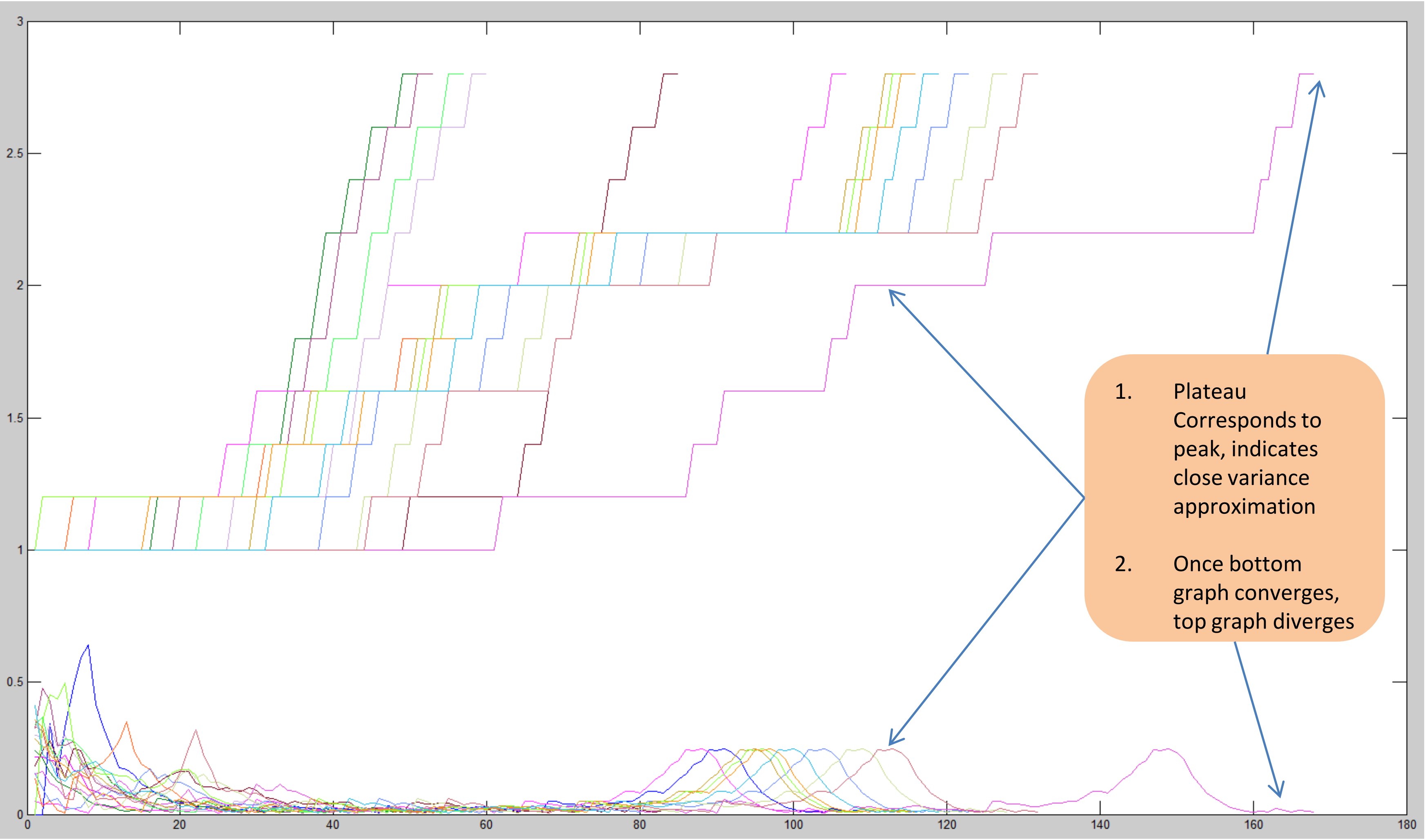


- Displacement Analysis:
  - Keep repeating until displacement vector experiences a slight divergence & re-convergence
  - This happens as window ingests statistical outliers to the distribution

Two Normal Distributions

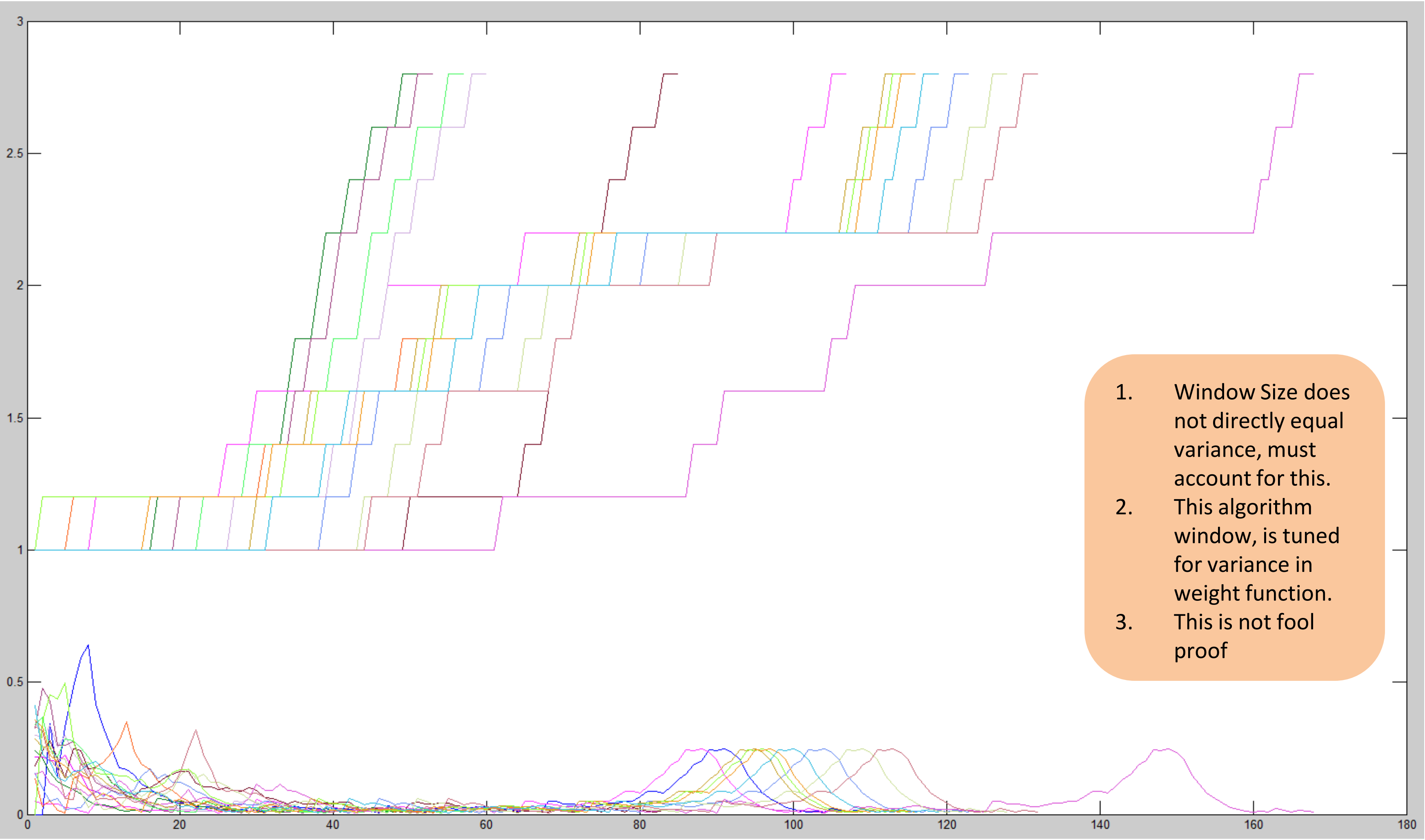


Displacement Analysis



1. Plateau Corresponds to peak, indicates close variance approximation
2. Once bottom graph converges, top graph diverges

Displacement Analysis

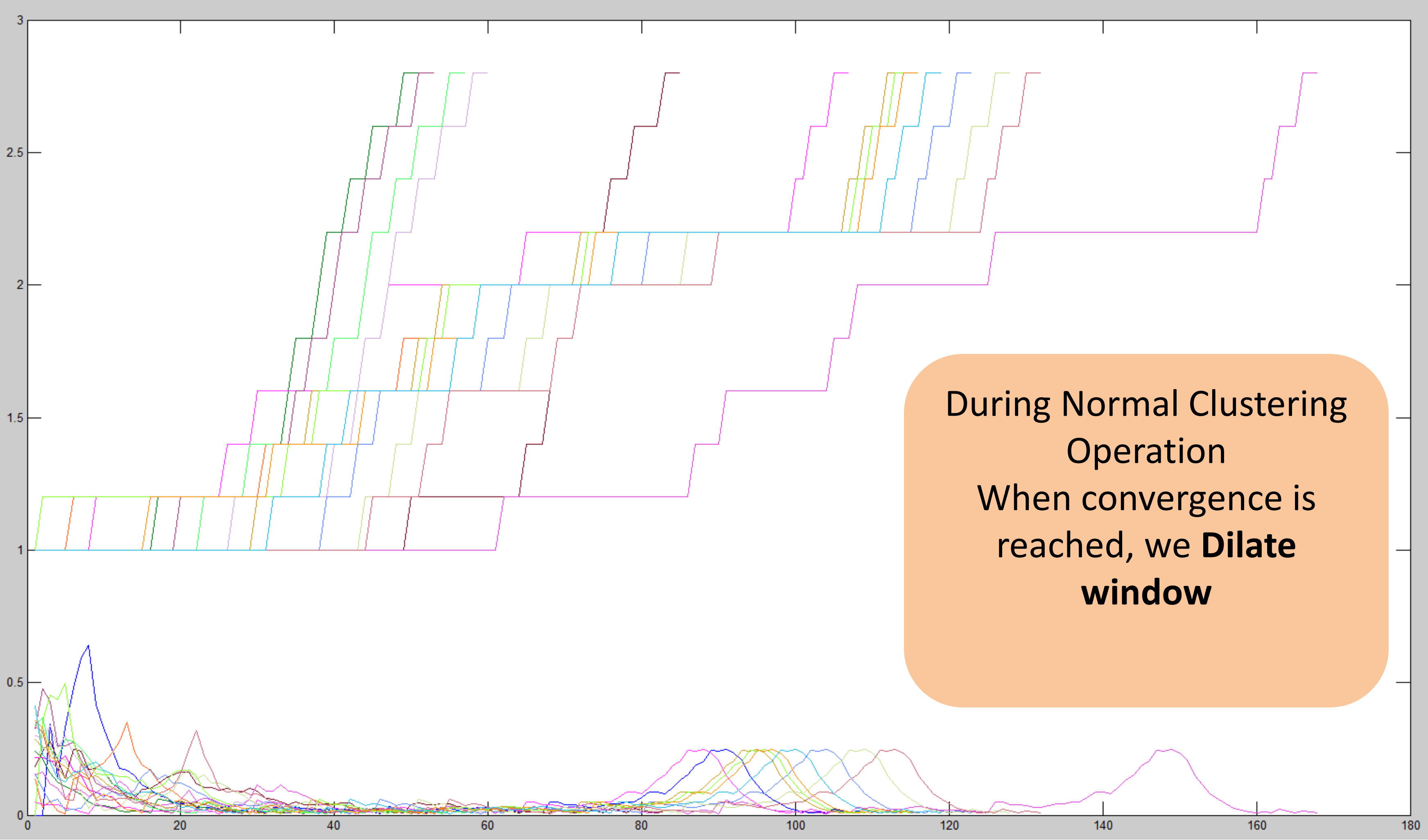


1. Window Size does not directly equal variance, must account for this.
2. This algorithm window, is tuned for variance in weight function.
3. This is not fool proof

The End

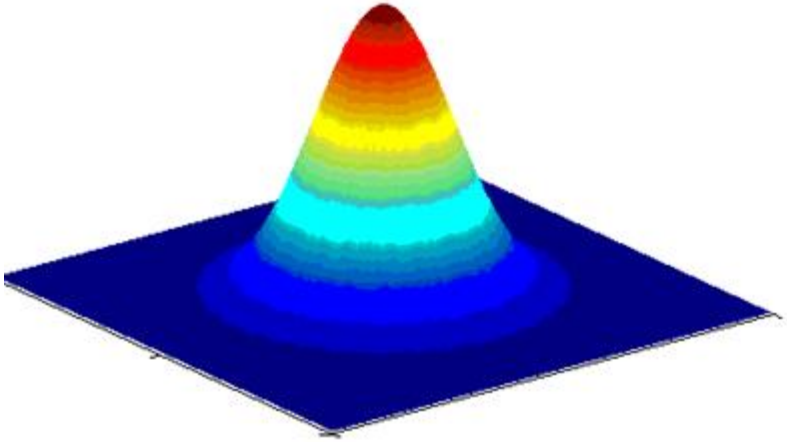
({[Back Up]})

Displacement Analysis

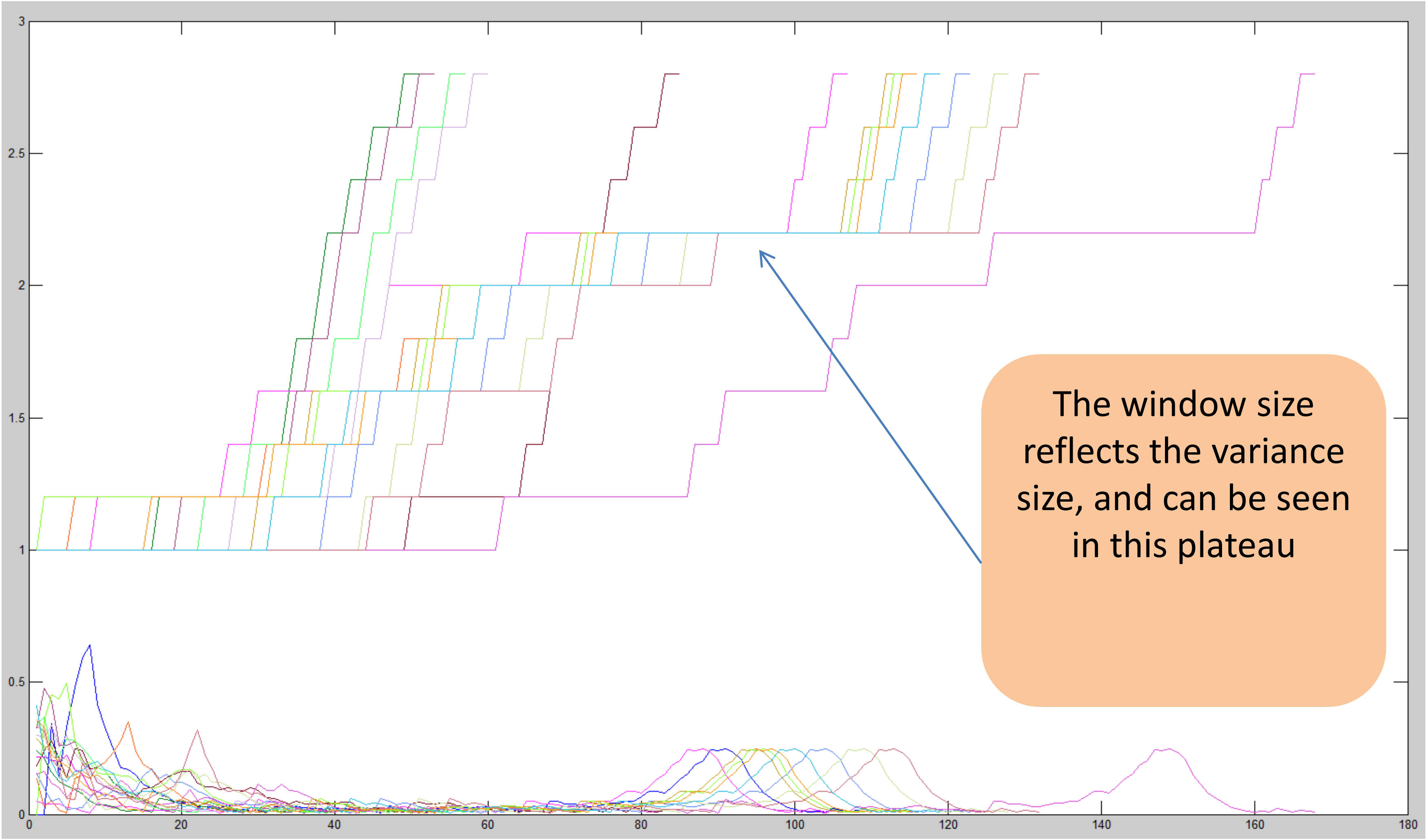


During Normal Clustering Operation
When convergence is reached, we **Dilate window**

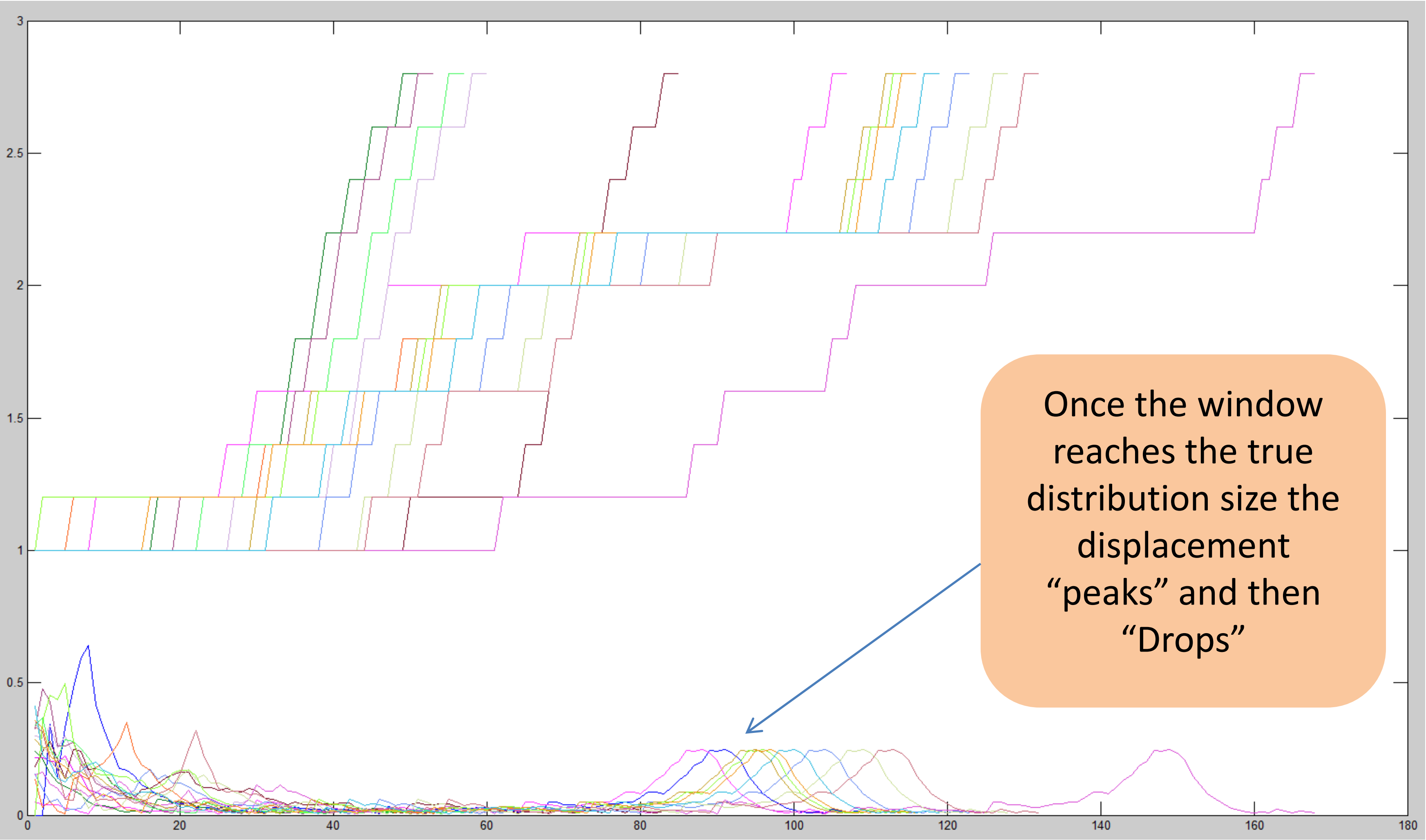
Back - Up



Displacement Analysis



Displacement Analysis



Method Two – Group Cluster Centers