On Clustering Stability

Department of Statistics,
University of Washington

Hanyu Zhang

Problem

- Given data D and one of the following:
  - **Loss-based clustering:** a partition C of data D that minimizes a loss function $L(C, D)$

Problem

- Given data D and one of the following:
  - **Loss-based clustering:** a partition C of data D that minimizes a loss function $L(C, D)$
  - **Model-based clustering:** a model P fitted to the data from some model class

Problem

- Given data D and one of the following:
  - **Loss-based clustering:** a partition C of data D that minimizes a loss function $L(C, D)$
  - **Model-based clustering:** a model P fitted to the data from some model class

Target: Decide whether C or P is meaningful.

Intuition: Loss-based clustering

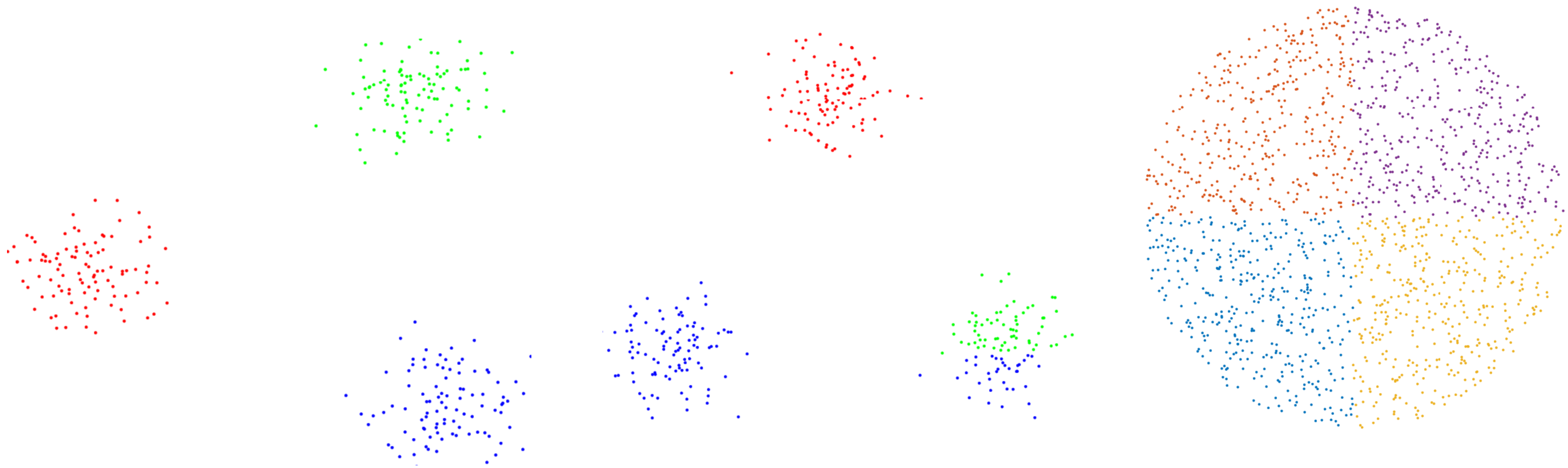
- Loss-based clustering:
    - A loss-based clustering seeks a partition C of Data D that minimizes a loss function $L(C, D)$

Intuition: Loss-based clustering

- Loss-based clustering:
  - A loss-based clustering seeks a partition C of Data D that minimizes a loss function $L(C, D)$
  - A meaningful clustering C should have a small loss and have a stability property

Example: K-means Clustering

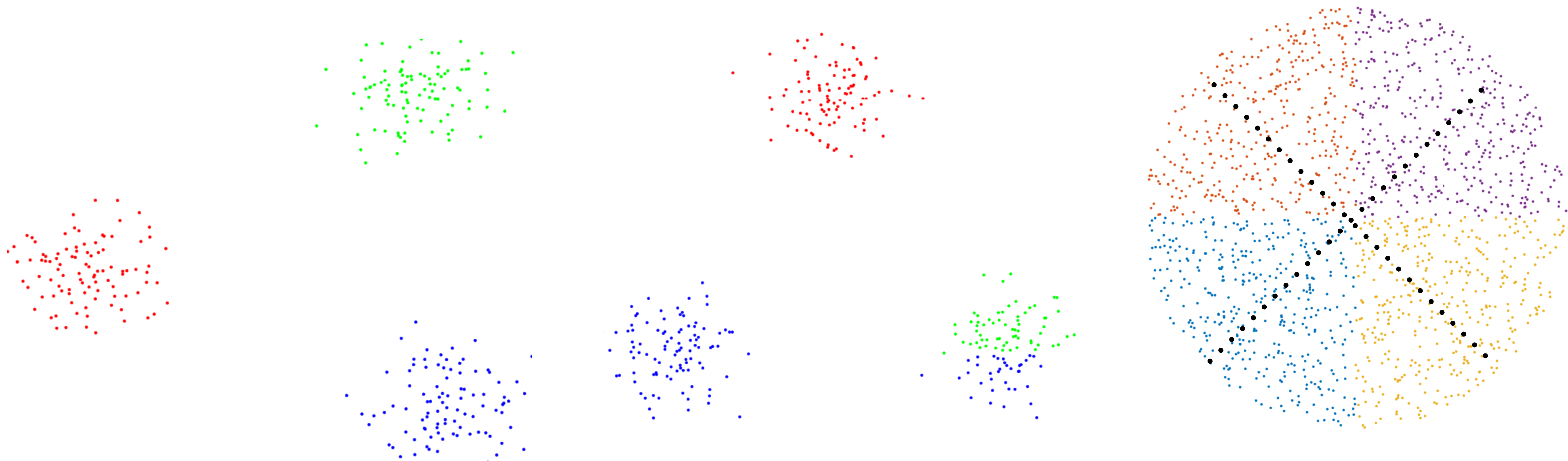
A meaningful clustering on data should be good and stable



Good, Stable **Bad, Not Stable** **Unstable, Good?**

Example: K-means Clustering

A meaningful clustering on data should be good and stable

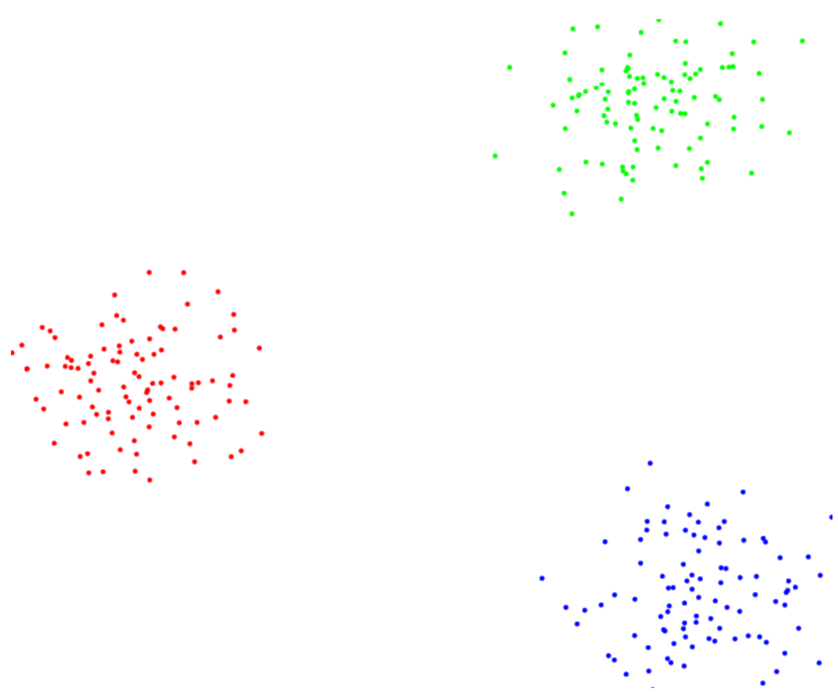


Good, Stable **Bad, Not Stable** **Unstable, Good?**

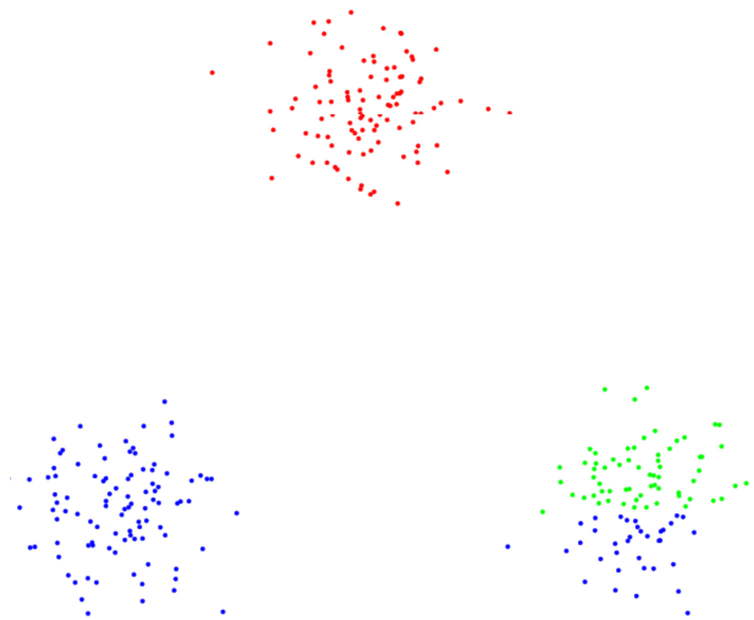
Main Result: loss-based clustering

- Given data D and a clustering $\hat{C}$ obtained by trying to minimize some loss $L(C, D)$:
    - Any other clustering C' such that $L(C', D) \leq L(\hat{C}, D)$ is close to $\hat{C}$ in earth mover's distance under some computable conditions of $\hat{C}$.

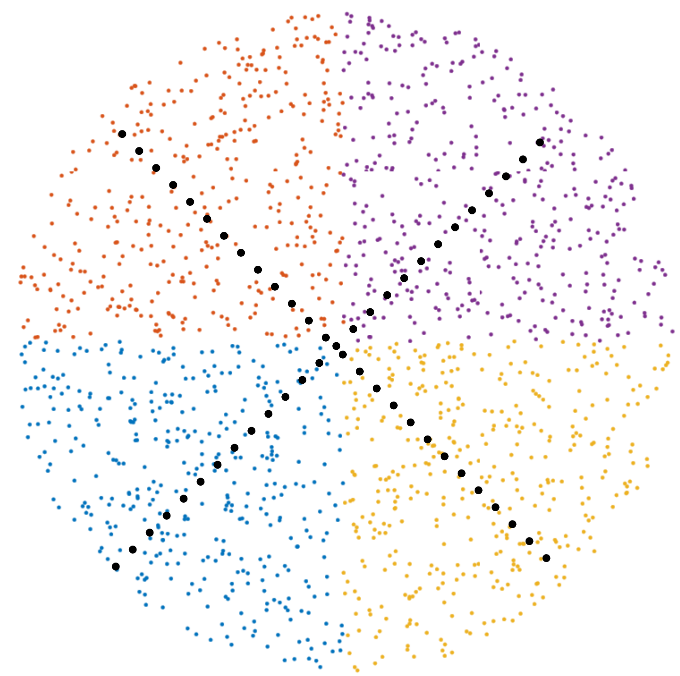
Example: K-means Clustering



Good, Stable, d=1e-4



Bad, Not Stable,
No Guarantee



Unstable, Good?
No Guarantee

Intuition: Model-based clustering

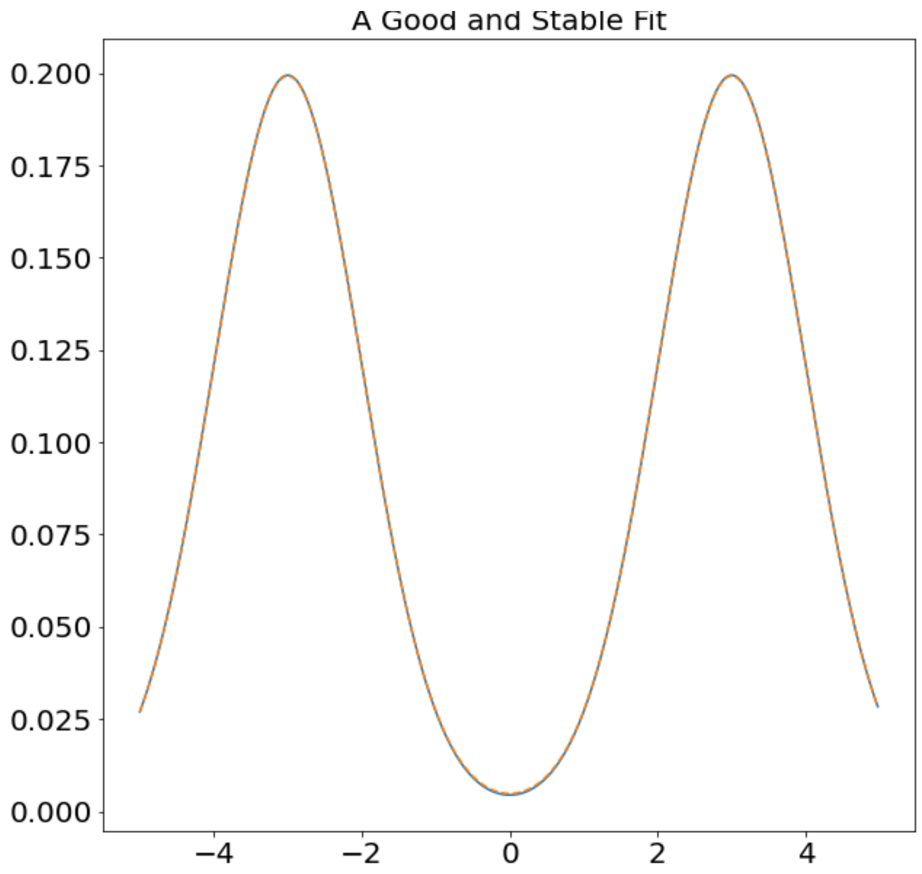
- Model-based clustering:
    - A model-based clustering fits the data to a model P

Intuition: Model-based clustering

- Model-based clustering:
  - A model-based clustering fits the data to a model P
  - A meaningful fitted model P for model-based clustering should be also be good and stable.

Example: Model-based Clustering

A meaningful fitted model should be good and stable



A Good and Stable Fit

True Density $P = ?$

Fitted Density $\hat{P} = 0.5N(-3,1) + 0.5N(3,1)$

Total Variation Distance: $d_{TV}(P, \hat{P}) \leq \epsilon = 0.001$

Example: Model-based Clustering

A meaningful fitted model should be good and stable

P: A gaussian mixture with 4 components, means at (-3,-1,1,3), each variance = 2.25

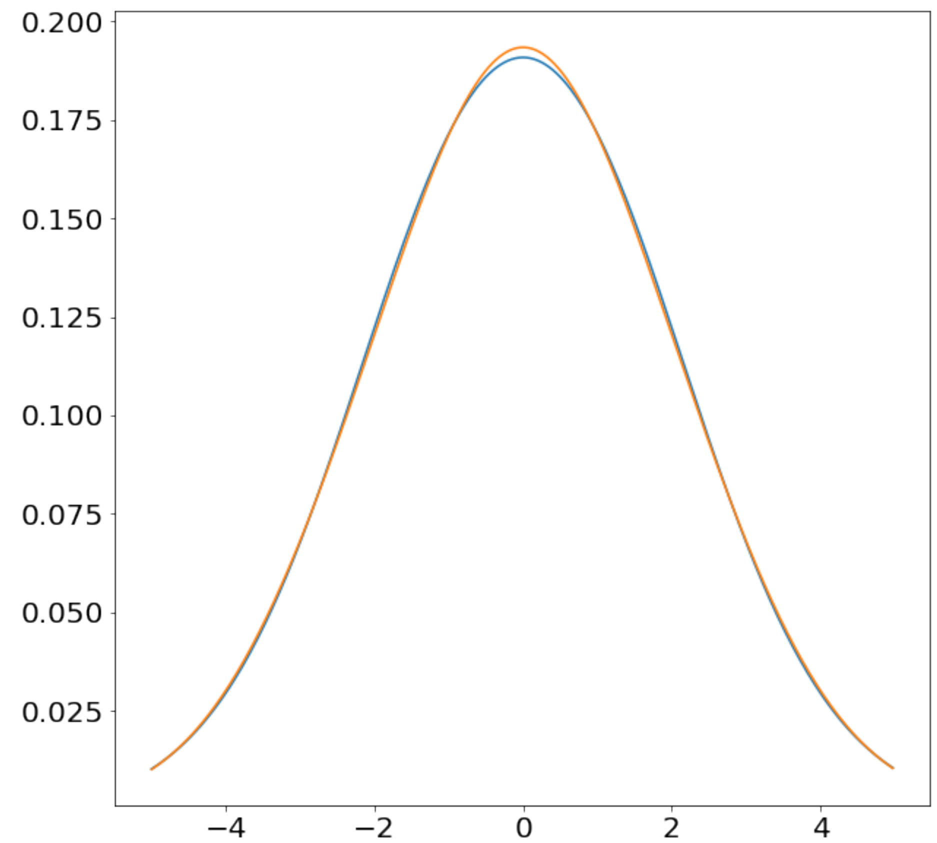
P': A gaussian mixture with 5 components, means at (-4,-2,0,2,4), each variance = 2.25

Example: Model-based Clustering

A meaningful fitted model should be good and stable

P: A gaussian mixture with 4 components, means at (-3,-1,1,3), each variance = 2.25

P': A gaussian mixture with 5 components, means at (-4,-2,0,2,4), each variance = 2.25



A Good but Unstable Fit

Main Result: Model-based clustering

P in Model class $M(K, \pi_{\min}, c)$ if:

- $P = \sum_{k=1}^{K} \pi_k N_d(\mu_k, \sigma_k^2 I_d), \quad \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0$

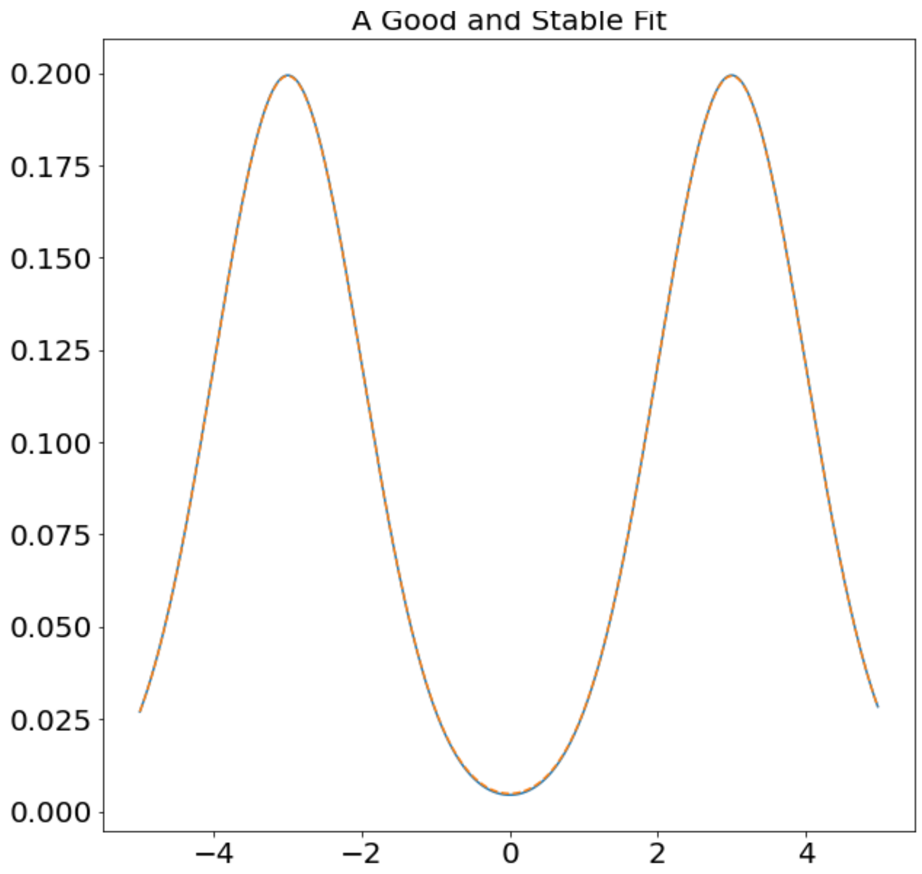
- Minimum weight $\min \pi_i \geq \pi_{\min}$

- Minimum separation $\min_{i \neq j} \frac{\|\mu_i - \mu_j\|}{\sigma_i + \sigma_j} \geq c$

Main Result: Model-based clustering

- Given data D from true density P and a model $\hat{P}$ in model class $M(K, \pi_{\min}, c)$ such that total variation distance $d_{TV}(\hat{P}, P) \leq \epsilon$
  - Under computable conditions on $K, \pi_{\min}, c, \epsilon$, any other $P' \in M(K', \pi_{\min}, c)$ such that $d_{TV}(P', P) \leq \epsilon$ must have $K' = K$, and close in parametric distance to $\hat{P}$.

Example: Model-based Clustering



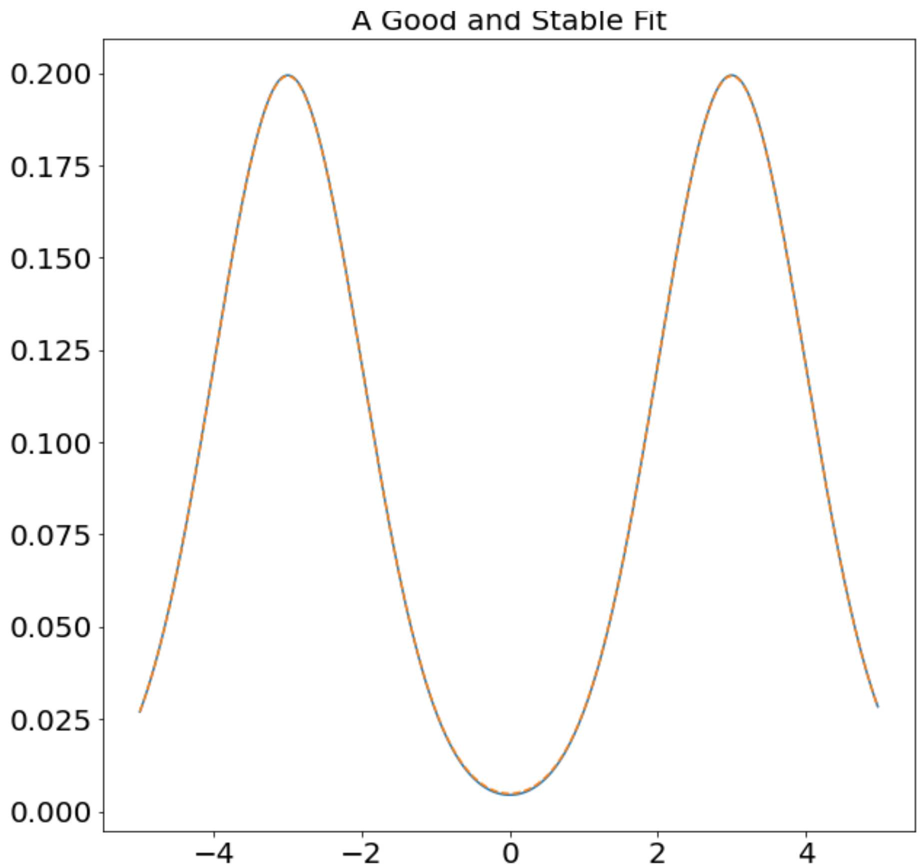
A Good and Stable Fit

True Density $P = ?$

Fitted Density $\hat{P} = 0.5N(-3,1) + 0.5N(3,1)$

Total Variation Distance: $d_{TV}(P, \hat{P}) \leq 0.001$

Example: Model-based Clustering



A Good and Stable Fit

True Density $P = ?$

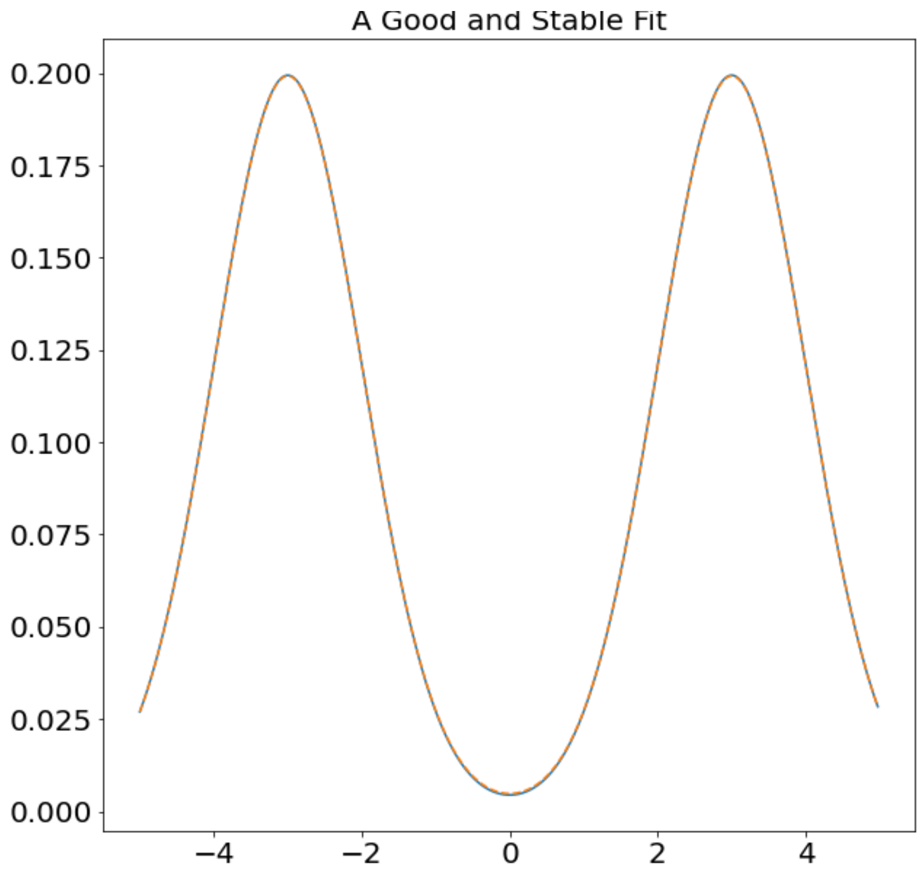
Fitted Density $\hat{P} = 0.5N(-3,1) + 0.5N(3,1)$

Any Density $P' \in M(2,0.45,3)$

Total Variation Distance: $d_{TV}(P, \hat{P}), d_{TV}(P, P') \leq 0.001$

$$P' = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$$
$$\mu_1 < \mu_2$$

Example: Model-based Clustering



A Good and Stable Fit

True Density $P = ?$

Fitted Density $\hat{P} = 0.5N(-3,1) + 0.5N(3,1)$

Any Density $P' \in M(2,0.45,3)$

Total Variation Distance: $d_{TV}(P, \hat{P}), d_{TV}(P, P') \leq 0.001$

$$P' = \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$$
$$\mu_1 < \mu_2$$

Mean: $|\mu_1 - (-3)| \leq 0.02, |\mu_2 - 3| \leq 0.02$

Variance: $\max\{\sigma_{1,2}^2, 1/\sigma_{1,2}^2\} \leq 1.034$

Weight: $\max\{|\pi_1 - 0.5|, |\pi_2 - 0.5|\} \leq 0.004$

Thank you!