Sequential Clustering of Data Streams from Unknown Distributions

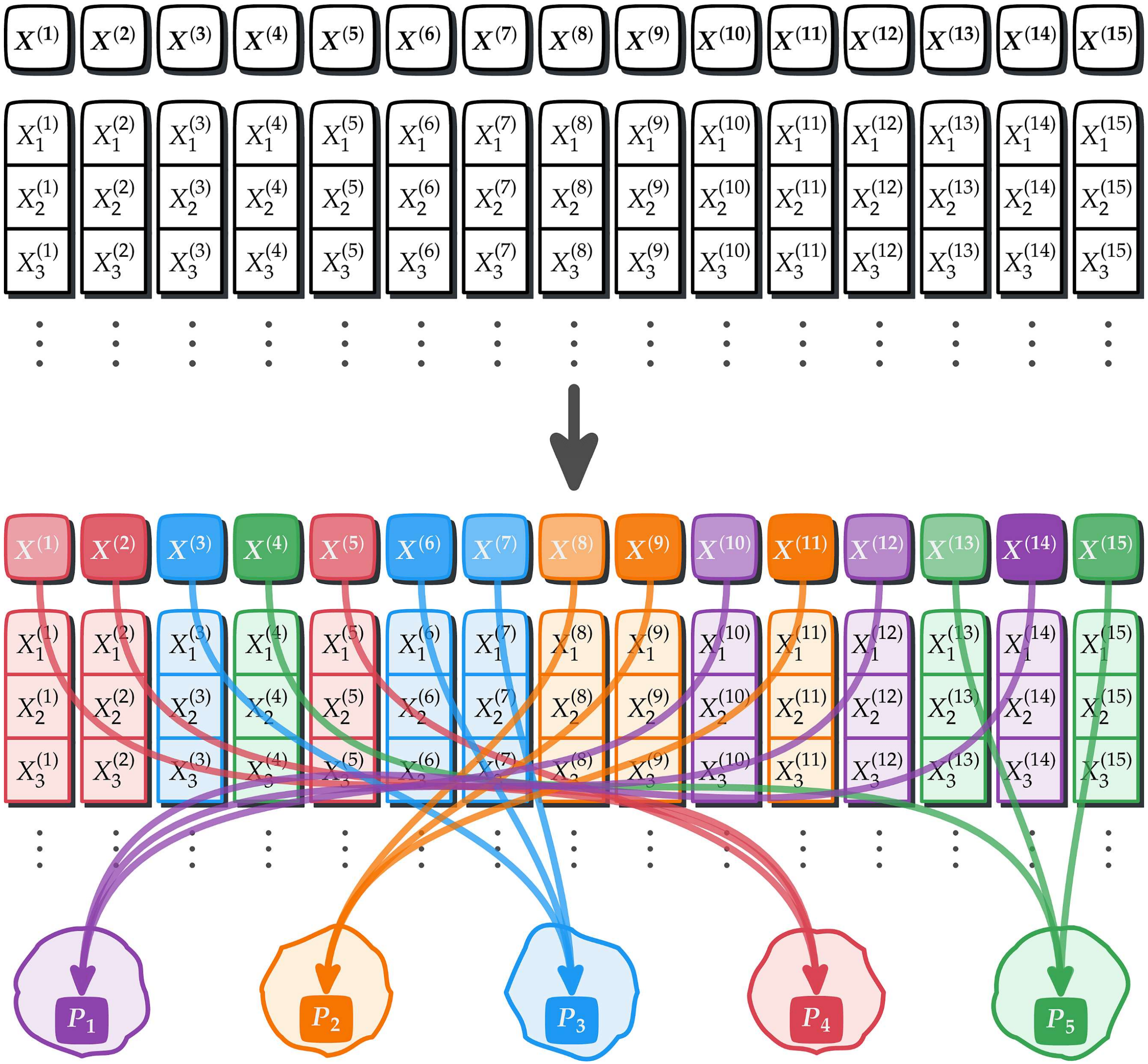
Srikrishna Bhashyam
IIT Madras

Joint work with Sreeram C. Sreenivasan

December 18, 2023

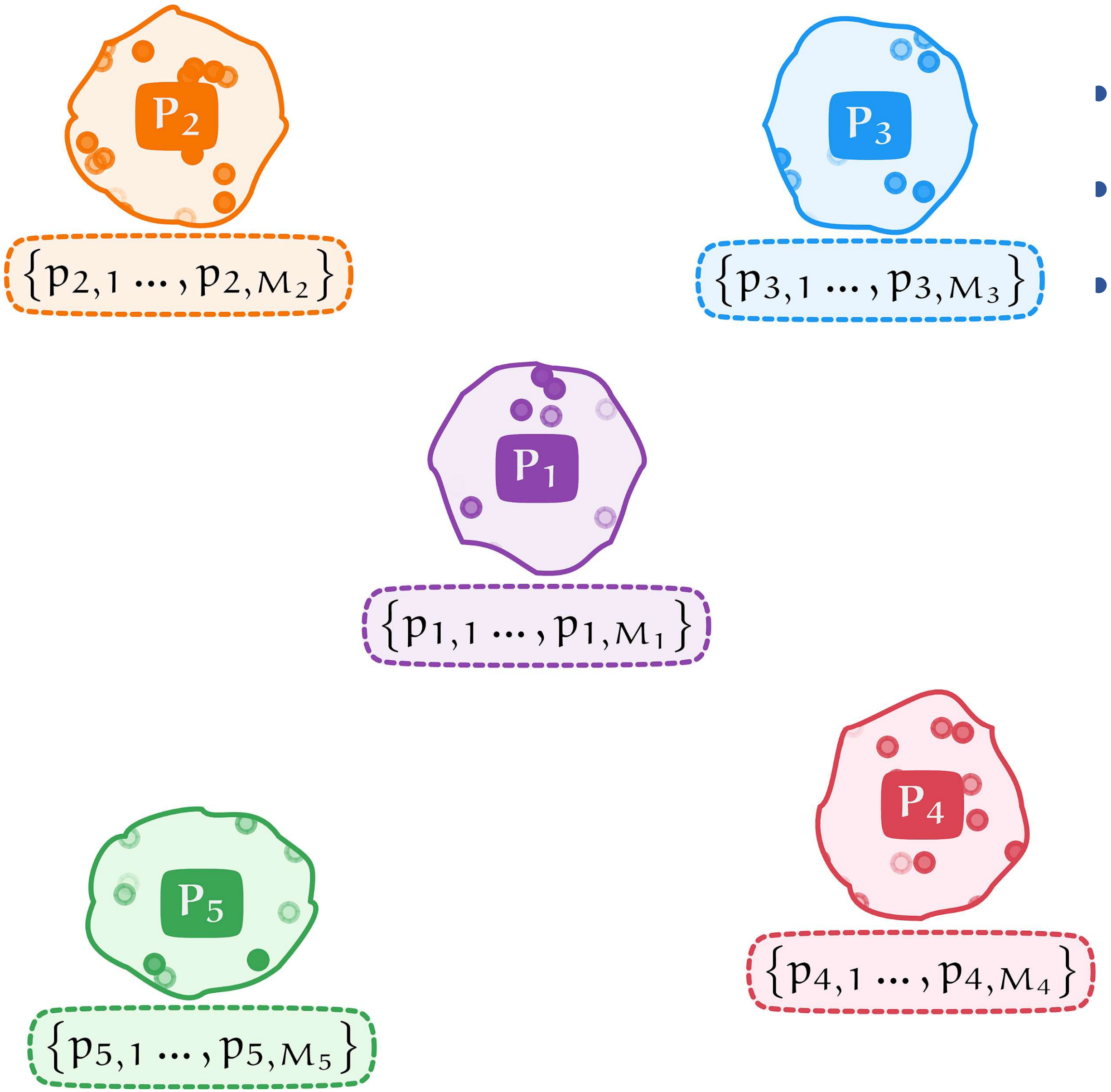
CNI Seminar, IISc Bangalore

Clustering of Data Streams



- Each stream of i.i.d. samples can be from a different distribution
- Need to cluster based on underlying distributions
- Unknown distributions

Clustering



- S data streams
- K clusters
- M_k distributions in cluster k

Comparing distributions/data streams

- Known set distributions, unknown indices
  - Likelihood-based rule
- Unknown distributions + Parametric model for distributions
  - Generalized likelihood instead of likelihood
  - Parameters estimated under each hypothesis and plugged into likelihood
- Unknown distributions, Nonparametric
  - Estimated distances
    - KL divergence
    - Maximum Mean Discrepancy (MMD)
    - Kolmogorov-Smirnov Distance (KSD)

Comparing distributions/data streams

- MMD and KSD
  - Estimates based on the observed samples
  - Estimates that converge to true distance
  - Sequential updates possible

Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD}(p,q) = \sup_{f \in F} \quad E_p[f(X)] - E_q[f(Y)]$$

- $X \sim p$ and $Y \sim q$,
- f a real − valued function from class F
- F: Unit ball in a Reproducing Kernel Hilbert Space (RKHS) with kernel $k(.,.)$
- Estimate with finite number of samples
- Convergence as number of samples grows

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. The Journal of Machine Learning Research, 13(1), 723-773.

MMD Estimate and Convergence

$$X_i^n = \{x_{i1}, x_{i2}, \dots, x_{in}\}$$
$$X_j^n = \{x_{j1}, x_{j2}, \dots, x_{jn}\}$$

Gaussian Kernel

$$k(x,y) = e^{-\frac{(x-y)^2}{2\sigma^2}}$$

$$M_b(i,j,n) = \left[\frac{1}{n^2}\sum_{l,m}\left(k(x_{il}, x_{im}) + k(x_{jl}, x_{jm}) - k(x_{il}, x_{jm}) - k(x_{jl}, x_{im})\right)\right]^{1/2}$$

$M_b(i,j,n)$ converges a.s. to $\mathrm{MMD}(p,q)$ as $n \to \infty$

$$P\left[|M_b(i,j,n) - \mathrm{MMD}(p,q)| > 4\sqrt{\frac{K}{n}} + \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{4K}\right)$$

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. The Journal of Machine Learning Research, 13(1), 723-773.

MMD sequential update

Sequential update with O(n) computations

$$M_b^2(i,j,n) = \left[\left(\frac{n-1}{n}\right)^2 M_b^2(i,j,n-1) + \frac{1}{n^2}\left(\sum_{l=1}^{n} h(x_{il},x_{in},x_{jl},x_{jn}) + \sum_{m=1}^{n-1} h(x_{in},x_{im},x_{jn},x_{jm})\right)\right]$$

$$h(x_{il},x_{im},x_{jl},x_{jm}) = k(x_{il},x_{im}) + k\left(x_{jl},x_{jm}\right) - k\left(x_{il},x_{jm}\right) - k(x_{jl},x_{im})$$

KS Distance

$$\mathrm{KS}(p,q) = \sup_{a \in R} \left| F_p(a) - F_q(a) \right|$$

Estimate

$$\mathrm{KS}(i,j,n) = \sup_{a \in R} \left| \hat{F}_i^{(n)}(a) - \hat{F}_j^{(n)}(a) \right|$$

Sequential update

$$\hat{F}_i^{(n)}(a) = \frac{n-1}{n} \hat{F}_i^{(n-1)}(a) + \frac{1}{n} I_{(-\infty,a]}(x_{in})$$

KS Distance: Convergence of estimate

$$\mathrm{KS}(p,q) = \sup_{a \in R} \left| F_p(a) - F_q(a) \right|$$

$$P[|\mathrm{KS}(i,j,n) - \mathrm{KS}(p,q)| > \epsilon] \leq 4 \exp\left(-\frac{n\epsilon^2}{2}\right)$$

T. Wang, Q. Li, D. J. Bucci, Y. Liang, B. Chen and P. K. Varshney, "K-Medoids Clustering of Data Sequences With Composite Distributions," in IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2093-2106, 15 April15, 2019.

Performance

- Fixed Sample Size (FSS) setting
  - Probability of error vs number of samples
- Sequential (SEQ) setting
  - Probability of error vs expected number of samples
- Performance metrics
  - Universal consistency
  - Universal exponential consistency
  - Error Exponent

FSS Non-parametric Clustering

- Use pairwise distances (MMD/KSD)
- Cluster based on k-medoid clustering
  - Number of clusters K known (K-MED)
  - Number of clusters K unknown

- Steps
  - Center and Cluster initialization
  - Update till convergence

- Universal exponential consistency ($n \to \infty$)

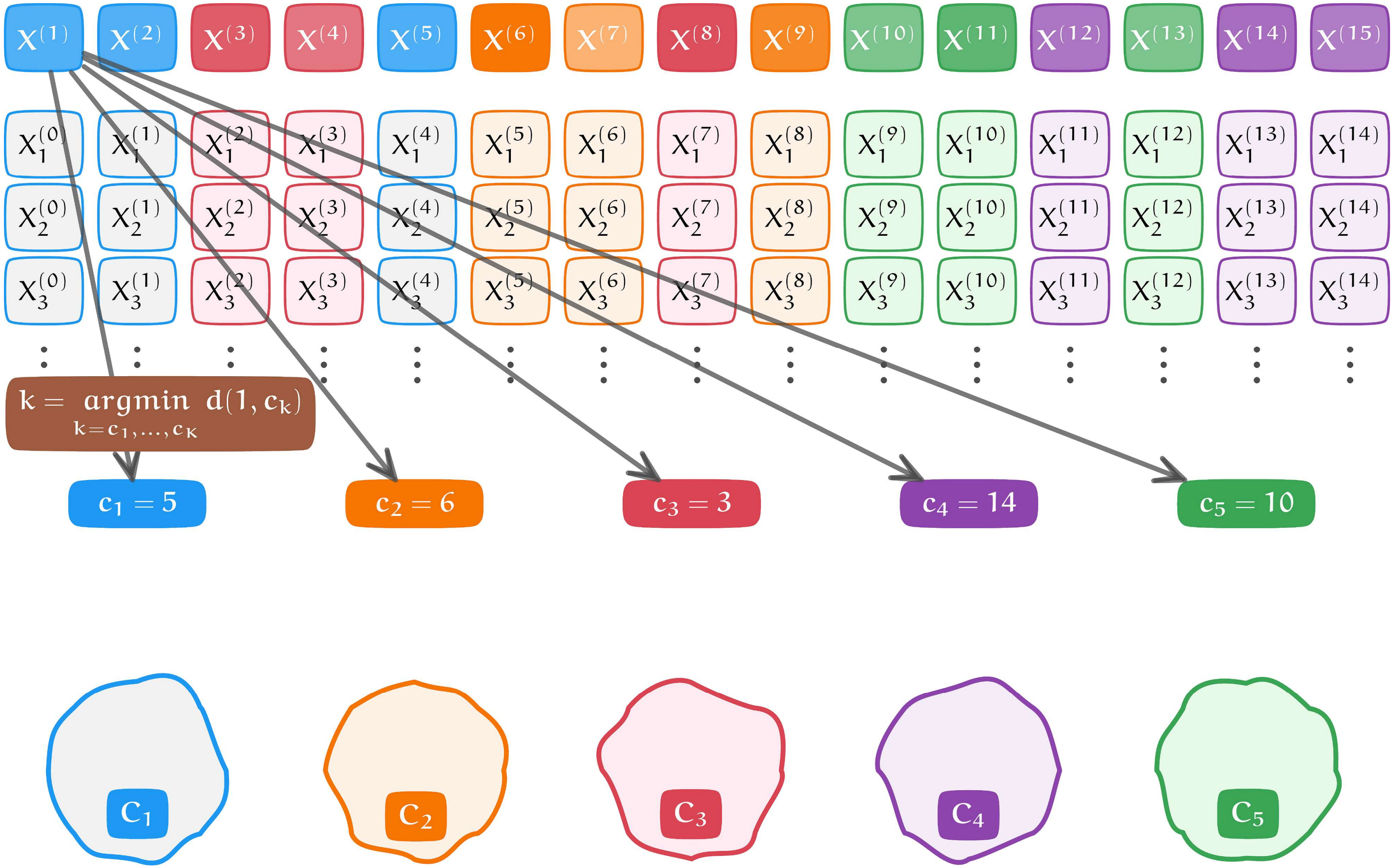
T. Wang, Q. Li, D. J. Bucci, Y. Liang, B. Chen and P. K. Varshney, "K-Medoids Clustering of Data Sequences With Composite Distributions," in IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2093-2106, 15 April15, 2019.

K-Medoids Clustering (KMED)

- Compute Pairwise distances
- Center initialization
  - First center: Pick a random stream initially
  - Pick next center that has maximum minimum distance to already chosen centers
  - Repeat

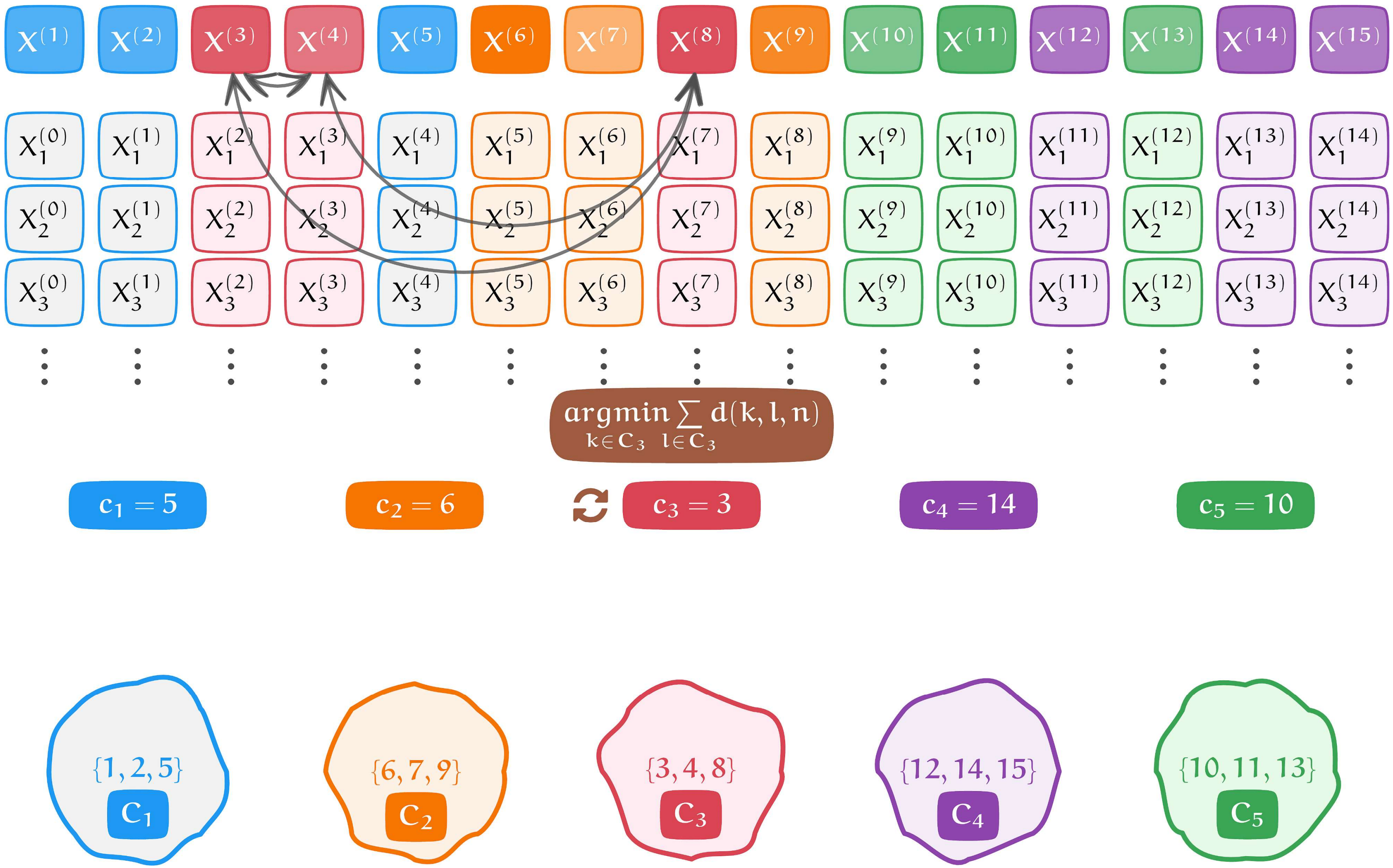
T. Wang, Q. Li, D. J. Bucci, Y. Liang, B. Chen and P. K. Varshney, "K-Medoids Clustering of Data Sequences With Composite Distributions," in IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2093-2106, 15 April15, 2019.

K-Medoids Clustering



- Assign each stream to cluster with closest center

K-Medoids Clustering



- Center update (Medoid)

K-Medoids Clustering

- Repeat cluster and center update until convergence
- Performance

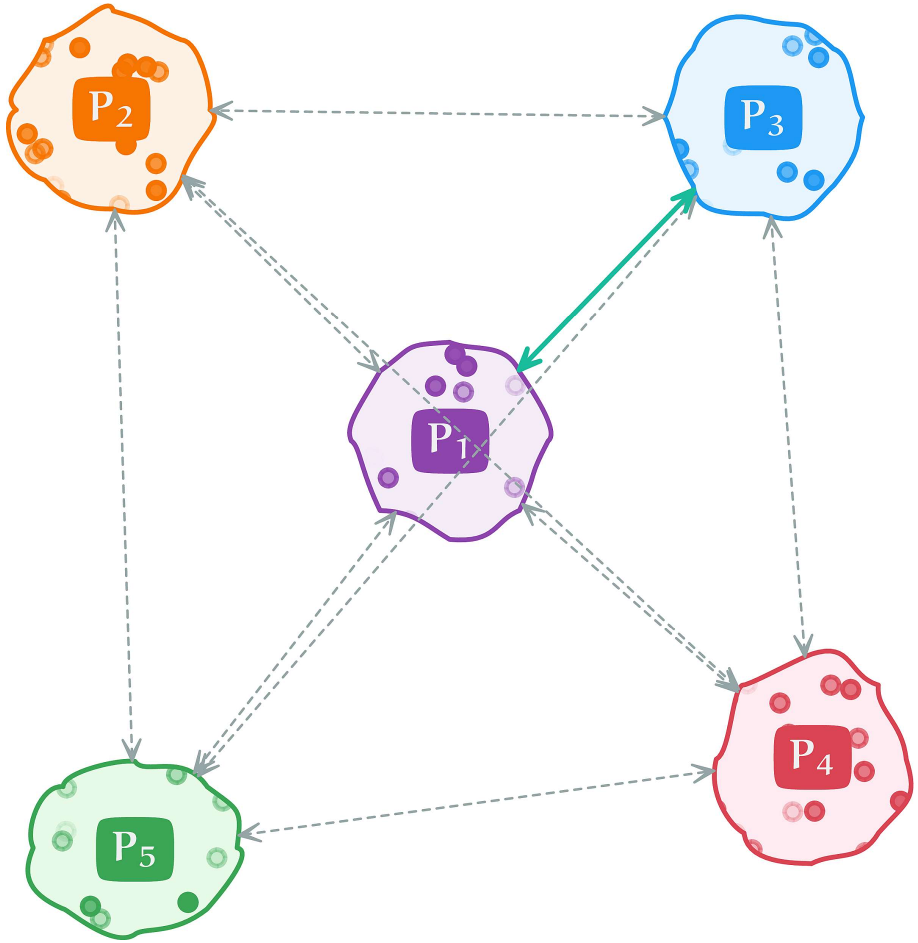
$$P_e \le M^2(4T+8)\exp\left(-\frac{n\Delta_{\mathrm{mmd}}^2}{64K}\right)$$

$$\Delta_{\mathrm{mmd}} = d_H - d_L$$

T. Wang, Q. Li, D. J. Bucci, Y. Liang, B. Chen and P. K. Varshney, "K-Medoids Clustering of Data Sequences With Composite Distributions," in IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2093-2106, 15 April15, 2019.

Assumptions (for the analysis)

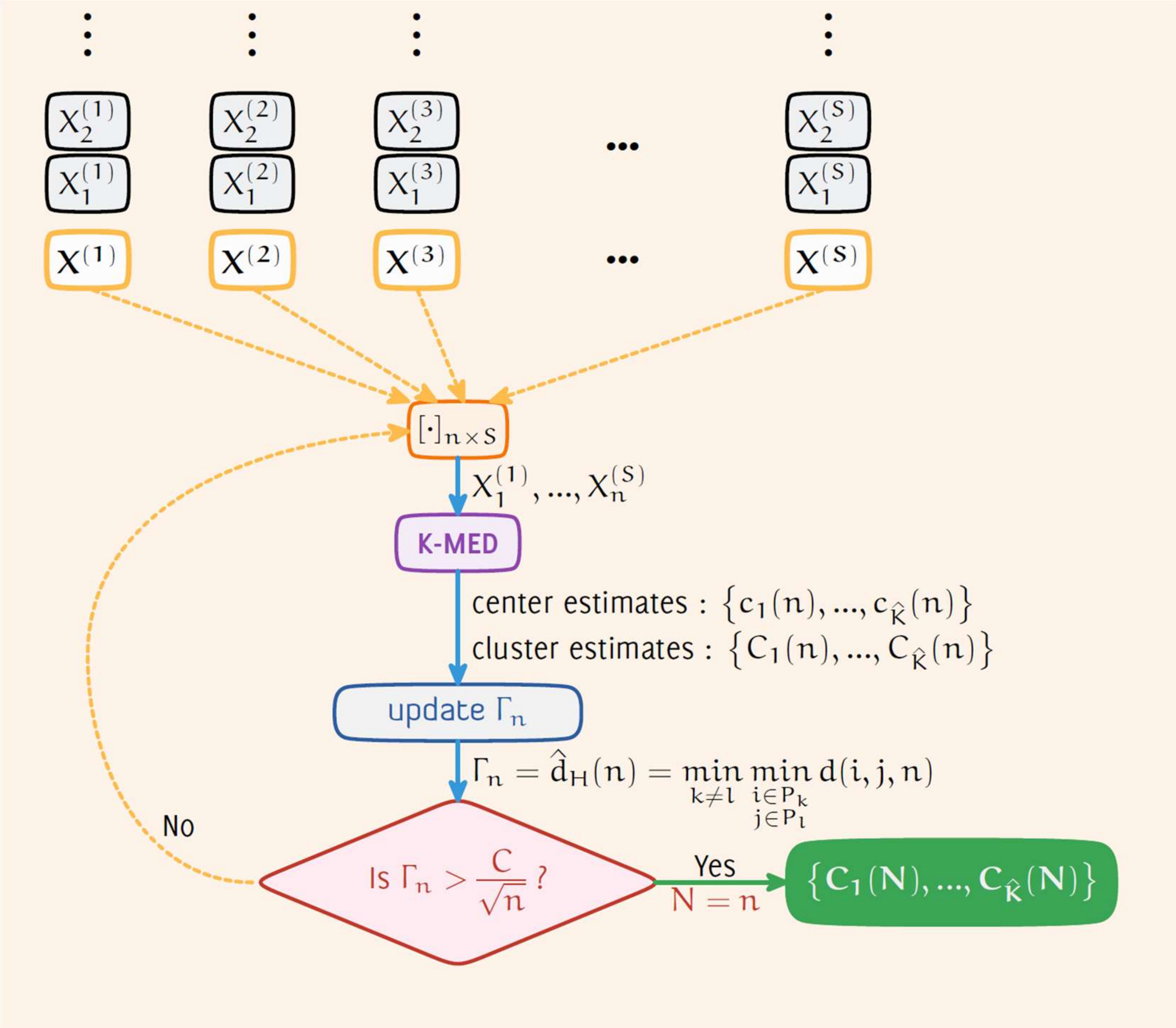
- Minimum inter-cluster distance d_H
- Maximum intra-cluster distance $d_L < d_H$



FSS Non-parametric Clustering

- Unknown number of clusters
- Need to know something about d_H and d_L
- Two variants of KMED
    - KMED-MERGE
        - Generate enough centers so that each stream is close enough to a center
        - Merge clusters whose centers are close
    - KMED-SPLIT
        - Begin with one cluster
        - Split cluster if a sequence has a large distance from center
- These variants are also exponentially consistent

Our Work: Sequential Clustering



- Need a stopping rule
- Threshold on empirical minimum inter-cluster distance
- Sequential updates for pairwise distances
- Analysis for consistency

Sreeram C. Sreenivasan, Srikrishna Bhashyam, Nonparametric Sequential Clustering of Data Streams with Composite Distributions, Signal Processing, Volume 204, March 2023.

Analysis: Sequential Clustering

Stops in finite time

$$N = \underset{n \geqslant 1}{\operatorname{argmin}} \left\{ \underbrace{\min_{k \neq l} \min_{i \in C_k} \min_{j \in C_l} d(i,j,n)}_{\Gamma_n} > T_n \right\}$$

$$\forall\, n \geqslant n_d : \mathbb{P}[N > n] \leqslant \,?\, \exp(-\,?\, n)$$

$$\mathbb{P}[N > n] = \mathbb{P}[\{N > n\} \cap E_n] + \mathbb{P}[\{N > n\} \cap E'_n]$$

$$\leqslant \mathbb{P}[E_n] + \mathbb{P}[\{\Gamma_n < T_n\} \cap E'_n]$$

$$\leqslant \mathbb{P}[E_n] + \mathbb{P}\left[\bigcup_{k \neq l} \underbrace{\bigcup_{i \in P_k} \bigcup_{j \in P_l}}_{\text{true clusters}} \{d(i,j,n) < T_n\}\right]$$

$$\leqslant \mathbb{P}[E_n] + \sum_{k \neq l} \sum_{i \in P_k} \sum_{j \in P_l} \mathbb{P}\left[\underbrace{d(i, j, n) < T_n}_{\text{i,j from different clusters}}\right]$$

Analysis: Error probability

$$E = \left\{ \underbrace{\left\{ C_1(N), \ldots, C_{\hat{K}(N)}(N) \right\}}_{\text{clustering output}} \neq \{P_1, \ldots, P_K\} \right\}$$

$P_e = \mathbb{P}[E]$ for a configuration $\{P_1, \ldots, P_K\}$

$$P_{max} = \max_{\{P_1, \ldots, P_K\}} P_e$$

$$\forall\, C \geqslant C_d : P_e \leqslant ?\, \exp\left(-?\, C^2\right)$$

Consistency $\lim_{C \to \infty} P_{max} = 0$

Analysis: Universal Consistency

$$\mathbb{P}[E] = \sum_{n=1}^{\infty} \mathbb{P}[N = n, E_n]$$

$$= \sum_{n=1}^{n_d} \mathbb{P}[N = n, E_n] + \sum_{n>n_d}^{\infty} \mathbb{P}[N = n, E_n]$$

$$\leqslant \sum_{n>n_d}^{\infty} \mathbb{P}[E_n] + \sum_{n=1}^{n_d} \mathbb{P}[N = n, E_n]$$

$$= \sum_{n>n_d}^{\infty} \mathbb{P}[E_n] + \sum_{n=1}^{n_d} \mathbb{P}\left[\left\{\underbrace{\min_{i \in C_k} \min_{j \in C_l}}_{\text{wrong clusters}} KS(i, j, n) > T_n \forall\, k, l\right\}\right]$$

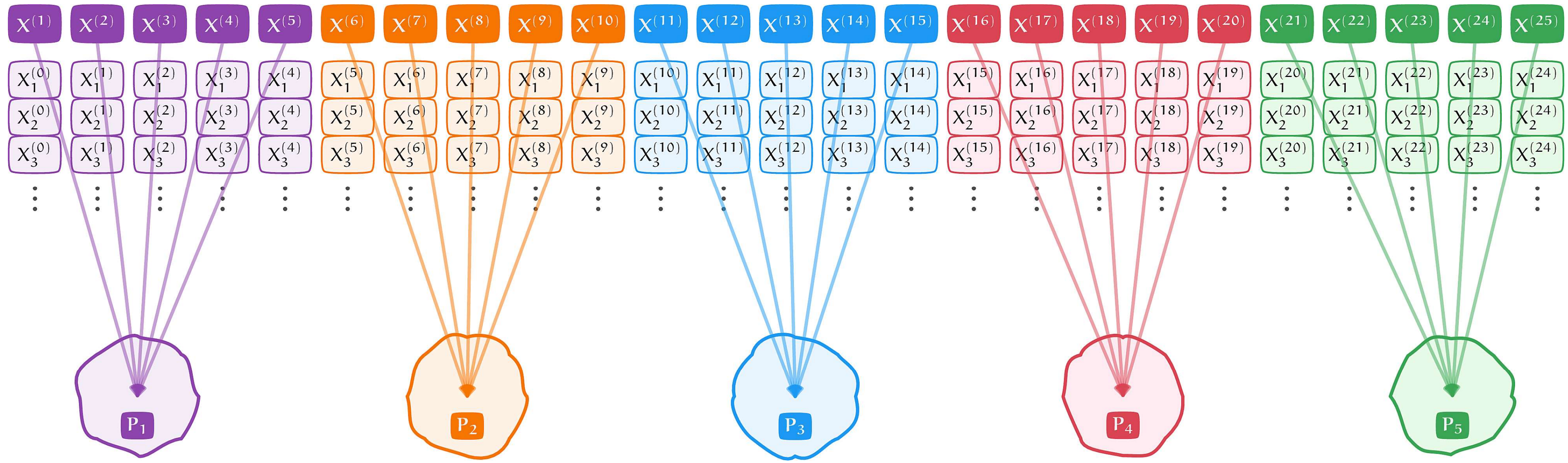
$$\leqslant \sum_{n>n_d}^{\infty} \mathbb{P}[E_n] + \sum_{n=1}^{n_d} \mathbb{P}\left[\underbrace{d(i, j, n) > T_n}_{\text{i,j from same cluster}}\right]$$

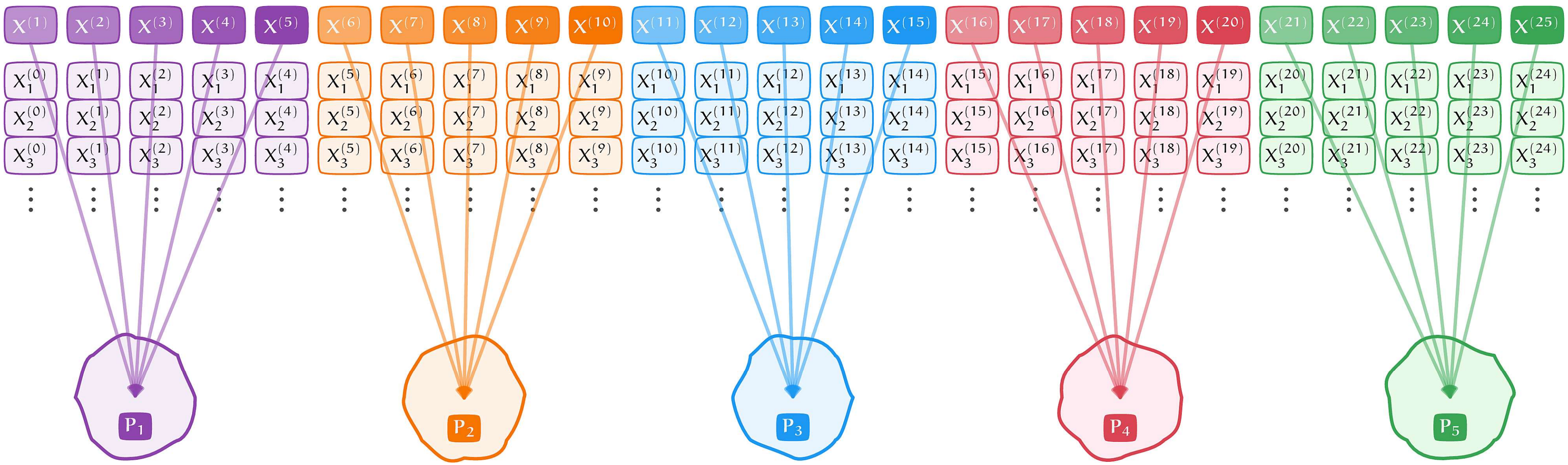
Analysis: Exponential Consistency

- Proper choice of n_d and c_d

$$\mathbb{E}[N] \leqslant -\frac{B^2}{d_H^2} \log P_e (1 + o(1))$$

Simulation Setting

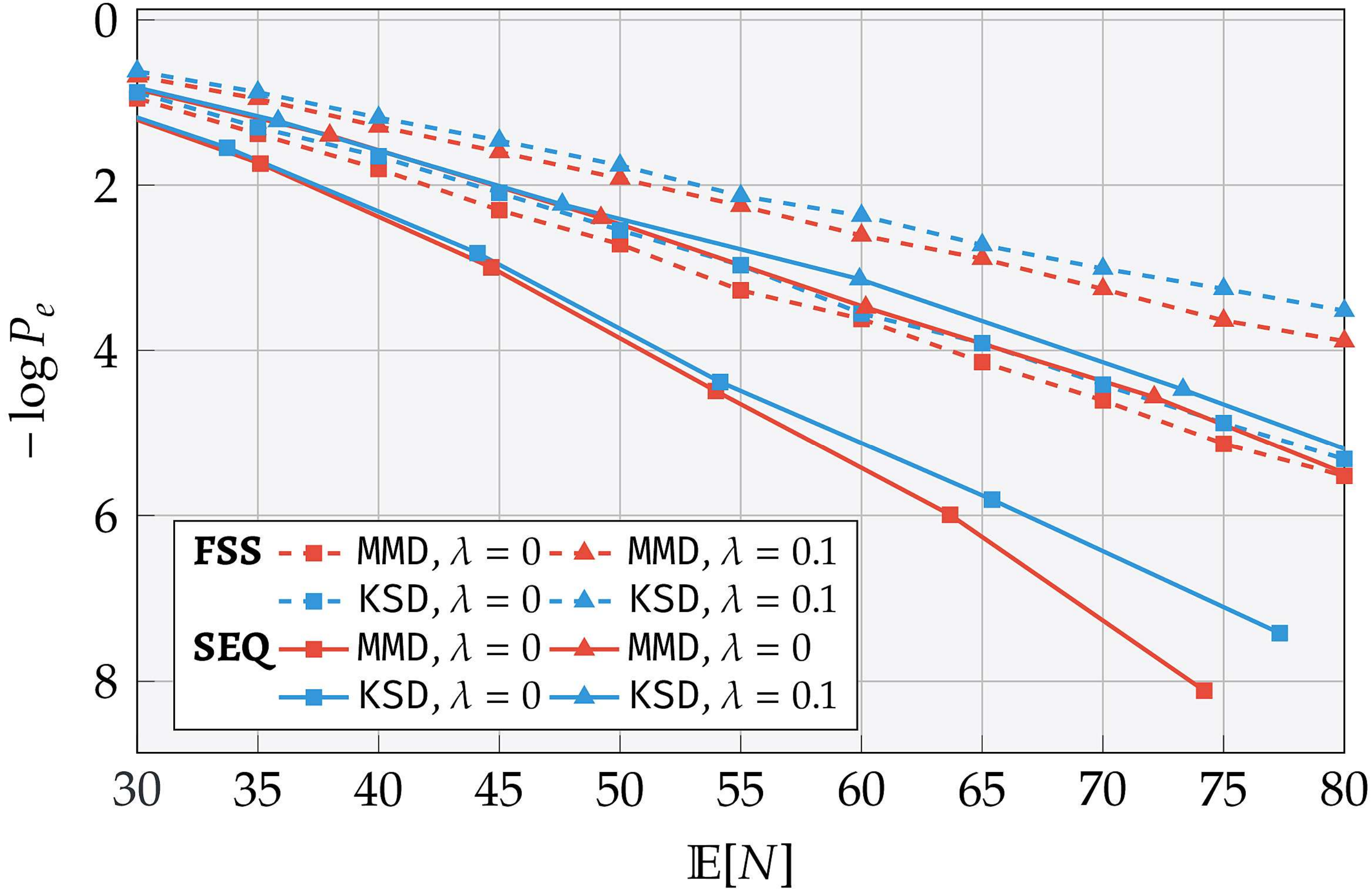




Simulation Setting

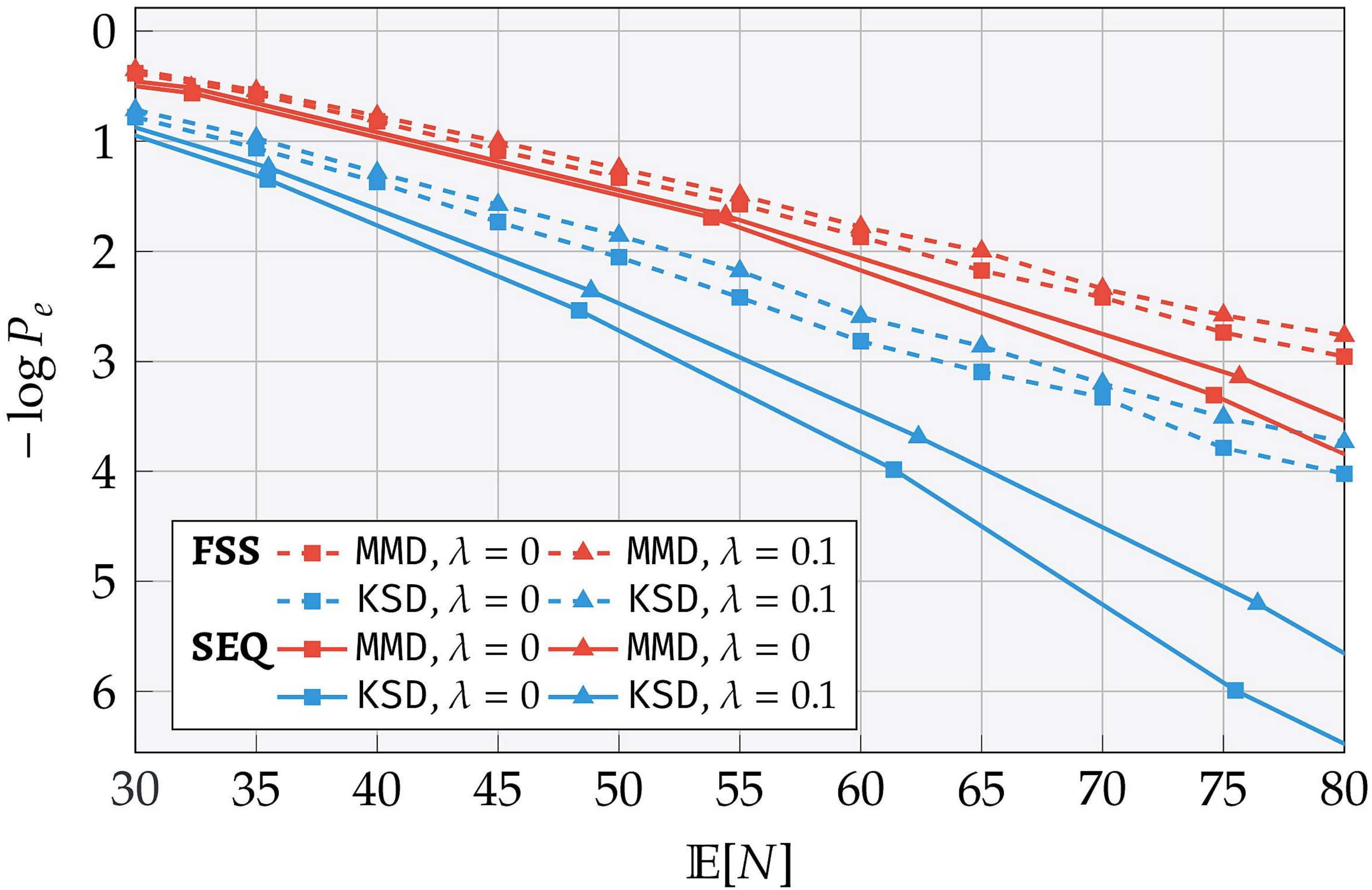
	Gaussian $\mathcal{N}(\mu, 1)$		Gamma $\Gamma(\mu, 1)$	
	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0$	$\lambda = 0.1$
P_1	$\{0\}$	$\{-0.1, -0.05, 0.0, 0.05, 0.1\}$	$\{1.0\}$	$\{0.9, 0.95, 1.0, 1.05, 1.1\}$
P_2	$\{1\}$	$\{0.9, 0.95, 1.0, 1.05, 1.1\}$	$\{3.5\}$	$\{3.4, 3.45, 3.5, 3.55, 3.6\}$
P_3	$\{2\}$	$\{1.9, 1.95, 2.0, 2.05, 2.1\}$	$\{6.0\}$	$\{5.9, 5.95, 6.0, 6.05, 6.1\}$
P_4	$\{3\}$	$\{2.9, 2.95, 3.0, 3.05, 3.1\}$	$\{8.5\}$	$\{8.4, 8.45, 8.5, 8.55, 8.6\}$
P_5	$\{4\}$	$\{3.9, 3.95, 4.0, 4.05, 4.1\}$	$\{11.0\}$	$\{10.9, 10.95, 11.0, 11.05, 11.1\}$

Results: Known number of clusters



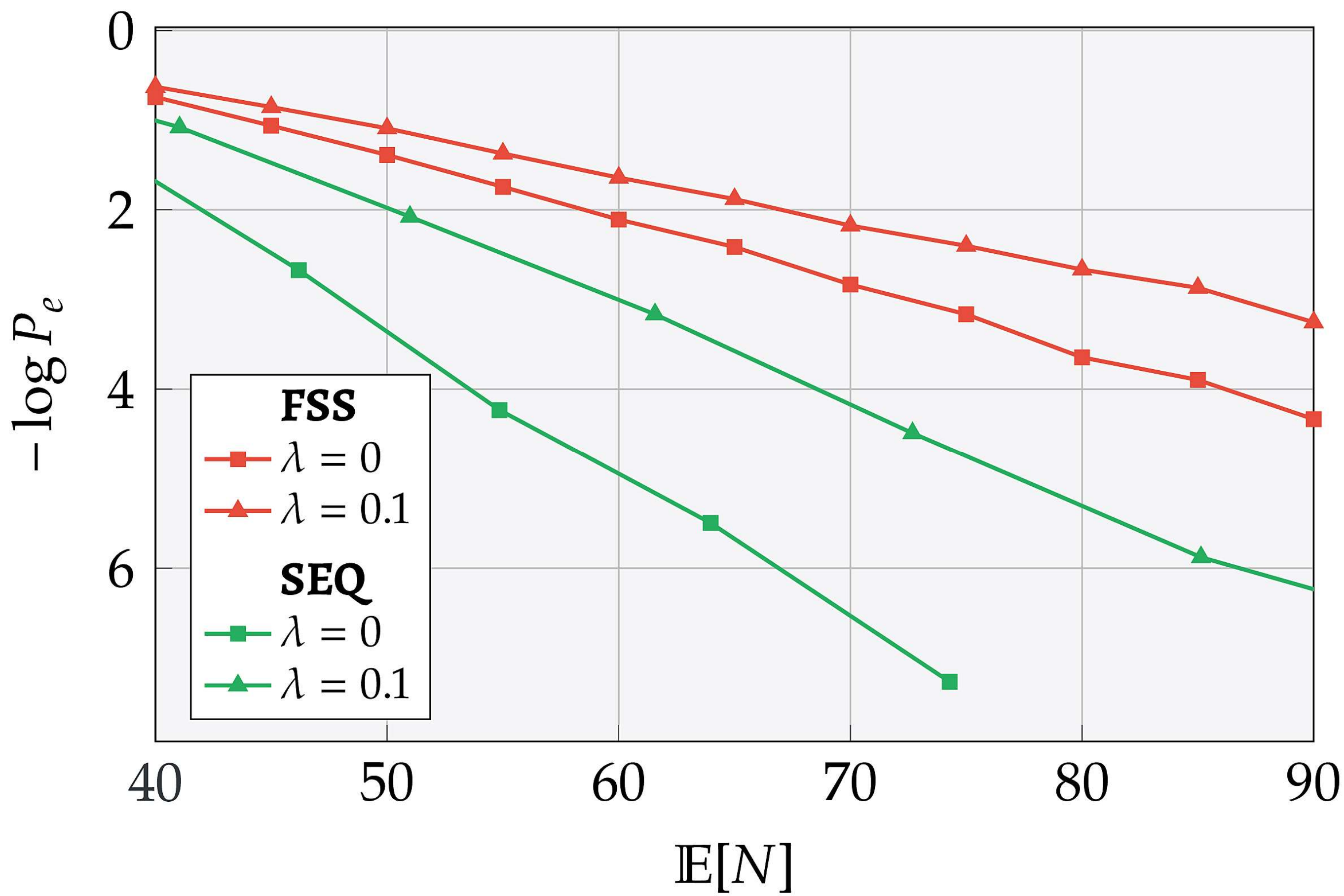
- Gaussian distributions case: MMD better than KSD
- Fewer samples required than FSS clustering on average

Results: Known number of clusters



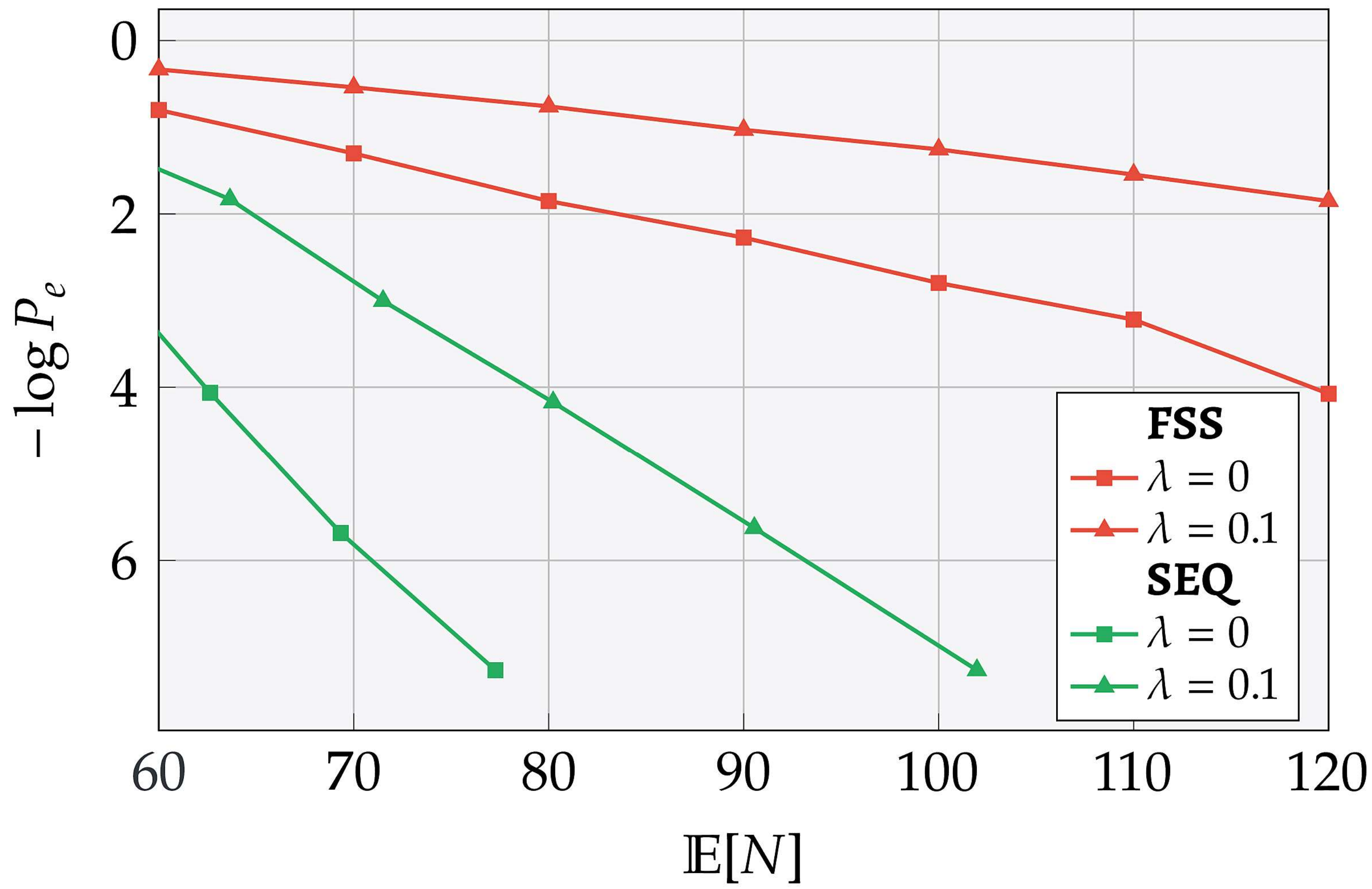
- Gamma distributions case: KSD better than MMD
- Fewer samples required than FSS clustering on average

Results: Unknown number of clusters



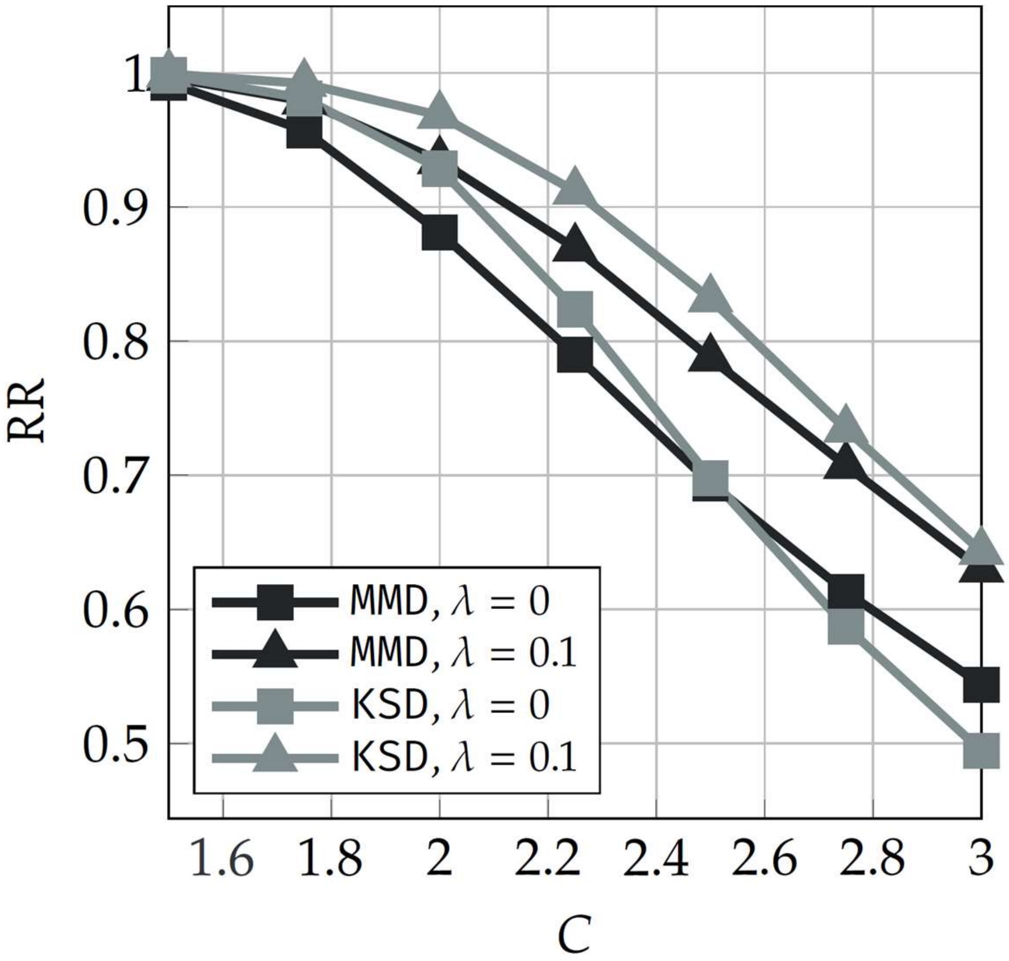
- Gaussian distributions case, K-MED + Merge
- Fewer samples required than FSS clustering on average

Results: Unknown number of clusters



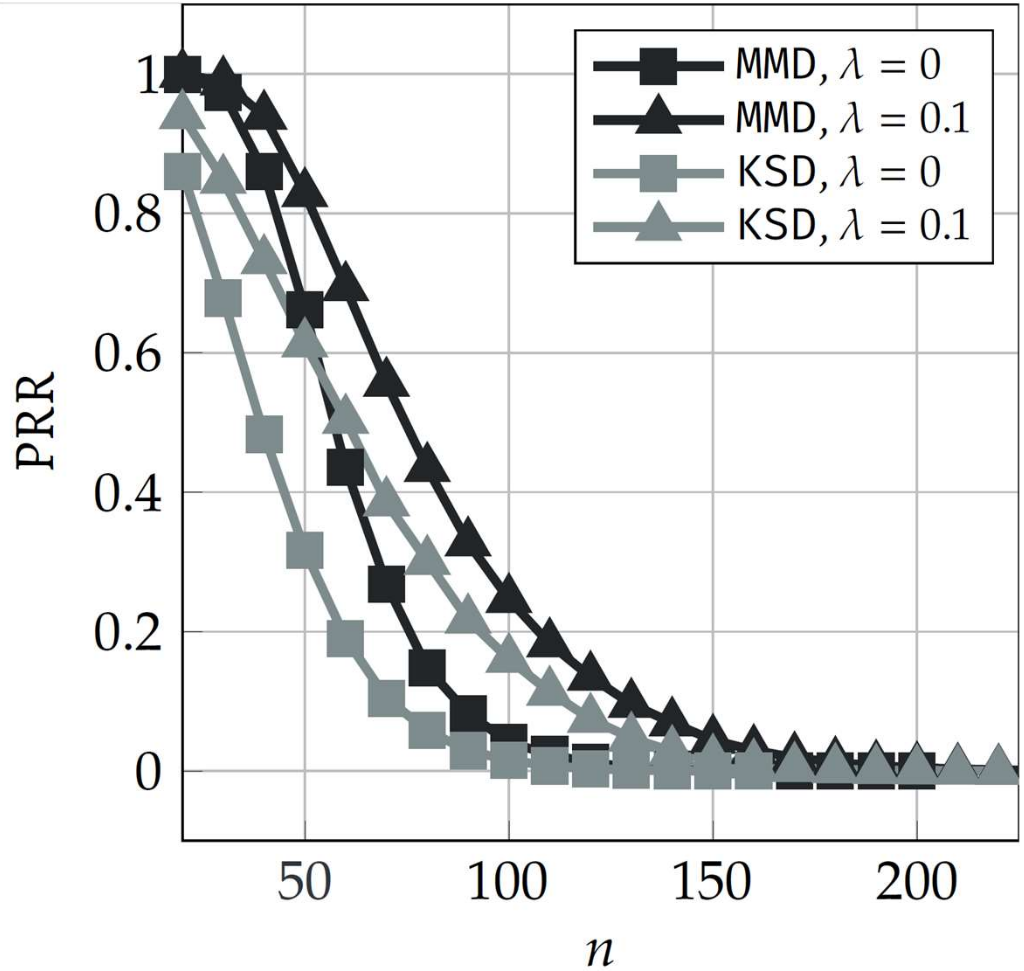
- Gaussian distributions case, K-MED + Split
- Fewer samples required than FSS clustering on average

Cluster initialization update



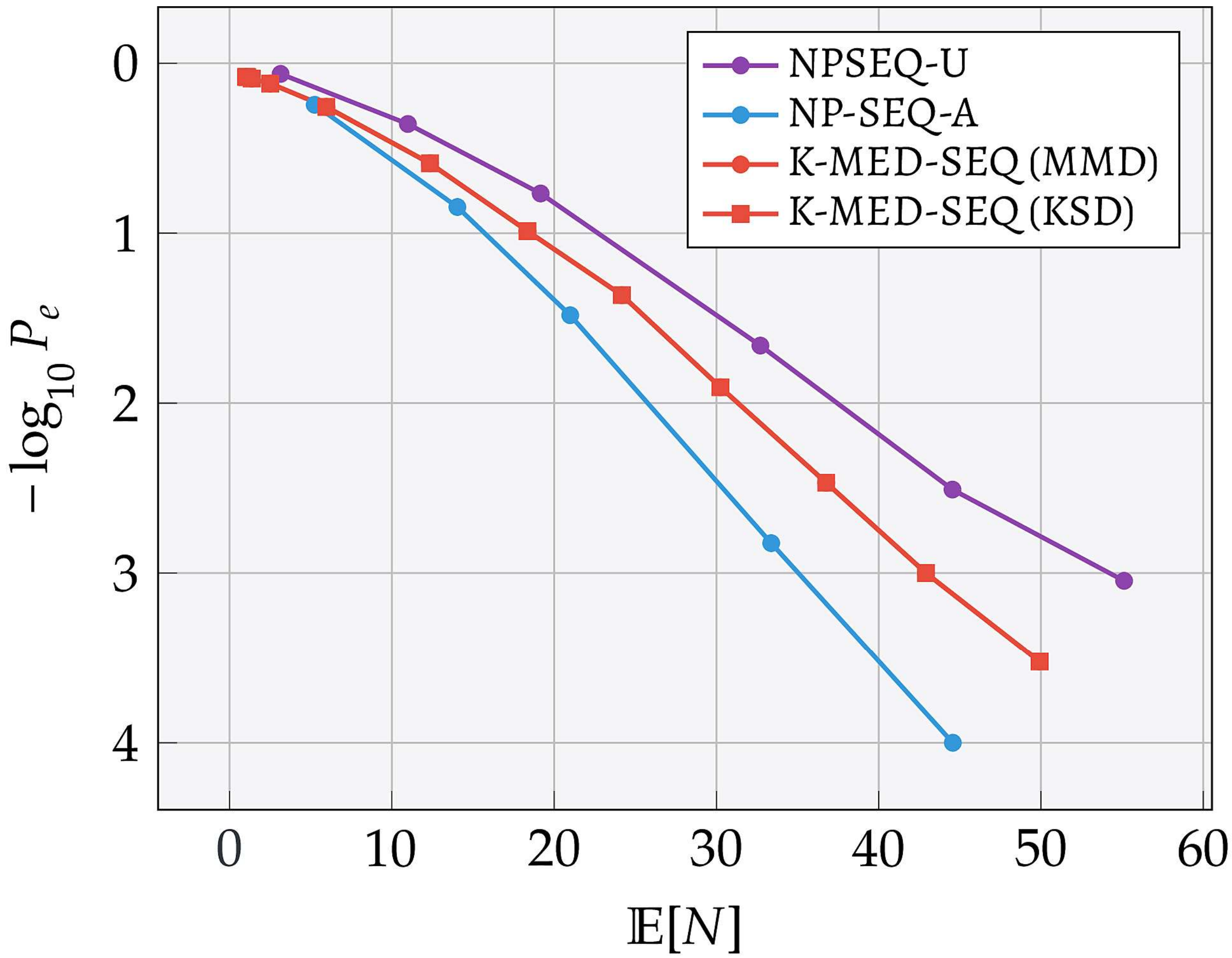
- Reuse cluster output from previous time as initialization
- Computational savings
- RR: Redo ratio

Cluster initialization update



- Fraction of realizations where re-initialization is done at time *n*

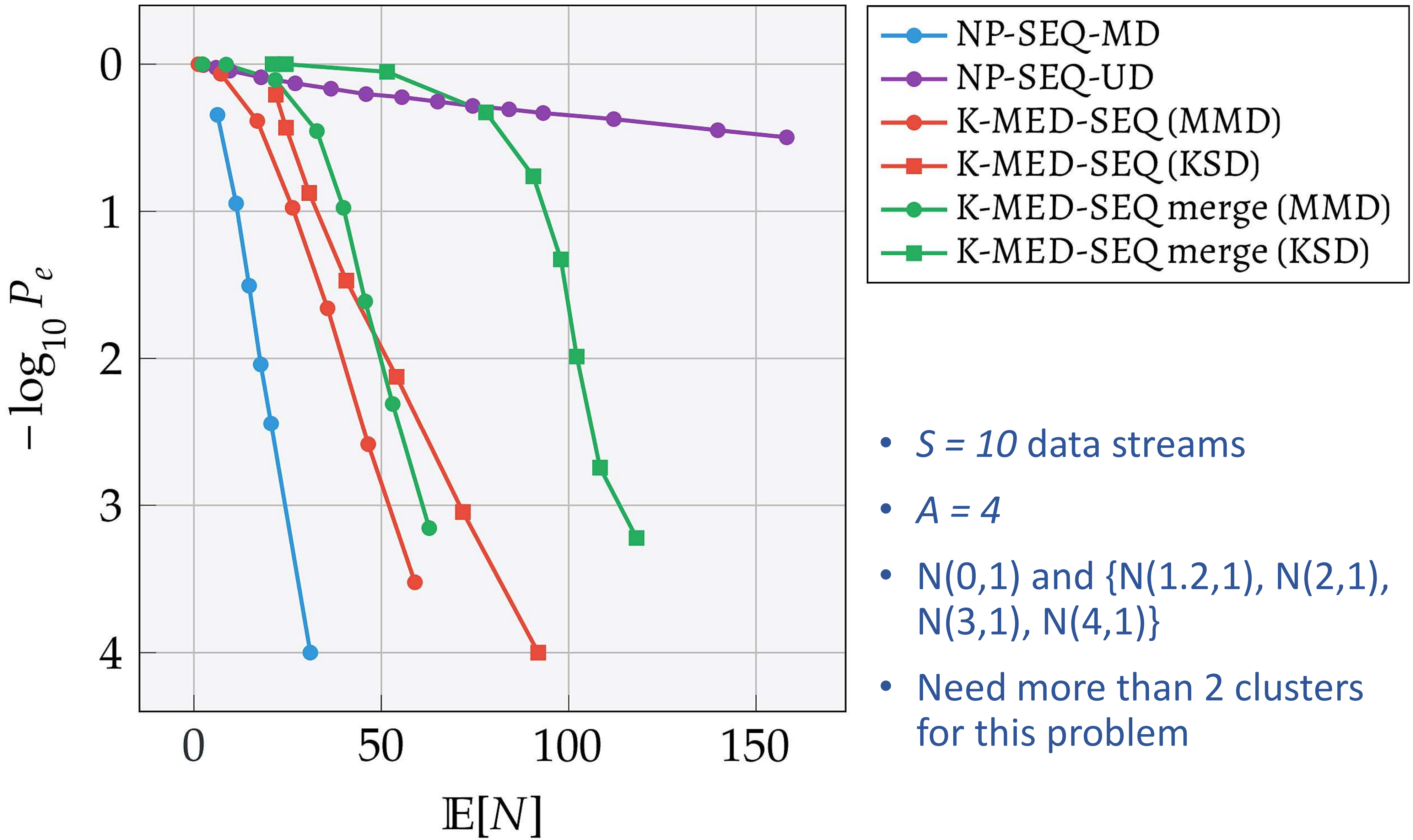
Special case: Multiple Anomalies



- $S = 5$ data streams
- $A = 2$
- N(0,1) and N(1.2,1)

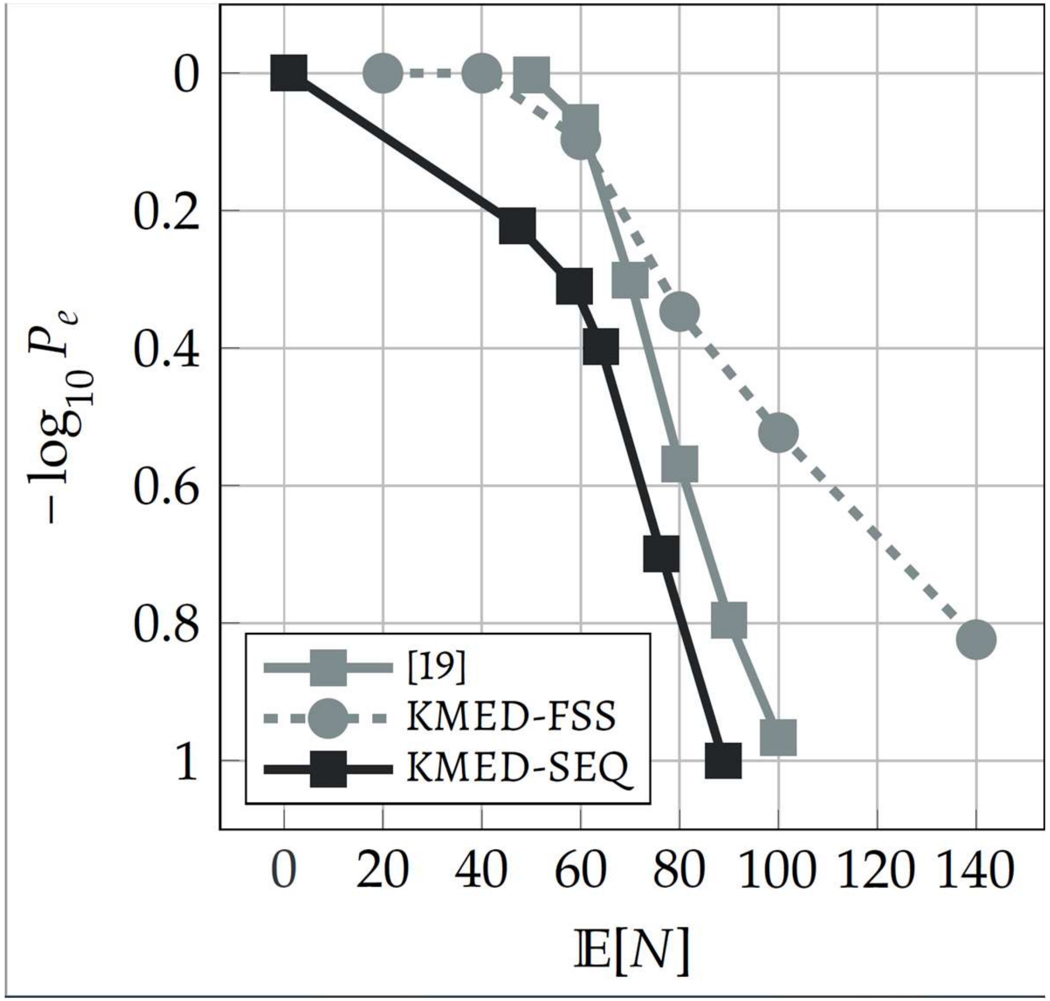
- NP-SEQ-A: Known A
- NP-SEQ-U: Unknown A

Special case: Multiple Distinct Anomalies



- *S = 10* data streams
- *A = 4*
- N(0,1) and {N(1.2,1), N(2,1), N(3,1), N(4,1)}
- Need more than 2 clusters for this problem

Discrete distributions



- MMD-based vs KL divergence-based

Y. Bu, S. Zou and V. V. Veeravalli, "Linear-Complexity Exponentially-Consistent Tests for Universal Outlying Sequence Detection," in IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2115-2128, 15 April15, 2019.

Linkage-based clustering

- Linkage-based hierarchical clustering algorithms
- Exponential consistency under the $d_L < d_H$ assumption
- Possible improvement
  - Maximum intra-cluster nearest neighbour distance instead of d_L

Tiexing Wang, Yang Liu, and Biao Chen, “On exponentially consistency of linkage-based hierarchical clustering algorithm using Kolmogrov-Smirnov distance,” in ICASSP 2020, pp. 3997–4001.

Summary

- Nonparametric sequential clustering of data streams
- Universal consistency