Introduction to Trajectory Clustering

By YONGLI ZHANG

Outline

1. **Problem Definition**
2. Clustering Methods for Trajectory data
3. Model-based Trajectory Clustering
4. Applications
5. Conclusions

1 Problem Definition

Trajectories:

- the path that a moving object follows through space as a function of time. -Wiki

- the sequence of spatial locations visited by the object, together with the time-stamps of such visits, form a *trajectory*. Namely, the whole history of a moving object is stored and available for analysis

1 Problem Definition

Trajectories:

- dynamical systems—a trajectory is the set of points in state space that are the future states resulting from a given initial state.

- In a discrete dynamical system—a trajectory is a set of isolated points in state space.

- In a continuous dynamical system—a trajectory is a curve in state space.

- In discrete mathematics—a trajectory is a sequence of values

$$(f^k(x))_{k\in\mathbb{N}}$$

calculated by the iterated application of a mapping f to an element x of its source.

1 Problem Definition

Trajectory clustering: Trajectories describe the movement behavior of objects, therefore clustering can be used to detect groups of objects that behaved in a similar way.

for example:

- by following similar paths (maybe in different time periods),
- by moving consistently together (i.e., keeping close to each other for long time intervals)
- by sharing other properties of movement.

Outline

1. Problem Definition

2. **Clustering Methods for Trajectory data**

3. Model-based Trajectory Clustering

4. Applications

5. Conclusions

2 Clustering Methods for trajectory data

2.1 (descriptive and generative) Model based clustering

Objective:

- Derive a global model capable of describing the whole dataset

- Each cluster modeled as a prototype function with some variability around the prototype, namely, it produced a descriptive and interpretable model for each cluster.

2 Clustering Methods for trajectory data

2.2 Distance-based clustering

Step1: Transform the complex data to features vectors-multidimensional vectors, each dimension represent 1 single characteristic of the object

Step2: Then use generic clustering algorithm, like K-means, to cluster them.

Problem: Most methods require all vectors to be of the same length

2 Clustering Methods for trajectory data

2.3 Density-based clustering and DBSCAN family

Objective: It uses a density threshold around each object to distinguish the interesting data items from the noise.

DBSCAN:

Step1: Visit the whole dataset and tag each object- core object, border object, noise.

(Noise means objects that are definitely outside any cluster)

Step2: The core objects that are close each other are joined in a cluster.

Density threshold defined by 2 parameters- maximum radius ***e*** around each object, A minimum number of objects within the interval, say ***MinPts***

2 Clustering Methods for trajectory data

2.3 Density-based clustering and DBSCAN family

- Density-based method strongly rely on an efficient implementation of the neighborhood query.
- How to choose a distance function?

—Temporal focusing method: cluster trajectories using all possible time intervals (time windows), evaluate the results and find the best clustering.

For example, two trajectories may be very different if the whole time interval is considered. However, if only a small sub-interval is considered, these trajectories may be found very similar.

2 Clustering Methods for trajectory data

2.4 Visual-aided approaches

Why introduce visualization techniques?

Automatic methods may discover interesting behavioral patterns with respect to the optimization function but it may happen that these patterns are trivial or wrong from the point of view of the phenomena/domain expert.

The visual analytics field tries to overcome this issue.

2 Clustering Methods for trajectory data

2.4 Visual-aided approaches

Advantages:

The analyst or domain expert can control the computational process by setting different input parameters, interpret the results and direct the algorithm towards the solution that better describes the underlying phenomena.

To be more specific, the analyst or domain expert can apply different distance functions that work with spatial, temporal, numerical or categorical variables on the spatio-temporal data to gain understanding of the underlying data in a stepwise manner.

2 Clustering Methods for trajectory data

2.5 Micro clustering methods

- the trajectories are represented as piece-wise segments, possibly with missing intervals.

- The proposed method try to determine a *close time interval*, i.e. a maximal time interval where all the trajectories are pair-wise close to each other.

- The similarity of trajectories is based on the amount of time in which trajectories are close.

- The mining problem is to find all the trajectory groups that are close within a given threshold.

Outline

1. Problem Definition
2. Clustering Methods for Trajectory data
3. **Model-based Trajectory Clustering**
4. Applications
5. Conclusions

3 Model-based Trajectory Clustering

Problem with Standard/traditional clustering algorithm (i.e. Kmeans):

Treat $\boldsymbol{y_j}$ trajectories as a set of n-dimensional vectors in an n-dimensional space and then use any of clustering methods which operate in vector spaces.

Not applicable:

- trajectories with different length,
- be measured at different time point,
- $\boldsymbol{y}$ may be multidimensional with no natural vector representation

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Standard mixture model clustering：

$$P(y_j|\theta) = \sum_{k}^{K} f_k(y_j|\theta_k) w_k$$

The generative model is a linear combination of component models

$\boldsymbol{K}$: the number of clusters, the paper assumes it is fixed

$\boldsymbol{w_k}$: The probability an individual assigned to cluster

$f_k(y_j|\theta_k)$:Given an individual belongs to cluster $\boldsymbol{k}$, this density function will generate the observed data $\boldsymbol{y_j}$ from individual $\boldsymbol{j}$

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Standard mixture model clustering：

$$P(y_j|\theta) = \sum_{k}^{K} f_k(y_j|\theta_k) w_k$$

If we observe y_j's, and we assume a particular functional form for the f_k components, then the problem becomes how to estimate the parameters: w_k and θ_k

—**EM (expectation maximization) algorithm** is used in this paper

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Experiment:

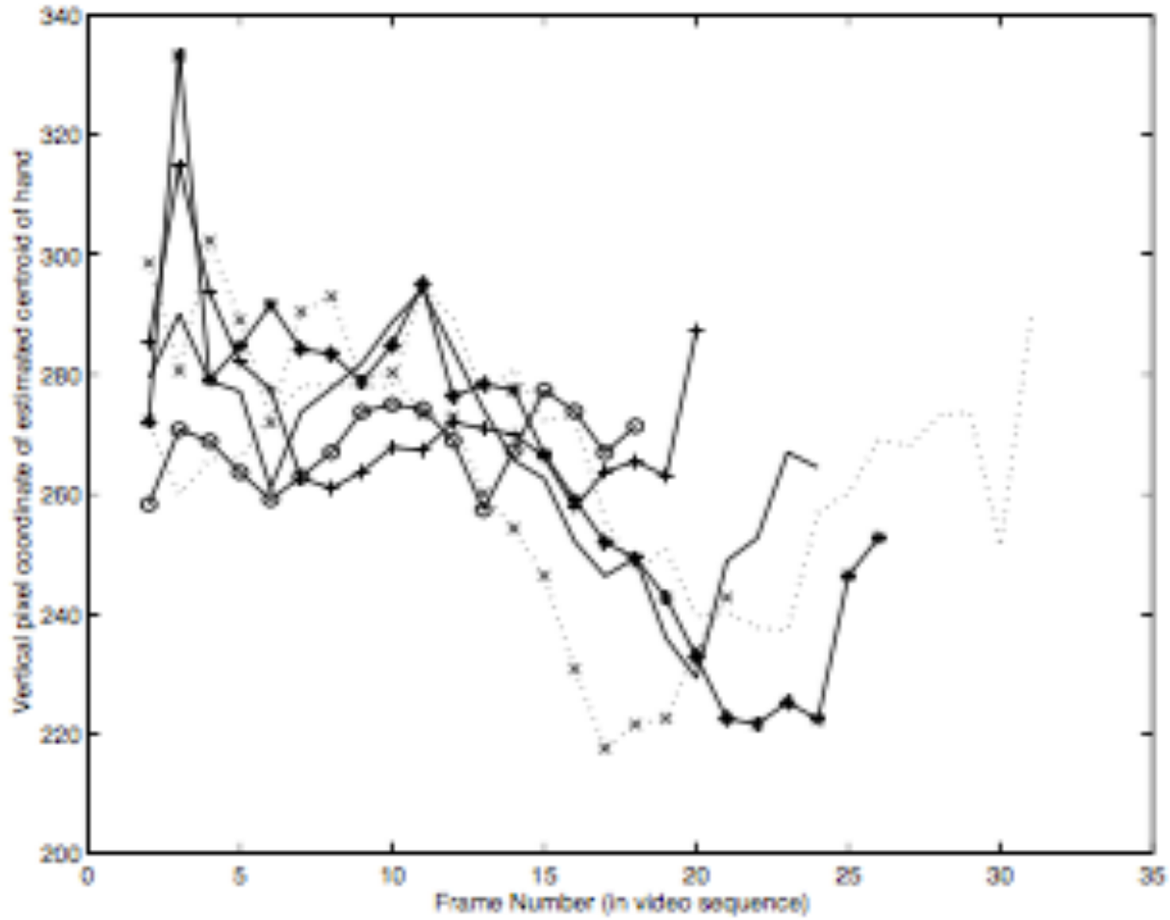


Figure 1: Trajectories of the estimated vertical position of a moving hand as a function of time, estimated from 6 different video sequences.

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Assume x,y 1-dimensional, we get standard regression relationship:

$$y = g_k(x) + e$$

Gaussian noise ***e****:* mean=*0,* stand deviation=$\sigma_{k.}$

$g_k(x)$ is a deterministic function of x

$f_k(y|x, \theta_k)$ *:* Given that y belongs to the **k**th group, has mean $g_k(x)$ and standard deviation $\sigma_{k.}$

y: vertical position of hand

x: time (t)

For simplicity of notation, assume ***e*** to be a constant

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Define the probability of a complete trajectory, given a particular model k:

$$P(y_j|x_j,\theta_k) = P(y_j(1),\ldots,y_j(n_j)|x_j(1),\ldots,x_j(n_j),\theta_k) = \prod_{i}^{n_j} f_k(y_j(i)|x_j(i),\theta_k) \quad (1)$$

y_j: trajectory of measurements for the jth individual

$y_j(i)$: ith measurement of y_j

x_j: measurement of y_j were taken at times x_j

θ_k is the set of parameters for component/cluster k

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Define cluster model for trajectories:

In practice for clustering, we don't know which component generate that trajectory, the conditional density of observed data $P(y_j|x_j)$ is a mixture density:

$$P(y_j|x_j,\theta) = \sum_k f_k(y_j|x_j,\theta_k)w_k \qquad (2)$$

$f_k(y_j|x_j,\theta_k)$: are the mixture models,

$\boldsymbol{w_k}$ are the mixing weights

$\boldsymbol{\theta_k}$ is the set of parameters for component/cluster k

$\boldsymbol{y_j}$: jth trajectory

$\boldsymbol{k}$: Cluster index-kth cluster

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Define cluster model for trajectories:

Combining (1) and (2), we get full joint density:

$$P(Y|X,\theta) = \prod_{j}^{M} \sum_{k}^{K} w_k \prod_{i}^{n_j} f_k(y_j(i)|x_j(i),\theta_k). \qquad (3)$$

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Define cluster model for trajectories:

The log-likelihood of the parameter $\boldsymbol{\theta}$ given the data set S can be directly defined from Eq. (3)

$$\mathcal{L}(\theta|\mathcal{S}) = \sum_{j}^{M} \log \sum_{k}^{K} w_k \prod_{i}^{n_j} f_k(y_j(i)|x_j(i), \theta_k) \qquad (4)$$

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Next, using EM algorithm to pull the mixture components out of the joint density.

Hidden data problem- the group membership of each trajectory is unknown

EM algorithm—a sketch:

estimate the hidden data—>work out the answer—>then re-estimate the hidden data again using the current answers we just computed—> repented until some stabilization occurs

The EM framework gives a consistent way to estimate the hidden data so that $\mathcal{L}(\theta|\mathcal{S})$ is guaranteed to never decrease

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Experiment:

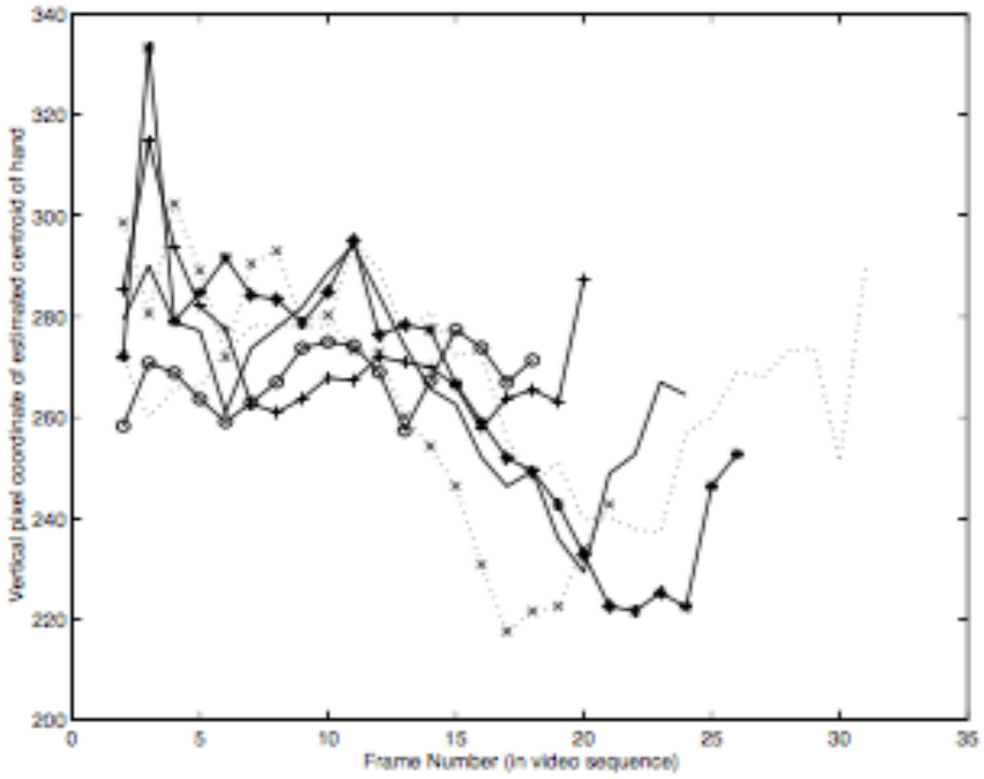


Figure 1: Trajectories of the estimated vertical position of a moving hand as a function of time, estimated from 6 different video sequences.

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Result:

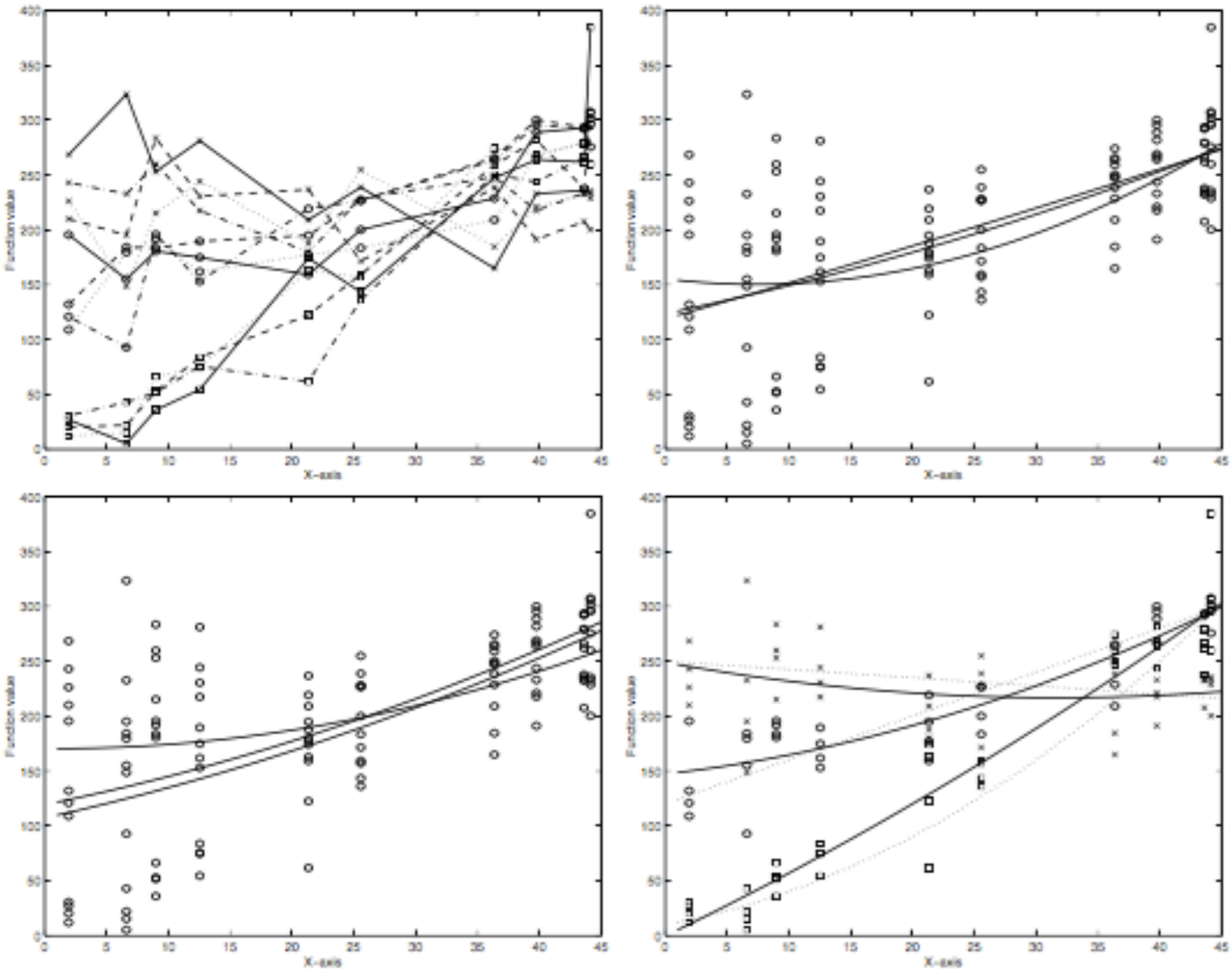


Figure 2: Trace of the EM algorithm as applied to a linear regression mixture model at various iterations. The upper left plot shows some of the original trajectories, the upper right shows the initial locations of the 3 cluster trajectories for EM, lower left shows the locations after 1 iteration of EM, and lower right shows the cluster locations (solid) after

https://en.wikipedia.org/wiki/Polynomial_regression

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Explanation:

Data set sampled from 3 underlying polynomial (3 clusters):

$y=120+x$; $y=10+2x+0.1x^2$; $y=250-0.75x$

From: $y = a0 + a1\ x + e$

to $y = a0 + a1x + a2x^2 + e$

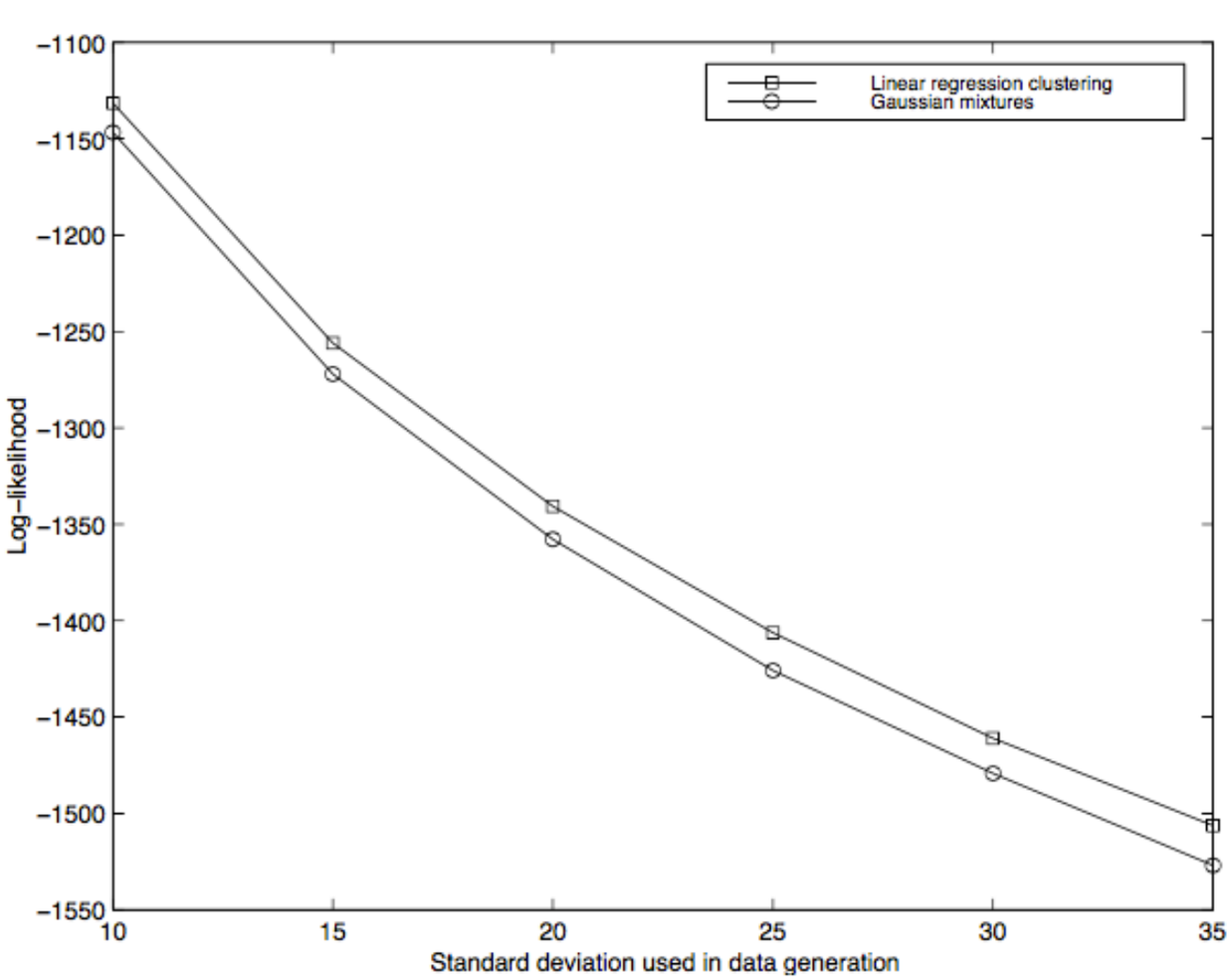
In this model, when the temperature is increased from x to x + 1 units, the expected yield changes by a1 + 2a2x. The fact that the change in yield depends on x is what makes the relationship nonlinear (this must not be confused with saying that this is nonlinear regression; on the contrary, this is still a case of linear regression).

— https://en.wikipedia.org/wiki/Polynomial_regression

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Testing



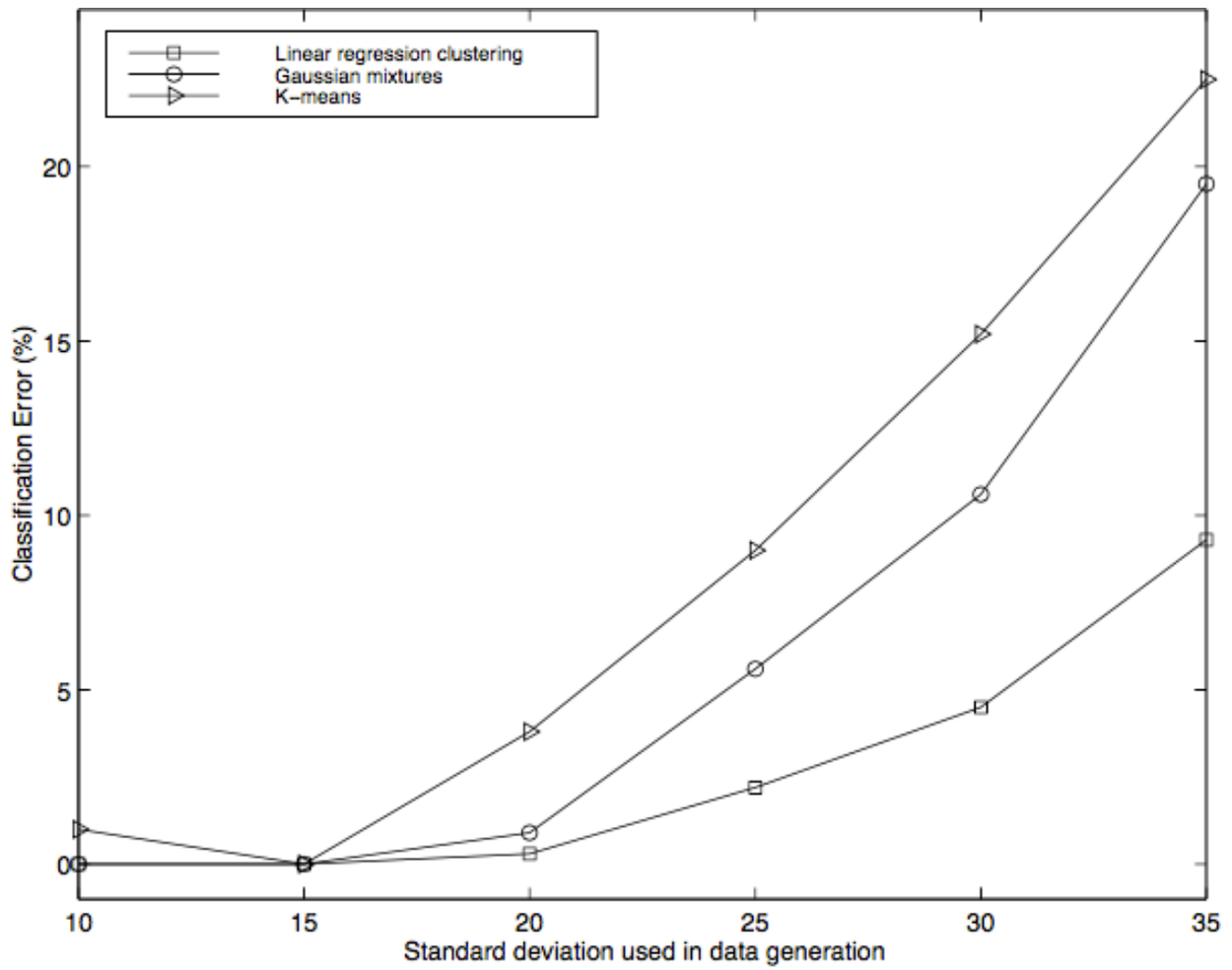


Figure 3: Mean log-likelihood and classification error rate performance on test data as the noise level increases from $\sigma = 10$ to $\sigma = 35$. $(l = 10, n = 15)$.

3 Model-based Trajectory Clustering

—Algorithm: Mixtures of Regression Model

Summary:

-Based on a principle method for probabilistic modeling of a set of trajectories as individual sequences of points generated from finite mixture model consisting of regression model components.

-Unsupervised learning is carried out using maximum likelihood principles

-EM algorithm for hidden data problem, (i.e. cluster membership)

Outline

1. Problem Definition
2. Clustering Methods for Trajectory data
3. Model-based Trajectory Clustering
4. **Applications**
5. Conclusions

4 Application

4.1 Environmental Data

i.e. Cyclones detection

- We can take the track of cyclones as trajectories.

- Apply trajectory clustering, to locate them and track where they go

4 Application

4.2 Flocks and convoy

In some application, there is a need in discovering group of objects that move together during a given period of time.

For example, migrating animals, flocks of birds or convoys of vehicles, march-detecting military activity

4 Application

4.3 Movement data

A specific example:

The track of commuters—trajectories

How to trace? —GPS based devices

If there are groups of commuters within a city that move from one area of the city to another one within a particular time frame.

This kind of information and analysis can give meaningful hints to city planners in order to avoid regular traffic jams.

Outline

1. Problem Definition
2. Clustering Methods for Trajectory data
3. Model-based Trajectory Clustering
4. Applications
5. **Conclusions**

5 Conclusion

- Trajectories represent the most complex and promising (from a knowledge extraction viewpoint) form of data among those based on point-wise information.

- Clustering is one of the general approaches to a descriptive modeling of a large amount of data, allowing the analyst to focus on a higher level representation of the data.

- Trajectory clustering has a lot of meaningful real world applications.

References

1. Kisilevich, Slava, et al. Spatio-temporal clustering. Springer US, 2009.

2. Gaffney, Scott, and Padhraic Smyth. "Trajectory clustering with mixtures of regression models." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.