Extension of fixed point clustering: A cluster criterion

A. Hutt[1], **F. Kruggel**
Max Planck-Institute of Cognitive Neuroscience,
Stephanstrasse 1a, 04103 Leipzig, Germany

The present report extends the method of fixed point clustering [7] by introducing an indirect criterion for the number of clusters. The derived probability function allows an objective distinction of clustered data and data in between clusters. Applications on simulated data illustrate the clustering method and the probability function.
PACS numbers: 02.60.-x, 45.10.-b

1 Introduction

The dynamics of spatially extended systems can be measured by sets of multi-detector arrays. Most spatio-temporal analysis methods fitting multi-dimensional dynamical models [1, 2, 3, 4, 5, 6] consider data over the full time range. In [7], a method was described for partitioning spatio-temporal signals into time segments, in which the signal can be modeled by deterministic ordinary differential equations near fixed points. Each dynamical system is determined by a nonlinear spatio-temporal analysis [6]. The earlier proposed algorithm in [7] works with an arbitrary number of clusters k. Since results are dependant on k, an objective criterion for the number of clusters is necessary. The present report introduces a different segmentation algorithm and aims to derive an indirect criterion for the number of temporal segments.
In the present report, we use the K-Means algorithm [9, 10] for segmenting data, which addresses each data point to a single cluster. Since K-Means works with an arbitrary number of clusters and this number is crucial to clustering results, we derive a probability function representing the degree of membership of a data point at time t to a cluster. It addresses data to clusters or transition parts between clusters and hence determines the number of necessary clusters. Applications to simulated non-stationary data illustrates the probability measure.

2 Fixed Point Clustering (FPC)

In the following, a signal trajectory is assumed as compound of a sequence of segments governed by saddle point dynamics. Under the hypothesis, that these segments comprise the main functionality of the underlying system, we

[1] e-mail:hutt@cns.mpg.de

aim to extract them from the signal. Trajectories approach saddle points along their stable manifolds whereas they leave the vicinity of the fixed points along the unstable manifolds. The signal points accumulate close to the fixed points if the signal is sampled at a constant rate. This accumulation may also be regarded as a point cluster in data space. Subsequently, stable manifolds in multi-dimensional signals lead to point clusters and their detection can be treated as a recognition problem in data space [7].

The clustering algorithm

A N-dimensional spatio-temporal signal can be described by a data vector $\mathbf{q}(t) \in \Re^N$, where the component $q_j(t_i)$ represents a data point at time i and detection channel j. The clustering algorithm aims at cluster centers $\{\mathbf{k}_k\}$, whose mean Euclidean distance to a set of data points $\mathbf{q}(t_i)$ is minimal. The presented implementation follows Moody et al.[10] and is sketched in Fig. 1.

Cluster centers $\mathbf{k}^0$ are initialized at random locations in the data and their Euclidean distances to each data point are calculated. K-Means defines memberships of data points to a cluster by the smallest Euclidean distance to its center. Thus, data are segmented into k clusters and new cluster centers $\mathbf{k}^1$ are calculated as means of clustered data points. Distances between data points and centers $\mathbf{k}^n$ are re-estimated until a convergence condition is fulfilled. This criterion can be set either as a upper Euclidean distance limit between sequential cluster centers $\mathbf{k}^n, \mathbf{k}^{n+1}$ or as number of iterations. We choose to limit the number of iterations to 25.

Simulated spatio-temporal data and results

Now, a low-dimensional simulated signal $\mathbf{A}(t)$ is introduced describing amplitudes of multi-dimensional spatial patterns $\mathbf{v}_i$ by

$$\mathbf{q}(t) = \sum_i A_i(t)\mathbf{v}_i.$$

This superposition describes a spatio-temporal signal $\mathbf{q}(t)$.

The dataset $\mathbf{A}(t)$ is generated by

$$\begin{aligned} \dot{A}_1 &= \epsilon A_1 - A_1[A_1^2 + (2+b)A_2^2 + (2-b)A_3^2] + \Gamma(t) \\ \dot{A}_2 &= \epsilon A_2 - A_2[A_2^2 + (2+b)A_3^2 + (2-b)A_1^1] + \Gamma(t) \\ \dot{A}_3 &= \epsilon A_3 - A_3[A_3^2 + (2+b)A_1^2 + (2-b)A_2^2] + \Gamma(t). \end{aligned} \tag{1}$$

Parameters are set to $\epsilon = 1$, $b = 2$ and $\Gamma(t) \in [-0.05...0.05]$ represents additive noise following a uniform deviate. Equations 1 describe the convection onset of a Rayleigh-Benard-experiment in the presence of rotation[7, 11, 12].

A 3-dimensional trajectory $\mathbf{A}(t)$ is calculated by 2200 integration steps with the initial condition $\mathbf{A}(t = 0) = (0.03, 0.2, 0.8)$, see Fig. 2. The trajectory passes the

saddle points $\mathbf{A}_3^0 = (0,0,1), \mathbf{A}_1^0 = (1,0,0)$ and $\mathbf{A}_2^0 = (0,1,0)$ in this sequence, and then returns to $\mathbf{A}_3^0$.
The K-Means algorithm is applied on the simulated data for different number of clusters $k = 2,..,7$. In Fig. 3, the Euclidean distances from each data point to the determined cluster centers are plotted in temporal sequence for each k. When a trajectory approaches or moves from a cluster center, its Euclidean distance to the center decreases resp. increases. These changes can be observed in Fig. 3. For fixed number of clusters, each data point is considered to be member of a cluster, whose center is closest to the data point.
Comparing obtained clustering results for different k, clustered time windows [0;∼350], [∼350;∼1050], [∼1160;∼1610] and [∼1740;2200] are recognized, which borders remain similar for different k.

3 The cluster criterion

Although there might be only a limited number of clusters $k_d < k$ in the data, K-Means determines k clusters also including void clusters. In Fig. 3, small clustered time windows are visible, whose occurences and temporal widths strongly depend on k. They are considered as invalid clusters. Conversely, a first qualitative criterion for valid clusters may be formulated as:

- cluster widths and locations in time remain independent of k and
- the Euclidean distances of clustered data points to centers is obviously smaller then the Euclidean distances of points to the next nearest cluster center and
- the width of the clustered time window is not too small.

Although these criteria are rather heuristic than formal, they proved to be useful in practice [7]. Now, we try to evolve them quantitatively. The first item can be formulated as a sum over all clustering results: valid contributions are additive if they occur for all k, others vanish in the sum as small contributions. Thus the contribution of a valid cluster to the sum should be large, not reliable clusters should contribute with small values. A good quantity for these contributions is the area between the curves of the signal-nearest cluster-distance and signal-next cluster-distance. This definition allows the analytical formulation of the second item and is outlined in Fig. 4. Each data point t_i obtains an index corresponding to the cluster j it is member of. The index is equal the relativ area $\frac{A_j^{(k)}(t_i)}{T\sum_j A_j^{(k)}}$, where T denotes the number of data points. By summing up the indices over K cluster realizations for every data point, a degree of membership

$M(t)$ for every data point is obtained:

$$\begin{aligned} A_r^{(k)}(t_i) &= \frac{A_j^{(k)}(t_i)}{\sum_j A_j^{(k)}} \\ M(t_i) &= \frac{\sum_{k=2}^{K\leq T} A_r^{(k)}(t_i)}{\sum_{i=1}^{T}\sum_{k=2}^{K\leq T} A_r^{(k)}(t_i)}. \end{aligned}$$

$M(t)$ represents a probability, that a data point at time t belongs to a cluster. The application to simulated data with $K = 30$ leads to results of $M(t)$ shown in Fig. 5. Clustered time windows can be recognized as regions of high values of $M(t)$. Four plateaus of $M(t)$ are recognized at [0;250], [340;1010], [1180;1560] and [1750;2200] with borders at drastic value changes. Regions between these plateaus are considered as non-functional transitions parts. Comparing time windows in the original signal(Fig. 2) and detected clustered time windows in Fig. 3 and Fig. 5, good accordance of time windows near fixed points and cluster results are recognized.

4 Conclusion

The present brief report extends the fixed point clustering method by introducing a probability function $M(t)$. High values of $M(t)$ indicate clustered data points. Fixed point clustering relates temporal dynamics near fixed points showing attractive and repelling properties with clusters in dataspace. By the presented extension, regions in data space near such fixed points can be determined independant of the number of clusters. Applications to spatio-temporal signals in hydrodynamics, metereology or brain science [13, 14] are possible.

References

[1] M. Kirby, Physica D **57**, 466-475 (1992).

[2] C. Uhl, R. Friedrich, H. Haken, Phys. Rev. E **51**, 5, 3890-3900 (1995).

[3] R. Friedrich and C. Uhl, Physica D **98**, 171-182 (1996).

[4] J.O. Ramsay, B.W. Silverman, *Functional Data Analysis*, Springer, New York (1997).

[5] S. Siegert and R. Friedrich, J. Peinke, Phys. Lett. A **243**, 275-280 (1998).

[6] A. Hutt, C. Uhl and R. Friedrich, Phys. Rev. E **60**, 2, 1350-1358(1999).

[7] A. Hutt, M. Svensen, F. Kruggel, R. Friedrich, Phys. Rev. E **61**, 5, R4691-R4693 (2000).

[8] C. Uhl, F. Kruggel, B. Opitz, D.Y. von Cramon, Hum. Brain Map. **6**, 137-149 (1998).

[9] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge (1997).

[10] J. Moody, C.J. Darken, Neur. Computation **1**, 2, 281-294 (1989).

[11] G. Küppers and D. Lortz, J. Fluid Mech. **35**, 609 (1969).

[12] F.H. Busse and K.E. Heikes, Science **208**, 173 (1980).

[13] A. Hutt, F. Kruggel, B. Opitz, Psychophysiology, submitted (2000).

[14] A. Hutt, F. Kruggel, R. Friedrich, Verhandl. DPG (**VI**)35, DY46.38, Physik-Verlag, Weinheim (2000).

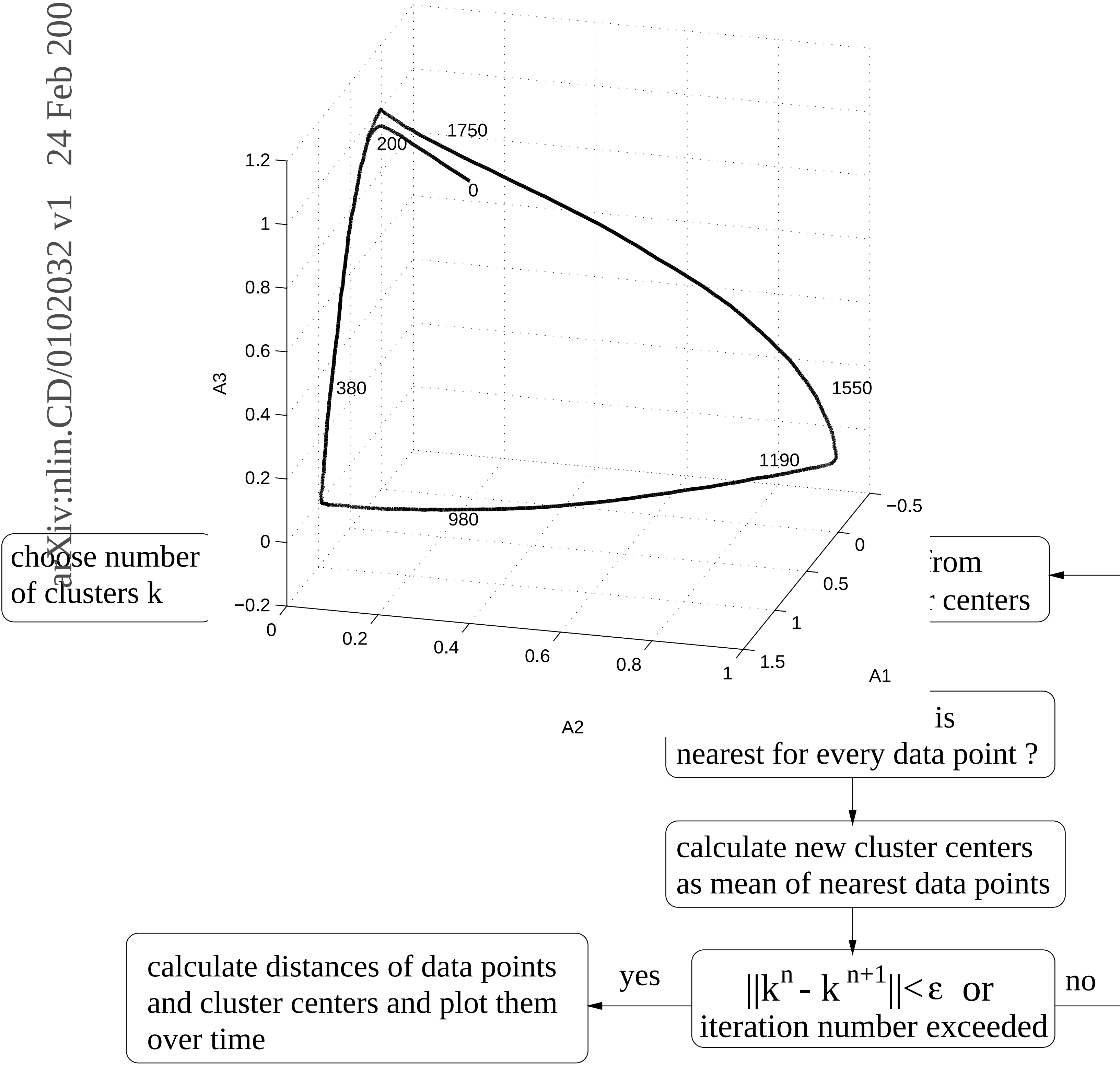
Figure 1: The implementation steps of the K-Means algorithm.

Figure 2: Trajectory of the 3-dimensional signal $\mathbf{A}(t)$. It starts near a saddle point and passes two others, before it returns to the initial saddle point. The numbers denote the timesteps of the trajectory at their locations.

Figure 3: Cluster results for $k = 2, .., 7$. The Euclidean distances between data points and detected clusters are shown.

Figure 4: Sketch to illustrate the introduced criterion of a clusters validity. Area A_j between two distance curves indexes the data points, which belong to cluster j. Large areas indicates at a high degree of membership.

Figure 5: Degree of membership $M(t)$ for every data point as a sum of $K = 30$ clustering results. Plateaus denote valid clusters, which are delimited by rapid changes.



choose number of clusters k

from

centers

is

nearest for every data point ?

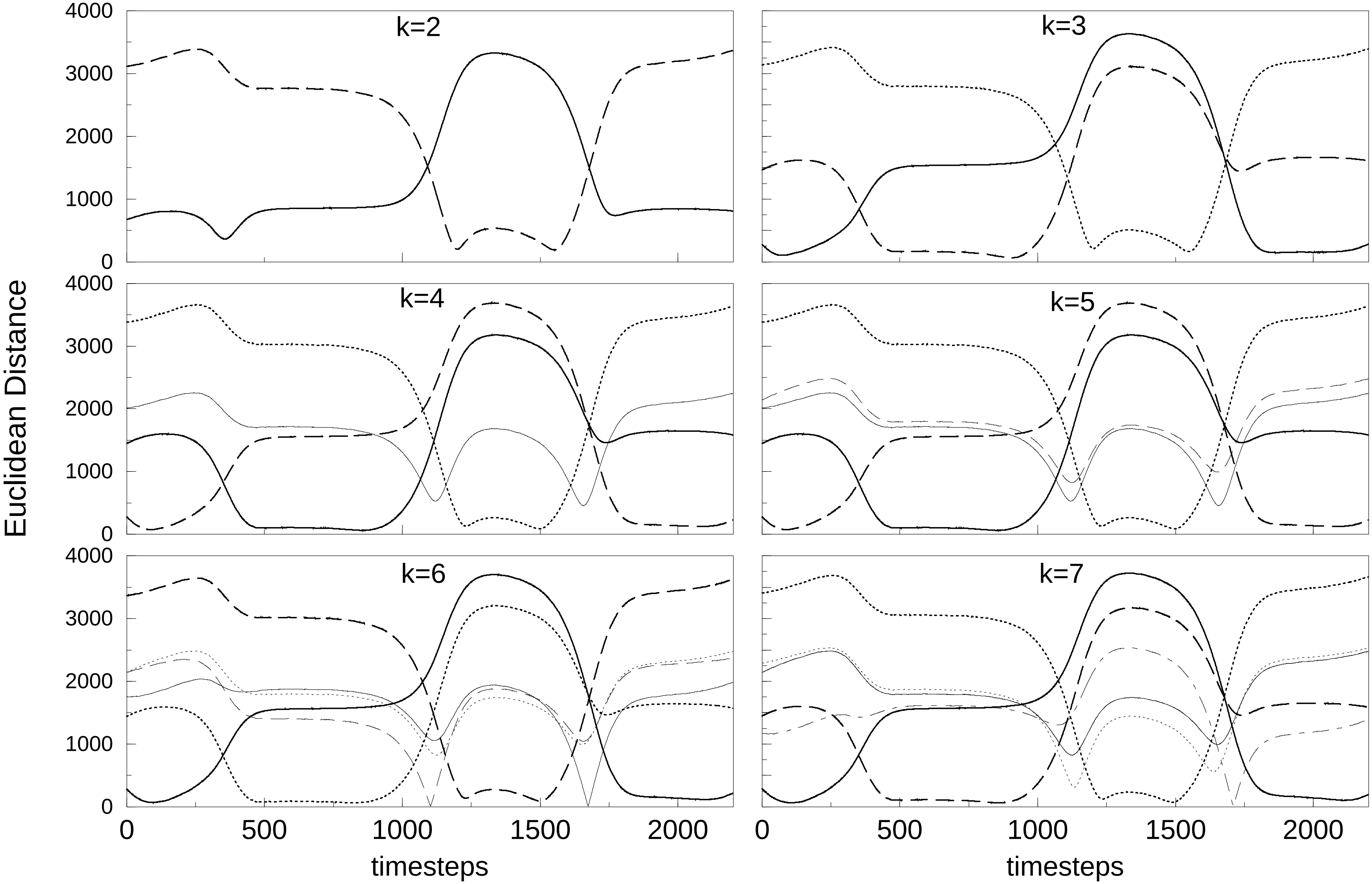
calculate new cluster centers as mean of nearest data points

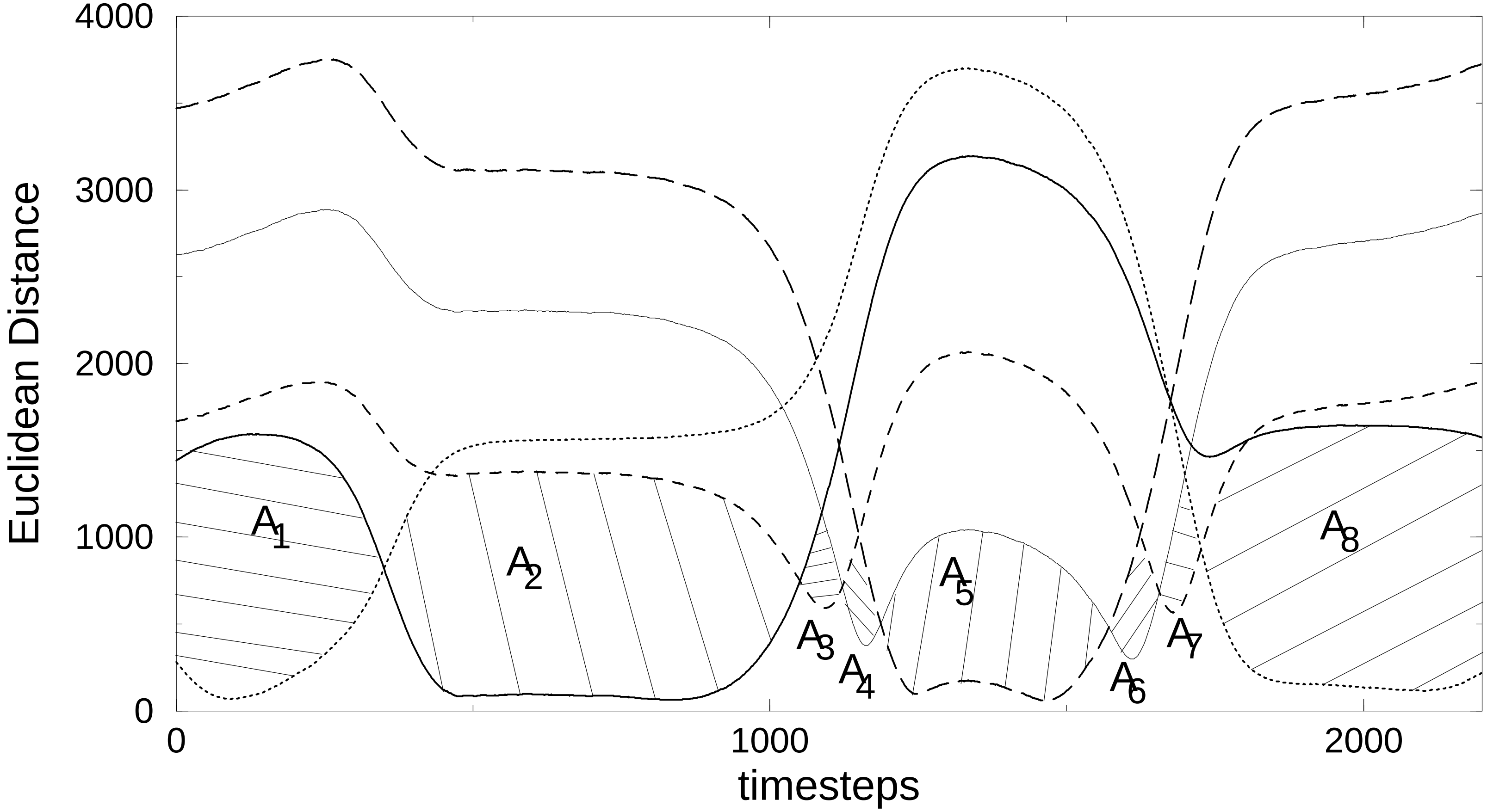
$\|k^n - k^{n+1}\| < \varepsilon$ or iteration number exceeded

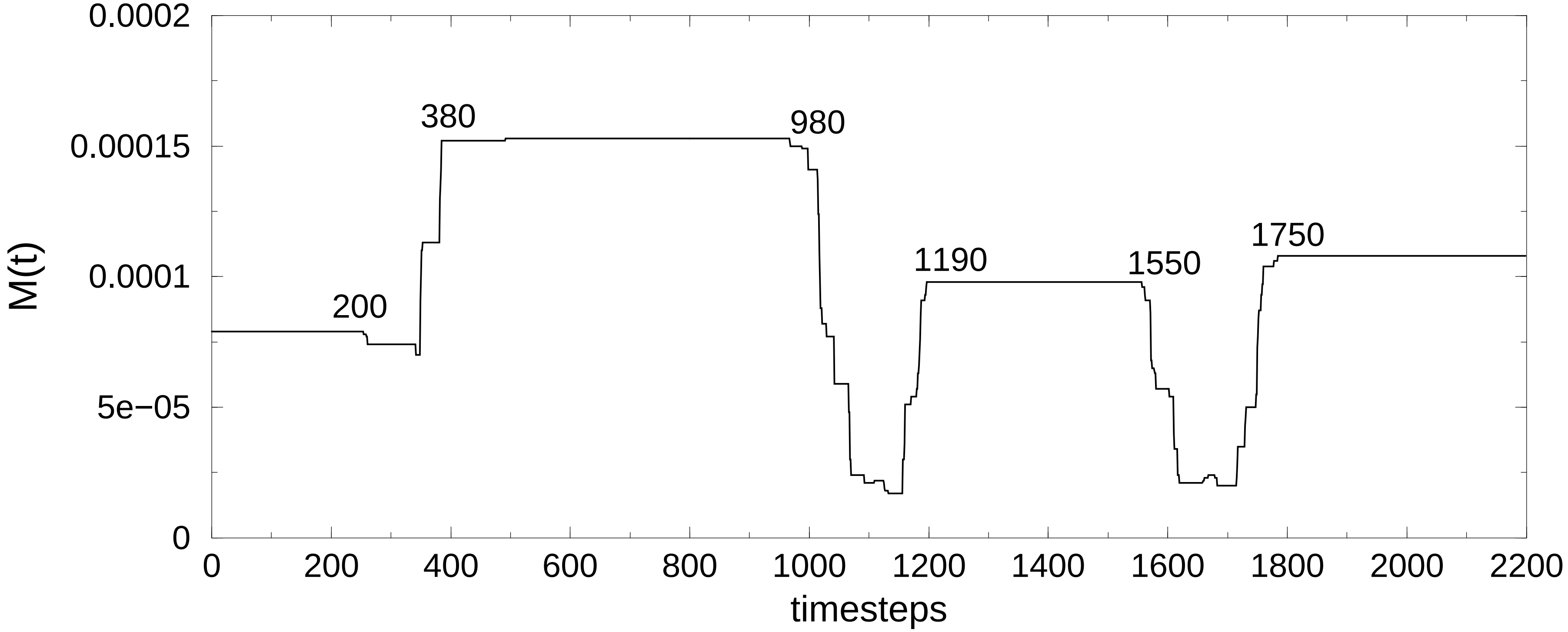
yes

no

calculate distances of data points and cluster centers and plot them over time


k=2
k=3
k=4
k=5
k=6
k=7
Euclidean Distance
0
1000
2000
3000
4000
0
500
1000
1500
2000
timesteps


Euclidean Distance
4000
3000
2000
1000
0
0
1000
2000
timesteps
A_1
A_2
A_3
A_4
A_5
A_6
A_7
A_8


0.0002
0.00015
0.0001
5e−05
0
M(t)
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
timesteps
200
380
980
1190
1550
1750