Machine Learning in Healthcare

Lab 2:
Point Processes

Jorge Chamorro Pedrosa – 100496527
Juan José Jiménez De Juan – 100496468
Mario Fernández Busto – 100496459

1. Initial clustering

1.1 Data processing

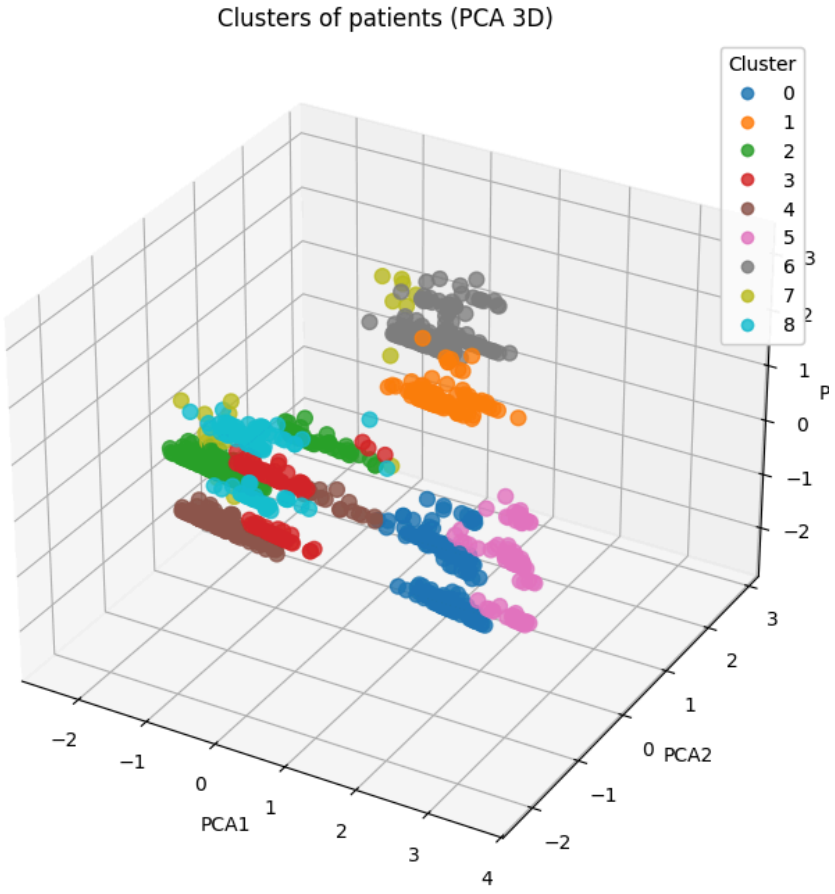
We applied a log-transform to F1 to reduce the impact of very large values. Then, we scaled all features using StandardScaler so that each one had a mean of 0 and a standard deviation of 1. This ensures that each feature contributes equally to the clustering process.

1.2 Patient clustering

We used KMeans, where we applied the Silhouette Score to determine the optimal number of clusters (k). From that values we obtained 9 different clusters, among which we performed cluster profiling, getting these huge insights from the data: F1 represents baseline intensity, F7 indicates self-excitation potential, and binary variables (F0–F6) helped form homogeneous groups with similar clinical characteristics.

Clusters with high F1 had patients with more frequent events.. Cluster 7 had a very high F7 → rare patients with explosive event patterns.

Cluster	F1_log	F7	Interpretation
2	High	Low	Frequent, regular events
7	Medium	Very high	Rare but clustered events
0, 6	Medium	Medium	Regular, stable events



We used 2D and 3D PCA to reduce the dimensionality, and we visualized the clusters.

Each point represents a patient, and colors distinguish each one of the clusters.

Point size was adjusted by F1, showing the baseline intensity per patient.

2. Hawkes Process Fitting per Cluster

2.1 Time data processing

As event times were given as relative differences (in months) from the previous event, we decided to convert them into absolute times by adding the differences. Then, we added a final event = last event + 1 month, to define the observation endpoint.
Any patients without events were removed.

2.2 Cluster Analysis

Each cluster was fitted with a univariate exponential Hawkes process with three parameters:

- **μ (mu) Baseline intensity:** frequency of spontaneous events.
- **η (eta/alpha) Excitation:** how much the probability increases after an event.
- **θ (theta) Decay:** how fast the excitation effect fades.

From this analysis, we obtained the following results:

Cluster	μ (basal)	η (excitation)	θ (decay)	Patients	Events	Interpretation
0	12.6	0.055	~0	70	219	Frequent, regular events
2	25.5	0.027	~0	103	329	High baseline frequency, low self-excitation
5	3.44	0.703	0.348	33	125	Few baseline events, strong clustering
7	4.53	0.604	0.157	17	101	Atypical cluster: explosive events due to high F7
8	4.38	0.422	0.138	22	69	Infrequent events, moderate excitation

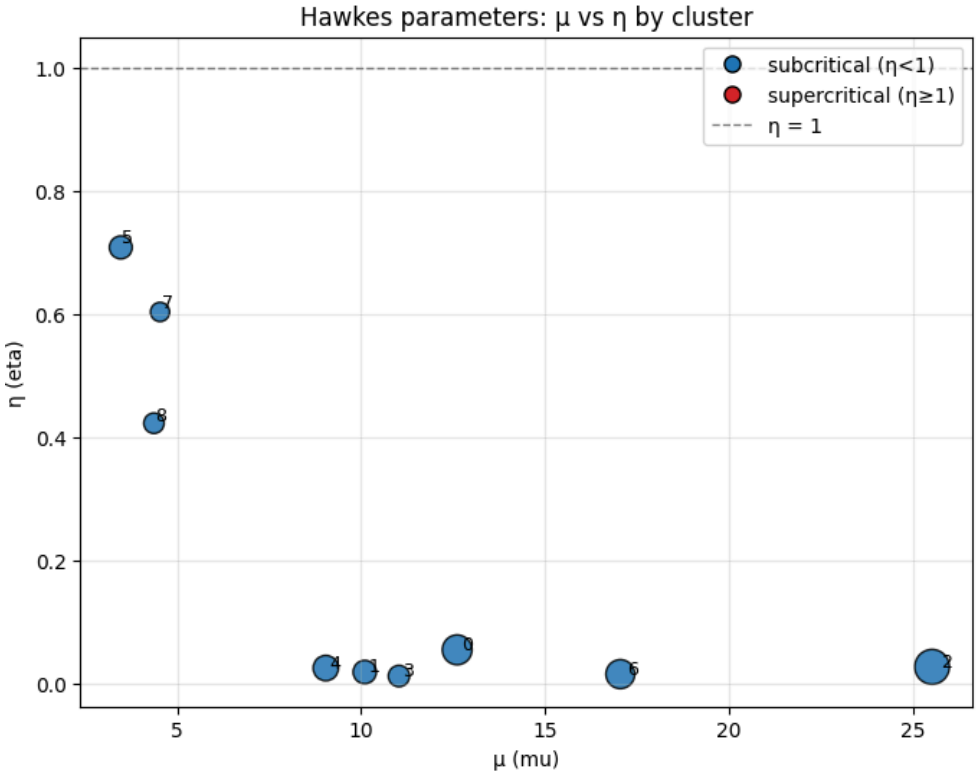
We differentiated three main groups:

➔ Large clusters with high μ, low η lead to frequent but regular events.
➔ Small/atypical clusters with low μ, high η lead to rare but clustered events.
➔ Cluster 7 is a rare group with explosive patterns, consistent with its high F7.

In terms of clusters, we can see: regular clusters (0, 2, 6): frequent, predictable events with low self-excitation (majority of patients), and explosive/rare clusters (5, 7, 8): few baseline events, but each triggers cascades (critical clinical phenomena).

We also realized how the relationship between some features and the Hawkes works:

➔ **μ correlates with F1:** patients with frequent events.
➔ **η correlates with F7:** patients with event bursts.



All clusters are subcritical (η<1), which indicates the fitted Hawkes processes are stable, without uncontrolled event growth, and most events are driven by the base rate rather than prolonged self-excitation.

- Three clusters show relatively high η values (0.4-0.7), these represent dynamics with stronger temporal dependence, where one event increases the likelihood of subsequent ones.
- The remaining clusters have very low η values (close to 0), so these events are almost independent of each other, and exogenous factors (μ) dominate, which may suggest that no significant contagion or feedback dynamics are present in those clusters.
- Clusters with high μ and low η (on the lower-right side) indicate processes with many spontaneous events but little dependency between them.
- Conversely, clusters with low μ but higher η show that a few exogenous events can generate longer or more dependent sequences.

3. Conclusions.

Combining clustering and Hawkes modeling enables segmentation of patients and understanding of how their events evolve. This integration of clinical features and temporal patterns reveals clear differences among patient groups, even within anonymized populations. So that it can help to:

- ❖ Identify patients at high risk of recurrent events.
- ❖ Detect atypical subgroups requiring special monitoring.
- ❖ Improve interpretation of clinical data in longitudinal studies.

https://github.com/Mario0fdez/Machine-Learning-in-Healthcare/tree/main/Lab%202