The notebook we created for this lab is available in the following GitHub repository: https://github.com/javierferna/point-processes-hawkes

METHODOLOGY

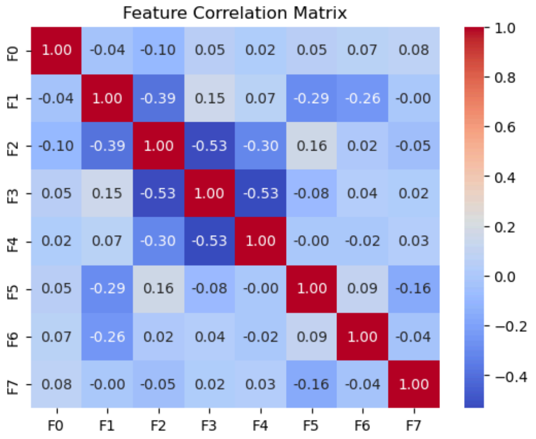
The dataset comes from a real clinical study, so it was anonymized, and we have downloaded it directly from Aula Global. The initial exploratory data analysis described the data and showed that no missing values were present. Only two of the features were of a non-binary nature. Out of those, only **F7** needed normalization to be in a 0-1 scale like the other variables. The main dataset `features.csv` included one categorical variable (**ID**) and 8 numerical ones. Groups of correlated features were identified. However, none of the pairwise correlations were high enough to warrant elimination. As for `timelines.csv`, we noticed that a significant number of patients didn't have any recorded events for the duration of the study (i.e., all **time_x** variables were 0.0). Because we don't know whether these patients voluntarily left the research or whether they didn't suffer any events, we decided to remove their respective rows altogether.

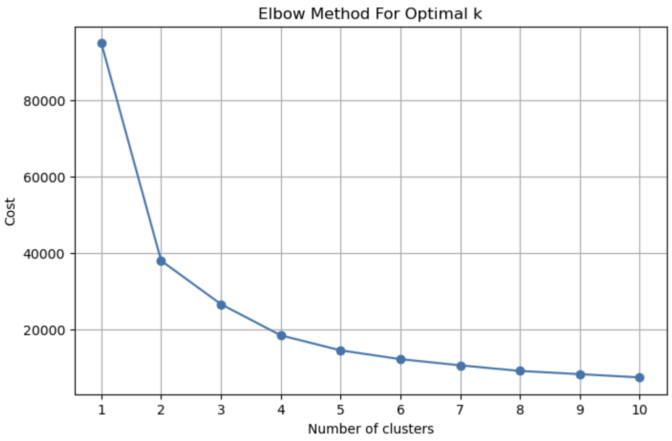
Clustering was performed with **k-means** on all the numerical features. The number of clusters was selected using the elbow method, and the resulting groups were visualized with PCA. Note that in this case, we only reached an explained variance of 73.6% with 3 principal components. Hence, the groups may not be perfectly representative.

For the Hawkes processes modeling, we create a new dataframe that contains the *absolute* time of events with respect to the start of the study. The relative time distance between events is not appropriate to fit a Hawkes process. Then, we set all **time_x** features after the last recorded event to 0.

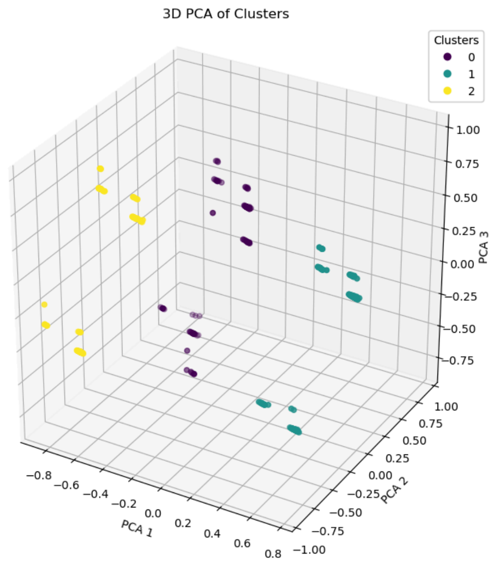
VISUALIZATIONS

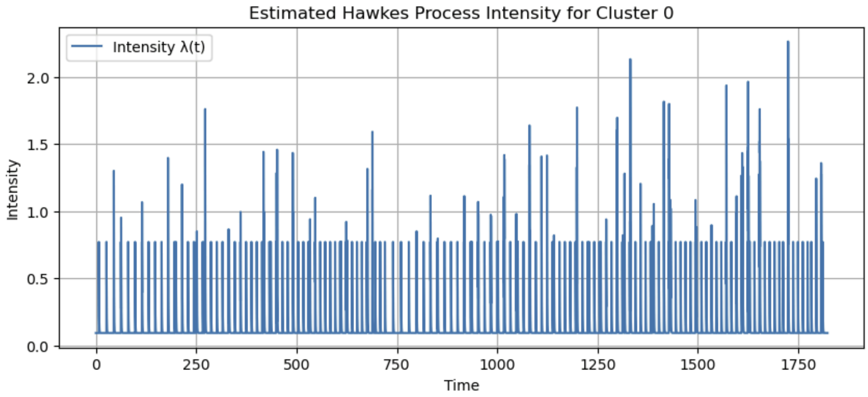
This correlation plot corresponds to the initial dataset after removing the **ID** variable. Conversely, the second plot shows the optimal number of clusters using the elbow method.

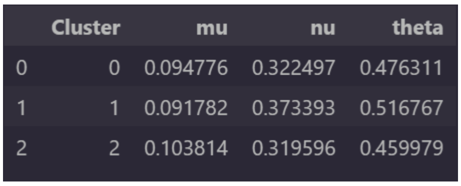




The third plot shows how clusters are spread across a 3D space with PCA, while the fourth one allows us to see the intensity rate of the Hawkes Process for cluster 0. The table below it displays the parameters obtained for each of the fitted models.







	Cluster	mu	nu	theta
0	0	0.094776	0.322497	0.476311
1	1	0.091782	0.373393	0.516767
2	2	0.103814	0.319596	0.459979

RESULTS DISCUSSION

The final features dataset included 8 predictors after preprocessing. **K-means** used three clusters, which may not be illustrative of the actual groups because of the low explained variance. The Hawkes process model highlighted three different parameters for each cluster: **mu (μ)** (baseline event rate), **eta (η)** (self-excitation level), and **theta (θ)** (decay rate of excitation). Based on the results, it seems that **cluster 1** has the lowest spontaneous rate of events. However, it has the highest **eta (η)**, which indicates that a bigger fraction of events influences future ones. The influence of these events lasts the shortest out of all groups, though. As for **cluster 2**, note that it has the highest baseline event rate and the lowest self-excitation level. This means that events happen more sparsely without too much influence from previous events. **Cluster 0** seems to be a middle ground between the other 2 clusters.

CHALLENGES

At first, we didn’t know how to deal with the fact that the timelines dataset had relative time distances for events. We didn’t know if those would be enough to fit our model. Additionally, we had problems obtaining good cluster groups that could be interpreted. This is due to the low explained variance of the principal components and the fact that we don’t know the nature of the anonymized features. Lastly, we had trouble modelling the Hawkes processes for each cluster. `ExpHawkesProcessInference` objects take a single array of timelines. Therefore, we had to add an offset of 12 months to our single array of timelines per cluster so that the model could understand events that came from different patients.

CONCLUSION

This lab has allowed us to learn more about point processes and how they can be applied to real-life situations and studies. In this case, we have discovered that different groups of patients suffer events more (or less) spontaneously than other groups. We have also learned that previous events that have happened can have a huge impact on the rate of events that may happen in the future.