

Mini-Project (ML for Time Series) - MVA 2025/2026

Maryam El Yaagoubi Maryam.El-Yaagoubi@eleves.enpc.fr
Homère Tandeau de Marsac Homere.TANDEAU-DE-MARSAC@eleves.enpc.fr

January 2, 2026

1 Method

Let $\{x_t\}_{t=1}^T$ denote a multivariate time series, where $x_t \in \mathbb{R}^d$ represents ideological measurements observed at time t . To construct a supervised learning problem, the original time series is transformed into a design matrix using past information only. For each time index t , a feature vector X_t is built from lagged values and rolling statistics computed over a finite historical window.

An example of a classification target $y_t \in \{0, 1\}$ is defined as the sign of a future variation of a reference series over a fixed horizon h . Concretely, $y_t = 1$ if the reference series increases between t and $t + h$, and $y_t = 0$ otherwise. This yields a binary time-series classification problem, where the task is to predict the direction of future movement using historical multivariate dynamics.

1.1 Laplacian Score for feature selection

Laplacian Score is an unsupervised feature selection method designed to preserve the local geometric structure of the data. Given a feature matrix $X \in \mathbb{R}^{n \times p}$, an affinity graph is constructed using a k -nearest neighbor rule. The affinity between observations x_i and x_j is defined by a heat kernel

$$S_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{t}\right)$$

whenever x_j belongs to the neighborhood of x_i , and zero otherwise. Let D denote the diagonal degree matrix with entries $D_{ii} = \sum_j S_{ij}$. The graph Laplacian is then defined as $L = D - S$.

For each feature f_i , the Laplacian Score is given by

$$\text{LS}(f_i) = \frac{\tilde{f}_i^\top L \tilde{f}_i}{\tilde{f}_i^\top D \tilde{f}_i},$$

where \tilde{f}_i denotes the D -centered version of f_i . Features with small Laplacian Scores vary smoothly along the data manifold and are considered representative of the intrinsic structure. Feature selection is performed by ranking features individually and retaining the top-ranked subset.

In the present context, Laplacian Score is applied to the expanded feature matrix prior to supervised learning, using only the training data.

2 Data

2.1 Dutch political orientation time series

The dataset consists of a multivariate time series with 988 observations and 13 variables, derived from a Dutch weekly survey on political orientation. Each observation corresponds to one survey period, ordered in time by a strictly increasing index. No explicit calendar dates are provided; the temporal structure is implicit.

The main variables ('Orr', 'Vvd', 'Cda', 'D66', 'Pvda', 'Glef') represent average left-right ideological self-placement scores, either at the aggregate population level or associated with specific political parties. Each ideological series is accompanied by a corresponding sample-size variable ('wn*'), indicating the number of respondents used to compute the weekly average. These variables provide information about the reliability of the ideological measurements.

All variables are fully observed, with no missing values and no duplicated rows. Visual inspection and basic diagnostics indicate strong temporal smoothness in the ideological series, suggesting dominant low-frequency dynamics. Cross-series dependence is expected due to shared political context and overlapping electorates. The sample-size variables exhibit non-negligible variability over time and may influence the stability of the ideological estimates.

To make feature selection non-trivial, the raw time series are expanded into a higher-dimensional representation using lagged values and rolling statistics computed over past windows. This transformation introduces redundancy and correlation among features, while preserving causality by relying exclusively on historical observations. We consider the problem of feature selection for multivariate time series in a classification setting. The objective is to identify a subset of features that captures the intrinsic structure of the data while reducing dimensionality prior to supervised learning.

3 Results

3.1 Laplacian Score

Laplacian Score is applied to the expanded feature matrix prior to classification. Since the method is unsupervised, it operates solely on the feature space X (training split), without access to the labels. An affinity graph is constructed using a k -nearest neighbor rule in the expanded feature space, and features are ranked according to their ability to preserve the local geometric structure of the data manifold.

In this context, Laplacian Score favors features that are smooth with respect to the intrinsic dynamics of the multivariate time series, while penalizing noisy or weakly structured features. The selected subset is then used as input to a supervised classifier, which makes it possible to assess how well manifold-preserving features support downstream prediction.

In our experiment, we build a $k_{NN} = 10$ nearest-neighbor graph and retain the top $k = 30$ features (smallest scores). The best-ranked features are dominated by rolling means (especially with a 12-week window), along with a smaller number of lagged variables. Figure 1 shows the resulting ranking.

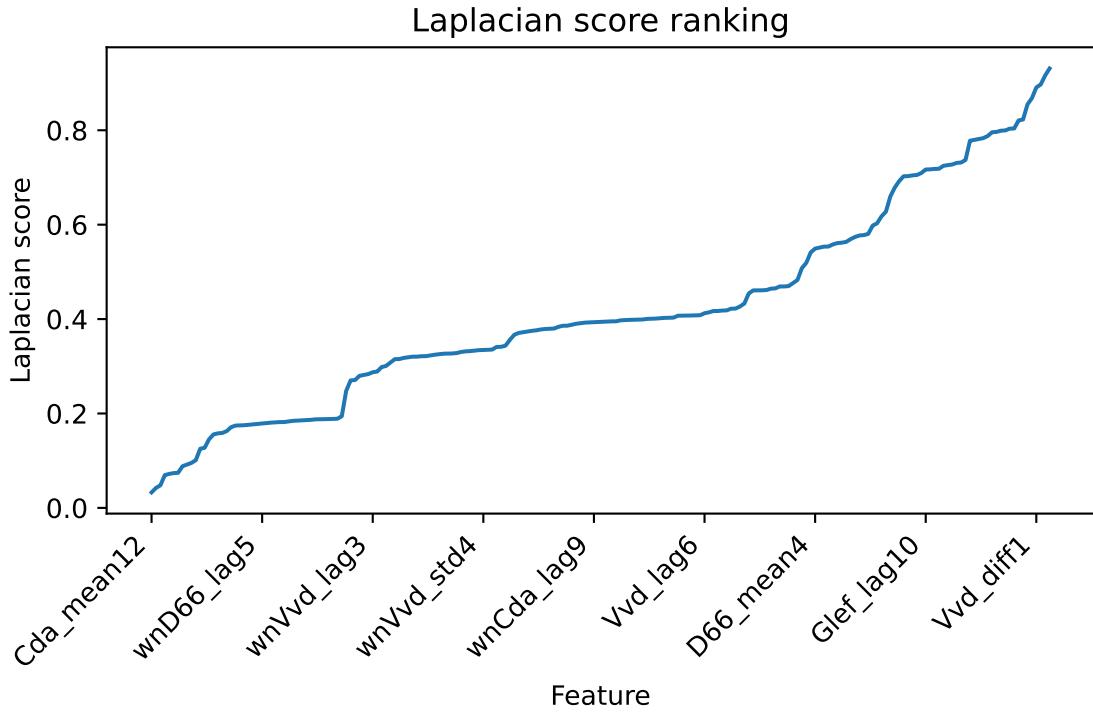


Figure 1: Laplacian Score ranking for the $k = 30$ selected features (lower is better).

Laplacian was presented as a way to preserve feature manifold geometry. Figure 2 compares 2D PCA embeddings of standardized observations computed from the full feature space and from the Laplacian-selected subset; the two representations are visually similar, indicating that geometry is effectively preserved.

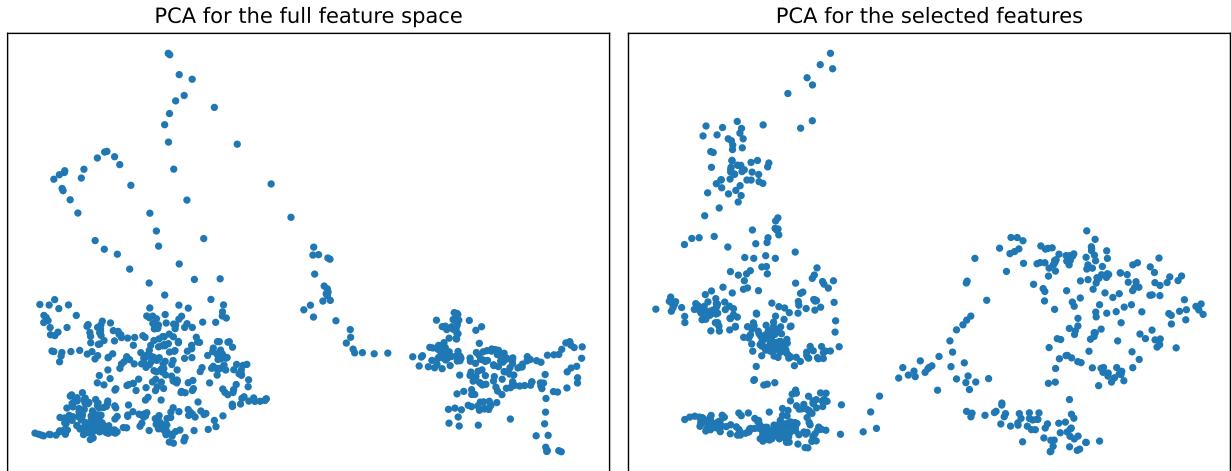


Figure 2: 2D PCA embedding of standardized observations using the full feature space (left) and the Laplacian-selected subset (right).

Finally, we evaluate the selected subset in a simple downstream task: predicting whether Orr in-

creases over the next $h = 4$ weeks using logistic regression. Using the full feature set yields an accuracy of 0.645, while restricting to the Laplacian-selected features reduces accuracy to 0.556. This value remains above chance but reflects a substantial loss of discriminative information relative to the full feature space. This behavior is consistent with the objective of Laplacian Score: the method selects features that preserve the intrinsic geometric structure of the data rather than those that are maximally informative for the supervised task.