# FEM-H1 Clustering

Lucas Kania
lucas.kania@usi.ch

## 1 Implementation

Given an input $x$ of nxT, and the following objective function

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{k,t}\|x_t - S_k\|^2 + \epsilon^2 \sum_{t=1}^{T}\sum_{k=1}^{K}(\gamma_{k,t+1} - \gamma_{k,t})^2 \qquad \text{s.t. } \forall_{1\leq t\leq T}\sum_{k=1}^{K}\gamma_{k,t} = 1 \tag{1}$$

We need to solve two optimization problems, getting the best centroids (S-problem) and determining the best possible allocation (Gamma-problem).

### 1.1 The S-problem

The S-problem is equivalent to solving K-means

$$\arg\min_{S}\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_{k,t}||x_t - S_k||^2 \tag{2}$$

which has for solution

$$S_k = \frac{\sum_{t=1}^{T}\gamma_{k,t}x_t}{\sum_{t=1}^{T}\gamma_{k,t}} \tag{3}$$

Using the vector $\gamma$ of (K*T)x1

$$\gamma := \begin{bmatrix}\gamma_1\\ \gamma_2\\ \vdots\\ \gamma_T\end{bmatrix} \tag{4}$$

where

$$\gamma_k := \begin{bmatrix} \gamma_{k,1} \\ \vdots \\ \gamma_{k,T} \end{bmatrix} \tag{5}$$

the computation of the centroids reduces to

$$S_k = \frac{\text{sum}(\gamma_k^T * x)}{sum(\gamma_k)} \tag{6}$$

where the sum is done across rows ($S_k$ is a nx1 vector), $*$ is the element-wise multiplication, and $\gamma_k^T$ must be broadcasted (i.e. n copies $\gamma_k^T$ are stacked vertically).

## 1.2 The Gamma-problem

The Gamma-problem can be stated as

$$\arg\min_{\gamma} \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_{k,t}||x_t - S_k||^2 + \epsilon^2 \sum_{t=1}^{T}\sum_{k=1}^{K} (\gamma_{k,t+1} - \gamma_{k,t})^2 \qquad \text{s.t. } \forall_{1\leq t\leq T} \sum_{k=1}^{K} \gamma_{k,t} = 1 \tag{7}$$

For this problem, we will express each one of the terms in matrix notation and then optimize the problem 7 with `quadprog`.

### 1.2.1 Constraint

Given $\gamma$, see definition 4, the constraint

$$\forall_{1\leq t\leq T}\sum_{k=1}^{K}\gamma_{k,t}=1 \tag{8}$$

can be rewritten as

$$B\gamma=c \tag{9}$$

where B is a matrix of dimension Tx(T*K) consisting of K concatenated TxT diagonal matrices

$$B:=\begin{bmatrix}1 & & & & 1 & \\ & \ddots & & \dots & & \ddots & \\ & & 1 & & & & 1\end{bmatrix} \tag{10}$$

and C is a vector of dimension Tx1 filled with ones.

### 1.2.2 Quadratic Term

The quadratic term can be rewritten as follows

$$\epsilon^2\sum_{t=1}^{T}\sum_{k=1}^{K}(\gamma_{k,t+1}-\gamma_{k,t})^2=\epsilon^2\sum_{k=1}^{K}\langle D\gamma_k,D\gamma_k\rangle=\epsilon^2\sum_{k=1}^{K}\gamma_k^TD^TD\gamma_k \tag{11}$$

where $D$ is the discrete derivate (T-1)xT matrix

$$D:=\begin{bmatrix}-1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1\end{bmatrix} \tag{12}$$

Thus, $H:=D^TD$ is the Hessian matrix.

Finally, building a block diagonal matrix of $H$s, named $\hat{H}$,

$$\hat{H} := \epsilon^2 \begin{bmatrix} H & & & \\ & H & & \\ & & \ddots & \\ & & & H \end{bmatrix} \tag{13}$$

allows us to rewrite the quadratic term as

$$\epsilon^2 \sum_{t=1}^{T} \sum_{k=1}^{K} (\gamma_{k,t+1} - \gamma_{k,t})^2 = \epsilon^2 \sum_{k=1}^{K} \gamma_k^T D^T D \gamma_k = \gamma^T \hat{H} \gamma \tag{14}$$

### 1.2.3 Linear Term

Defining the vector g of dimensions (K*T)x1

$$g := \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_K \end{bmatrix} \tag{15}$$

where

$$g_k := \begin{bmatrix} \|x_1 - S_k\|^2 \\ \vdots \\ \|x_T - S_k\|^2 \end{bmatrix} \tag{16}$$

enables us to rewrite the linear term as follows

$$\begin{aligned} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{k,t} ||x_t - S_k||^2 &= \sum_{t=1}^{T} \left( \sum_{k=1}^{K} \gamma_{k,t} ||x_t - S_k||^2 \right) \\ &= \sum_{t=1}^{T} g_k^T \gamma_k \\ &= g^T \gamma \end{aligned} \tag{17}$$

### 1.2.4 Matrix form of the Gamma-problem

Given the matrix forms of the constraint, linear term, and quadratic term, the Gamma-problem can be rewritten as (using Lagrangian multipliers)

$$\arg\min_{\gamma,\lambda} \gamma^T \hat{H}\gamma + g^T\gamma + \lambda^T(B\gamma - c) \tag{18}$$

which can be solved using `quadprog`.

# 2 Experiments

The dataset $X$ had dimensions 5x500. In figure 1.a and 1.b, we can observe the *L-curves* for the regularization parameters (from $10^{-10}$ to $10^{1}$) for fix K. The lack of L-shape for the figure 1.a is due to the solution provided by `quadprog` (an effect that can be also observed in the implementation provided in icorsi). However, see figure 1.b, it is clear that the increase of regularization is positively correlated with an increase in the modeling error since this effect occurs for all settings of the number of clusters.

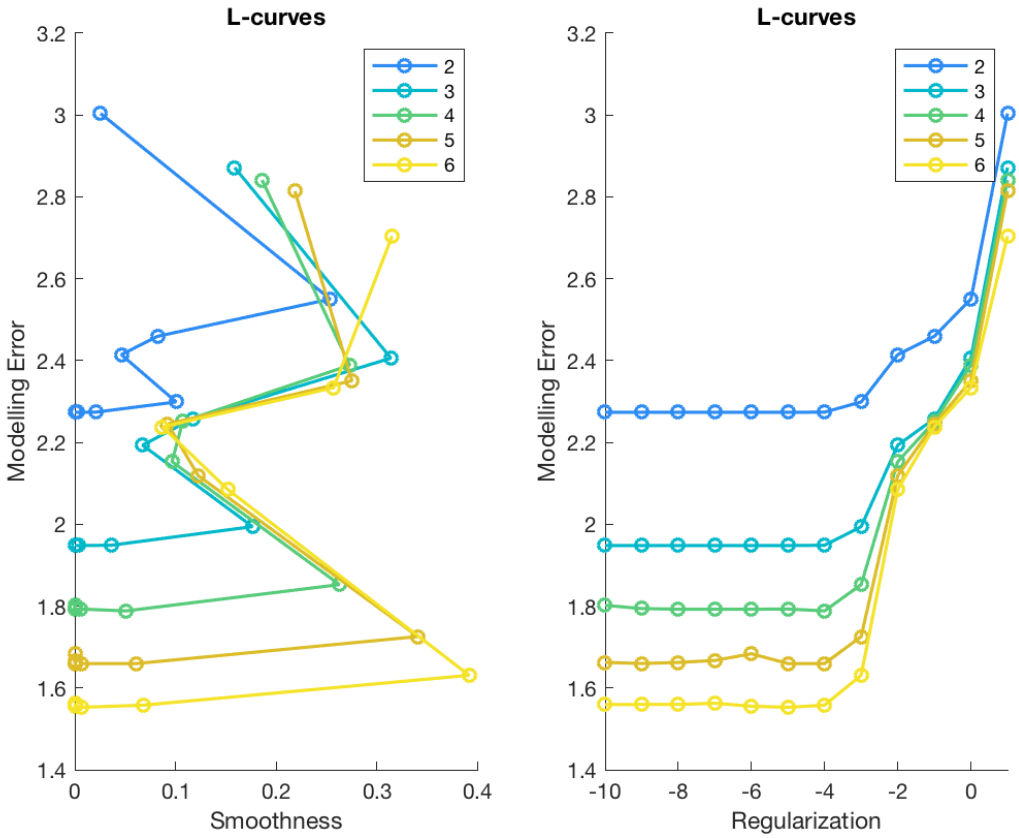


Figure 1: *L-curves* for fixed number of clusters. Each line denotes a different number of clusters. (a)(Left) Modelling error vs. Smoothness. (b)(Right) Modelling error vs. Regularization parameters. The x-axis is logarithmic.

Figure 1.b shows that increasing the number of clusters enhances the fitting of the data, which is expected. However, figure 2 shows that the improvement for a fixed $\epsilon$ (i.e. the regularization factor), as the number of clusters is increased, is marginal from 3 clusters onwards.

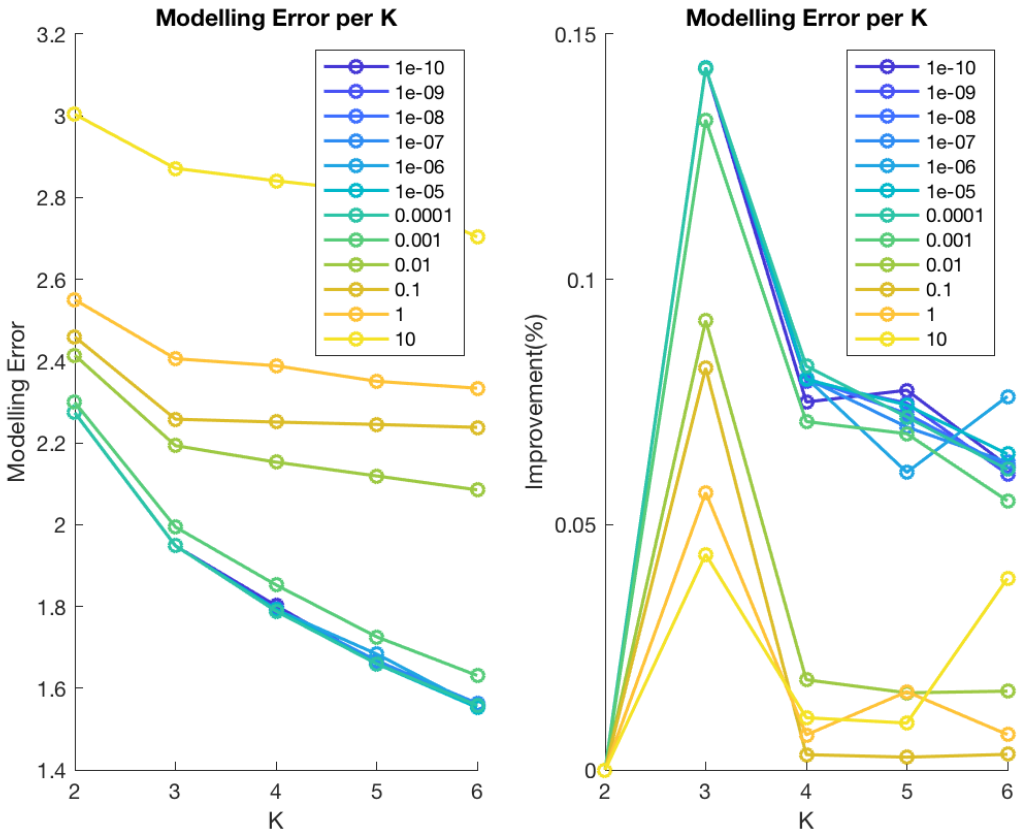


Figure 2: Modelling Error vs. number of clusters. Each line is represents a different fixed regularization factor. (Left) Absolute values. (Right) Percentual improvement over the previous number of clusters.

Therefore, we select 3 clusters as the optimal number of clusters. Moreover, looking at figure 1.b , we can observe that there is a marginal improvement in the modeling error, for 3 clusters, when the $\epsilon$ is lower than $10^{-2}$ Thus, we choose a regularization factor of $10^{-2}$ .

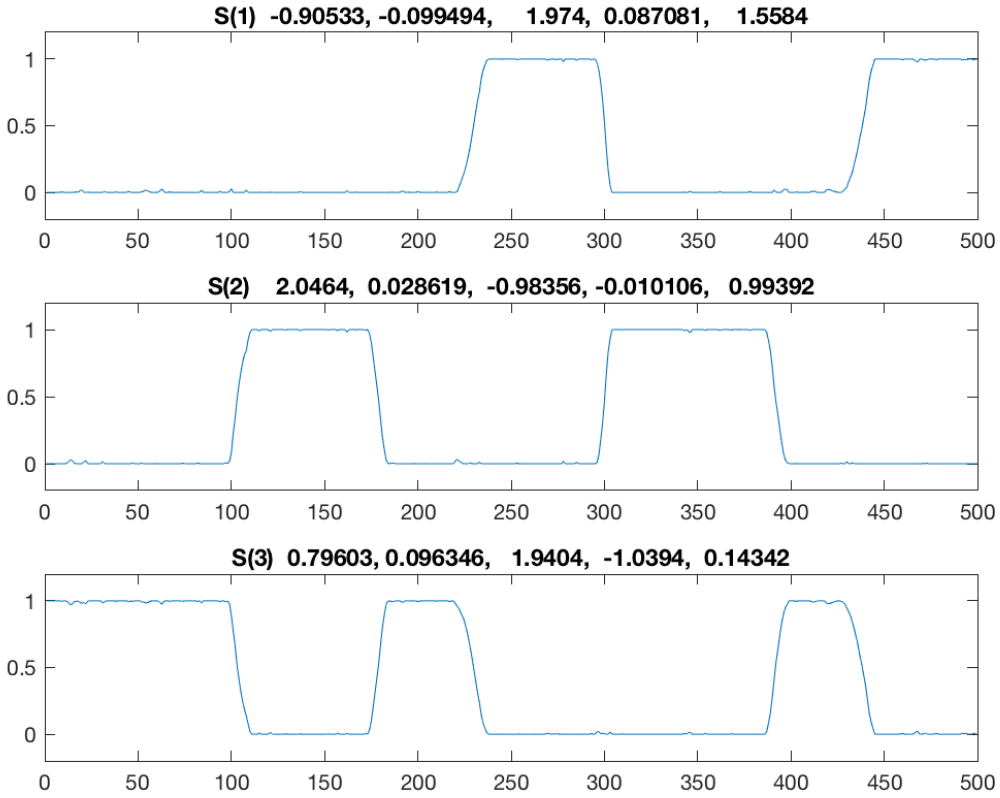


Figure 3: Clusters switching for $K = 3$ and $\epsilon = 10^{-2}$. The coordinates of the clusters are shown over each on of the plots.

Figure 3 shows the resulting clusters for the chosen parameters.