Supplementary Material

Penalized Model-Based Clustering of Complex Functional Data

Nicola Pronello[1], Rosaria Ignaccolo[2], Luigi Ippoliti[3], Sara Fontanella[4]

[1]Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Italy

[2]Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Italy

[3]Department of Economics, University of Chieti-Pescara, Italy

[4]National Heart and Lung Institute, Imperial College London, United Kingdom

October 9, 2022

1 Simulation

The selection of cluster-relevant discriminant basis functions appears an important and challenging issue within functional clustering. Since the inclusion of unnecessary components might mask the cluster structure, the aim of this simulation is twofold: i) show that the L_1 penalization on the set of the expansion coefficients of equation (2) favours an adaptive regularization that removes noise while preserving dominant local features, accommodating for possible heterogeneity along the domain of the curves; ii) investigate how different sources of variability affect the clustering performance (*accuracy*) of PFC-L_1, measured as the percentage of correctly classified functions.

In the following scenarios, the statistical analysis is illustrated for data simulated by means of a finite mixture of specific distributions. In particular, based on equations (1) and (2), the curves are simulated (except for scenario E) using a combination of $p = 25$ Fourier basis functions defined over a one-dimensional regular grid with 100 discrete points (i.e. $\mathcal{T} = \{0, 1, \ldots, 99\}$). Cluster-relevant basis functions are defined as those with specific frequency content (from 1 to 12 Hz) and for which the corresponding expansion coefficients appear "significantly" larger than others and particularly relevant for discriminating between groups. In all cases, the measurement noise is considered as independent with mean zero and variance $\sigma_\epsilon^2 = 1$, and the estimate of the latent/true function $X(t)$ follows the procedure described in Section 3.1 using the MAD estimator for σ_ϵ^2. The estimate of the number of groups K is performed using both the BIC and the ICL indices.

To generate different patterns along the domain, and define the axes of group separability, the Gaussian distributions of the expansion coefficients are specified according to the following five scenarios:

1) ***Scenario A***: considers a mixture of four ($K = 4$) multivariate Gaussian distributions with isotropic covariance

matrices, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \mathbf{I}_k), \quad k = 1, \ldots, 4.$$

With the exclusion of 3 entries per group, the means $\boldsymbol{\mu}_k$ are all zero mean vectors. The amplitude of the non-zero expansion coefficients identify the cluster-relevant variables and, at the same time, highlight the dominant frequency contents of the sine and cosine functions per each group. The number (n) of simulated curves is 200, with 50 functions per group.

2) ***Scenario B***: considers a mixture of two $(K = 2)$ multivariate Gaussian distributions with diagonal covariance matrices, $\boldsymbol{\Sigma}_k$, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k) \text{ with } \boldsymbol{\Sigma}_k = diag\{\sigma^2_{j,k}\}, \sigma^2_{j,k} \sim U(1,3), \quad j = 1, \ldots, 25, \; k = 1, 2.$$

As in the previous scenario, with the exclusion of 3 entries per group, the means $\boldsymbol{\mu}_k$ are all zero mean vectors. However, differently from scenario A, the expansion coefficients for the two groups are defined at the same frequencies such that the difference is only in their amplitude. The number (n) of simulated curves is 200, with 100 functions per group.

3) ***Scenario C***: considers a mixture of two $(K = 2)$ multivariate Gaussian distributions with non-diagonal covariance matrices, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k), \; k = 1, 2.$$

In this case, apart from 5 entries per group, the means $\boldsymbol{\mu}_k$ are all zero mean vectors and the covariance matrices are sparse. The expansion coefficients of the two groups are specified at the same frequencies and the group covariance matrices are obtained by rescaling a given correlation matrix by a diagonal matrix of standard deviations drawn from a Uniform distribution with interval $[1, 3]$. The covariance matrices result sparse with maximum correlations of 0.6 between coefficients. The number (n) of simulated curves is 200, with 100 functions per group.

4) ***Scenario D***: considers a mixture of four $(K = 4)$ multivariate Skew Normal distributions, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{SN}(\boldsymbol{\mu}_k; \mathbf{I}_k; \alpha_k) \text{ with } \alpha_k \sim \mathcal{U}(-4; 4), \quad k = 1, \ldots, 4.$$

With the exclusion of 3 entries per group, the location parameters $\boldsymbol{\mu}_k$ are all zero vectors. The amplitude of the non-zero expansion coefficients identify the cluster-relevant variables and, at the same time, highlight the dominant frequency contents of the sine and cosine functions per each group. Note that here, differently from *Scenario A*, the parameters $(\boldsymbol{\mu}_k; \mathbf{I}_k)$ are not the mean and the covariance in each cluster, while the vector parameter α_k regulates the skewness of the distribution in each cluster (see, for example, Azzalini and Capitanio, 1999 for more details). The number (n) of simulated curves is 200, with 50 functions per group.

5) ***Scenario E***: considers a mixture of four ($K = 4$) multivariate Gaussian distributions with spheric covariance matrices, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \mathbf{I}_k), \quad k = 1, \ldots, 4.$$

Here the main difference from the previous scenarios is that the functions are generated from a series of $p = 128$ *Daubechies extremal phase* wavelets basis functions (Daubechies, 1999) but reconstructed using Fourier basis functions.

In the following, for each scenario, estimation results are shown over 100 replicated datasets, each consisting of 200 curves. For each scenario we report graphically an example of the simulated curves with the relative mixture parameters (μ_k Σ_k), the basis functions selected in each cluster and the reconstructed curves using the cluster specific basis.

1.1 Scenario A

Under *Scenario A* we consider 50 simulated curves in each of 4 clusters which differ just in terms of means. The simulated curves (50 per group) and the non-zero group expansion coefficients are represented in Figure 1.

By considering the Frobenius norm $F(p)$, as defined in Section 3, it is possible to choose the number of basis p_0 to reconstruct the data. In Figure 2 we report the pattern of $F(p)$ as a function of p.

For this simple simulation setting, estimation results suggest that grid search procedure on the triplet $(K, \lambda_1, \lambda_2)$ (see, Section 3) is always able to correctly select the cluster-relevant basis functions. This is confirmed by Figure 3 which shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases that, as it can be noticed, clearly highlight the identification of 4 clusters. Each barplot represent how many time over 100 replications a determinated basis function (expressed in terms of Hz) is selected by the LASSO penalization.

Under this scenario, the quality of the estimated clusters thus appears very good as the analysis of the misclassification rate suggests an 100% of accuracy in all the replicated datasets.

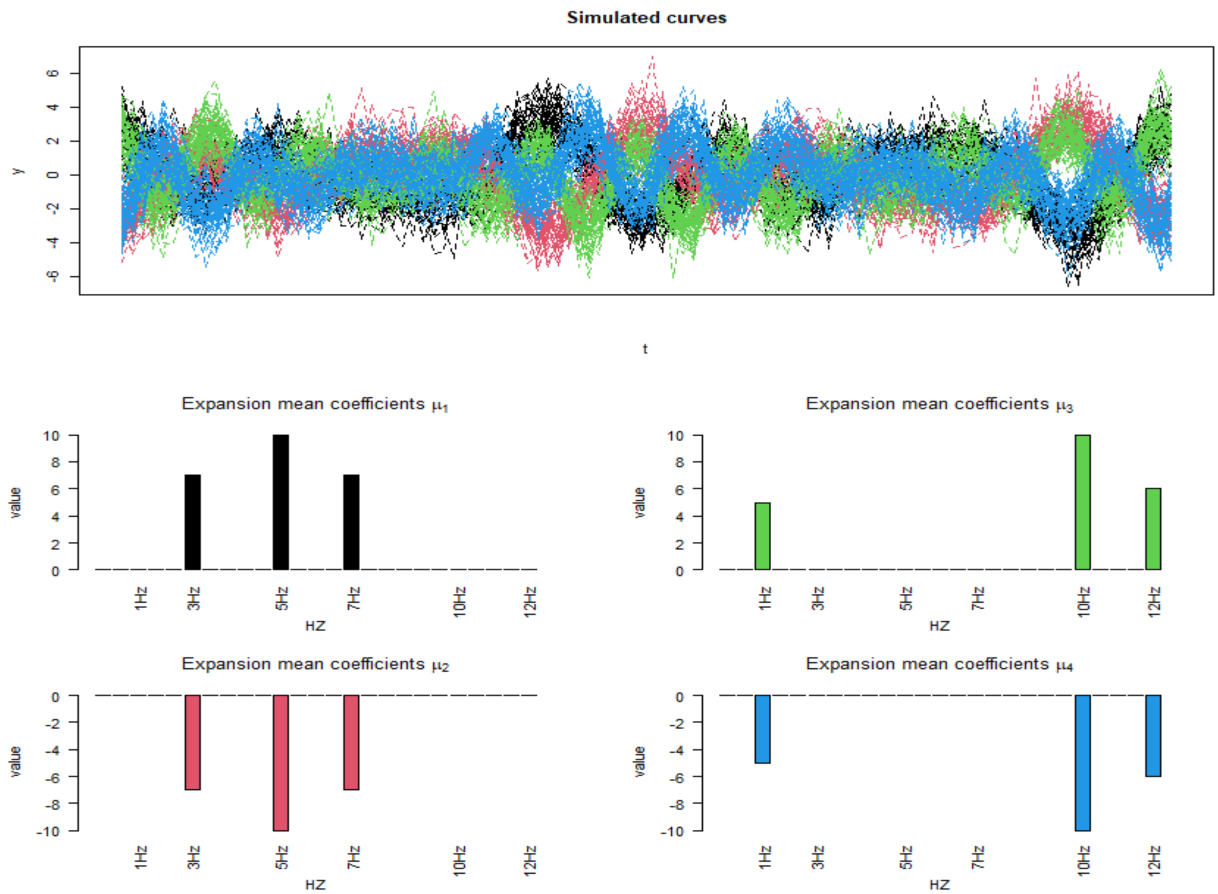

Figure 1: *Scenario A. Top: Curves simulated using* 25 *Fourier basis functions. Bottom: The vector of expansion coefficients. Each group has only three non-zero coefficients corresponding to basis functions with specific periodicities.*

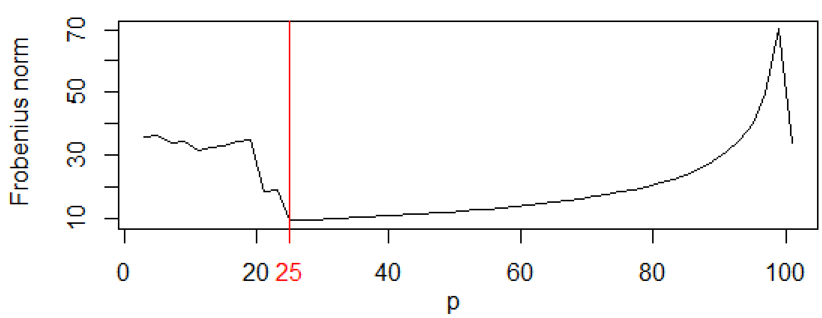

Figure 2: *The $F(p)$ function at different values of number of basis functions p ($p_0 = 25$ chosen at the minimum).*

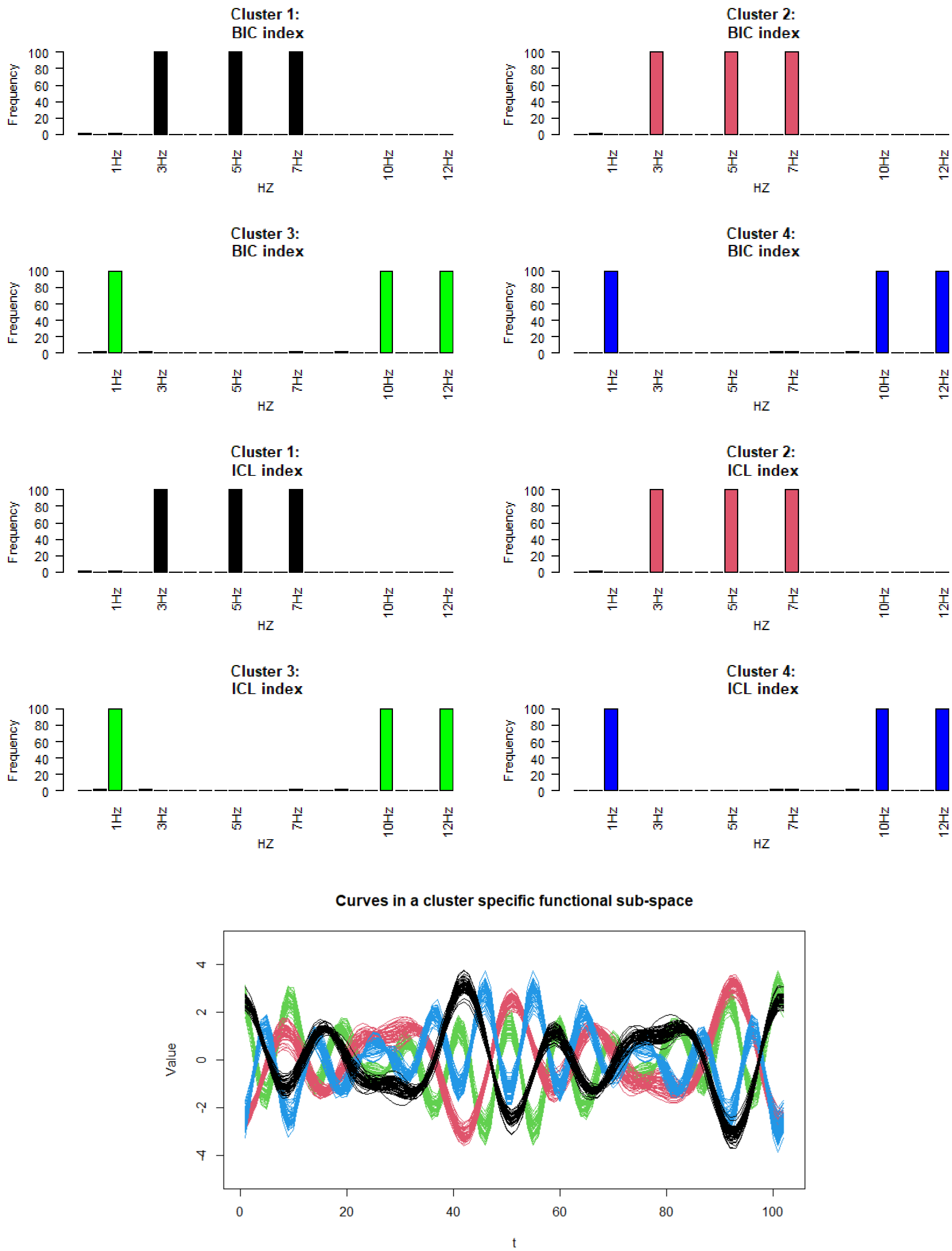

Figure 3: *Scenario A. Top panels: Barplot of the basis functions selected over 100 replications, both for BIC (top row) and ICL index (bottom row). The distribution is shown for pairs of sine and cosine basis functions. Bottom panel: Data projected on the selected basis functions.*

1.2 Scenario B

Considering *Scenario B*, here in each of the two cluster there are 100 curves, the non-zero group expansion coefficients and the heatmap of the group diagonal covariance matrices are shown in Figure 4. Differently from the previous scenario, the expansion coefficients for the two groups are defined at the same frequencies such that the difference is only in their amplitude. Moreover here we consider a cluster specific covariance structure.

Under this scenario, there do not seem to be particular problems in retrieving the cluster-relevant variables at the dominant frequencies. This seems to be confirmed by Figure 5 which shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases. Considering the quality of the estimated clusters, we have that, on average, the accuracy is 97% (std 0.01) using both information indices.

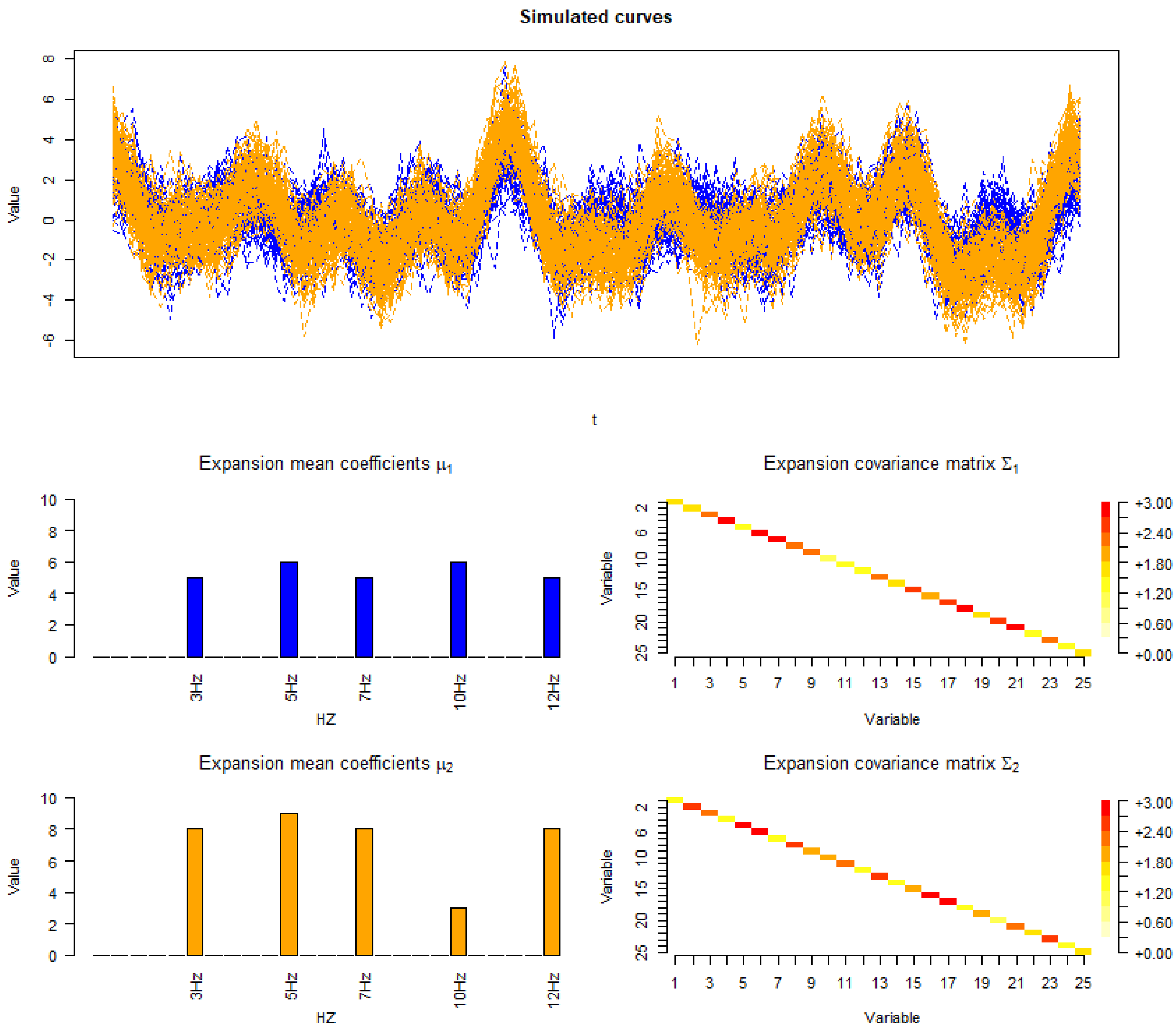

Figure 4: *Scenario B. Top: Curves simulated using* 25 *Fourier basis functions.Bottom Left: Non-zero group expansion coefficients; Bottom Right: Group diagonal covariance matrices.*

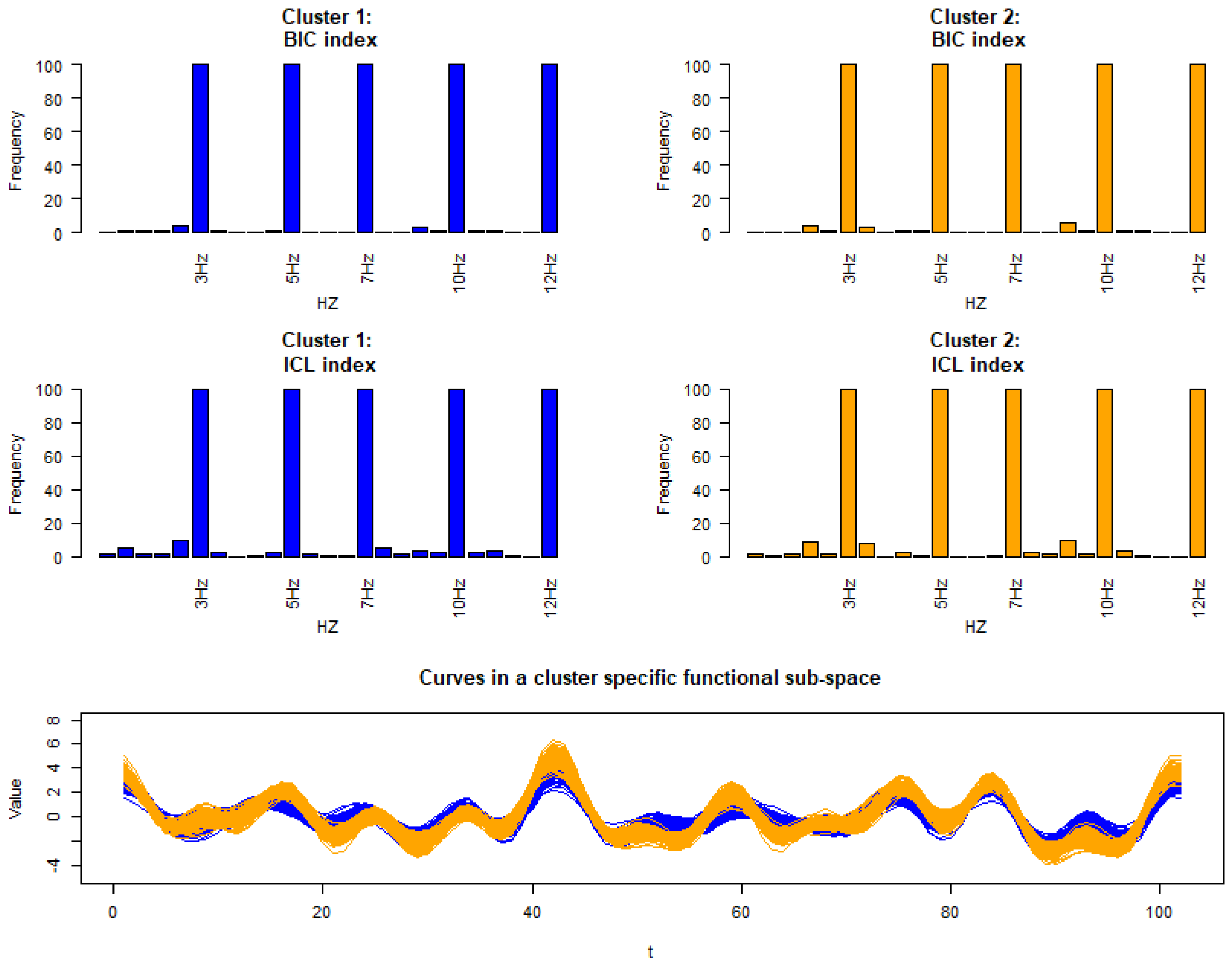

Figure 5: *Scenario B. Top panels: Barplot of the basis functions selected over 100 replications, both for BIC (top row) and ICL index (bottom row). The distribution is shown for pairs of sine and cosine basis functions. Bottom panel: Data projected on the selected basis functions.*

1.3 Scenario C

Regarding the third scenario, Figure 6 shows the simulated curves (100 per group), the non-zero group expansion coefficients and the heatmap of the group covariance matrices. Here, the number of expansion coefficients different to zero is the same as in *Scenario B*. Differently from the previous case, the group covariance matrices of the coefficients are sparse.

As shown in Figure 7, which gives the distribution of the selected basis functions as well as the corresponding projected data, it appears that it is a bit more difficult to retrieve the cluster-relevant basis functions under this scenario. However, looking at the quality of the clustering procedure, the accuracy remains significantly high being 0.86% on average both for the BIC (std 0.05) and ICL index.

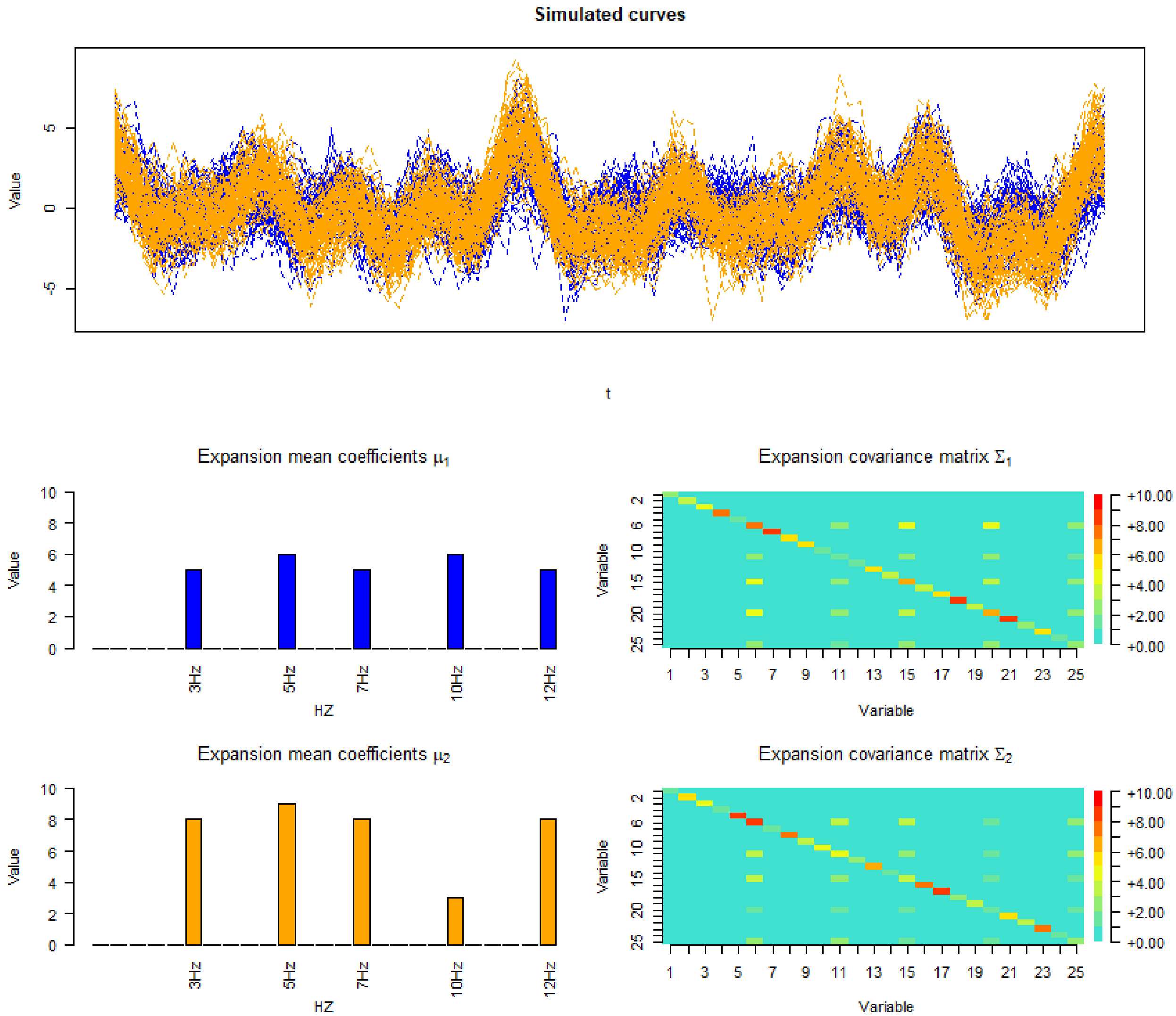

Figure 6: *Scenario C. Top: Curves simulated using* 25 *Fourier basis functions. Bottom Left: Non-zero group expansion coefficients; Bottom Right: Group sparse covariance matrices.*

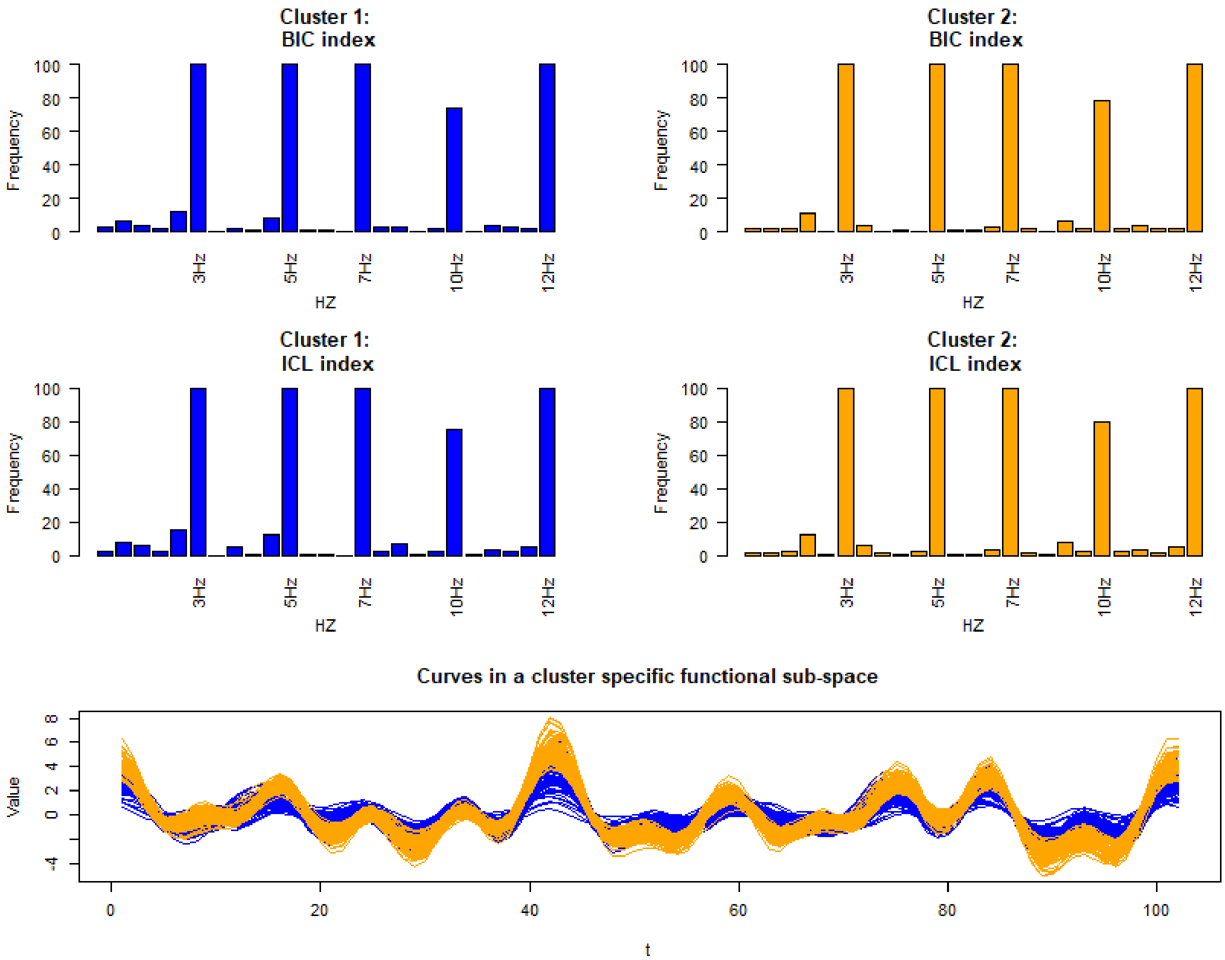

Figure 7: *Scenario C. Top panels: Barplot of the basis functions selected over 100 replications, both for BIC (top row) and ICL index (bottom row). The distribution is shown for pairs of sine and cosine basis functions. Bottom panel: Data projected on the selected basis functions.*

1.4 Scenario D

Under *scenario D*, we consider 50 simulated curves in each of 4 cluster. The non-zero group expansion coefficients are represented in Figure 8. Differently from *scenario A* for this simulation setting the expansion coefficients are simulated from a multivariate Skew-Normal distribution in each cluster.

Estimation results suggest are similar to the ones obtained under *Scenario A*, although the selection of the most discriminative basis is not that clear. This is confirmed by Figure 9 which shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases that, as it can be noticed, clearly highlight the identification of 4 clusters. Under this scenario, the quality of the estimated clusters thus appears very good as the analysis of the misclassification rate suggests an 100% of accuracy in all the replicated datasets.

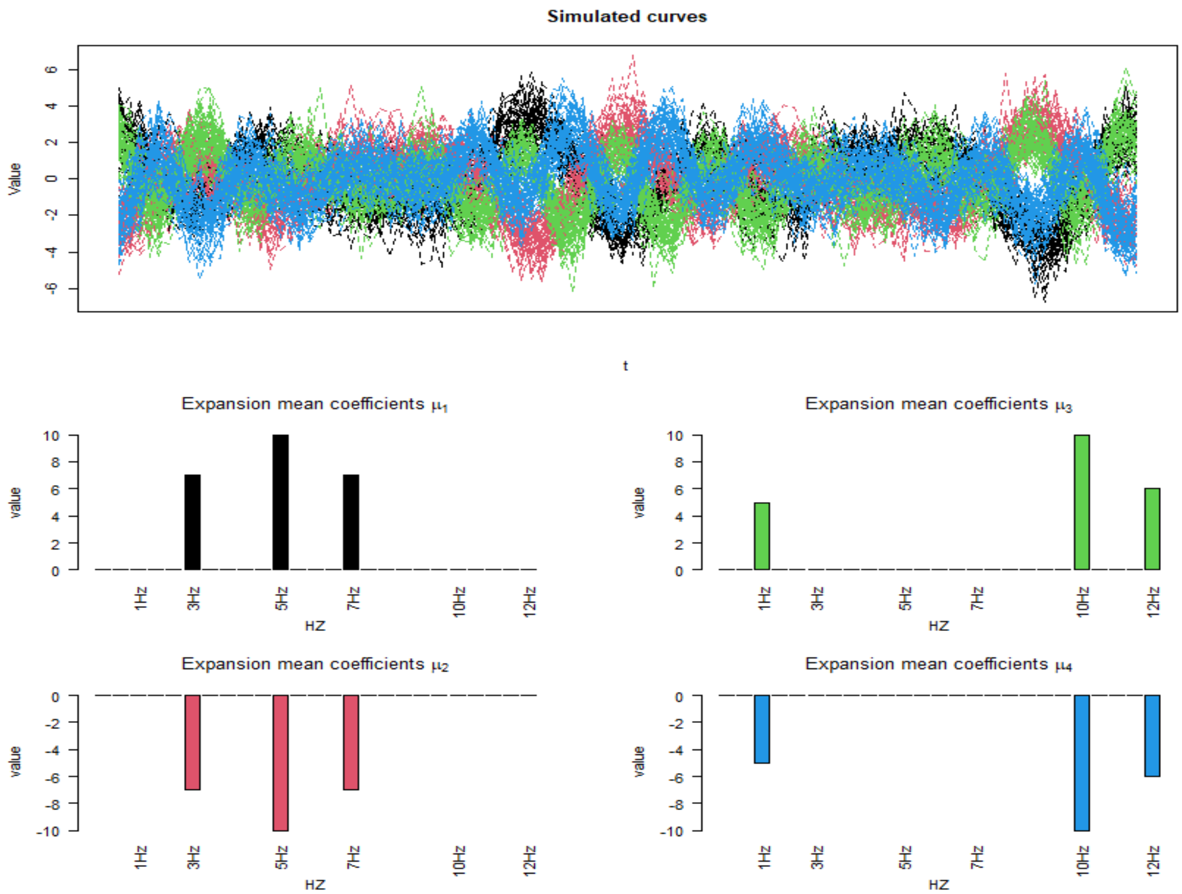

Figure 8: *Scenario D. Top: Curves simulated using* 25 *Fourier basis functions. Bottom: The vector of expansion coefficients. Each group has only three non-zero coefficients corresponding to basis functions with specific periodicities.*

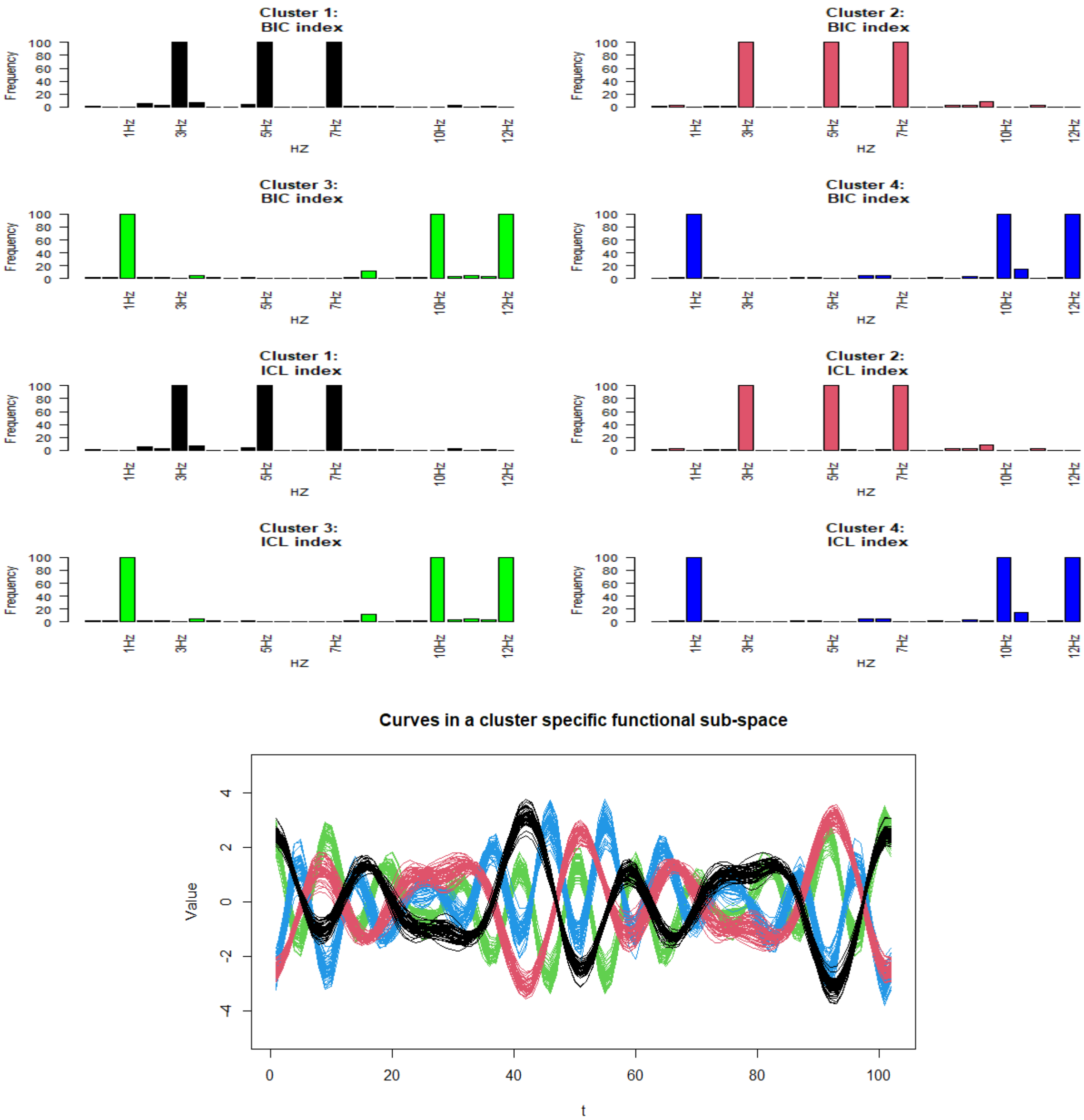

Figure 9: *Scenario D. Top panels: Barplot of the basis functions selected over 100 replications, both for BIC (top row) and ICL index (bottom row). The distribution is shown for pairs of sine and cosine basis functions. Bottom panel: Data projected on the selected basis functions.*

1.5 Scenario E

Finally considering *Scenario E*, the simulated curves (50 per group) and the non-zero group expansion coefficients are represented in Figure 10. Here to generate the data we consider a mixture of multivariate normal distribution on 128 wavelet basis functions, while to reconstruct the curves before clustering we consider a Fourier basis with 35 basis functions. Figure 11 shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases. Under this scenario, the quality of the estimated clusters thus appears very good as the analysis of the classification performance suggests an 99% accuracy (with standard deviation 0.009) using BIC index versus an 98.9% accuracy (with standard deviation 0.013) with ICL index.

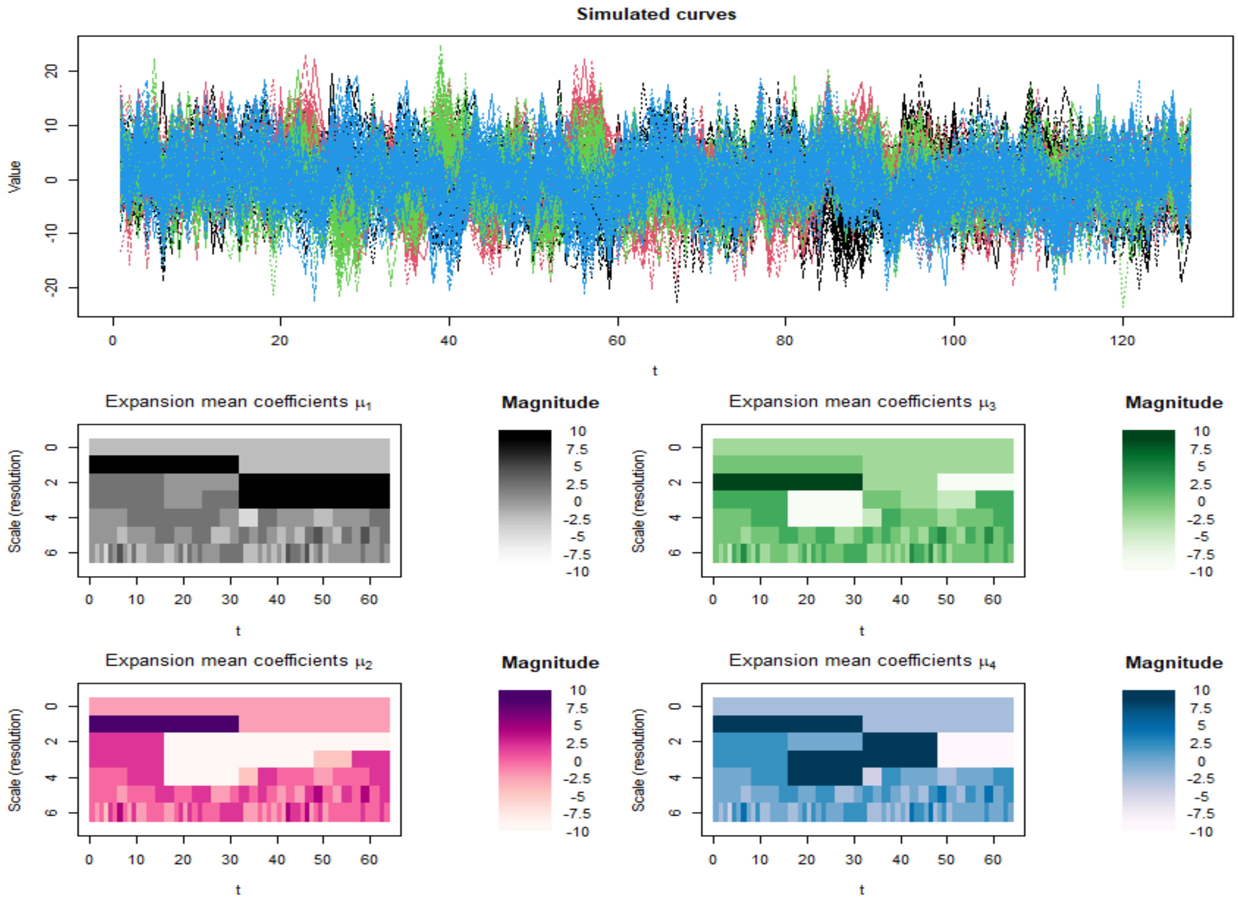

Figure 10: *Scenario E. Top: Curves simulated using* 128 *Wavelets basis functions. Down: The vector of expansion coefficients. Each group has only three non-zero coefficients corresponding to basis functions with specific periodicities.*

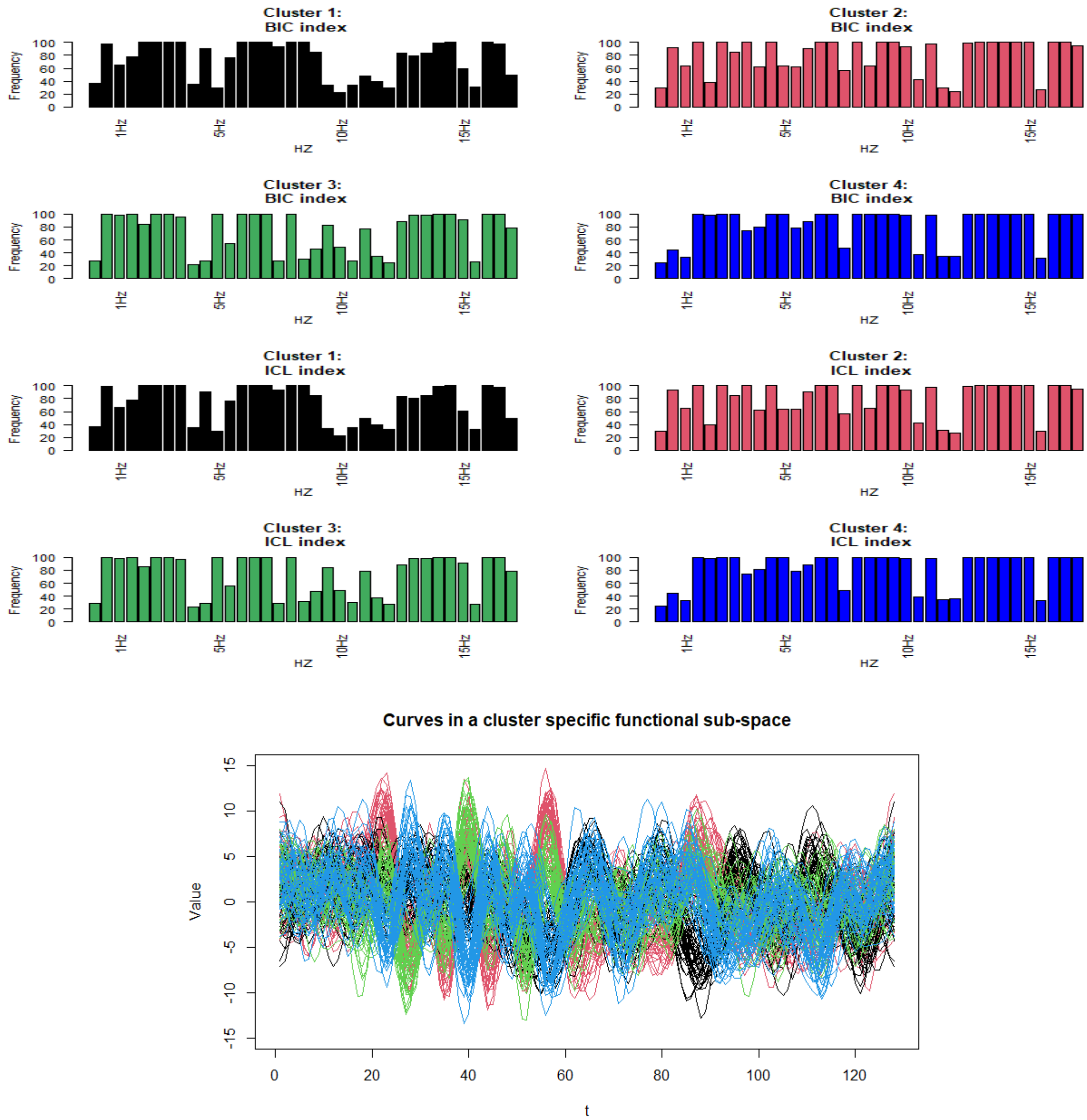

Figure 11: *Scenario E. Top panels: Barplot of the fourier basis functions selected over 100 replications, both for BIC (top rows) and ICL index (bottom rows). The distributions are shown for pairs of sine and cosine basis functions. Bottom panel: Data projected on the selected basis functions.*

1.6 Model selection

In this section, we present a simulation study in which we evaluate to which extent the indices presented in Section 3.3 succeed in indicating the correct underlying numbers of clusters for data generated according to the five scenarios described above. The accuracy of the indices is investigated over 100 replicated datasets for which Gaussian mixture models have been estimated for a number of clusters from 2 to 8. The results are shown in Table 1 which, for each scenario, gives the frequencies of the selected K using both the BIC and the ICL index.

Simulation setup	Index	K=2	K=3	K=4	K=5	K=6	K=7	K=8
Scenario A	BIC index	0	0	**100**	0	0	0	0
	ICL index	0	0	**100**	0	0	0	0
Scenario B	BIC index	**100**	0	0	0	0	0	0
	ICL index	**100**	0	0	0	0	0	0
Scenario C	BIC index	**100**	0	0	0	0	0	0
	ICL index	**100**	0	0	0	0	0	0
Scenario D	BIC index	0	0	**100**	0	0	0	0
	ICL index	0	0	**100**	0	0	0	0
Scenario E	BIC index	0	4	**96**	0	0	0	0
	ICL index	0	4	**96**	0	0	0	0

Table 1: *Frequencies of the selected numbers (K) of clusters using both BIC and ICL index for the five scenarios - from A to E - over* 100 *replicated datasets.*

2 Mandibular shape changes

We consider a dataset which consists of Computed Tomography (CT) images processed to obtain 3D surface meshes of the human mandible. Understanding the morphological changes of the mandible is important for orthodontic treatment planning. To decide on a proper treatment plan, the events of growth and development must be correlated with the maturational level of individuals. In particular, the biological question of interest is to understand when and where there exist localized differences between the maturational levels. The following two figures shows the means shapes of groups 2 and 3 (Figure 12) and groups 3 and 4 (Figure 13) estimated by means of the FPC-L_1 procedure.

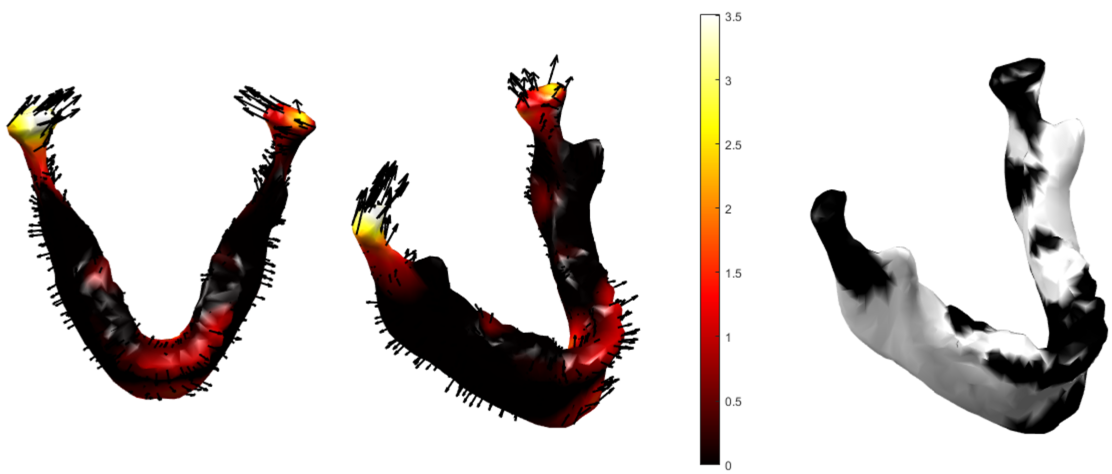

Figure 12: *Pairwise local differences between the mean displacements of groups 2 and 3. The differences are displayed by arrows and colors indicate their lengths in mm. Longer arrows imply larger mean displacement. The arrows have been blown up by a factor of 5 for clarity. On the right, regions of significant changes between the two groups are coloured in black and correspond to point-wise confidence intervals which do not include zero.*

References

A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society, Series B, 61:579–602, 1999.

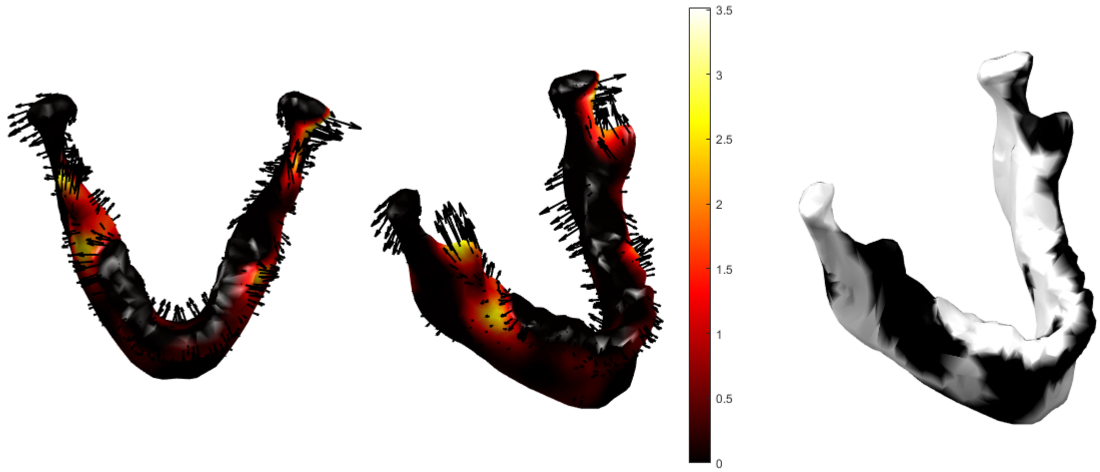

Figure 13: *Pairwise local differences between the mean displacements of groups 3 and 4. The differences are displayed by arrows and colors indicate their lengths in mm. Longer arrows imply larger mean displacement. The arrows have been blown up by a factor of 5 for clarity. On the right, regions of significant changes between the two groups are coloured in black and correspond to point-wise confidence intervals which do not include zero.*

I. Daubechies. Orthonormal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics, 41:909–996, 1999.