Simulation study

Contents

A simulation study was conducted to evaluate the clustering performance of multivariate projection-based methods such as PCA, and the ability to interpolate time points in LMMS.

Twenty reference time profiles were generated on 9 equally spaced time points and assigned to 4 clusters (5 profiles each). These ground truth profiles were then used to simulate new profiles. We generated 500 simulated datasets.

Clustering performance. We first compared profiles simulated then modelled with or without LMMS:

For each of the reference profiles, 5 new profiles (corresponding to 5 individuals) were sampled to reflect some inter-individual variability as follows: Let x be the observation vector for a reference profile r, r = 1 . . . 20, for each time point t (t = 1, . . . , 9), 5 measurements were randomly simulated from a Gaussian distribution with parameters µ = xt,r and σ^2, where $\sigma = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, 3$ to vary the level of noise. This noise level was representative of the data described below in Section gut microbiota development. The profiles from the 5 individuals were then modelled with LMMS (section 3.2.1, resulting in 500 matrices of size (9 × 20).

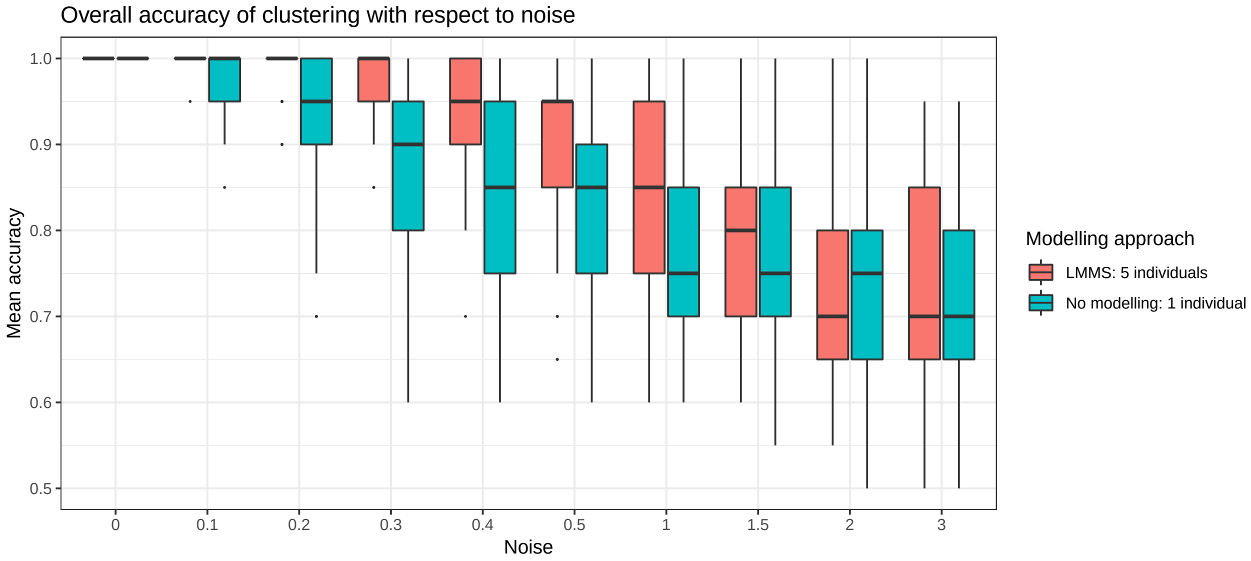
For each of the reference profiles, one new profile was simulated as described in step A, but no LMMS modelling step was performed, resulting in 500 matrices of size (9 × 20).

Clustering was obtained with PCA and compared to the reference cluster assignments in a confusion matrix. The clustering was evaluated by calculating the accuracy of assignment Embedded Image from the confusion matrix, where for a given cluster, TP (true positive) is the number of profiles correctly assigned in the cluster, FN (false negative) is the number of profiles that have been wrongly assigned to another cluster, TN (true negative) is the number of profiles correctly assigned to another cluster and FP (false positive) is the number of profiles incorrectly assigned to this cluster. Besides accuracy, we also calculated the Rand index (Rand, 1971) as a similarity metric to the clustering performance of PCA. The clustering results from fPCA were poor, even for a low level of noise (Suppl. Figure S6), thus fPCA was not compared against PCA.

Interpolation of missing time points.

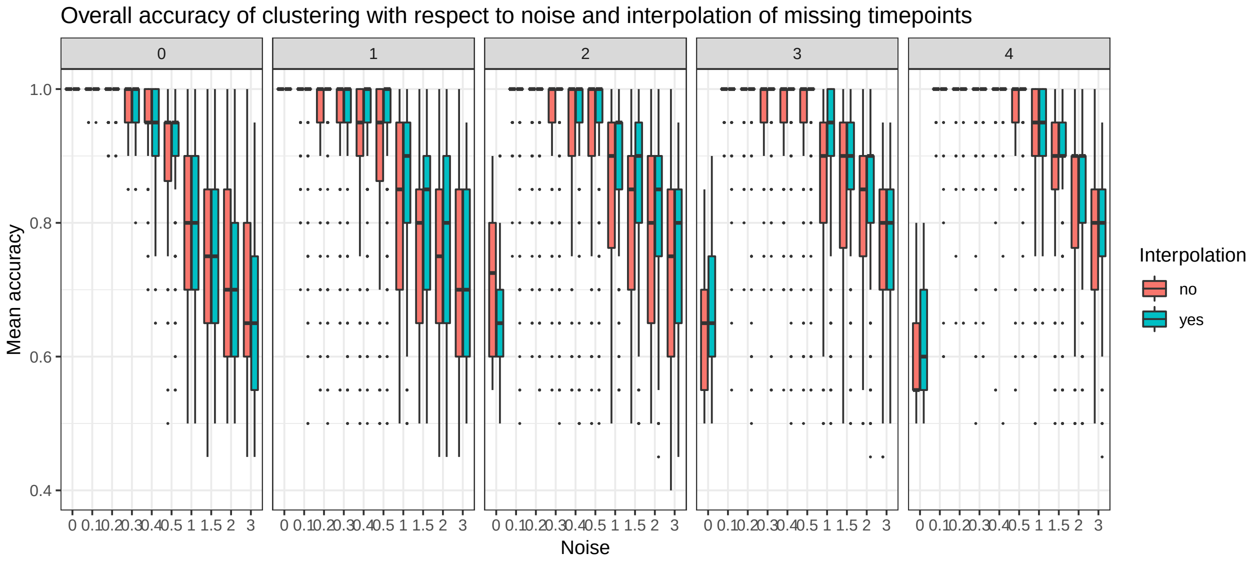
To evaluate the ability of LMMS to predict the value of a missing time point for a given feature over time, we randomly removed 0 to 4 measurement points in the simulated datasets described above in step A. We compared the PCA clustering performance with or without LMMS interpolation.

1 Performance



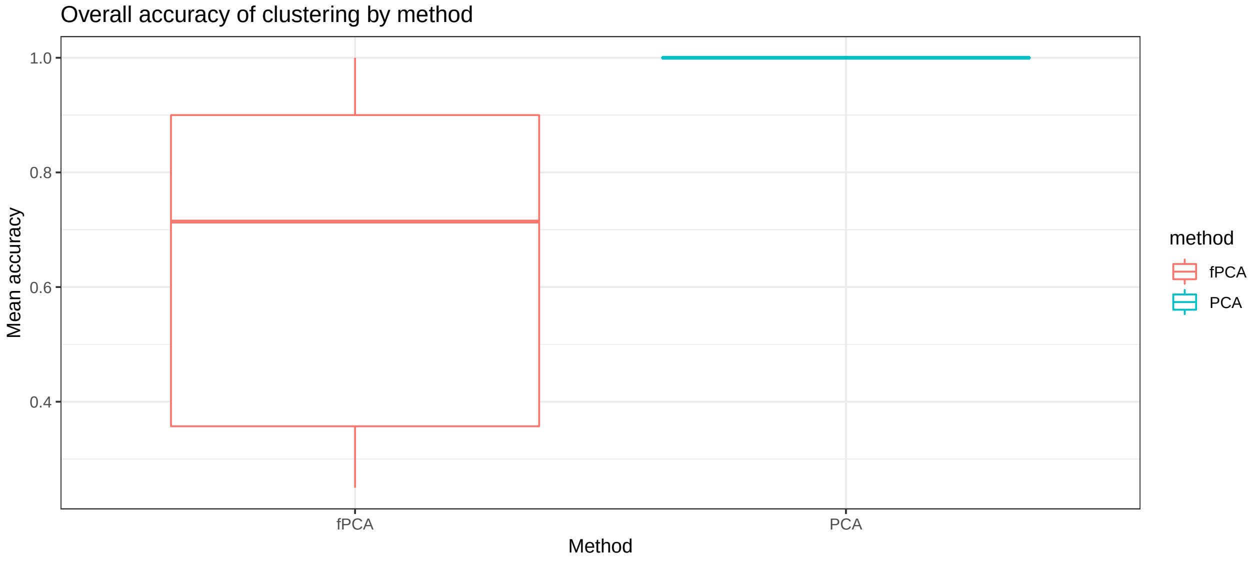
Overall accuracy of clustering with respect to noise. Twenty reference profiles, grouped into 4 clusters were used as a basis for simulation and each of the new simulated profiles were generated with random noise. We compared two approaches: with LMMS modelling: 5 new profiles were generated per reference, and without modelling: only one profile was simulated per reference. We evaluated the ability of PCA clustering to correctly assign the simulated profiles in their respective reference clusters based on mean accuracy: without noise, both approaches lead to a perfect clustering, with Noise < 1, LMMS modelling acts as a denoising process with better performance than no modelling, and with a high level of noise ≥ 1 the performance of both approaches decrease.

2 Interpolation / No interpolation



Overall accuracy of clustering when time points are missing. The simulation scheme is described in 3.7.1, however, here some time points were removed. We compared the ability of LMMS to interpolate missing time points. When there are no time points missing, both interpolated and non-interpolated approaches gave a similar performance. When the number of time points increases, the classification accuracy decreases. Without noise and with several timepoints removed, LMMS tended to model straight lines, resulting in poor clustering..

3 fPCA / PCA


Overall accuracy of clustering by method
Mean accuracy
fPCA
PCA
Method
method
fPCA
PCA