A brief description of the phoneme data

This dataset is a part of the original one which can be found at *http://www-stat.stanford.edu/ElemStatLearn*. We observe $n = 2000$ pairs $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,n}$ where the $\boldsymbol{x}_i$'s correspond to the discretized log-periodograms ($\boldsymbol{x}_i = (\chi(f_1), \chi(f_2), \ldots, \chi(f_{150}))$ is the ith discretized functional data) whereas the y_i's give the class membership (five phonemes):

$$y_i \in \{1,2,3,4,5\} \text{ with } \begin{cases} 1 & \longleftrightarrow \text{ “sh”} \\ 2 & \longleftrightarrow \text{ “iy”} \\ 3 & \longleftrightarrow \text{ “dcl”} \\ 4 & \longleftrightarrow \text{ “aa”} \\ 5 & \longleftrightarrow \text{ “ao”} \end{cases}$$

The phoneme dataset “npfda-phoneme.dat” contains the pairs $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,2000}$ and is organized as follows:

	Col 1	$\cdots$	Col j	$\cdots$	Col 150	Col 151
Row 1	$\chi_1(f_1)$	$\cdots$	$\chi_1(f_j)$	$\cdots$	$\chi_1(f_{150})$	y_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Row i	$\chi_i(f_1)$	$\cdots$	$\chi_i(f_j)$	$\cdots$	$\chi_i(f_{150})$	y_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Row 2000	$\chi_{2000}(f_1)$	$\cdots$	$\chi_{2000}(f_j)$	$\cdots$	$\chi_{2000}(f_{150})$	y_{2000}

The first 150 columns correspond to the 150 frequencies whereas the last column contains the categorical responses (class number). Note that the size of each class is the same (400).