Discriminating spectrometric curves from the fat content

- **Statistical aim**

  We recall that for each unit i (among 215 pieces of finely chopped meat), we observe one spectrometric curve ($\boldsymbol{x}_i$) which corresponds to the absorbance measured at 100 wavelengths (i.e. $\boldsymbol{x}_i = (\chi_i(\lambda_1), \ldots, \chi_i(\lambda_{100}))$. Moreover, for each unit i, we have at hand its fat content y_i obtained by an analytical chemical processing. The file "npfda-spectrometric.dat" contains the pairs $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,215}$. But, in a discrimination setting, we have to consider a categorical response instead of a scalar one. Therefore, the observed responses $y_1, \ldots, y_{215}$ (column 101) are replaced with $y_1^*, \ldots, y_{215}^*$ where

$$\forall i = 1, \ldots, 215, \quad y_i^* = \begin{cases} 1 \text{ if } y_i < 20 \\ 2 \text{ else.} \end{cases}$$

  The curves in the group labelled "1" (resp. "2") correspond to a fat content smaller (larger) than 20 %. Given a new sctrometric curve $\boldsymbol{x}$, our main task is to predict the corresponding class membership.

- **Measuring performance**

  For measuring the performance of the functional discrimination procedure, we follow the same methodology than the one used in the speech recognition problem. More precisely, we built 50 learning and testing samples (the ratio between groups being preserved) which allow us to get 50 misclassification rates. The smooth shape of the curves allows to use semi-metrics based on the derivatives. We give here directly the results with the semi-metric d_2^{deriv}.

Remark: the commandlines for R or S+ are the same.

- **Entering spectrometric data**

```
SPECDAT <- as.matrix(read.table("npfda-spectrometric.dat"))
attributes(SPECDAT)$dimnames[[1]]_character(0)
SPECURVES <- SPECDAT[,1:100]           # sample of curves
Dichotomous <- SPECDAT[,101]           # sample of scalar responses
Dichotomous[Dichotomous < 20] <- 1     # Class 1: fat content < 20%
Dichotomous[Dichotomous >= 20] <- 2    # Class 2: fat content >= 20%
```

- **Computing predicted class number over 50 samples**

```
Misclas.lcv.of.spec.over.50.samples <- 0
subject1 <- (1:215)[Dichotomous==1]
subject2 <- (1:215)[-subject1]
for(i in 1:50){
        set.seed(sample(0:1000,1))
        learn1 <- sample(subject1,77)
        learn2 <- sample(subject2,43)
        learning <- c(learn1,learn2)
        testing <- (1:215)[-learning]
        SPECURVESL <- SPECURVES[learning,]
        SPECURVEST <- SPECURVES[testing,]
        Specdichl <- Dichotomous[learning]
        Specdicht <- Dichotomous[testing]
        res <- funopadi.knn.lcv(Specdichl,SPECURVESL,SPECURVEST,
          2,20,c(0,1),kind.of.kernel = "quadratic",semimetric="deriv")
        Misclas.lcv.of.spec.over.50.samples[i] <-
          sum(res$Predicted.classnumber != Specdicht)/length(Specdicht)
}
```

- **Plotting misclassification rates over 50 runs**

The following commandline allows to obtain Figure 1 which summarizes the results. The semi-metric d_2^{deriv} leads clearly to good discrimination (the mean of the misclassification rates equals to 2 %).

```
boxplot(Misclas.lcv.of.spec.over.50.samples, ylab="Misclassification rates",
cex=1)
```

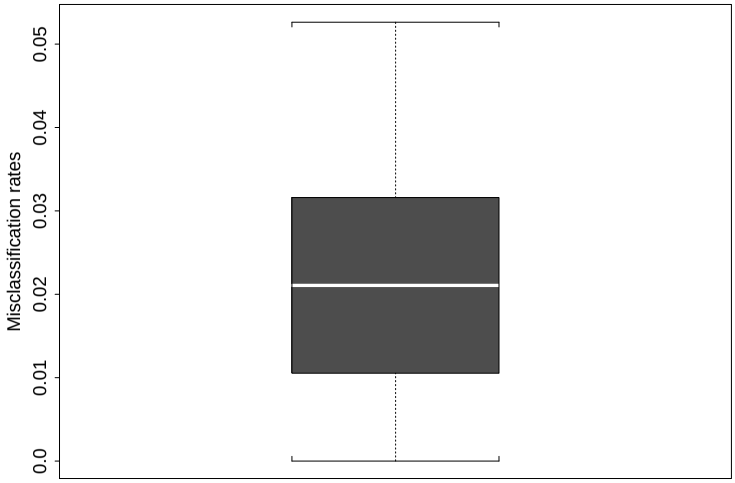

Figure 1: Results over 50 runs