Exam

Nonparametric Statistics, AY 2022/23

July 06, 2023

Instructions

- For all computations based on permutation/bootstrapping, use $B = 1000$ replicates, and $seed = 2022$ every time a permutation/bootstrap procedure is run.
- For Full Conformal prediction intervals, use a regular grid, where, for each dimension, you have $N = 20$ equispaced points with lower bound $\min(data) - 0.25 \cdot range(data)$ and upper bound $\max(data) + 0.25 \cdot range(data)$. Moreover, do not exclude the test point when calculating the conformity measure. Be advised that, except for the number of points, these are the default conditions of the `ConformalInference` R package.
- Both for confidence and prediction intervals, as well as tests, if not specified otherwise, set $\alpha = 0.05$.
- When reporting univariate confidence/prediction intervals, always provide upper and lower bounds.
- For solving the exam, you must use one of the templates previously provided and available here. Particularly, **for each question** you are required to report:
  - *Synthetic description of assumptions, methods, and algorithms*: which methodological procedure you intend to use to answer the question, succinctly describing the main theoretical characteristics of the chosen approach, and why it is suitable for the analytical task at hand,
  - *Results and brief discussion.* the actual result of the procedure applied to the data at hand, including any requested comment, output and plot.
- Data for the exam can be found at this link.

Exercise 1

An Irish farmer, Andrew O'Cappor, owns $N = 370$ cows and he aims at becoming the best quality milk-maker in the whole Ireland. Knowing that he must rely on the most sophisticated analytical techniques to achieve his goal he collects $N = 370$ milk samples, one for each cow, measuring two milk quality traits, namely κ-casein (grams per liter) and Milk pH. The resulting samples are contained in the `df_1.Rds` file. His first aim is to identify whether some cows in the herd produce anomalous milk, he therefore asks you to:

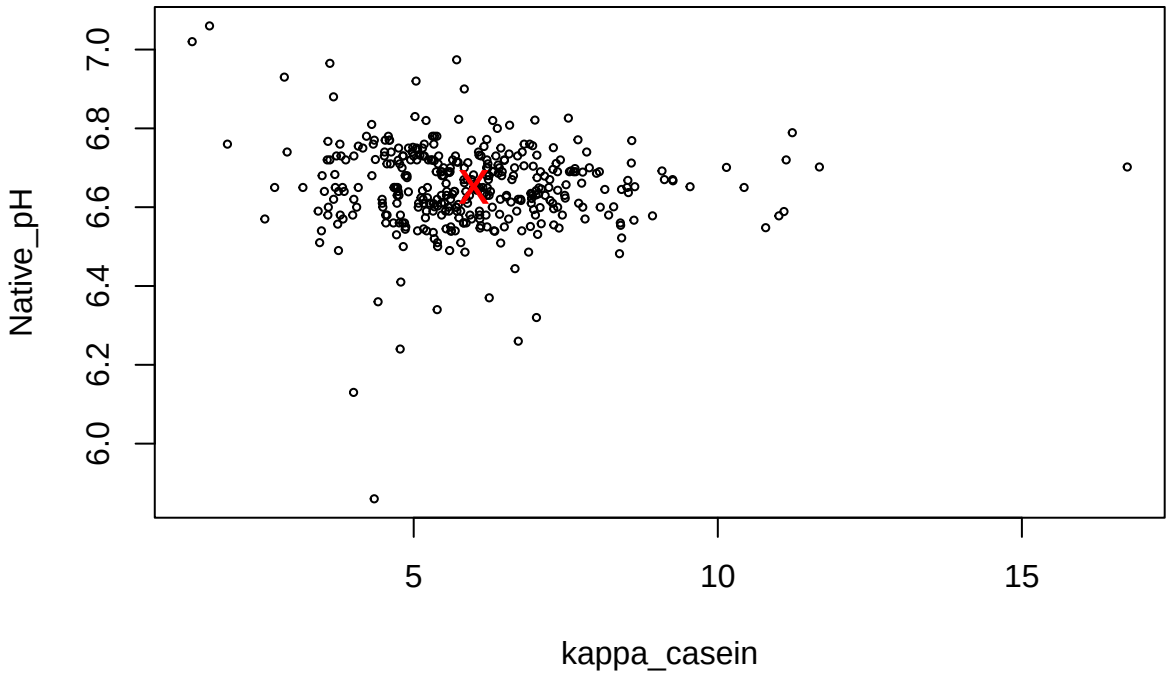
1. Provide the sample Mahalanobis median of the milk samples and superimpose it to the scatterplot of κ-casein vs Native pH

```
df_1 <- readRDS(here("2023-07-06/data/df_1.Rds"))
N <- nrow(df_1)
p <- ncol(df_1)
(maha_median <- depthMedian(df_1,depth_params = list(method='Mahalanobis')))
```

```
## kappa_casein    Native_pH
##      5.98748      6.65100
```

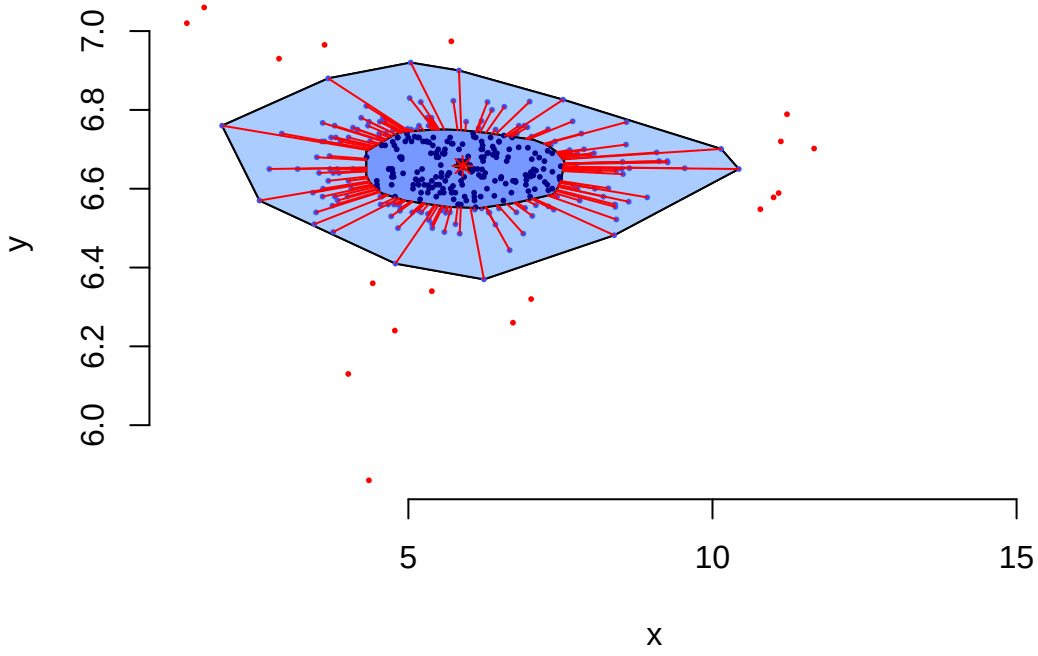
```
plot(df_1, cex = .5)
points(
```

```
  x = maha_median[1],
  maha_median[2],
  col = "red",
  cex = 2,
  pch = "x"
)
```



2. Provide a bagplot of the collected variables, determining a vector of row indexes identifying the milk samples that are outliers according to the procedure.

```
bagplot_milk <- aplpack::bagplot(df_1)
```

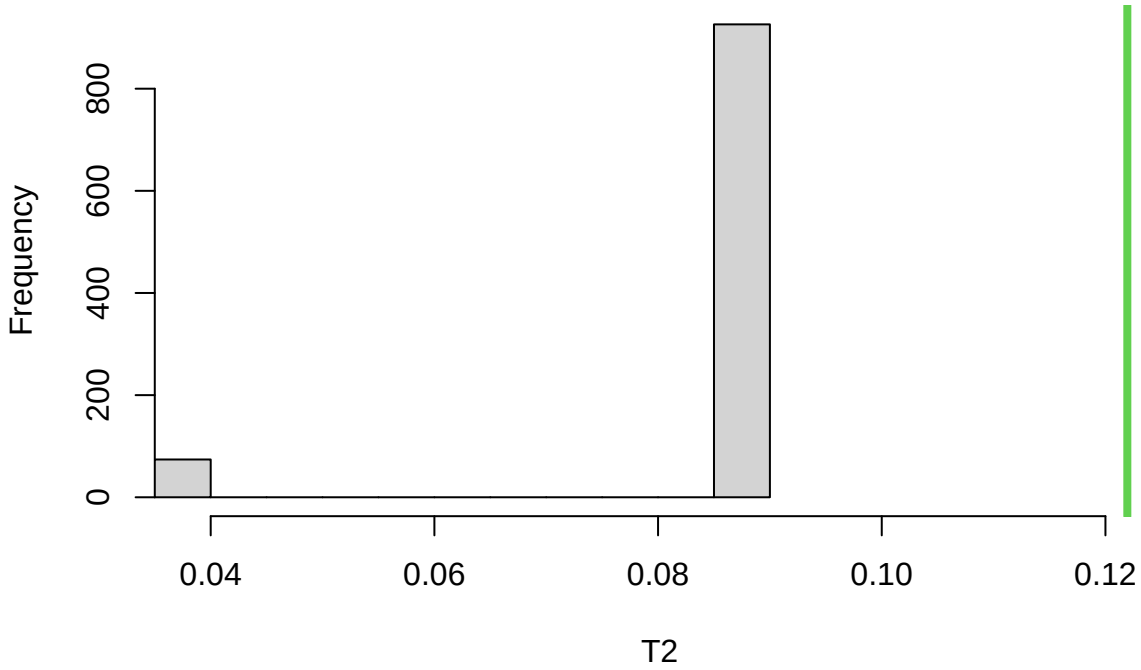


```r
ind_out <-
  which(apply(df_1, 1, function(x)
    all(x %in% bagplot_milk$pxy.outlier)))
ind_out
```

```
##  [1]   1   2   6  19  36  47  50  80 139 179 272 293 362 363 364 365 366 367 368
```

3. Test whether the 370 milk samples comply with the gold standard in terms of milk quality, for which κ-casein must be equal to 6 grams per liter and Milk pH to 7. Perform a permutation test using as test statistic the squared Euclidean distance between the sample Mahalanobis median and the gold standard. Provide the histogram of the permuted distribution of the test statistic and the p-value, commenting the results.

```r
standard_cow <- c(6,7)

T_20 <- norm(maha_median-standard_cow,"2")^2

B <- 1000
T2 <- numeric(B)

set.seed(2022)

pb=progress::progress_bar$new(total=B, format = " Processing [:bar] :percent eta: :eta")

for(perm in 1:B) {
  # Permuted dataset
  signs.perm <- rbinom(N, 1, 0.5) * 2 - 1
  df_perm <-
  sweep(sweep(df_1,2,standard_cow) * matrix(signs.perm,nrow=N,ncol=p,byrow=FALSE),2,-standard_cow)
  x.maha_perm <- depthMedian(df_perm,depth_params = list(method='Mahalanobis'))
  T2[perm]  <- norm(x.maha_perm - standard_cow, type = "2") ^ 2
  pb$tick()
}

hist(T2,xlim=range(c(T2,T_20)))
abline(v=T_20,col=3,lwd=4)
```



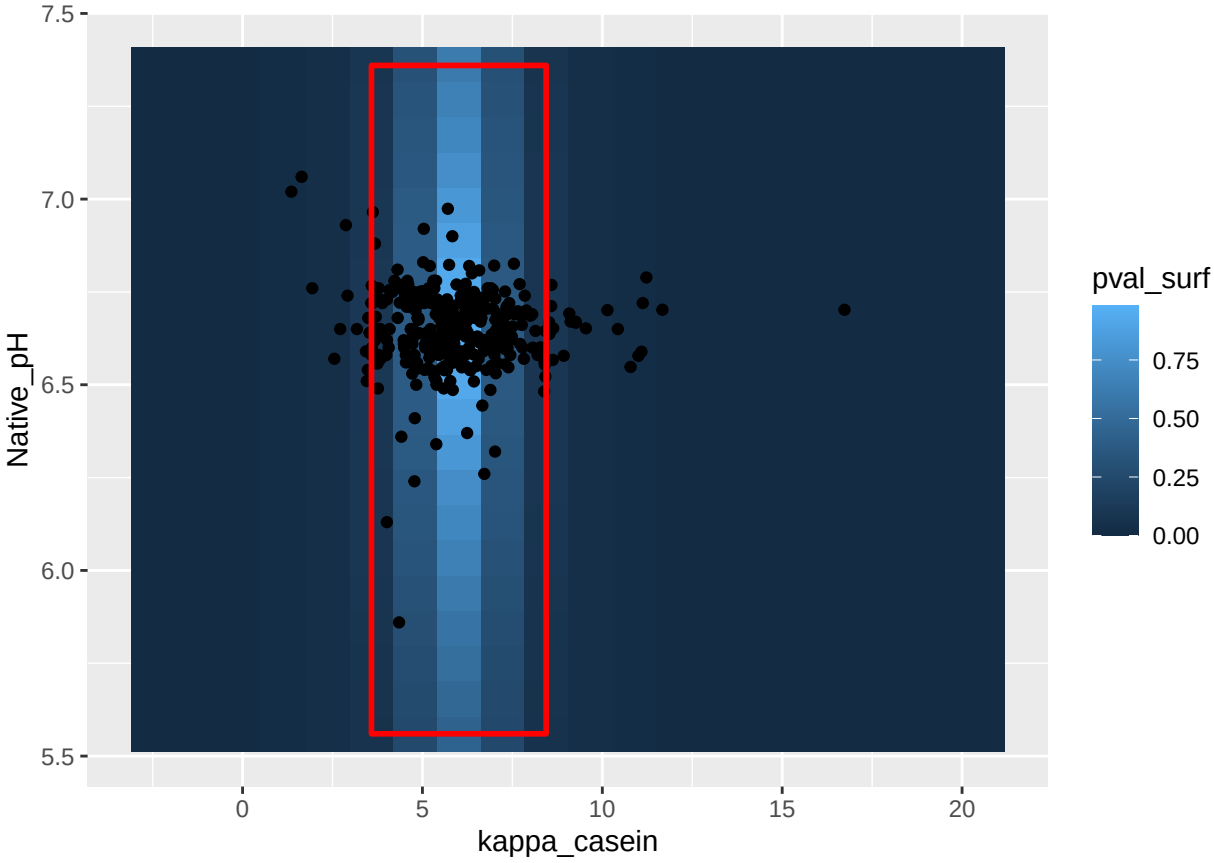

```r
# p-value
p_val <- sum(T2>=T_20)/B
p_val
```

```
## [1] 0
```

4. Provide a Full Conformal $1-\alpha = 95\%$ prediction region for the κ-casein and Milk pH of a new milk sample, using the Euclidean distance between the new data point and the sample Mahalanobis median of the augmented data set as non-conformity measure, providing a plot of it

```r
data_predict = df_1
n_grid = 20
grid_factor = 0.25
alpha = .05
n = nrow(data_predict)

range_x = range(data_predict[, 1])[2] - range(data_predict[, 1])[1]
range_y = range(data_predict[, 2])[2] - range(data_predict[, 2])[1]


test_grid_x = seq(
  min(data_predict[, 1]) - grid_factor * range_x,
  max(data_predict[, 1]) + grid_factor * range_x,
  length.out = n_grid
)
test_grid_y = seq(
  min(data_predict[, 2]) - grid_factor * range_y,
  max(data_predict[, 2]) + grid_factor * range_y,
  length.out = n_grid
)
xy_surface = expand.grid(test_grid_x, test_grid_y)
```

```r
colnames(xy_surface) = colnames(data_predict)

wrapper_multi_conf = function(test_point) {
  newdata = rbind(test_point, data_predict)
  newmedian = depthMedian(newdata, depth_params = list(method = 'Mahalanobis'))

  norm_tmp <- sweep(newdata,
                    MARGIN = 2,
                    STATS = newmedian,
                    FUN = "-")
  depth_surface_vec <- apply(norm_tmp, 1, norm, type = "2")
  sum(depth_surface_vec[-1] >= depth_surface_vec[1]) / (n + 1)
}


pval_surf = pbapply::pbapply(xy_surface, 1, wrapper_multi_conf)
data_plot = cbind(pval_surf, xy_surface)


p_set = xy_surface[pval_surf > alpha, ]
poly_points = p_set[chull(p_set), ]

ggplot() +
  geom_raster(data = data_plot, aes(kappa_casein, Native_pH, fill = pval_surf)) +
  geom_point(data = data.frame(data_predict), aes(kappa_casein, Native_pH)) +
  geom_polygon(
    data = poly_points,
    aes(kappa_casein, Native_pH),
    color = 'red',
    size = 1,
    alpha = 0.01
  )
```



Exercise 2

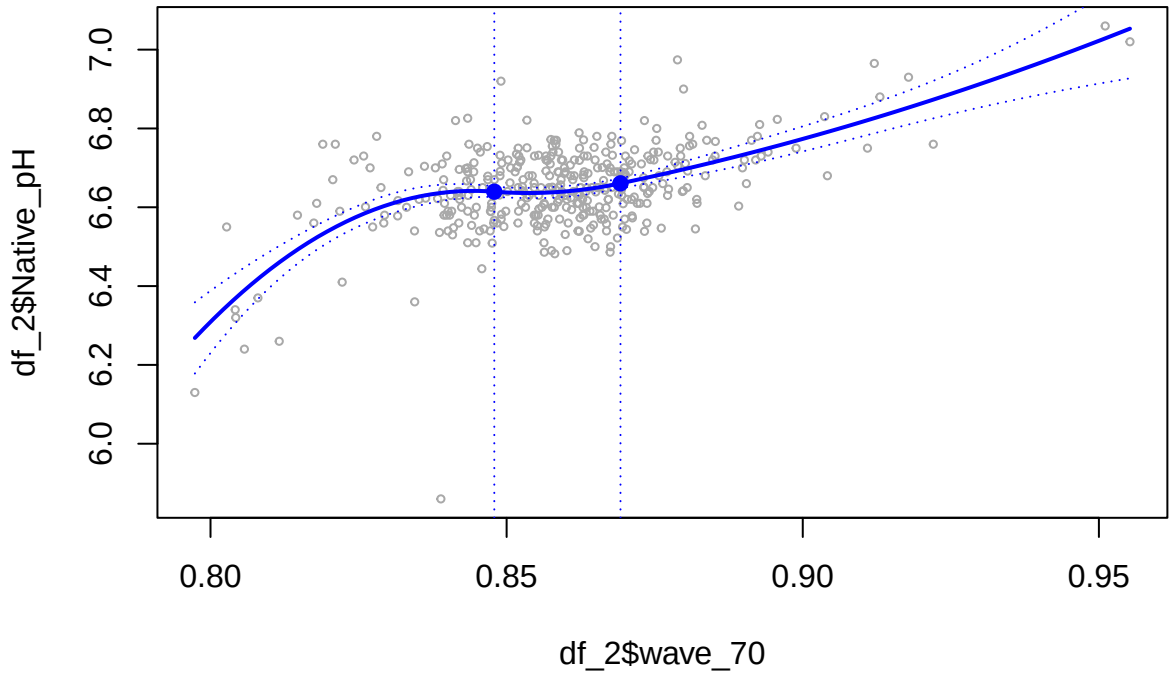
A friend of Andrew O'Cappor, Dr. Alexander House, statistician and chemometrician, advises the farmer that spectroscopy is the state-of-the-art technology to employ when it comes to evaluate milk quality. Motivated by this, and advised by Dr House, Andrew O'Cappor is interested in building a nonparametric model to predict Milk pH by means of the absorbance values at wavenumbers 70 and 300 cm^{-1}, contained in the `df_2.Rds` file. He therefore asks you to:

1. Build a degree 2 b-spline model to regress Milk pH on the milk absorbance at wavenumber $70cm^{-1}$. Set knots at the first and third quartiles of the explanatory variable. Provide a plot of the regression line with standard errors for the prediction, a table summarizing the coefficients and comment the results.

```
df_2 <- readRDS(here("2023-07-06/data/df_2.Rds"))

knots <- quantile(df_2$wave_70,probs = c(.25, .75))
fit_spline <- lm(Native_pH~bs(wave_70,knots = knots,degree = 2),
                 data = df_2)

knitr::kable(broom::tidy(summary(fit_spline)))
```

term	estimate	std.error	statistic	p.value
(Intercept)	6.2684164	0.0451687	138.777897	0
bs(wave_70, knots = knots, degree = 2)1	0.4007899	0.0618274	6.482399	0
bs(wave_70, knots = knots, degree = 2)2	0.3586227	0.0439991	8.150685	0
bs(wave_70, knots = knots, degree = 2)3	0.5283311	0.0607857	8.691701	0
bs(wave_70, knots = knots, degree = 2)4	0.7848049	0.0760754	10.316143	0

```
new_data_seq <-
  seq(min(df_2$wave_70),
      max(df_2$wave_70),
      length.out = 100)

preds=predict(fit_spline, newdata = list(wave_70=new_data_seq),se=T)
se.bands=cbind(preds$fit +2* preds$se.fit ,preds$fit -2* preds$se.fit)

plot(y=df_2$Native_pH,x=df_2$wave_70  ,cex =.5, col =" darkgrey ")
lines(new_data_seq,preds$fit ,lwd =2, col =" blue")
matlines(new_data_seq, se.bands ,lwd =1, col =" blue",lty =3)

knots_pred=predict(fit_spline,list(wave_70=knots))
points(knots,knots_pred, col='blue',pch=19)
abline(v = knots, lty=3, col="blue")
```



2. Build an additive model for regressing Milk pH on the milk absorbance at wavenumbers $70cm^{-1}$ and $300cm^{-1}$ using penalized regression splines. Write in proper mathematical form the additive model you estimate, report the adjusted R^2 and the parametric p-values of the tests. Comment on the results.

```
fit_gam <-
  gam(Native_pH ~ s(wave_70, bs = "cr") + s(wave_300, bs = "cr"), data = df_2)

table_fit_gam <- summary(fit_gam)
(r_2_squared <- table_fit_gam$r.sq)
```

```
## [1] 0.3458303
```

```
table_fit_gam$p.pv
```

```
## (Intercept)
##           0
```

```
table_fit_gam$s.table
```

```
##                   edf   Ref.df         F     p-value
## s(wave_70)  6.092183 6.893283 11.189118 0.000000000
## s(wave_300) 2.494018 3.286952  4.760114 0.002208255
```

3. Build a semi-parametric model for regressing Milk pH on the milk absorbance at wavenumbers $70cm^{-1}$ and $300cm^{-1}$, assuming the effect of wavenumber $300cm^{-1}$ to be linear on the response. Using a bootstrap approach, provide a reverse percentile confidence interval for the parametric effect of level 95%.

```
fit_semi <- gam(Native_pH~s(wave_70,bs = "cr")+wave_300,data = df_2)

fitted.obs <- fit_semi$fitted.values
res.obs <- fit_semi$residuals

wave300_obs = summary(fit_semi)$p.table[2,1]
T.boot.wave300 = numeric(B)

set.seed(2022)

for(b in 1:B) {


  Native_pH_boot <- fitted.obs + sample(res.obs, replace = T)
  fit_semi_boot <-
    gam(Native_pH_boot ~ s(df_2$wave_70, bs = "cr") +
          df_2$wave_300)

  fit_semi_boot_table = summary(fit_semi_boot)

  T.boot.wave300[b] = fit_semi_boot_table$p.table[2,1]

}

alpha <- 0.05
right.quantile.wave300 <- quantile(T.boot.wave300, 1 - alpha/2)
left.quantile.wave300  <- quantile(T.boot.wave300, alpha/2)

CI.RP.wave300 <-
  c(
    wave300_obs - (right.quantile.wave300 - wave300_obs),
    wave300_obs - (left.quantile.wave300- wave300_obs))
names(CI.RP.wave300)=c('lwr','upr')
CI.RP.wave300
```

```
##       lwr       upr
## 0.7406131 3.2112981
```

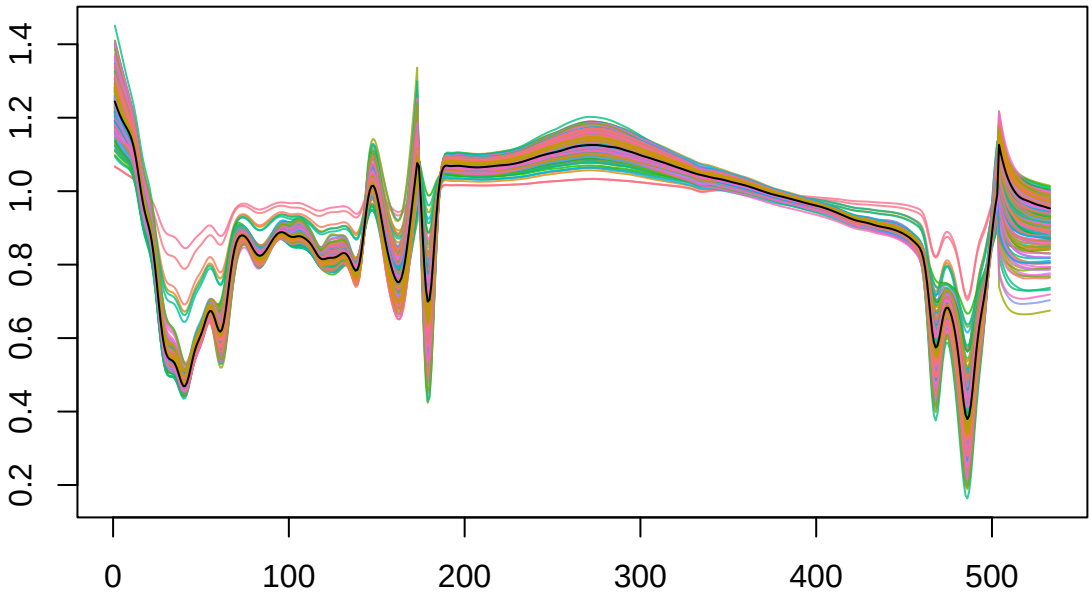
Exercise 3

Andrew O'Cappor has recently become passionate about functional data analysis (FDA): Dr. Alexander House is then ready to unleash the power of the MilkoScan FT6000 milk analyzer, producing MIRS transmittance spectra in the entire mid-infrared light region (wavenumbers 70 and 300 cm^{-1} of the previous exercise were just two points of this region). The MIRS for the $N = 370$ milk samples are contained in the `df_3_A.Rds`

file. Andrew O'Cappor would like to exploit these modern statistical methods to further analyze his milk samples. To this extent, he requires you to:

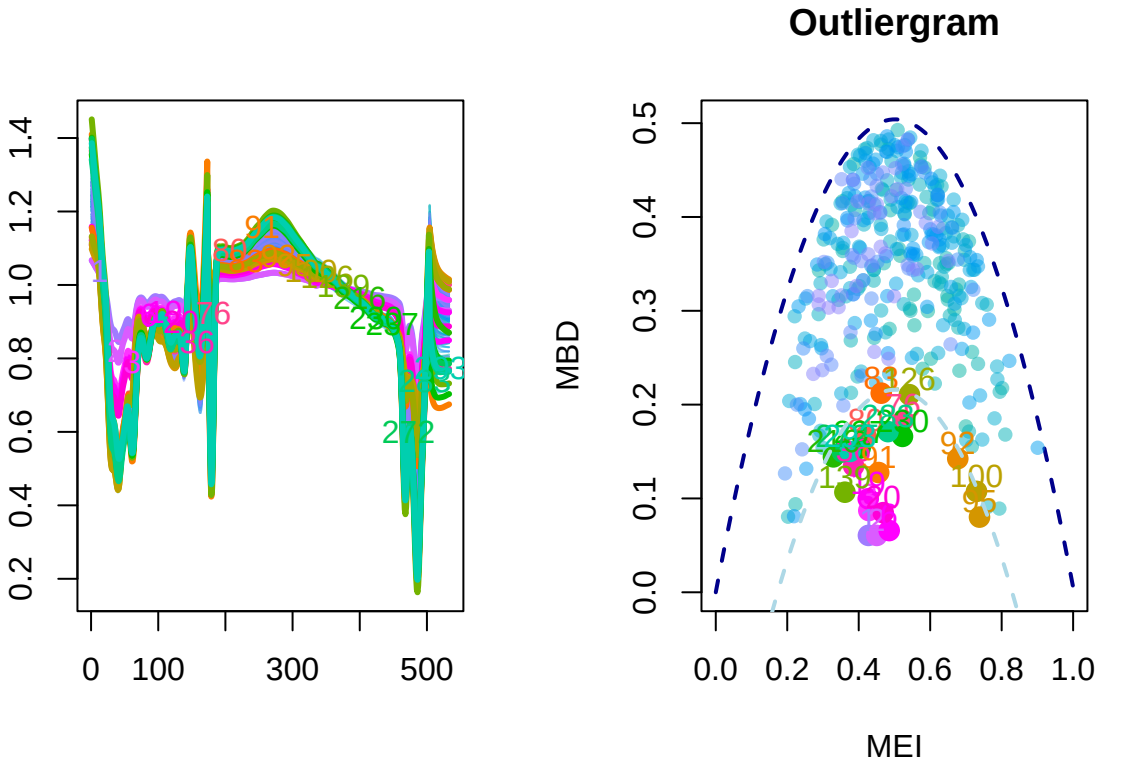
1. By suitably defining a functional data object, plot the resulting spectra for the $N = 370$ milk samples. Then, compute the median spectrum using the modified band depth and superimpose it to the previous plot.

```
df_3_A <- readRDS(here("2023-07-06/data/df_3_A.Rds"))
grid <- 1:ncol(df_3_A)
f_data <- fData(grid,df_3_A)
plot(f_data)

band_depth <- MBD(Data = f_data)
median_curve <- median_fData(fData = f_data, type = "MBD")
plot(f_data)
lines(grid,median_curve$values)
```



2. Produce an outliergram and use it to identify potential outlying curves present in the sample, setting the inflation factor F equal to 1.3. Report a vector of row indexes identifying the samples that are outliers according to the procedure.

```
outgr_plot <- outliergram(f_data,Fvalue = 1.3)
```




```
FDA_out <- c(outgr_plot$ID_outliers)
FDA_out
```

```
##  [1]   1   2   3   6  19  20  36  76  80  83  91  92  95 100 126 139 216 230 237
## [20] 272 283 293
```

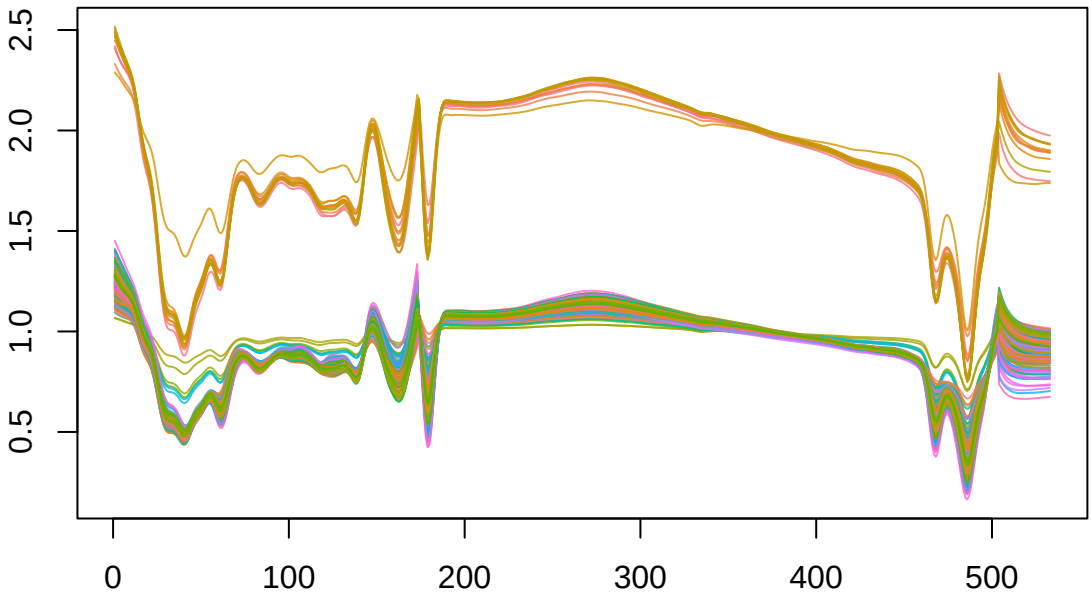
3. Andrew O'Cappor has received 12 extra milk samples from a neighboring farmer, Matthew O'Fountain. These samples originated from O'Fountain's cows and are contained in the `df_3_B.Rds` file. Matthew would like to compute the Relative Modified Band Depth of his milk with respect to O'Cappor samples[1], to figure out whether the milk analyzer he uses is correctly calibrated. Report the Relative Modified Band Depth of the 12 extra milk samples, what can you conclude about the calibration of Matthew's milk analyzer?

```
df_3_B <- readRDS(here("2023-07-06/data/df_3_B.Rds"))
f_data_B <- fData(grid,df_3_B)
MBD_f_data_B <- MBD_relative(Data_target = f_data_B, Data_reference = f_data)
MBD_f_data_B
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0
```

```
plot(append_fData(f_data_B,f_data))
```

[1] the `MBD_relative` function from the `roahd` package may serve the purpose




```
# For sure there is a calibration problem in O'Fountain's milk analyzer
```