Assigment: Non-linear dimensionality reduction.

Principal curves, local MDS, Isomap and t-SNE

Pedro Delicado

PART A. Principal Curves

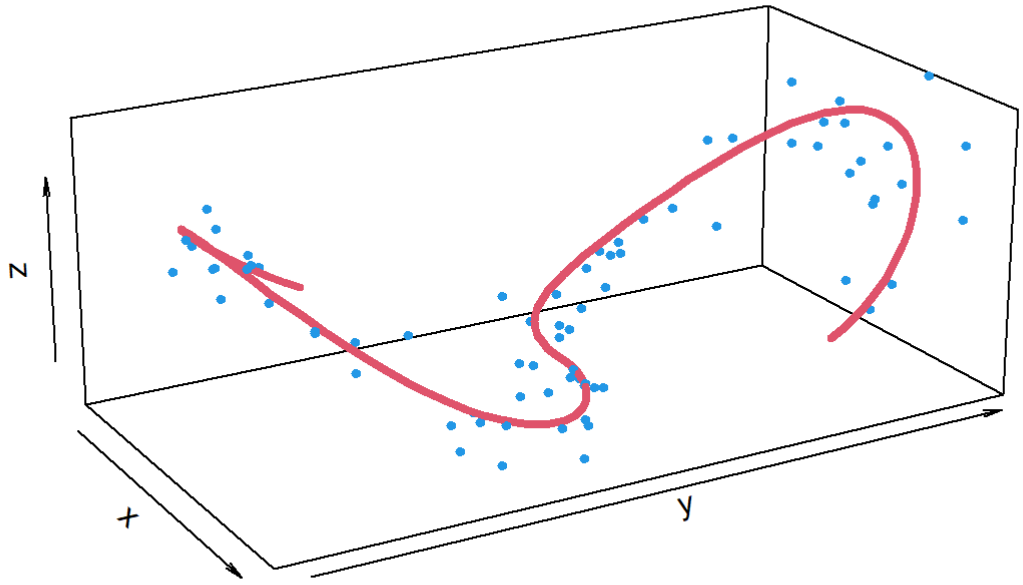
1. Choosing the smoothing parameter in Principal Curves (Hastie and Stuetzle 1989)

Consider the 3-dimensional data set generated by the following code.

```
t <- seq(-1.5*pi,1.5*pi,l=100)
R<- 1
n<-75
sd.eps <- .15

set.seed(1)
y <- R*sign(t) - R*sign(t)*cos(t/R)
x <- -R*sin(t/R)
z <- (y/(2*R))^2
rt <- sort(runif(n)*3*pi - 1.5*pi)
eps <- rnorm(n)*sd.eps
ry <- R*sign(rt) - (R+eps)*sign(rt)*cos(rt/R)
rx <- -(R+eps)*sin(rt/R)
rz <- (ry/(2*R))^2 + runif(n,min=-2*sd.eps,max=2*sd.eps)
XYZ <- cbind(rx,ry,rz)


require(plot3D)
lines3D(x,y,z,colvar = NULL,
         phi = 20, theta = 60, r =sqrt(3), d =3, scale=FALSE,
         col=2,lwd=4,as=1,
         xlim=range(rx),ylim=range(ry),zlim=range(rz))
points3D(rx,ry,rz,col=4,pch=19,cex=.6,add=TRUE)
```



When fitting principal curves to these data, use the function `princurve::principal_curve` with the following options:

- `smoother="smooth_spline"` . This is the default, so you do not need to use it explicitely.
- The only additional argument that you will pass to `smooth_spline` will be the *degrees of freedom* `df` (see `help(smooth.spline)` if you want)

For instance, the following sentence

```
principal_curve(XYZ, df=6)
```

fits the required principal curve with degrees of freedom `df` equal to 6.

Questions

a. Choose the value of the degrees of freedom `df` by leave-one-out cross-validation.
Restrict the search of `df` to `seq(2,8,by=1)` .
(Hint: The function `project_to_curve` should be used. See the element `dist` of the object it returns).

b. Give a graphical representation of the principal curve output for the optimal `df` and comment on the obtained results.

c. Compute the leave-one-out cross-validation error for `df=50` and compare it with the result corresponding to the optimal `df` value you found before.

- Before fitting the principal curve with `df=50` and based only on the leave-one-out cross-validation error values, what value for `df` do you think that is better, the previous optimal one or `df=50` ?
- Fit now the principal curve with `df=50` and plot the fitted curve in the 3D scatterplot of the original points.
- Now, what value of `df` do you prefer?

- The overfitting with `df=50` is clear. Nevertheless leave-one-out cross-validation has not been able to detect this fact. Why do you think that `df=50` is given a so good value of leave-one-out cross-validation error?

PART B. Local MDS, ISOMAP and t-SNE

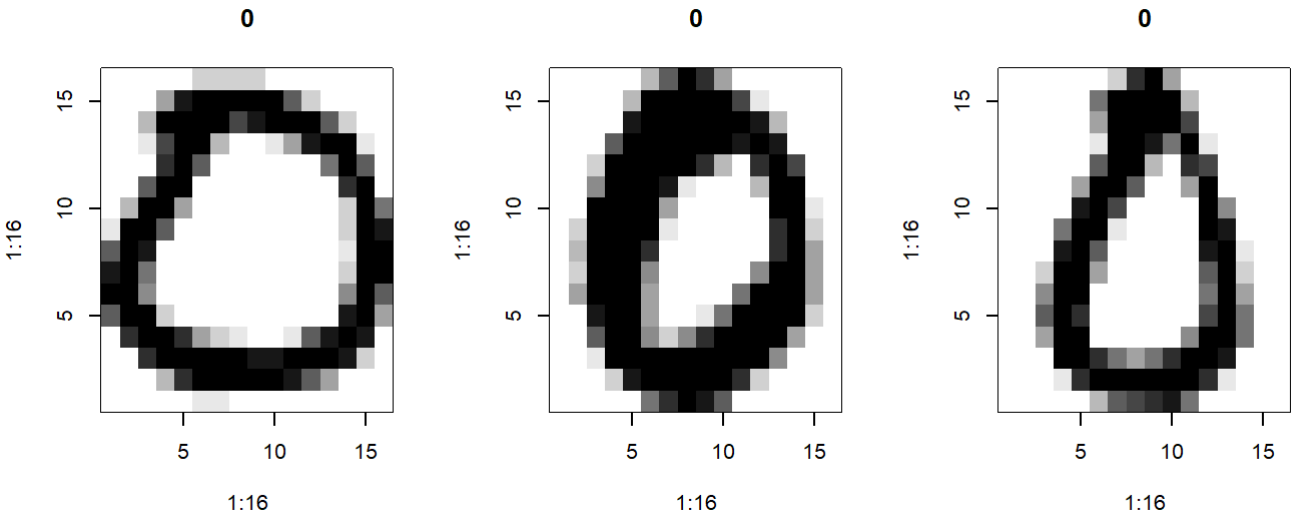
Consider the ZIP number data set, from the book of Hastie et al. (2009). Read the training data set (in the file `zip.train`) and select only the *ZEROs*.

There are $n = 1194$ digits corresponding to ZEROs in the data set.

The following function plots a digit:

```
# ploting 1 digit
plot.zip <- function(x,use.first=FALSE,...){
  x<-as.numeric(x)
  if (use.first){
    x.mat <- matrix(x,16,16)
  }else{
    x.mat <- matrix(x[-1],16,16)
  }
  image(1:16,1:16,x.mat[,16:1],
        col=gray(seq(1,0,l=12)),...)
  invisible(
    if (!use.first){
      title(x[1])
    }else{
    }
  )
  #col=gray(seq(1,0,l=2)))
}
```

Here you have several examples of these ZERO digits:




```
FALSE NULL
```

2. Local MDS for ZERO digits

You must apply Local MDS to reduce the dimensionality of this dataset using the function `lmds` from package `stops`. You have to install the library `stops` from this link (https://rdrr.io/rforge/stops/man/lmds.html) and then to attach the library:

```
if (!require(stops, quietly=TRUE, warn.conflicts=FALSE)){
  install.packages("stops", repos="http://R-Forge.R-project.org",INSTALL_opts="--no-test-load")
}

library(stops)
# help(lmds)
```

a. Look for a 2-dimensional ($q = 2$) configuration of the data using parameters $k = 5$ and $\tau = 0.05$ in `lmds` function. Do the scatterplot of the obtained 2-dimensional configuration.

b. In the previous scatterplot, select a few points (9 points, for instance) located in such a way that they *cover* the variability of all the points in the scatterplot. Then use the function `plot.zip` to plot the ZERO digits corresponding to the selected points. The images you are plotting should allows you to give an interpretation of the 2 coordinates obtained by Local MDS (observe how the shape of ZEROs changes when moving along each directions of the scatterplot).

c. Use the local continuity meta criteria to select the tuning parameters k and τ in Local MDS for ZERO digits. Then describe graphically the low dimensional configuration corresponding to the optimal parameters. *Indication:* As tentative values for k use `c(5,10,50)`, and for τ use `c(.1,.5,1)`.

3. ISOMAP for ZERO digits

a. Look for a 2-dimensional ($q = 2$) configuration of the data using parameter $k = 5$ in function `isomap` from package `vegan`. Do the scatterplot of the obtained 2-dimensional configuration.

b. In the previous scatterplot, select a few points (9 points, for instance) located in such a way that they *cover* the variability of all the points in the scatterplot. Then use the function `plot.zip` to plot the ZERO digits corresponding to the selected points. The images you are plotting should allows you to give an interpretation of the 2 coordinates obtained by ISOMAP (observe how the shape of ZEROs changes when moving along each directions of the scatterplot).

c. Use the local continuity meta criteria to select the tuning parameter k in ISOMAP for ZERO digits. Then describe graphically the low dimensional configuration corresponding to the optimal parameter. *Indication:* As tentative values for k use `c(5,10,50)`.

4. t-SNE for ZERO digits

You must apply t-SNE to reduce the dimensionality of this dataset using the function `Rtsne` from package `Rtsne`.

```
library(Rtsne)
# help(Rtsne)
```

a. Look for a 2-dimensional ($q = 2$) configuration of the data using parameters `perplexity` $= 40$ and `theta` $= 0$ in `Rtsne` function. Do the scatterplot of the obtained 2-dimensional configuration.

b. In the previous scatterplot, select a few points (9 points, for instance) located in such a way that they *cover* the variability of all the points in the scatterplot. Then use the function `plot.zip` to plot the ZERO digits corresponding to the selected points. The images you are plotting should allows you to give an interpretation of the 2 coordinates obtained by t-SNE (observe how the shape of ZEROs changes when moving along each directions of the scatterplot).

c. Use the local continuity meta criteria to select the tuning parameter `perplexity` in t-SNE for ZERO digits (use $q = 2$ and `theta` $= 0$). Then describe graphically the low dimensional configuration corresponding to the optimal parameter.
Indication: As tentative values for `perplexity` use `c(10,20,40)`.

5. Compare Local MDS, ISOMAP and t-SNE for ZERO digits

a. Compare graphically the dimensions of the 2-dimensional configurations you have obtained by Local MDS, ISOMAP and t-SNE for ZERO digits.
Indication: Use the function `pairs` applied to a 6-dimensional matrix.

b. Which method have produced the 2-dimensional configurations with the largest value of the local continuity meta criteria?