Density estimation. GMM. DBSCAN (Assignment)

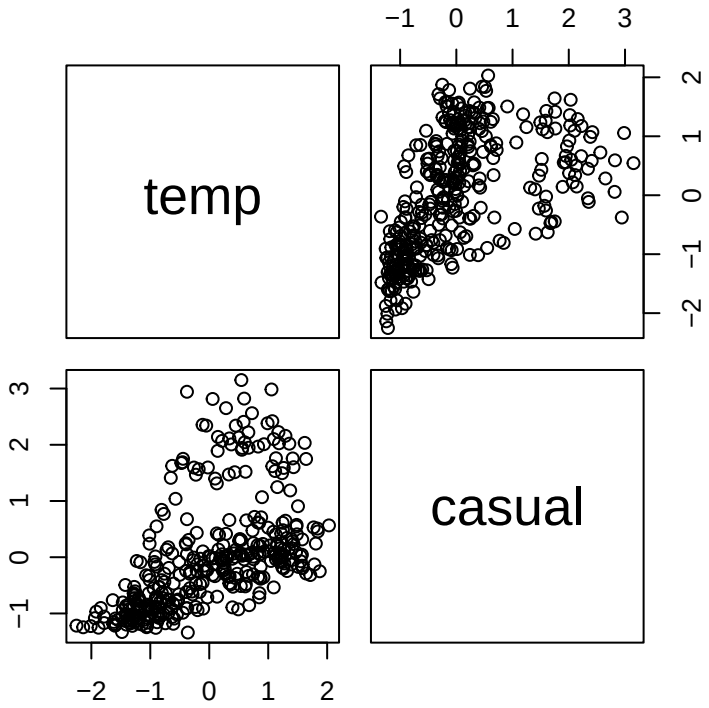
Pedro Delicado

The file `BikeDay.Rdata` contains information on the bike-sharing rental service in Washington D.C., USA, corresponding to years 2011 and 2012. This file contains only one data frame, `day`, with 731 rows (one for each day of years 2011 and 2012, that was a leap year) and 16 columns:

- `instant` row index, going from 1 to 731
- `dteday` date
- `season` (1:springer, 2:summer, 3:fall, 4:winter)
- `yr` year (0: 2011, 1:2012)
- `mnth` 1 for January, until 12 for December
- `holiday` weather day is holiday or not
- `weekday` day of the week (0 Sunday to 6 Saturday)
- `workingday` if day is neither weekend nor holiday is 1, otherwise is 0
- `weathersit`:
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp` Normalized temperature in Celsius. The values are divided to 41 (max)
- `atemp` Normalized feeling temperature in Celsius The values are divided to 50 (max)
- `hum` Normalized humidity. The values are divided to 100 (max)
- `windspeed` Normalized wind speed. The values are divided to 67 (max)
- `casual` count of rental bikes by causal users (not registered)
- `registered` count of rental bikes by registered users.
- `cnt` count of total rental bikes (casual + registered)

In particular we are interested in the joint distribution of the centered and scaled variables `temp` and `casual` for year 2012:

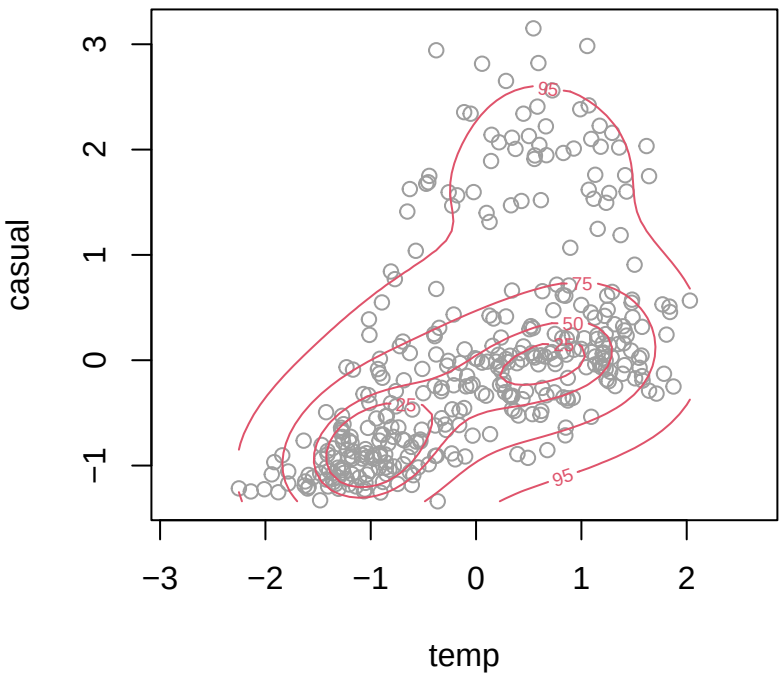
```
load("BikeDay.Rdata")
X <- scale(day[day$yr==1,c(10,14)])
pairs(X)
```



Questions

1. We want to estimate the joint bivariate density of (`temp`,`casual`) using a kernel estimator with the same bandwidth in both dimensions: $h = (a, a)$. For instance, the following code performs this estimation for $a = 0.5$:

```
library(sm)
plot(X,as=1,col=8)
sm.density(X,h=.5*c(1,1),display="slice",
           props=c(25,50,75,95),col=2,add=TRUE)
```



Use the *maximum log-likelihood cross-validation method* for choosing the value of a, when a takes values in the vector `seq(0.05,0.5,by=0.05)`. Then repeat the previous density estimation using the chosen value of a.

2. Use library `mclust` for the following task. Do a model based clustering of these data assuming a Gaussian Mixture Model, allowing varying volume, shape, and orientation for different components in the mixture. Choose the best number of clusters $k \in \{2, \ldots, 6\}$ according to BIC. Plot the resulting object from `Mclust` (do 3 different graphics: `BIC`, `classification`and '`density`).

3. Use library `fpc` to check if it is possible to merge some of the components in the Gaussian Mixture Model previously estimated. Let k^* be the final number of clusters after the merging process. Do the scatterplot of (`temp`,`casual`) with colors according to the new k^* clusters. *Indication:* Use the function `mergenormals` with the option `method="bhat"`.

4. For each one of the k^* clusters obtained above, do the following tasks *(A unique plot should be done, at which the k densities are represented simultaneously)*:

- Consider the bivariate data set of the points in this cluster.
- Estimate non-parametrically the joint density of (`temp`,`casual`), conditional to this cluster *(Use the optimal bandwidth found in the first point).*
- Represent the estimated bivariate density using the level curve that covers the 75% of the points in this cluster.

5. Use DBSCAN to find clusters (and outliers) in the data set (`temp`,`casual`). Try $\varepsilon \in \{0.25, 0.5\}$ and `minPts` $\in \{10, 15, 20\}$. Which combination of the tuning parameters do you consider the *best one*? Compare the DBSCAN clustering corresponding to your favorite combination of tuning parameters with the results of `mergenormals` (print their cross-table).

6. Give an interpretation (or explanation, or description) of the clusters your have found before.