Startup Research

Alessia Caponera, Francesco Denti, Tommaso Rigon, Andrea Sottosanti and Alan Gelfand

Data description

Using the software R, we loaded in memory the raw data, for a total of 2 `data.frame` and 3 `arrays`.

- `D` is a 70, 70, 24, 2 dimensional `array` comprising the 70×70 structural connectivity networks collected for the 24 subjects in each of the 2 scan-rescan imaging session. Each value in `D` is a **count**.
- `Y` is a 70, 404, 24, 2 dimensional `array` comprising the 70×404 multivariate (equally spaced) time-series activity data collected for the 24 subjects of the 2 scan-rescan imagin session. Each value in `Y` is a **real number**. The time lag between observations is 1400ms.
- `W` is a 70, 70, 24, 2 dimensional `array` obtained from `Y`. In particular each `W[v,u,i,k]` $\in (-1, 1)$ is equal to `cor(Y[v„i,k],Y[u„i,k])`, that is, the contemporary cross-correlation between time series for each brain region, subject and scan-rescan. Each value in W is a **number in between** -1 **and** 1.
- `SUBJ` is a 24, 7 dimensional `data.frame` containing information about each subject. For each subject is known the `age`, the `handedness` has been diagnosed a current and/or lifetime mental disorder. **Each row is a subject**.
- `ROI` is a 70, 6 dimensional `data.frame` containing information about brain regions. The `emisphere` (left or right) and the `lobes` (e.g. frontal, occipital, etc.) are reported, as well as the 3-D coordinates of the `centroids` of each region. **Each row is a brain region**.

Missing data

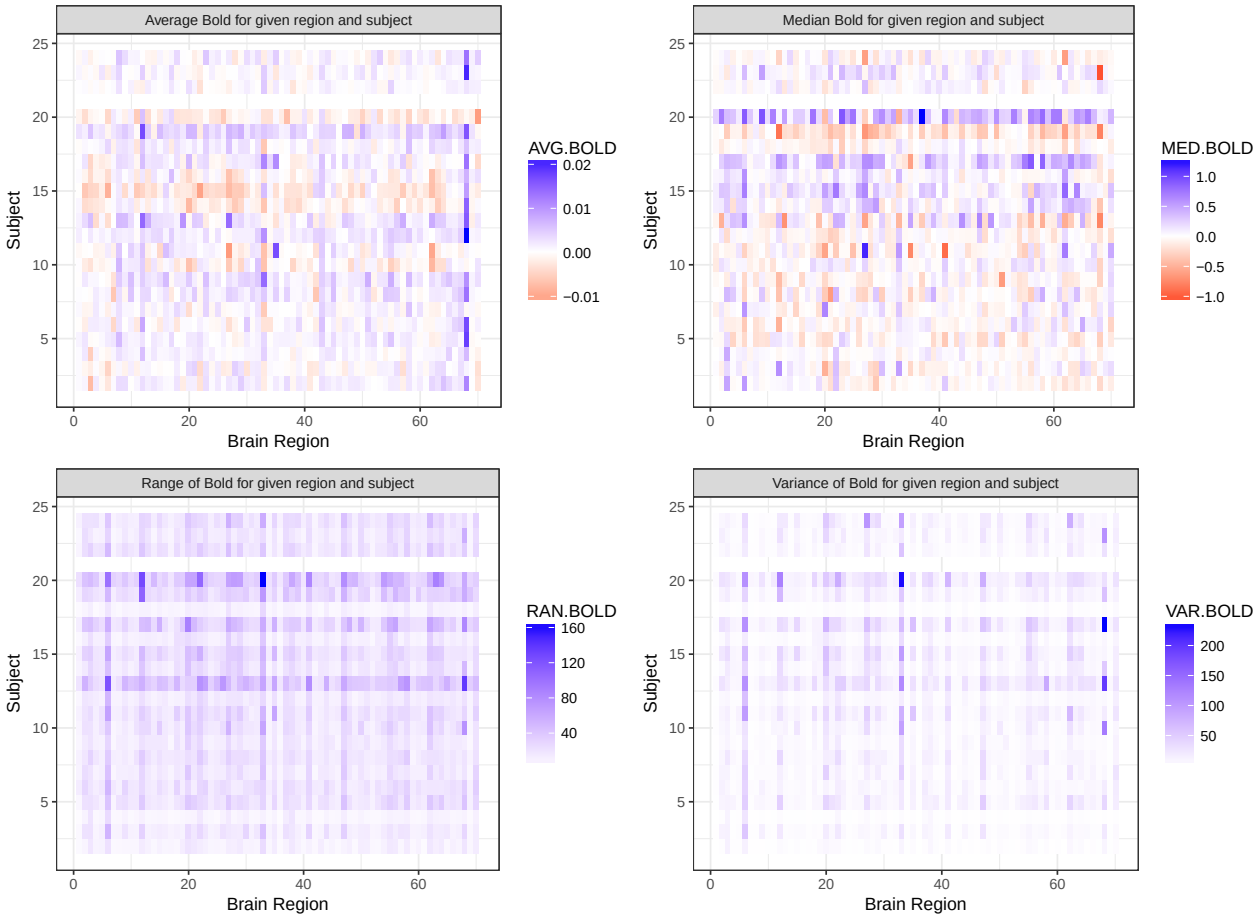
Many values of each dataset are missing. The 18% of values of `D` is not available, and similarly the 31% for `Y` and the 31% for `W`. In fact

- About `D`: Subjects $i = 6, 17, 20, 22$ are completely missing. Some values of `D` however, are not really missing. For each scan and individual the self-connection of a brain region is meaningless and coded as a missing value in the dataset.
- About `W`: the missingness is mainly due to the lack of rescan session for some subjects. For subjects $i = 4, 10, \dots, 18, 20$ we do have complete observations for the first scan, but we do not have any data for the rescan ($j = 2$). Moreover, subjects $i = 1, 21$ are completely missing.
- About `Y`: the situation is similar to that of `W`. For subjects $i = 4, 10, \dots, 18, 20$ we do not have the rescan ($j = 2$) and for $i = 1, 21$ data are completely missing. Moreover, for some individuals (e.g. $i = 2$), the last observation $T = 404$ is missing for each brain region.

Global summary statistics

In this section we try to understand if there are peculiarities among subjects and zones. We study the summary statistics, computed across time, of the brain activity index (BOLD) provided in the `Y` data structure. First of all, we focus our attention only on the first scan. Moreover, subject 1 and 21 are ignored since no data are available for them. As we have already noticed, there are some subjects for which there are missing values at time $t = 404$. We do not consider them, setting the option `na.rm=TRUE` in the computations.

Broadly speaking, we compute for each time series the mean, the median, the range and the variance. For each summary statistic, since we aggregate the data collapsing the time dimension, we collect a 22×70 matrix. To obtain immediate results, we use raster plots.



Even if in these four graphs the behavior of the four statistics seems to be quite erratic, we can notice some interesting patterns:

- Focusing on the **average** BOLD across time, the region 68 is the one which scores the highest indexes among various individuals. This does not hold anymore when we analyze the median of the brain activities.
- There are few patients that experience **very low range and variance** of BOLD among all the brain zones. Subject 4 is an example of this claim.
- The most interesting subject is number 20: in quite all his brain regions he has **negative** average value of BOLD, meanwhile his median BOLD levels are really high compared to the other subjects (the converse is true for subject 19). Moreover, his variability indexes seems to be greater than the others' ones. Subject 20 is the only one with `Generalized Anxiety` as `current_diagnosis`: could this mean something?

Empirical isotropic 2D semi-Variograms - Subject 10

A function that plots **empirical isotropic 2D semi-Variograms** is implemented. The brain regions are treated as sites, which coordinates are given by the centroids contained in the `ROI` dataset.

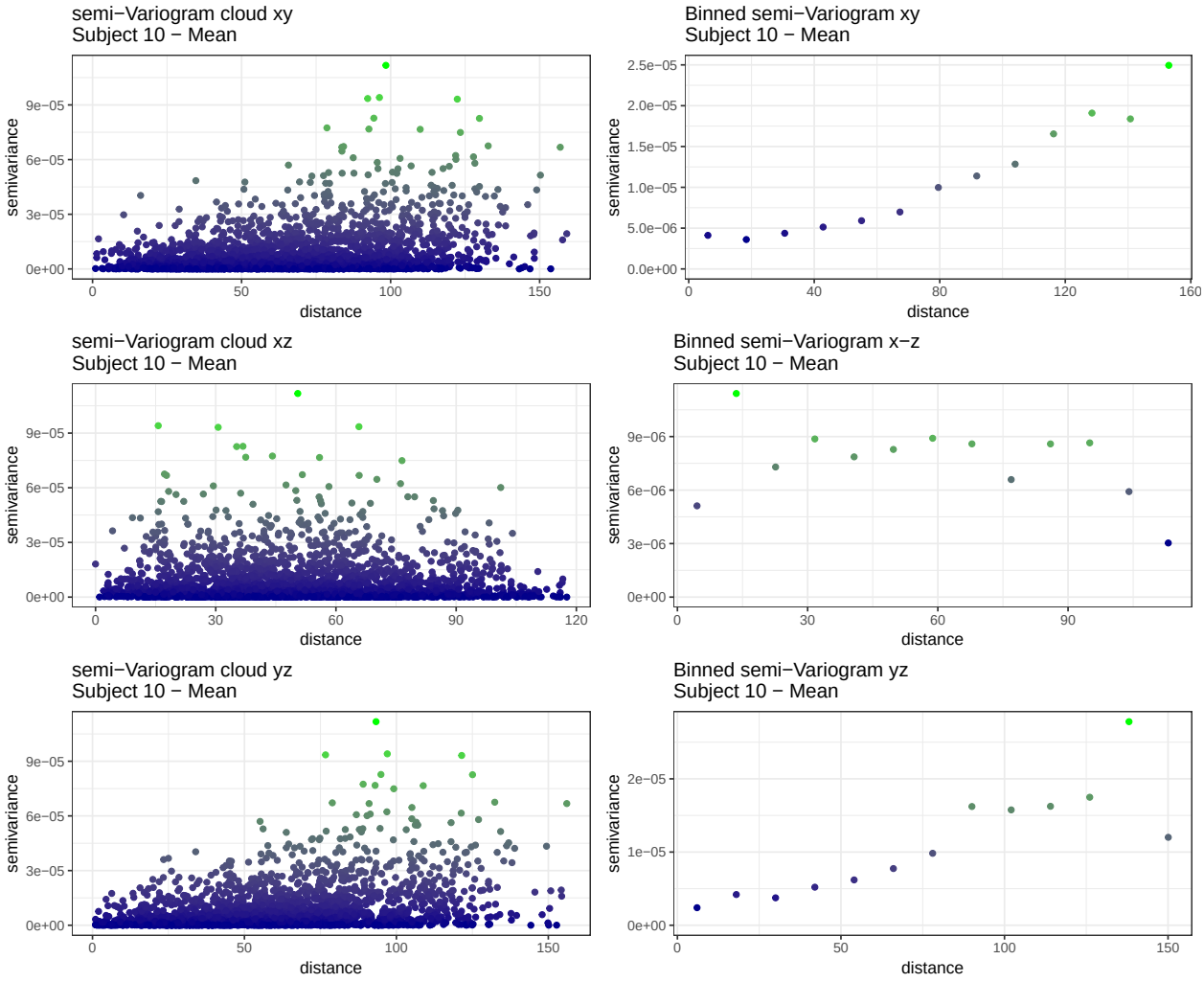
The output graph is composed by 6 plots, disposed on 3 rows, each of them representing a possible combination of the 3 coordinates (`xy`,`xz`,`yz`). The reported graphs are the semi-Variogram **cloud** and the **binned** semi-Variogram. We leave the default number of bins of the `geoR::variog` function.

For the semi-Variogram cloud, the empirical method of moments estimate is computed according to the formula

$$\hat{\gamma}(h) = \frac{1}{2N_h} \sum [Y(x_i + h) - Y(x_i)]^2 ,$$

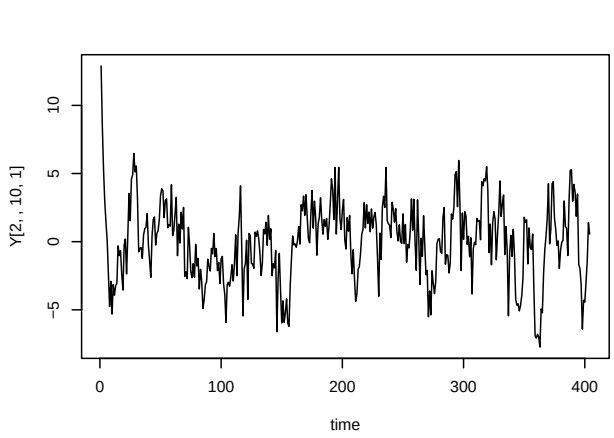
and then it is plotted against the correspondent distance h. See the help page of `variog` function for further details. On the other hand, to obtain the binned semi-Variogram, we divided the distance range $(0, \max(h))$ into K intervals and we computed for each bin the mean of $\hat{\gamma}(h)$, called $\text{avg}_k \hat{\gamma}$, for $k = 1, \ldots, K$. Plotting the average $\text{avg}_k \hat{\gamma}$ against the midpoint of each interval leads to the binned semi-Variogram.

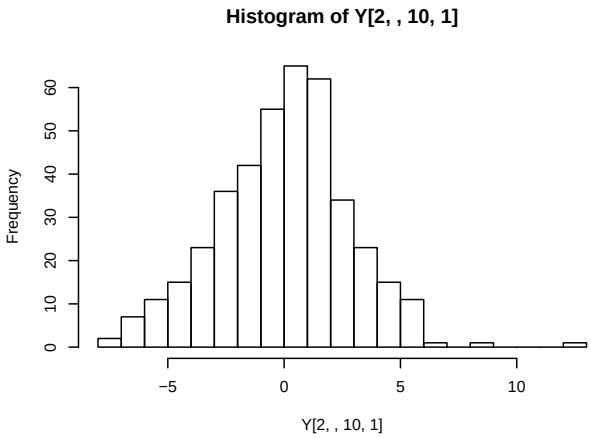
This type of analysis needs a dimensionality reduction. Starting from the **first scan** contained in `Y`, we compute the `mean` of each time series. An example for subject 10 is reported below.

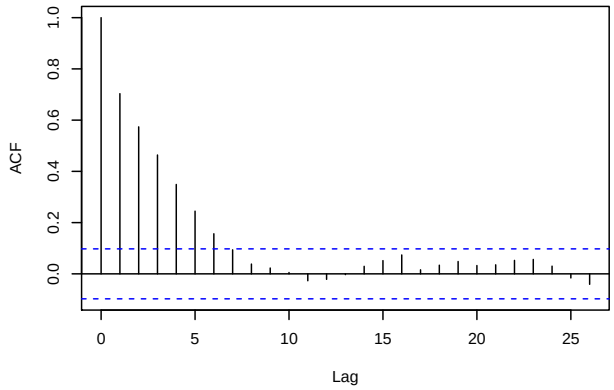


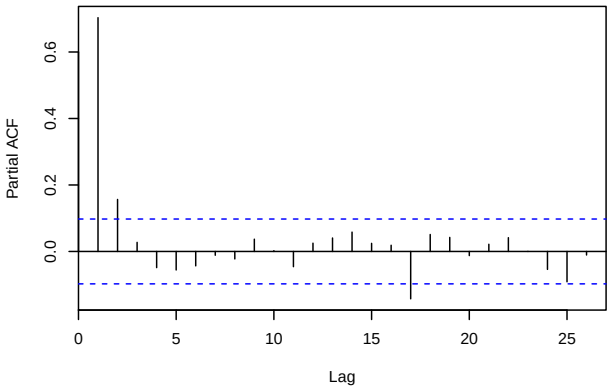
Univariate time series analysis - Subject 10

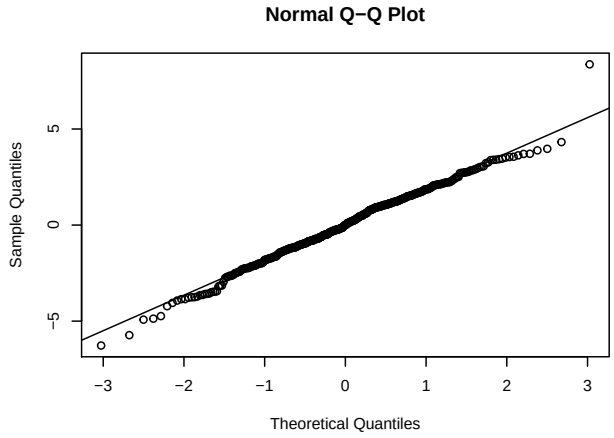
In this section we will consider the the first scan $k = 1$ for the individual $i = 10$. Also, we consider the second brain region, which is called `lh-bankssts`. The raw time series, its marginal distribution, the ACF and the PACF are reported below. For this time-series we fitted an arima model using the `auto.arima` command from the `forecast` R package, which identified the "best" model according to the AIC index, within a list of candidates. The selected model for `Y[2,,10,1]` was an ARIMA(2,0,2). Below, we reported a QQ-plot and the hisogram for the residuals of such a model.

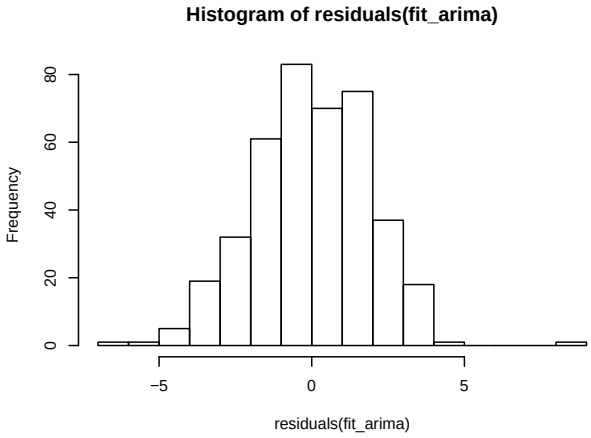












- Although not reported here for space reasons, we computed these 5 graphs for several combinations of individuals and brain regions. The vast majority of the time series presented a similar behaviour: high autocorrelation, low partial autocorrelation and a rough symmetricity around 0.
- The autocorrelation graph seems to be always exponentially decreasing, but the rate seems to be different across individuals and brain regions.
- The Gaussianity assumption seems to be not a huge issue, although in some cases —not the one displayed— the residual distribution presents **heavy tails**.

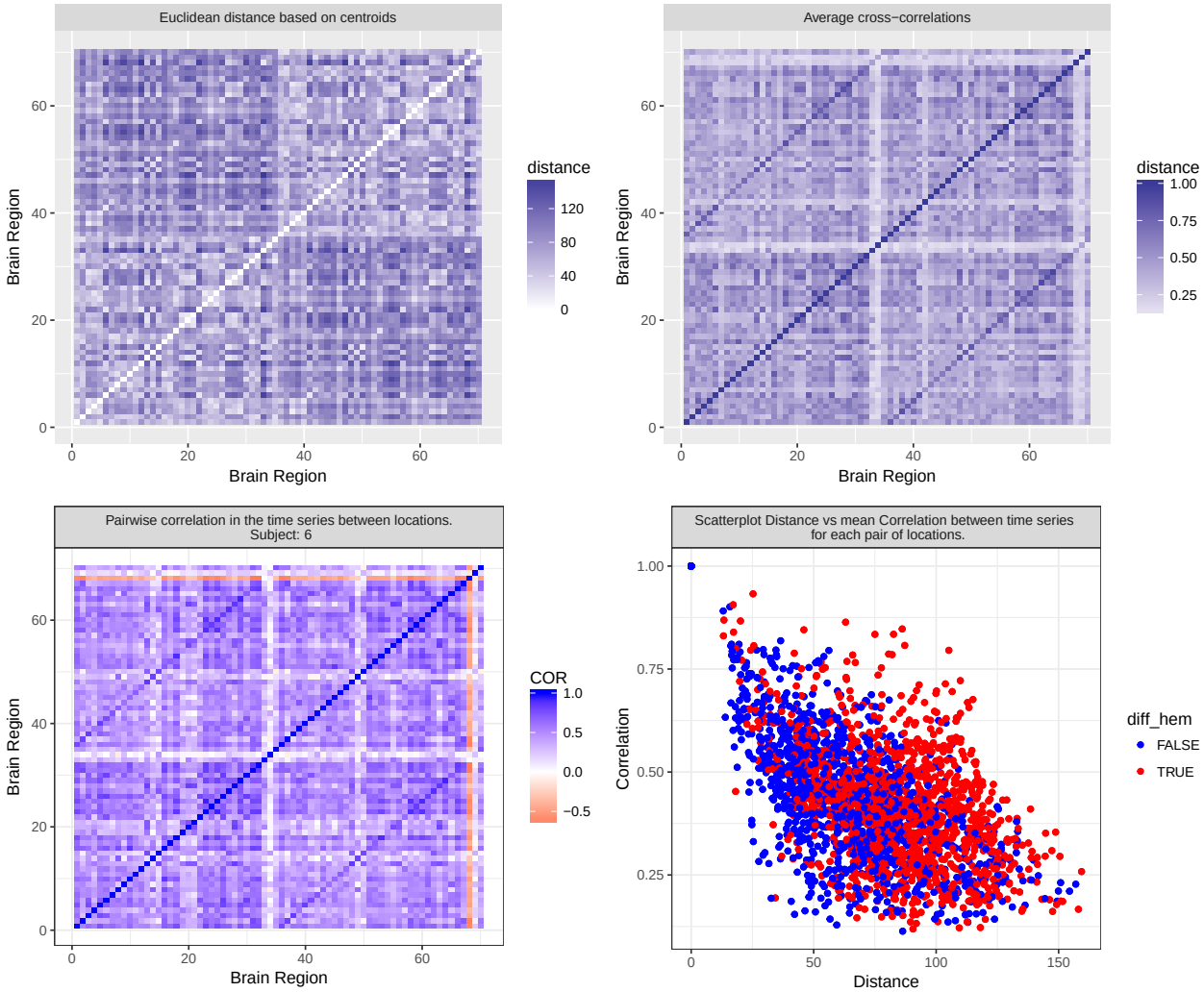
Correlations structure among time-series

The aim of this paragraph is to explore the correlation structure among time series. We compared such a structure with the physical distance between regions, measured as the euclidean distance between the centroids of each brain region.

Since we have replicates over individuals, we computed the **average correlation** among time series. Since many individuals in `Y`, and therefore also in `W` do not have the rescan, we considered only the first scan `W[,,1]`.

In the `R` file we saved these distances in the objects `euclidean_dist` and `correlation_dist`. These two quantities are **negatively correlated** (the correlation is about -0.5), thus closer regions are characterized by more correlated time-series.

Moreover, we create a function which plots two graphs for a specific subject $i = 1, \dots, 24$. The first one is a raster plot of the particular 70×70 correlation matrix among time series in different zones. The second one is a simple scatterplot: the euclidean distance between a pair of brain zones is plotted against the amount of correlation correspondent to the the same pair. The latter can be obtained using the ***average*** correlation among subjects too. Each point is then colored: if the pair of locations are in the same hemisphere the point will be `red`, otherwise `blue`. An example is given on the second row of the following graph.



From the graph displayed above, we noticed the following facts:

- Cross-correlations between brain regions are **all positive**, on average. On the other hand, there are some subjects, e.g. number 6, 12, 13, 14 and 17, for which zone 68 seems to be negatively correlated with all the others.
- Brain regions $34, 35$ and $68, 69$ are quite peculiar. They seems to be not associated among them or with the other regions, although they are physically close each other.
- The highlighted subdiagonals in the second plot seem to show strong association between some regions in the left hemisphere and regions in the right hemisphere: (1,36), (2,37), ... On average, regions in the same hemisphere are more correlated. Given two brain zones, there is no clear pattern between being in the same hemisphere or not and their correlation.
- In the left hemisphere the pair with higher average cross-correlation [0.9] is (23,25), which is composed of quite close regions. Specularly, in the right hemisphere the pair with higher average cross-correlation [0.89] is (58,60), which corresponds to the pair of closest regions.
- Correlations between `euclidean_dist` and `correlation_dist` specific for hemispheres are of the same order (about -0.6).