2. Exploring and displaying longitudinal data

Sonja Greven

Summer Term 2016

Overview Chapter 2 - Exploring and displaying longitudinal data

2.1 **Graphical display of longitudinal data**

2.2 Exploring mean and correlation

The graphical display of longitudinal data is important for building appropriate models and should always be the first step!

Notation again

- N is the number of subjects.
- n_i is the number of observations for the ith subject, $i = 1, \ldots, N$. Remember, balanced data have $n_1 = \cdots = n_N$.
- $n = \sum_{i=1}^{N} n_i$ is the total number of observations across all subjects.
- Response: $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ is the vector of n_i observations for the ith subject (random vector).
- We observe y_{ij}, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$.

Graphical display of longitudinal data

The display used depends on the data at hand and the questions of interest, but some general recommendations - wherever possible - are:

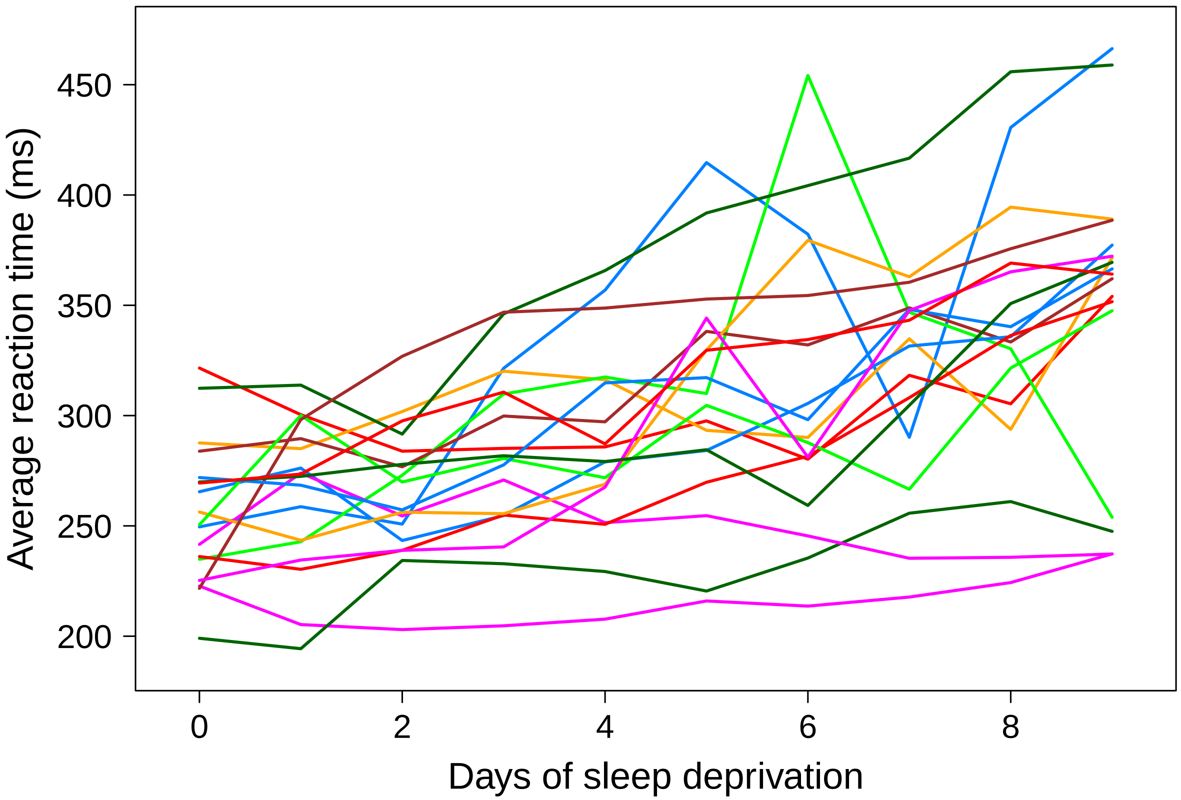
1. show the original data instead of aggregate measures as much as possible

2. also make general trends in the data visible

3. make it easy to pick out individuals and extreme or outlying observations/subjects

4. highlight cross-sectional as well as longitudinal patterns.

Display of individual profiles - Ex. sleep deprivation data



This data set

- is balanced
- has few subjects ($N = 18$)

Display of individual profiles: Standardization

It can be useful to display centered and/or standardized profiles. For balanced data, one shows

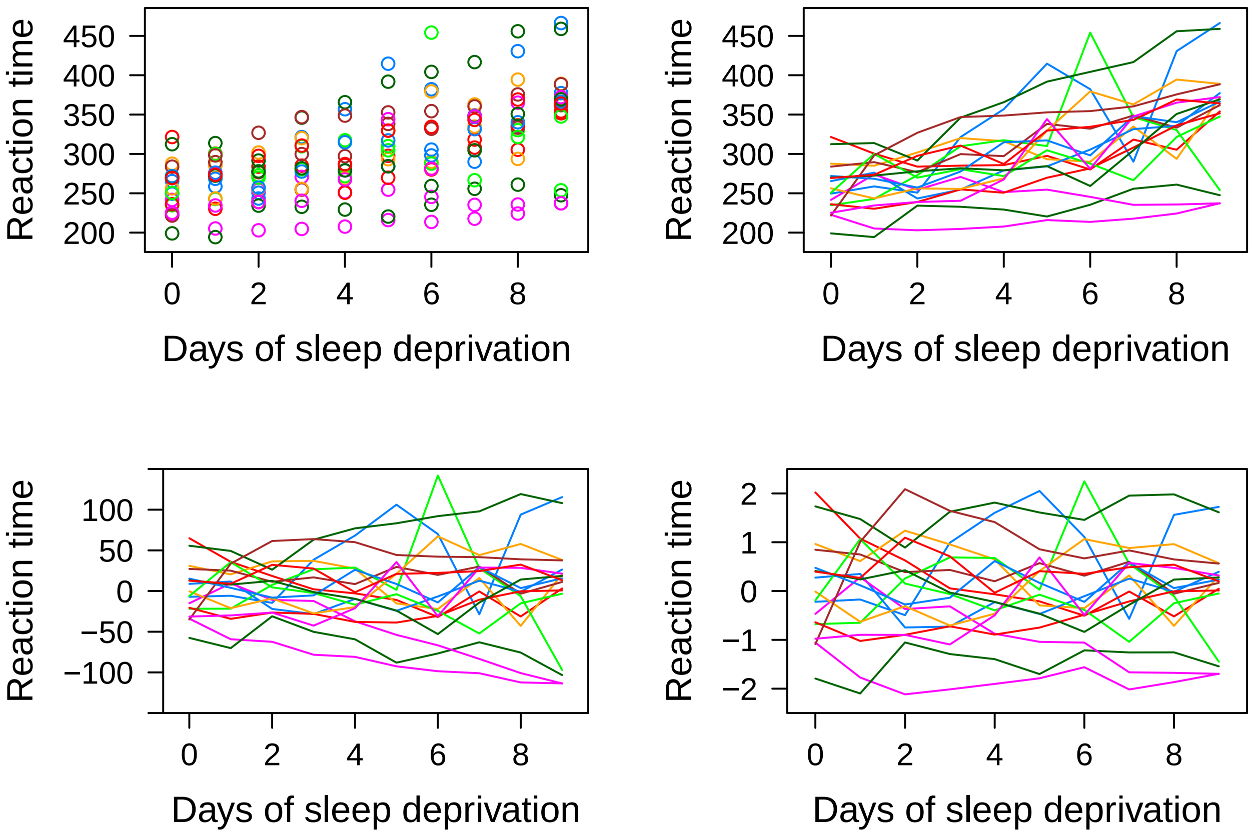
$$y_{ij}^c = (y_{ij} - \overline{y}_j), \qquad \text{or} \qquad y_{ij}^s = (y_{ij} - \overline{y}_j)/s_j,$$

where $\overline{y}_j = \sum_{i=1}^N y_{ij}$ is the arithmetic mean and s_j is the empirical standard deviation at t_j. (E.g. subtract a smooth mean, see 2.2, for unbalanced data.)

- Standardization can be helpful if the variance changes with time (zooming in for areas with low variance).

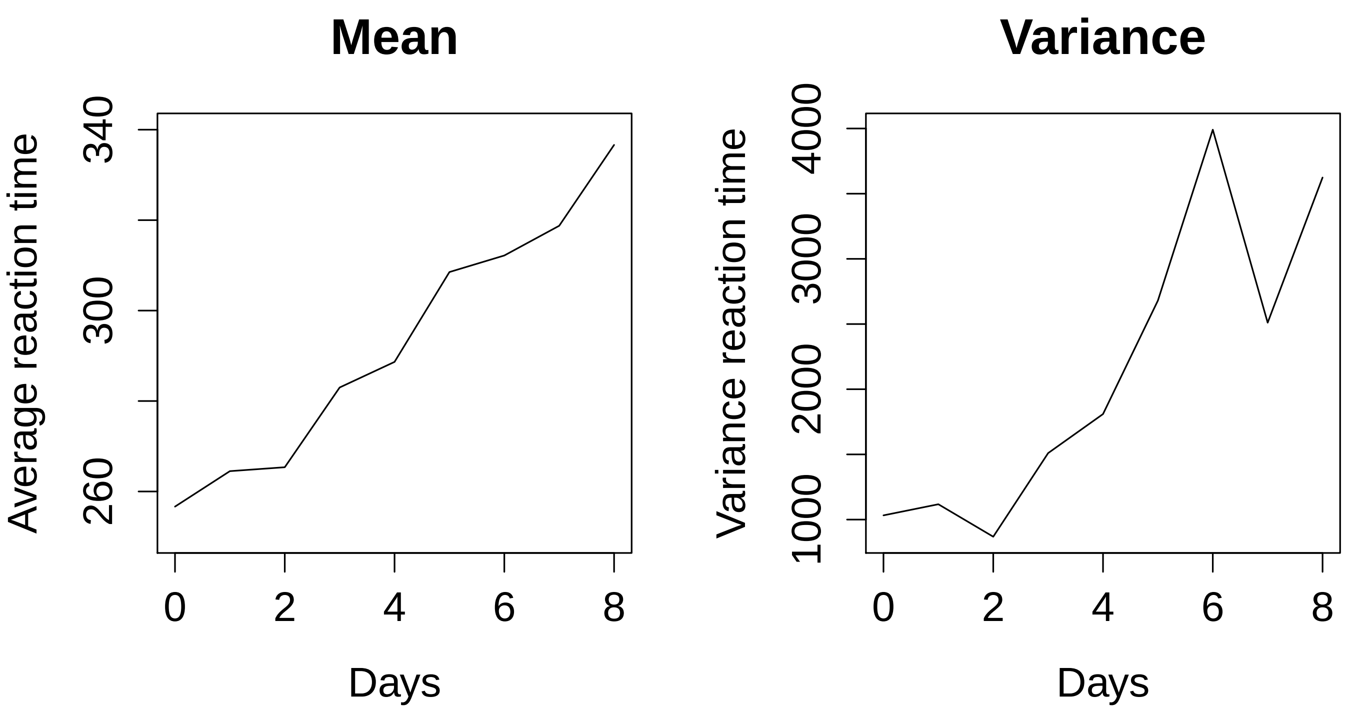
- Easier 'tracking' of individuals and whether they keep their relative positions.

Display of individual profiles

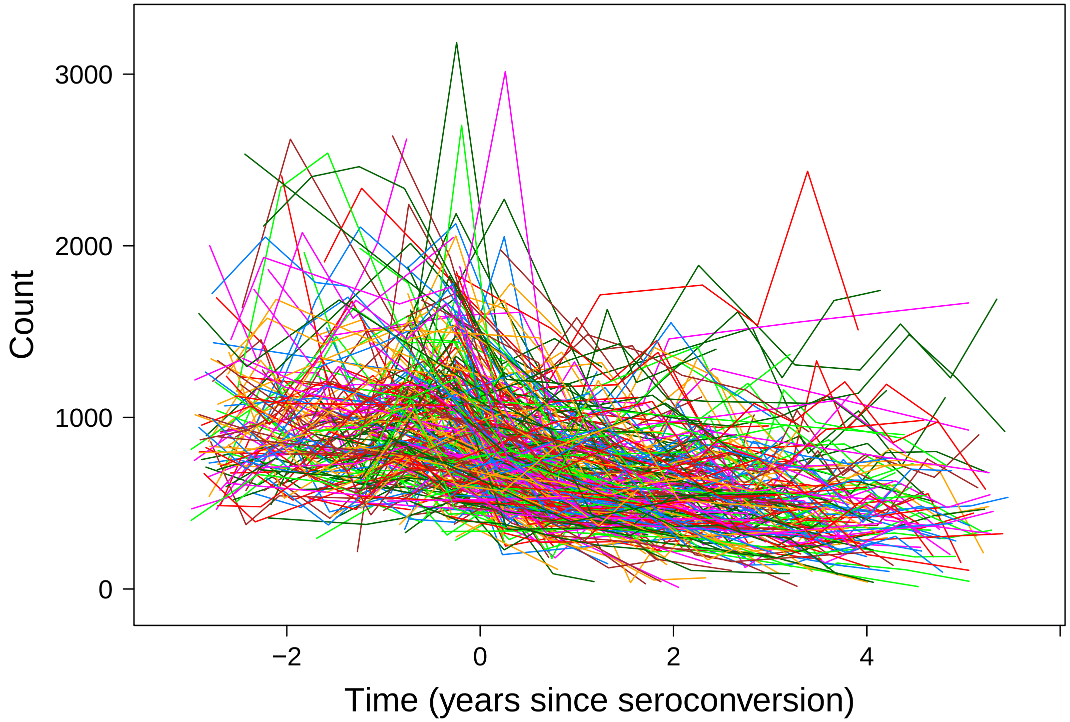


Top row: Raw data. Bottom left: centered. Bottom right: standardized.

Mean and variance curves over time



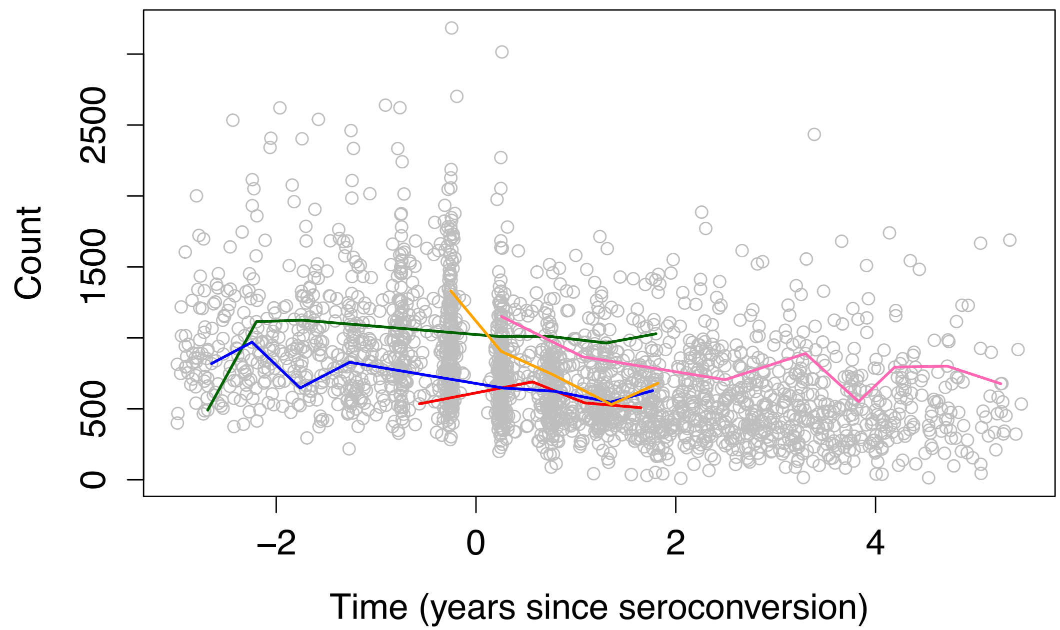
Display of large longitudinal data sets - Ex. CD4+ counts



Graphs with all individual curves can be hard to distinguish for large N.
It can then be useful to not show all individual curves.

Individual curves only for some subjects

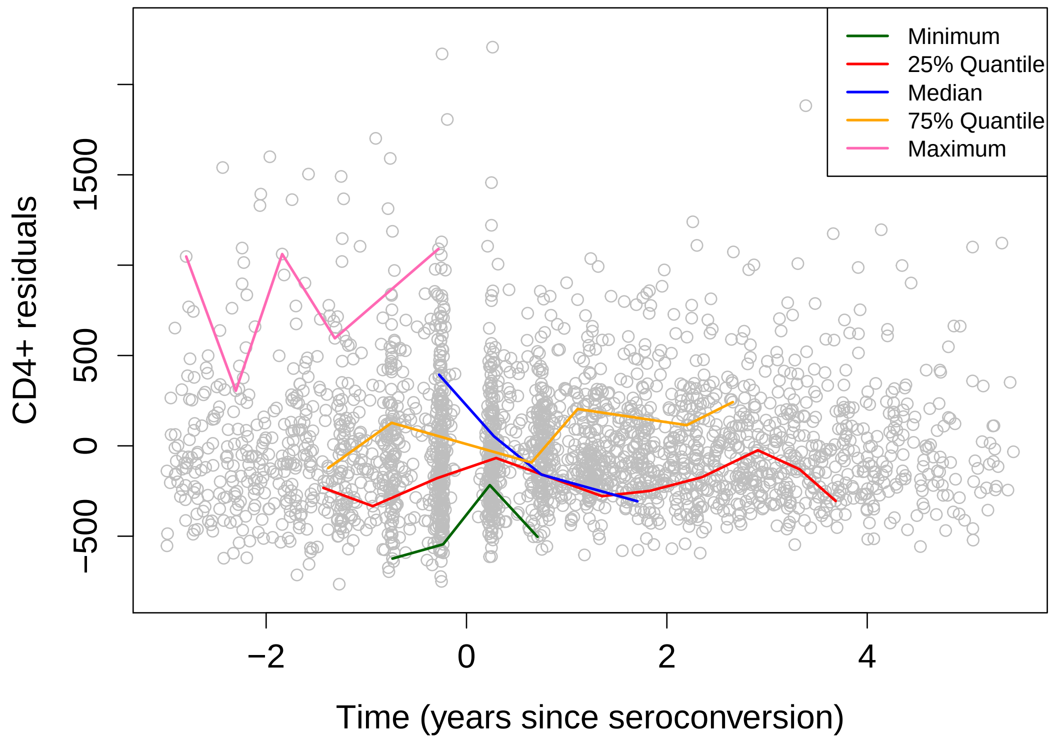
Alternative 1: Only show individual curves for randomly chosen subjects:



Disadvantage: The randomly drawn subjects need not be representative. Extreme curves are unlikely to be shown.

Individual curves only for some subjects - by quantiles

Alternative 2: Show individual curves for subjects chosen using quantiles of a statistic, e.g. average level or variability over time (here: median residual values after subtracting smooth mean curve).

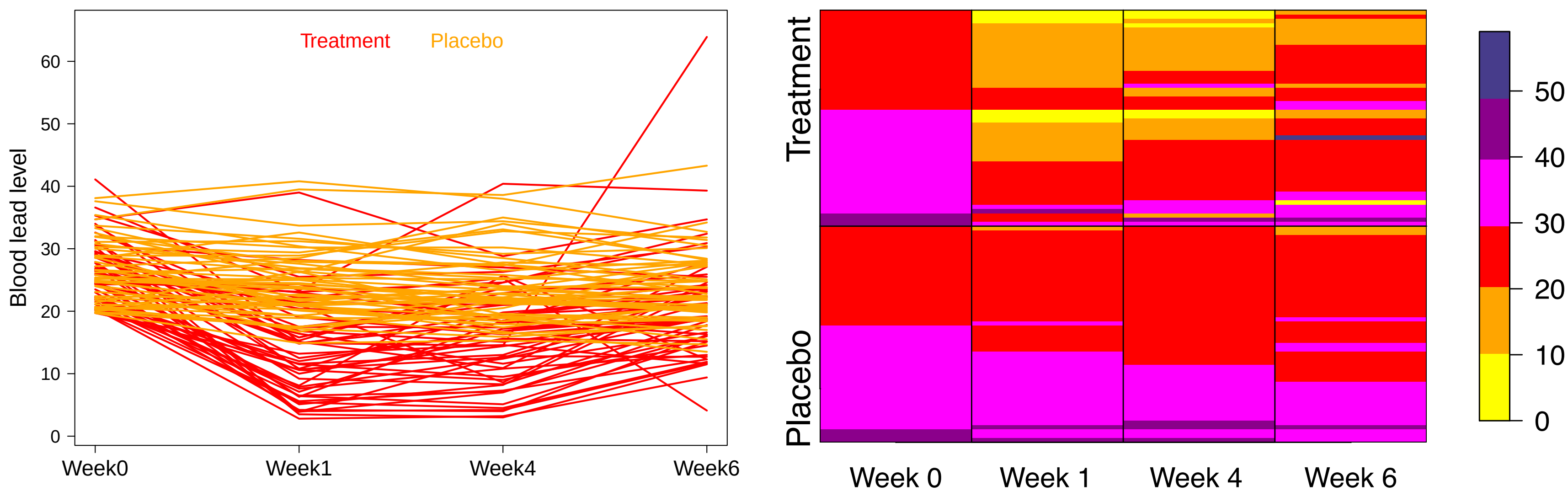


The Lasagna plot

Plots with individual curves are also called **spaghetti plots**. Swihart et al., 2010 propose the **lasagna plots** as an alternative (also for large N).

- The data is plotted as heat map with each column corresponding to one time point and each row to a subject (the 'layers').
- Subjects are ordered for better visisual distinction, e.g. grouped by treatment groups and then ordered by ascending average response value.
- Best suited to data with equal time points, $t_{ij} \equiv t_j$, i.e. balanced data or data with some missings, which are left white. (Or use binning of t_{ij}.)

Spaghetti and Lasagna plots for the TLC data



Overview Chapter 2 - Exploring and displaying longitudinal data

2.1 Graphical display of longitudinal data

2.2 **Exploring mean and correlation**

Fitting smooth means

- For balanced data we can display the mean at each time point.
- For unbalanced data one can use smoothing methods to estimate $\mu(\cdot)$ in

$$Y_{ij} = \mu(t_{ij}) + \epsilon_{ij}$$

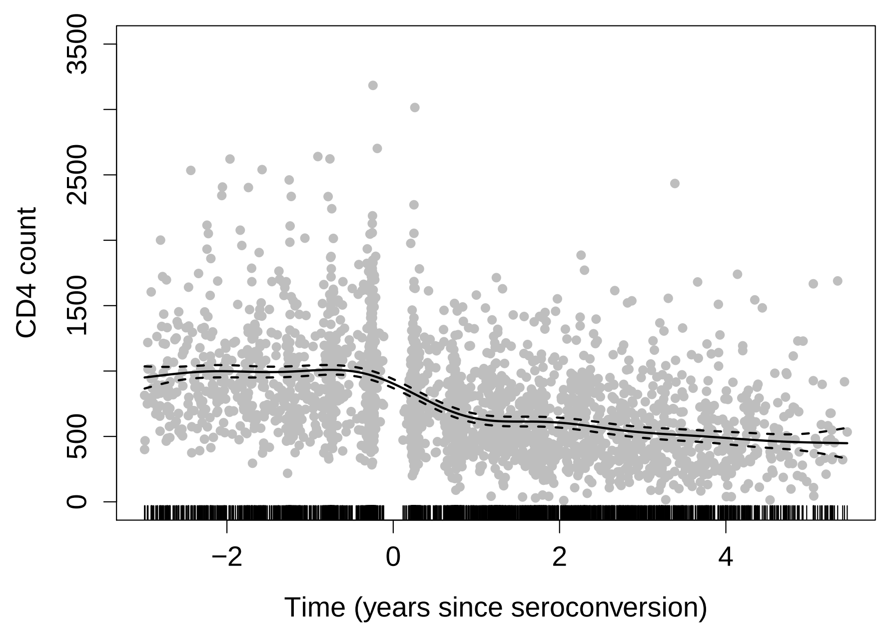
from data $(t_{ij}, y_{ij}),\;\; j = 1, \ldots, n_i,\; i = 1, \ldots, N$.

Three common nonparametric regression techniques are

- Kernel methods
- Spline smoothing
- Lo(w)ess

Fitting smooth means

- These smoothing methods (and the criteria for choosing smoothing parameters) assume independent and identically distributed (i.i.d.) ϵ_{ij}.
- The temporal correlation and unequal n_i for different subjects are not taken into account. They can be used as **exploratory** tools.
- We will learn how to incorporate smooth mean functions in mixed models accounting for repeated measurements in Ch. 6.2.



Exploring the correlation

- Data on the same subject is **correlated**, with correlation often decreasing with time distance.

- For visualization, consider the residuals $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T\widehat{\boldsymbol{\beta}}$, where $\mathbf{x}_{ij}$ is the covariate vector for the jth measurement of the ith subject and $\widehat{\boldsymbol{\beta}}$ is estimated by a linear regression ignoring the correlation.

- Alternative 1: display the correlation as scatterplot of r_{ij} vs. r_{ik} for each i, j, k (for equidistant and equal time points t_j, or binned time points)

- Alternative 2: plot products $r_{ij}r_{ik}$ - as estimates of the residual covariance - against their time distance $|t_{ij} - t_{ik}|$.

- Alternative 3: the (semi)variogram, see Ch. 6.1.

Conclusion

- The data should always be displayed graphically before beginning with the analysis.
- Graphics should be chosen appropriately to the data and questions at hand!
- Exploring the mean and correlation is helpful for model building.