2. Exploring and displaying longitudinal data

Sonja Greven

Summer Term 2015

Overview Chapter 2 - Exploring and displaying longitudinal data

2.1 **Graphical display of longitudinal data**

2.2 Exploring the mean: semiparametric smoothing

2.3 Exploring the correlation

2.4 Useful R commands

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 $n_1 = \ldots = n_N$.

- $n = \sum_{i=1}^{N} n_i$ 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:

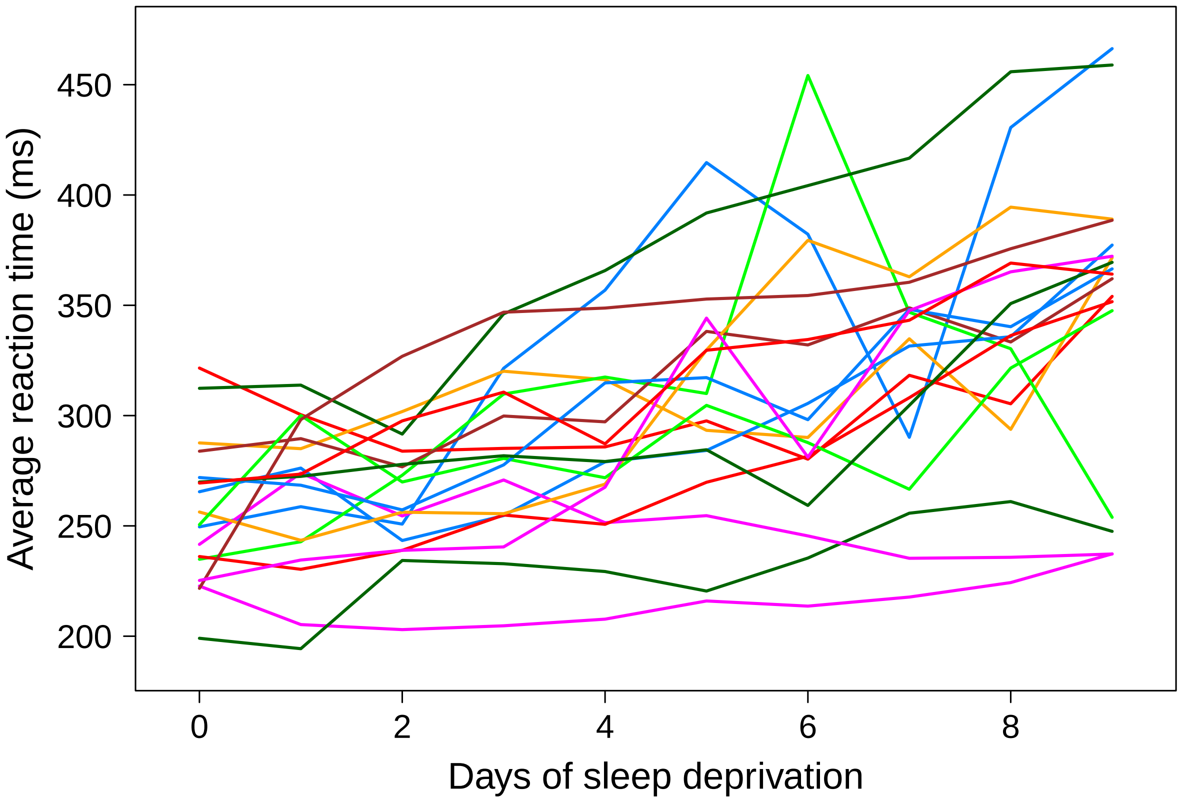
1. show the original data instead of aggregate measures as much as possible

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 - 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

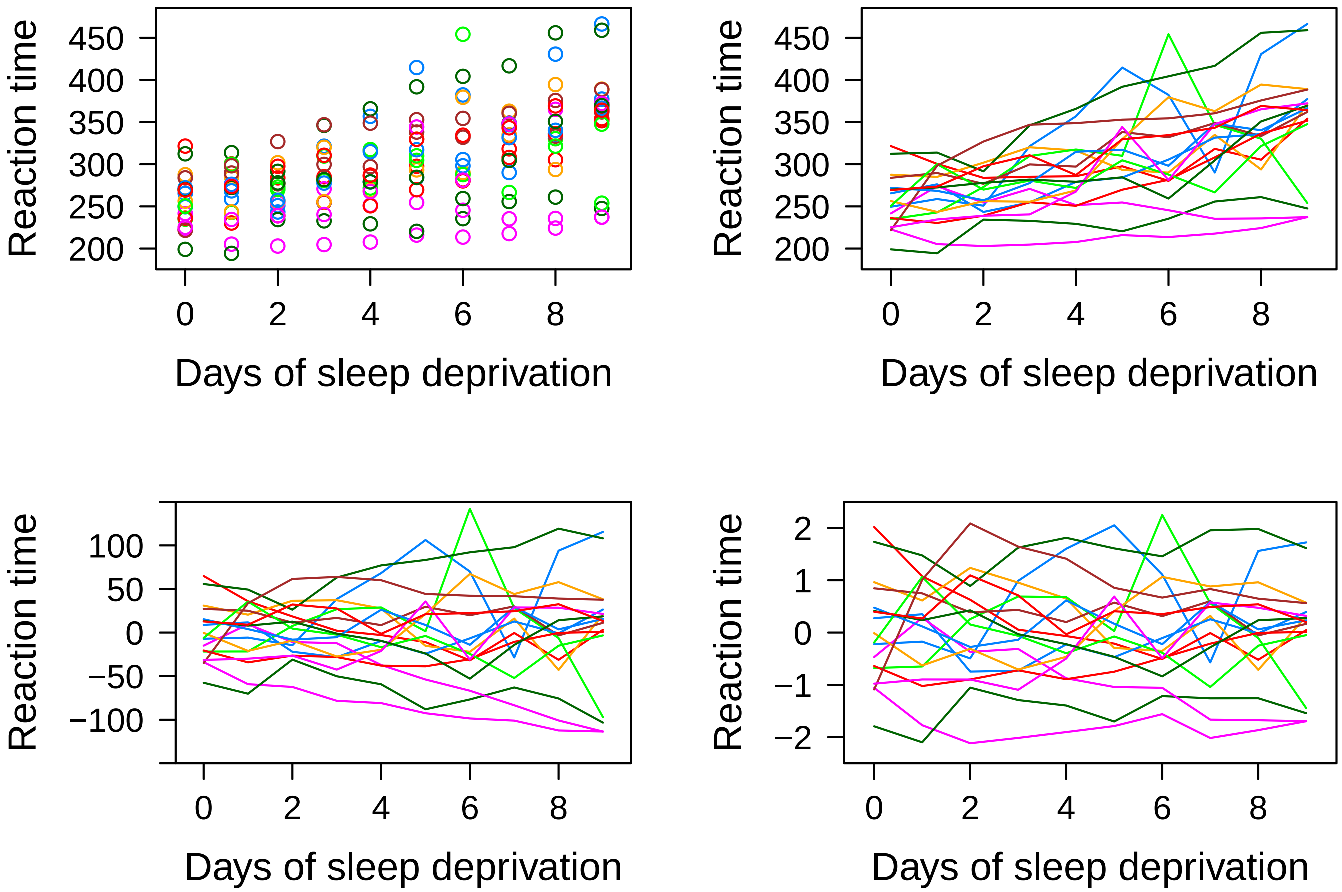
$$y_{ij}^c = (y_{ij} - \overline{y}_j), \qquad \text{or} \qquad y_{ij}^s = (y_{ij} - \overline{y}_j)/s_j,$$

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.)

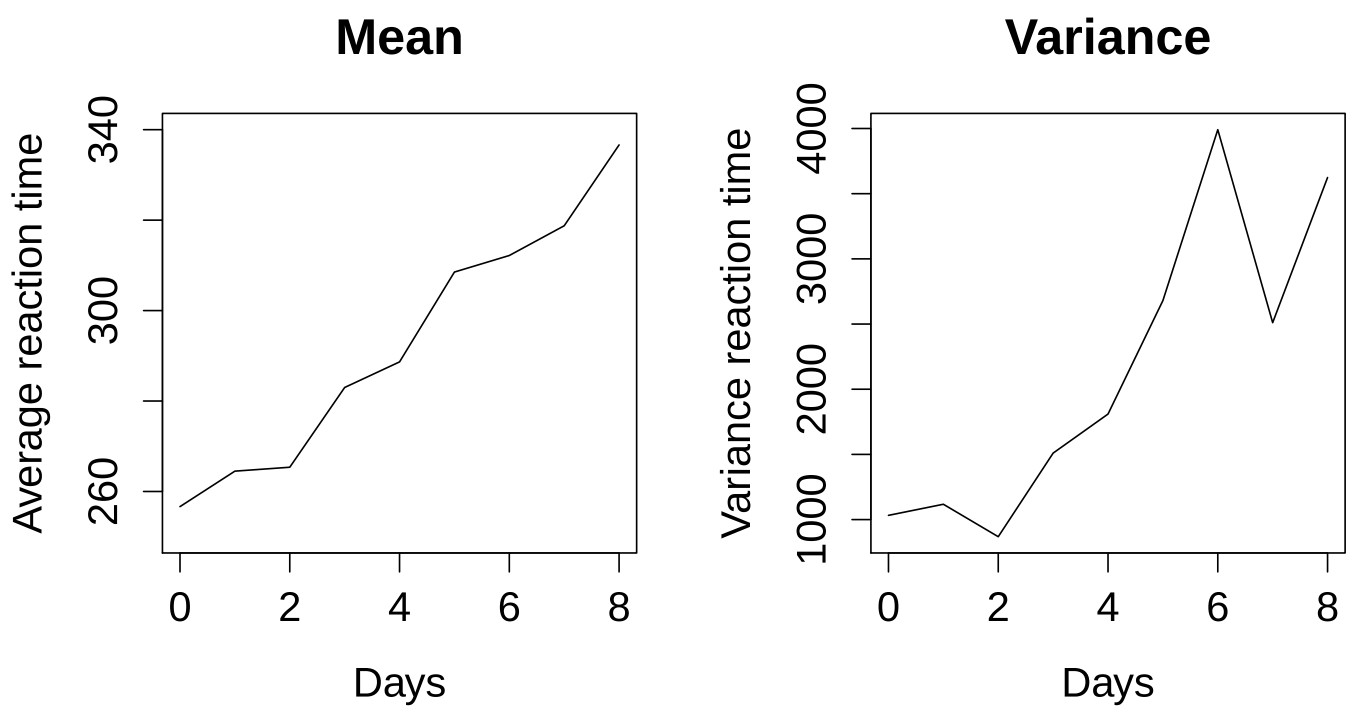
- Standardization can be helpful if the variance changes with time (zooming in for areas with low variance).

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

Display of individual profiles

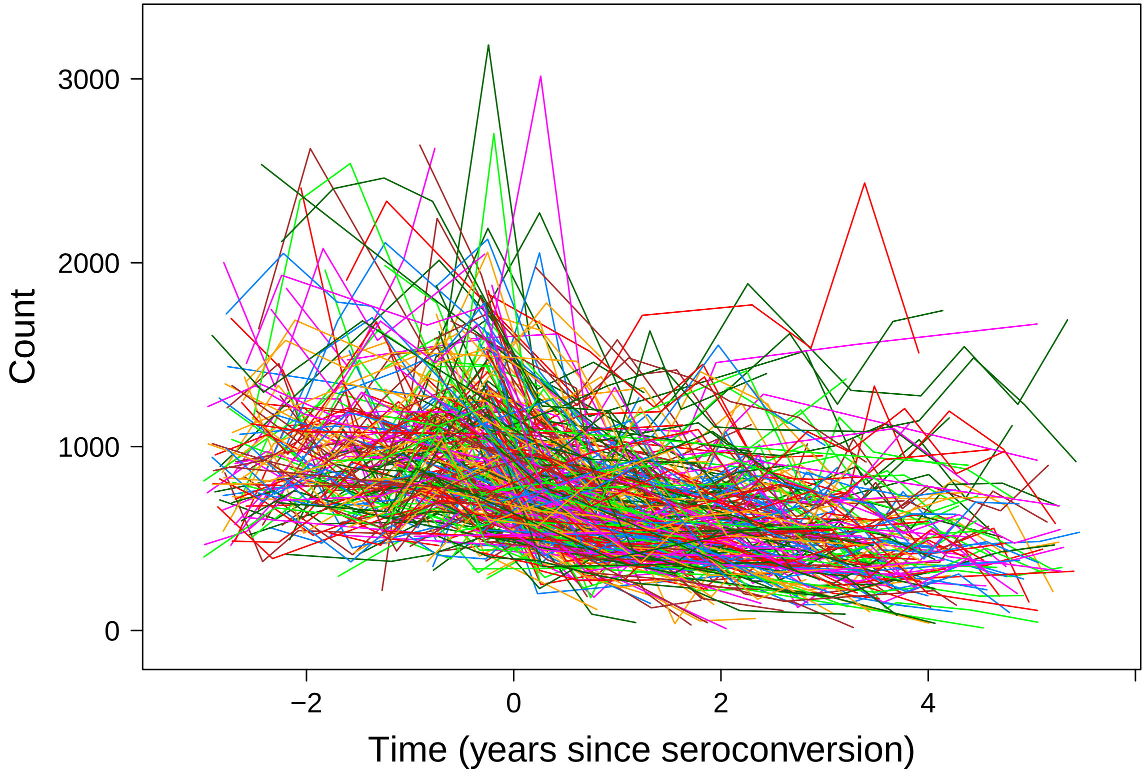


Mean and variance curves over time



Display of large longitudinal data sets - CD4+ counts

Graphs with all individual curves can be hard to distinguish for large N.

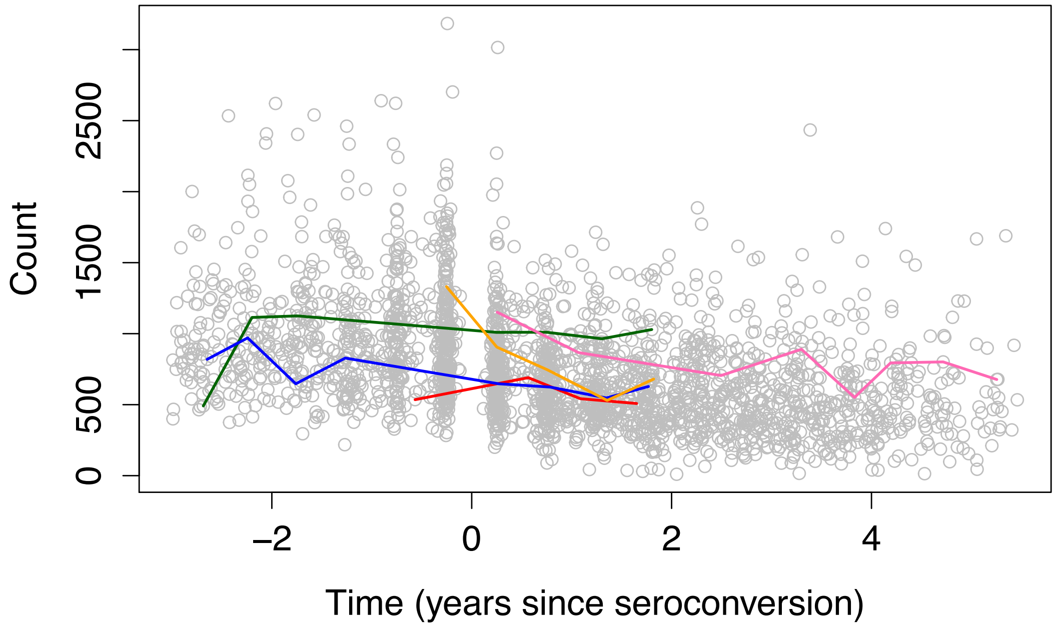


Display of large longitudinal data sets

- It can then be useful to not show all individual curves.
- Alternatives:
  - only show individual curves for some subjects (the others e.g. as dots or thin grey lines),
  - only show observations and a smooth mean (see 2.2)

Individual curves only for some subjects

Randomly chosen subjects:



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

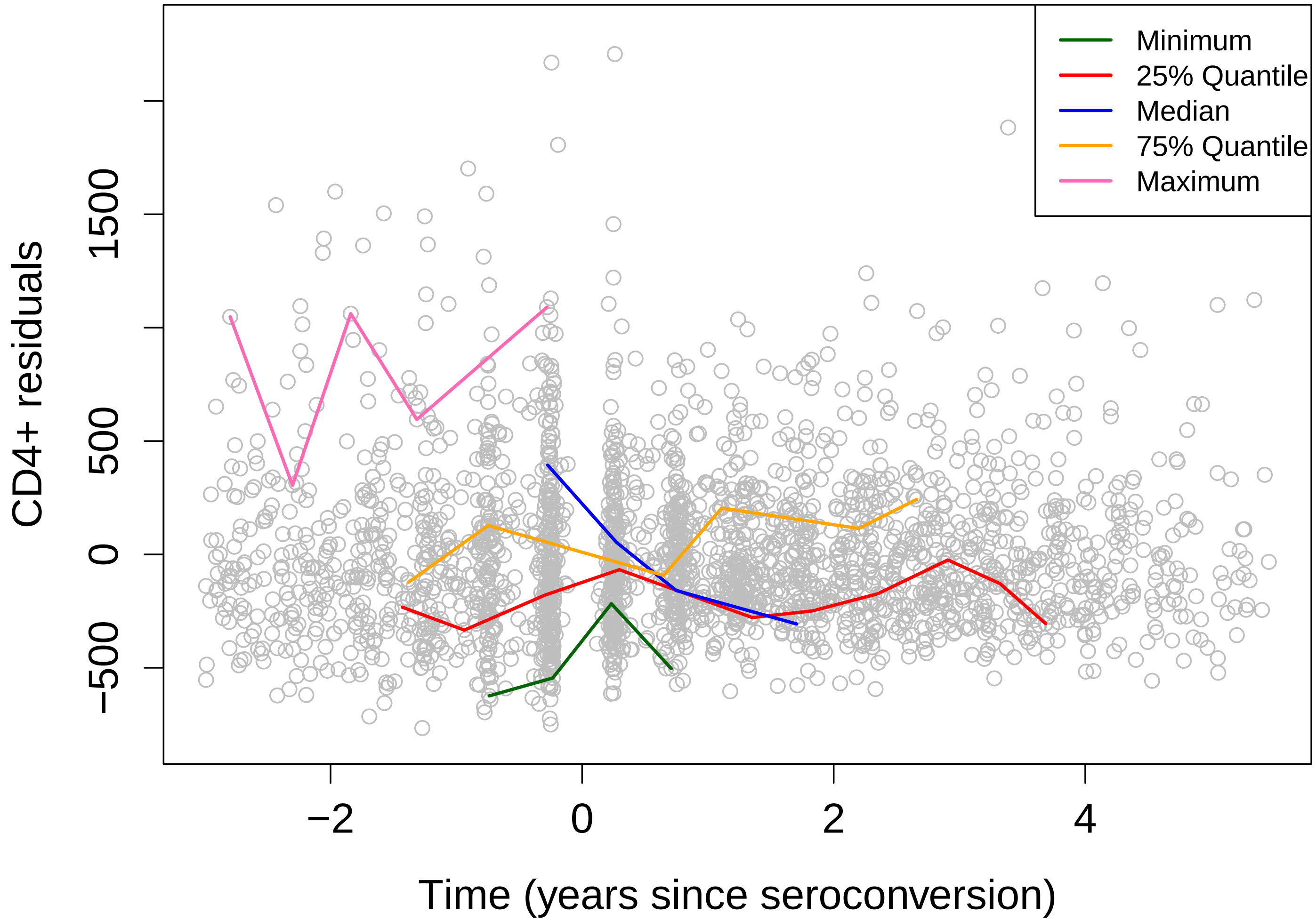
Individual curves only for some subjects

Alternatives: Choose subjects using a statistic, e.g. measuring

- the average level
- variability over time
- etc.

One option is to plot individuals with median residual values (after subtracting a mean curve, see 2.2) corresponding to certain quantils, e.g. minimum, 25% quantile, median, 75% quantile, maximum.

Individual curves only for some subjects - by quantiles



The Lasagna plot

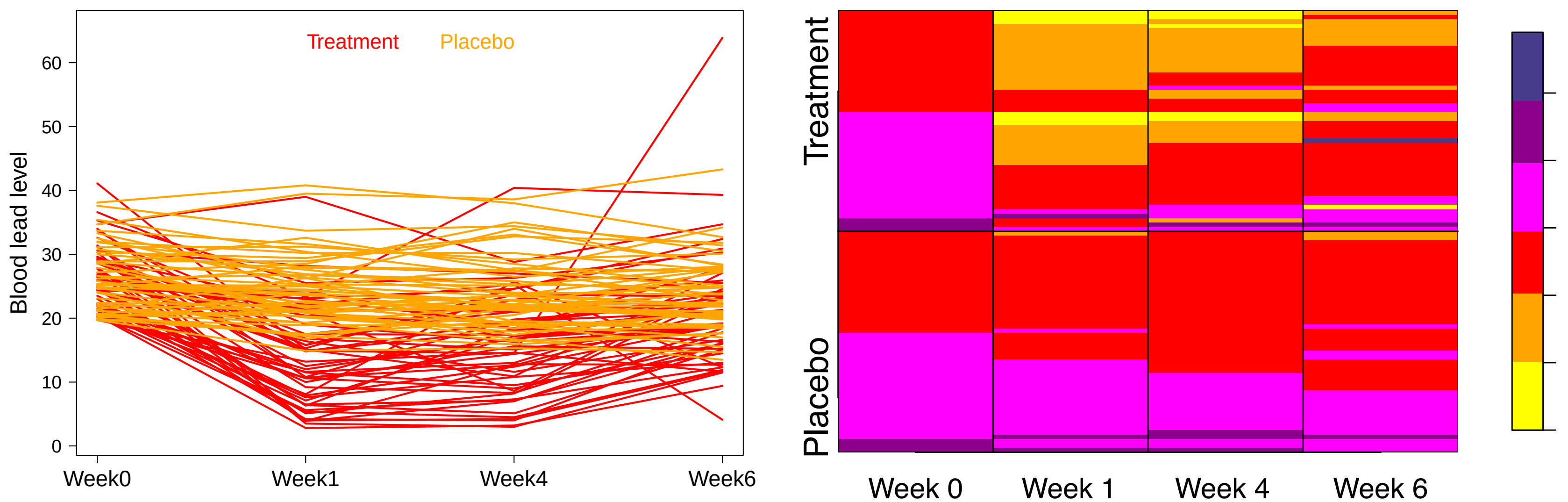
Plots with individual curves are also called **spaghetti plots**. Swihart et al., 2010 propose an alternative (also for large N) they term **lasagna plots**.

- 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 by some criterion that makes distinctions easier to see, 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. (Otherwise, need to handle time axis differently or use binning.)

Spaghetthi 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 the mean: semiparametric smoothing**

2.3 Exploring the correlation

2.4 Useful R commands

Fitting smooth curves

- For balanced data one can display the arithmetic mean at each time point.

- For unbalanced data one can use smoothing methods. Three common nonparametric regression techniques are
  - Kernel methods
  - Splines
  - Lo(w)ess

Semiparametric smoothing methods

- **Assumptions:** Only one observation y_i per subject at time point t_i.
- Data are thus of the form

$$(t_i, y_i), \quad i = 1, \dots, N.$$

- **Goal:** Estimation of the unknown mean curve $\mu(t)$ in the model

$$Y_i = \mu(t_i) + \epsilon_i,$$

where the ϵ_i are independent with mean 0.

Kernel methods: “Sliding window”

- Consider a window around time point t_1.
- Let $\widehat{\mu}(t_1)$ be the average of all y_i corresponding to t_i in that window.
- Analogously for $\widehat{\mu}(t_2), \widehat{\mu}(t_3), \ldots$.

$\longrightarrow$ Sliding window for the estimation.

Kernel methods: “Sliding window”

The width of the window is important:

- If the width is chosen very small, the window can include only one observation at the one extreme $\rightarrow$ interpolation instead of smoothing!

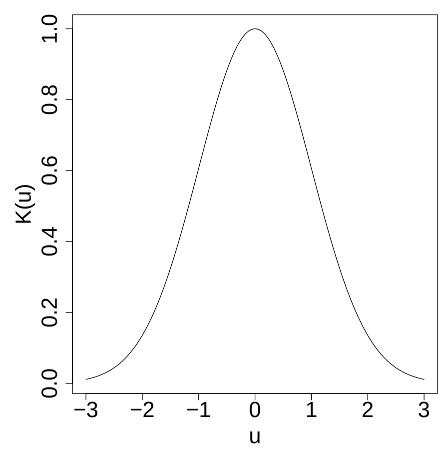
- If the width is chosen very wide, the window can include all observations at the other extreme. This yields a constant:

$$\widehat{\mu}(t) = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Kernel methods in general

- With the sliding window method, each observation gets the weight 1 ("in the window") oder 0 ("outside the window").
- This method is a special case of kernel smoothing methods.
- More generally, choose a smooth weight function that gives more weight to observations nearer in time than to observations further away.
- Common choice: Gaussian kernel

$$K(u) = \exp(-0.5u^2).$$



Kernel methods in general

- Definition of the kernel estimator:

$$\widehat{\mu}(t) = \sum_{i=1}^{N} \frac{w(t, t_i, h)}{\sum_{i=1}^{N} w(t, t_i, h)} y_i,$$

where $w(t, t_i, h) = K\left((t - t_i)/h\right)$ are the weights and h is the bandwidth.

- Larger values for h yield smoother curves.

- We'll discuss the choice of h in a few slides.

- How is the kernel K defined for the sliding window method?

Smoothing splines (Silverman, 1985)

- If we assume $\mu(t)$ can be well approximated by a twice continuous differentiable function $s(t)$ with second derivative $s''(t)$, consider minimizing

$$J(\lambda) = \sum_{i=1}^{N} (y_i - s(t_i))^2 + \lambda \int \{s''(t)\}^2 dt.$$

- The solution can be shown to be a natural cubic **spline** (a two times differentiable function consisting of piecewise cubic polynomials) with knots at the t_i and can be obtained from (relatively simple) linear equations.

- Penalized splines are an alternative that is computationally less demanding and can be incorporate into more complex models, see Chapter 6.2.

Lo(w)ess smoothing (Cleveland, 1979)

- LOWESS $=$ **LO**cally **WE**ighted regression **S**catterplot **S**moothing
- Function `lowess` in R
- Lo(w)ess can be seen as an extension of kernel methods: at each point t_i, a local polynomial regression is fitted using weighted least squares, giving more weight to observations closer by.
- There is an iterative version that is more robust to outliers, giving them smaller weight.

Choice of smoothing parameters

- In all three approaches (kernel, splines, lowess), the smoothness of the estimated curves is controlled by one **smoothing parameter** (e.g. h, λ). This parameter is typically chosen to optimize a criterion.

- **Goal:** compromise between bias and variance.

- A common criterion that combines bias and variance is the mean squared error, MSE (analogously for h instead of λ):

$$MSE(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\{y_i^* - \widehat{\mu}(t_i;\lambda)\}^2,$$

where y_i^* is a new observation at time point t_i.

Choice of smoothing parameters

$$MSE(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\{y_i^* - \widehat{\mu}(t_i;\lambda)\}^2$$

Observations y_i which were used for estimation of μ should not be compared to $\widehat{\mu}(t_i)$: This would lead to always choosing the smallest band width h or penalty λ and to interpolation instead of a smooth curve (overfitting).

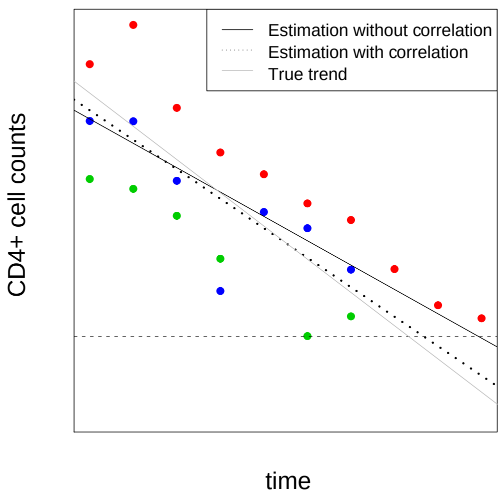
Solution: cross-validation (analogously for h instead of λ)

$$CV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\{y_i - \widehat{\mu}^{-i}(t_i;\lambda)\}^2,$$

where $\widehat{\mu}^{-i}(t_i;\lambda)$ is obtained without observation i. See Chapter 6.2 for mixed model-based estimation of smoothing parameters.

Note

- Please note that these smoothing methods (and the criterion for the choice of the smoothing parameter) assume independent and identically distributed (i.i.d.) errors.
- Also, dropout and missing values are not taken into account.
- They can still be useful **exploratory** tools.
- Example CD4 data: See lab.
- For how to incorporate smooth mean functions in mixed models accounting for repeated measurements, please see Chapter 6.2.



Overview Chapter 2 - Exploring and displaying longitudinal data

2.1 Graphical display of longitudinal data

2.2 Exploring the mean: semiparametric smoothing

2.3 **Exploring the correlation**

2.4 Useful R commands

Exploring the correlation

- Data from the same subject tend to be more similar than data from different subjects; longitudinal data are **correlated data**.

- Often observations closer in time are more similar than observations taken further apart, i.e. the correlation is decreasing with the time difference.

- This correlation can be visualized with scatterplots.

- 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.

Display of the correlation

- For equidistant time points that are the same across subjects, the correlation can be displayed as scatterplot of r_{ij} vs. r_{ik} for each i, j, k.

- For non-equidistant time points, this would require first binning the time points.

- Alternatively, one can plot the pair-wise products $r_{ij}r_{ik}$ - as estimates of the residual covariance - against their time distance $|t_{ij} - t_{ik}|$.

- Another alternative that does not require binning time points is the (semi)variogram. More in Chapter 6.1.

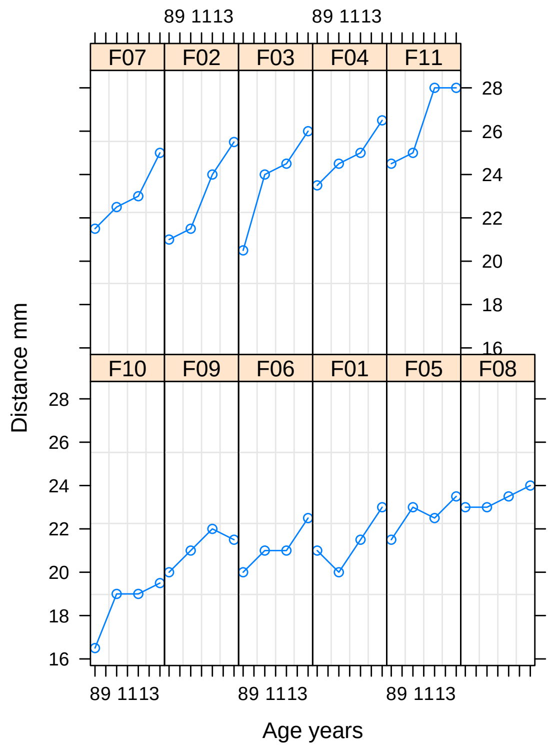
Overview Chapter 2 - Exploring and displaying longitudinal data

Useful R commands

- `reshape` - reshapes longitudinal data between 'wide' and 'long' format
- `groupedData`
- `plot` for `groupedData` objects
- `xyplot` from the package `lattice` for data frames

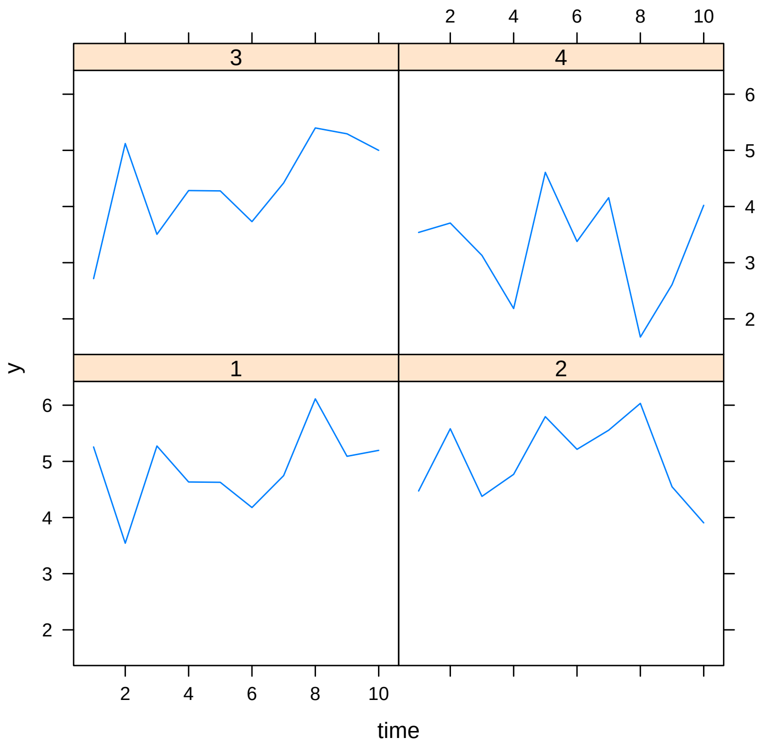
More in the lab session.

`plot` for `groupedData` objects



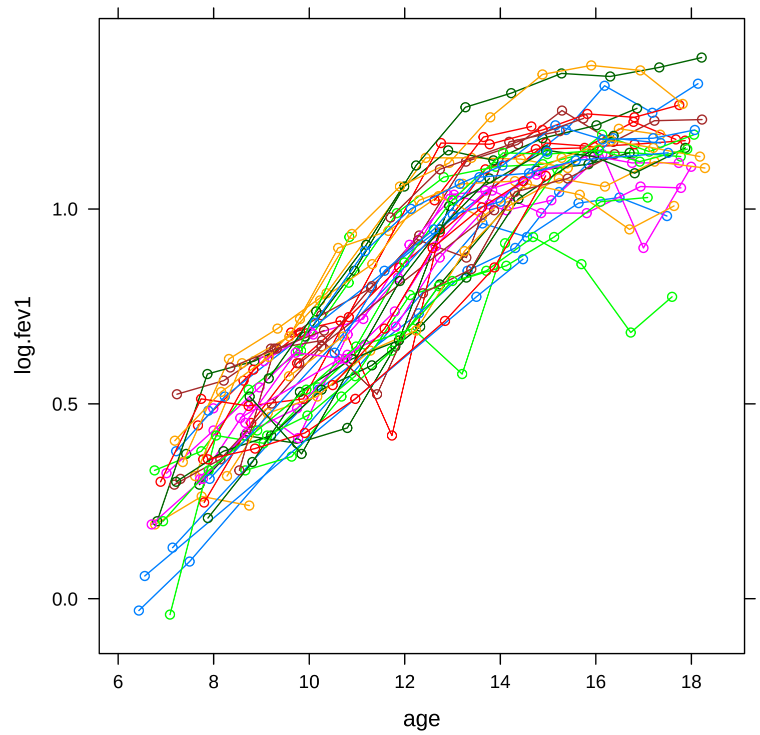
Only suitable for a limited number of subjects!

`xyplot` **for data frames** `xyplot(y~t|id,...)`



Only suitable for a small number of subjects!

`xyplot` **for data frames** `xyplot(..., groups=id,...)`



Also for somewhat larger numbers of subjects.

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!

- `R` offers functions for the display of longitudinal data.

- Exploring the (smooth) mean and correlation is helpful for model building.