Exploratory Data Analysis

Goals of EDA

- Relationship between *mean response* and covariates (including time).
- Variance, correlation structure, individual-level heterogeneity.

Guidelines for graphical displays of longitudinal data

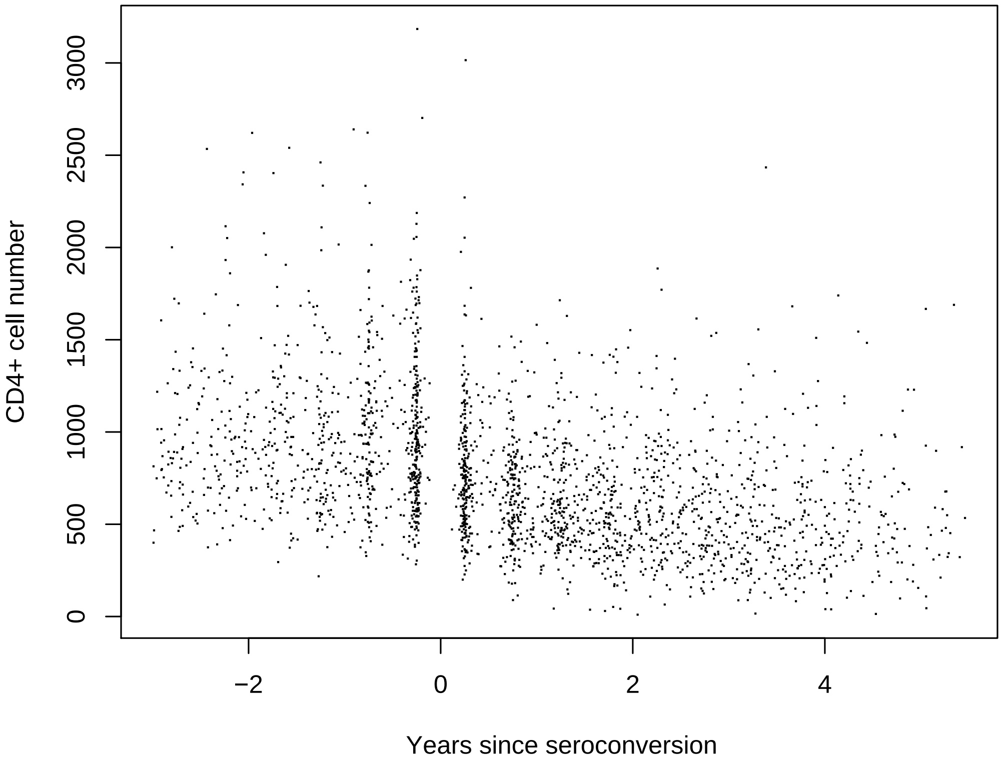
- Show relevant raw data, not just summaries.
- Highlight aggregate patterns of scientific interest.
- Identify both cross-sectional and longitudinal patterns.
- Identify unusual individuals and observations.

General Techniques

- Scatter plots
- Use smooth curves to reveal mean response profile, at the population level.
  - Kernel estimation
  - Smoothing spline
  - Lowess
- "Spaghetti" plot: use connected lines to reveal individual profiles.
  - Displays of the responses against time.
  - Displays of the responses against a covariate (with/without time trend being removed)
- Exploring correlation structure
  - Autocorrelation function: most effective for equally spaced data.
  - Variograms: for irregular observation times.
  - Lorelogram: for categorical data.

Scatter-plots

```
CD4 <- read.table ("data/cd4.dat", header = TRUE)
plot (CD4 ~Time,data = CD4, pch = ".", xlab = "Years since seroconversion",
ylab="CD4+ cell number")
```



Fitting Smooth Curve to Longitudinal Data

Nonparametric regression models can be used to estimate the mean response profile as a function of time.

- Assuming that we have a single observation y_i at time t_i, we want to estimate an unknown mean response curve $\mu(t)$ in the underlying model:

  $$Y_i = \mu(t_i) + \epsilon_i, i = 1, \ldots, m,$$

  where ϵ_i are independent errors with mean zero.

- Common smoothing techniques include:
  1. Kernel smoothing.
  2. Smoothing spline.
  3. Loess (local regression).

Kernel Smoothing

1. Select a window centered at time t.
2. $\hat{\mu}(t)$ is the average of Y values of all points within that window.
3. To obtain an estimator of the smooth curve at every time point, slide a window from the extreme left to the extreme right, calculating the average of the points within the window every time (moving average).
4. This "boxcar" approach is equivalent to computing $\hat{\mu}(t)$ as a weighted average of the y_i's with weights equal to zero or one. This may yield curves that are not very smooth.

- Alternatively we can use a smooth weighting function that gives more weights to the observations closer to t, eg. using Gaussian kernel $K(u) = \exp(-u^2/2)$.
- The kernel estimate is defined as:

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

  where $w(t, t_i, h) = K((t - t_i)/h)$ and h is the *bandwidth* of the kernel.
- Larger values of h produce smoother curves.

Loess

A “robust” version of kernel smoothing.

- Instead of computing a weighted average for points within the window, a low-degree polynomial is fitted using weighted least squares.
- Once the line has been fitted, the residual (distance from the line to each point) is determined.
- Outliers (points with large residuals) are down-weighted, through several iterations of refitting the model.
- The result is a fitted line that is insensitive to the observations with outlying Y values.

[1] Cleveland, W.S. (1979). ”Robust Locally Weighted Regression and Smoothing Scatterplots”. Journal of the American Statistical Association 74 (368): 829-836.
[2] Cleveland, W.S.; Devlin, S.J. (1988). ”Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting”. Journal of the American Statistical Association 83 (403): 596-610.

Smoothing Spline

- A cubic smoothing spline is the function $s(t)$ with two continuous derivatives which minimize the penalized residual sum of squares

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

- The $s(t)$ that satisfies the criterion is a piece-wise cubic polynomial.
- The larger the λ, the more it penalizes the “roughness” of the curve, so the smoother the $s(t)$ is.

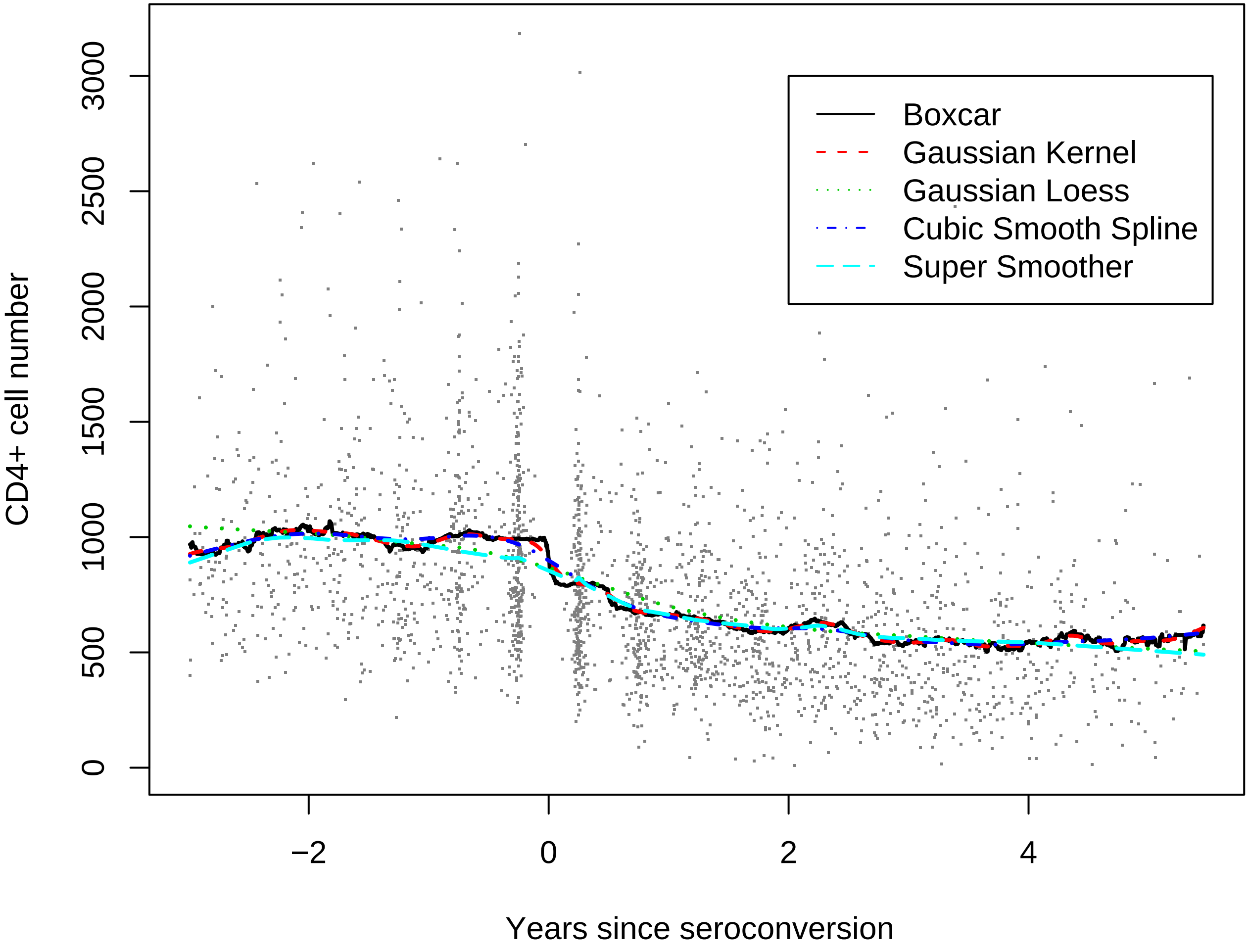
[1] Hastie, T. J.; Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall.

Example

```
plot (CD4 ~ Time, data = CD4, col = "gray50", pch = ".",
      xlab = "Years since seroconversion",
      ylab = "CD4+ cell number")

with (CD4, {
    lines (ksmooth (Time, CD4, kernel = "box"), lty = 1,
           col = 1, lwd = 2)
    lines (ksmooth (Time, CD4, kernel = "normal"), lty = 2,
           col = 2, lwd = 2)
    lines (loess.smooth (Time, CD4, family = "gaussian"),
           lty = 3, col = 3, lwd = 2)
    lines (smooth.spline (Time, CD4), lty = 4, col = 4, lwd = 2)
    lines (supsmu (Time, CD4), lty = 5, col = 5, lwd = 2)
})

legend (2, 3000,legend = c("Boxcar", "Gaussian Kernel", "Gaussian Loess",
        "Cubic Smooth Spline", "Super Smoother"),lty = 1:5, col = 1:5)
```


Boxcar
Gaussian Kernel
Gaussian Loess
Cubic Smooth Spline
Super Smoother
CD4+ cell number
0
500
1000
1500
2000
2500
3000
−2
0
2
4
Years since seroconversion

- Plain kernel (boxcar) smoothing `ksmooth` is generally not recommended.
- Loess and smoothing spline can give very similar results (the difference is usually greater at the boundary).
- The "super" smoother allows more flexible choice of the smoothing parameters and is also robust to outliers (For more details, see the documentation of R `supsmu` function and references therein.)
- Bias and variance trade-off. A generally accepted criterion for selecting the smoothing parameter is using *average predictive squared error* (PSE) estimated by *cross-validation* (leave-one-out):

$$\mathrm{PSE}(\lambda) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\{Y_i^* - \hat{\mu}(t_i; \lambda)\}^2,$$

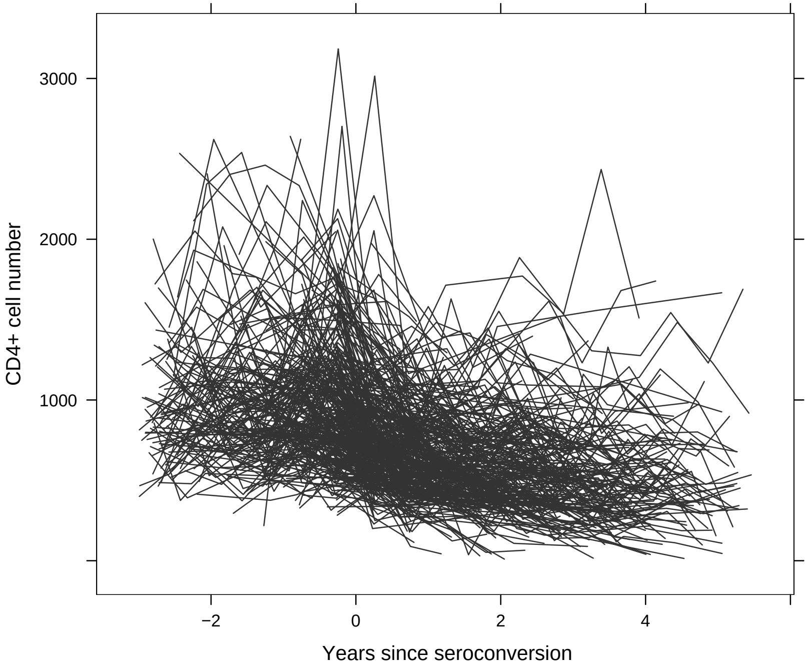
where Y_i^* is a new observation at t_i.

$$\mathrm{CV}(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\{y_i - \hat{\mu}^{-i}(t_i; \lambda)\}^2.$$

Why the observed y_i cannot replace Y_i^* in PSE(λ)?

“Spaghetti” Plot

```
xyplot (CD4 ~ Time, data = CD4, type = "l", group = ID,  xlab = "Years since
          seroconversion", col.line = "gray20", ylab = "CD4+ cell number")
```



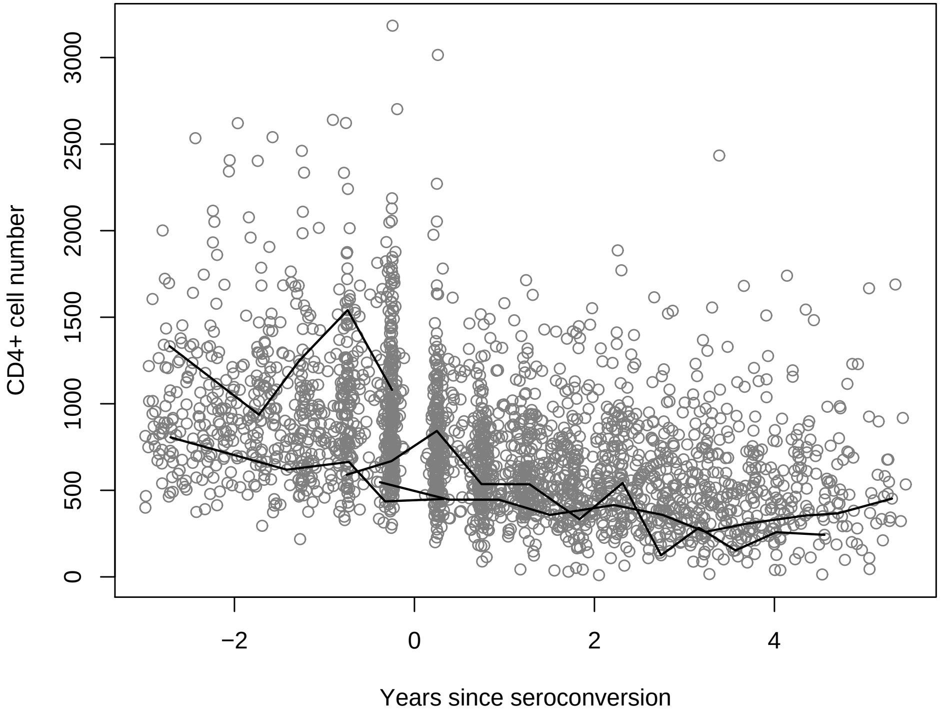
- Too busy.
- Using thin gray lines helps a bit.

Random Individuals

```
s <- sample (unique (CD4$ID), size = 4)

plot (CD4 ~ Time, data = CD4, col = "gray50",
      xlab = "Years since seroconversion",
      ylab = "CD4+ cell number")

for (i in 1:4) {
    lines (CD4$Time[CD4$ID == s[i]],
           CD4$CD4[CD4$ID == s[i]],  lwd = 1.5)
}
```



- Using circles instead of solid points make it easier to see overlapping points.
- Random individuals may be too “random” to be informative and unlikely to reveal “outliers”.

Display Individuals With Selected Quantiles

- Order individual curves with respect to some univariate characteristic of the curves, e.g., the average level, the median level, the variability of the curve, and the slope of the fitted line, etc.
- Then highlight those curves at particular quantiles.
- When there are multiple treatment groups, a separate plot can be made for each group.

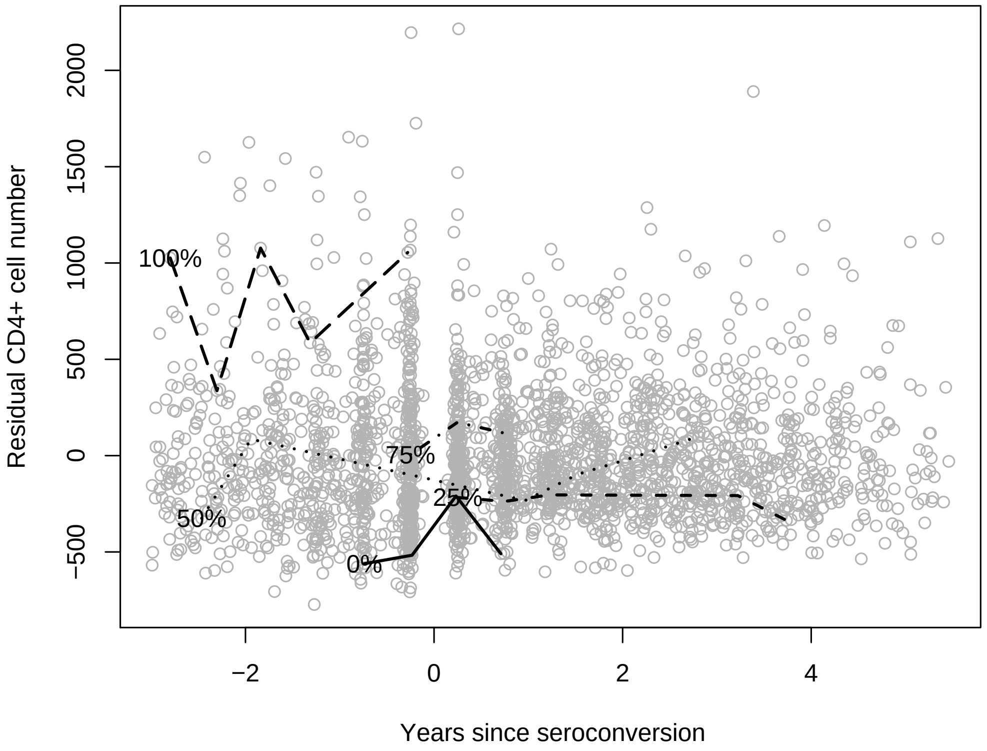
An example

- Regress y_{ij} on t_{ij} and get residuals r_{ij}.
- Choose one dimensional summary of the residuals, for example $g_i = \text{median}(r_{i1}, \ldots, r_{i,n_i})$;
- Plot r_{ij} versus t_{ij} using points.
- Order units by g_i.
- Add lines for selected quantiles of g_i.

```
cd4.lo <- loess (CD4 ~ Time, span = 0.25, degree = 1, data = CD4)
CD4$res <- resid (cd4.lo)

aux <- sort (tapply(CD4$res, CD4$ID, median))
aux <- aux[c(1, round ((1:4) * length (aux) / 4))]
aux <- as.numeric (names (aux))

leg <- c("0%", "25%", "50%", "75%", "100%")
plot (res ~ Time, data = CD4, col = "gray70",xlab = "Years since seroconversion",
      ylab = "Residual Sqrt Root of CD4+ cell number")
with (CD4,
      for (i in 1:5) {
          subset <- ID == aux[i]
          lines (Time[subset], res[subset],
                 lwd = 2, lty = i)
          text (Time[subset][1],
                res[subset][1],
                labels = leg[i])
      })
```



- Here the residuals from fitted lowess line are used. This "detrending" is helpful to explore the deviation to the trend (our eyes are better at comparing vertical distances).

Relationship with Covariates

To investigate the relationship between CD4+ cell number (Y) and CESD score (X), a measure of depressive symptoms, we use this model

$$Y_{ij} = \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n,$$

which implies

1. $Y_{i1} = \beta_c x_{i1} + \epsilon_{i1}$, and
2. $Y_{ij} - Y_{i1} = \beta_L(x_{ij} - x_{i1}) + (\epsilon_{ij} - \epsilon_{i1})$

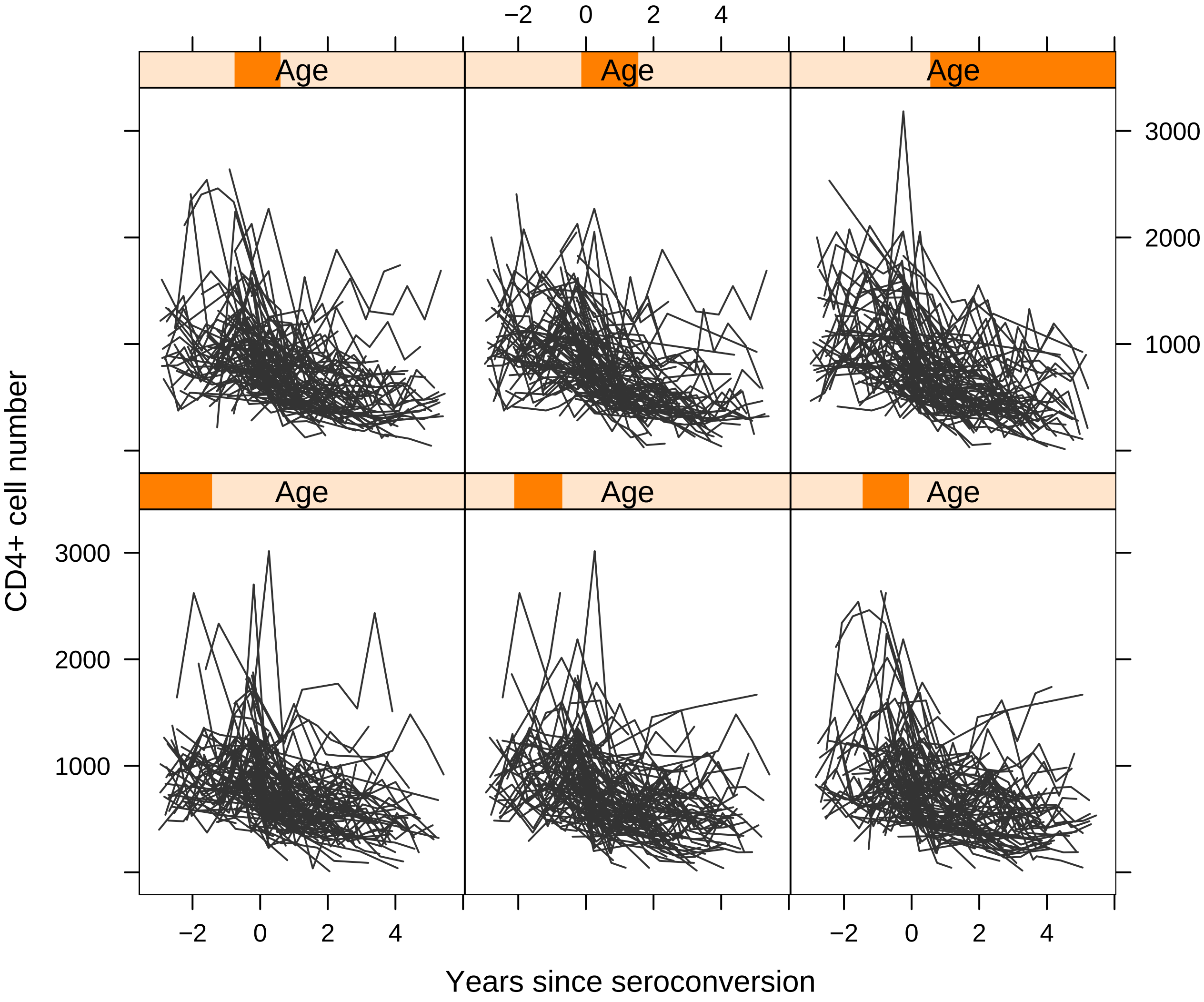
We can plot

- baseline CD4+ cell number y_{i1} against baseline CESD score x_{i1}; and
- $y_{ij} - y_{i1}$ against $x_{ij} - x_{i1}$,

to separate the cross-sectional and longitudinal effects.
See Fig. 3.8. in the textbook.

Conditional plots can be used to explore more than one covariates:

```
xyplot (CD4 ~ Time | equal.count (Age, 6), data = CD4,type = "l",  group = ID,
        xlab = "Years since seroconversion",col.line = "gray20",
        ylab = "CD4+ cell number",strip = strip.custom (var.name = "Age"))
```


Age
Age
Age
Age
Age
Age
CD4+ cell number
Years since seroconversion
−2
0
2
4
1000
2000
3000

Exploring Correlation Structure

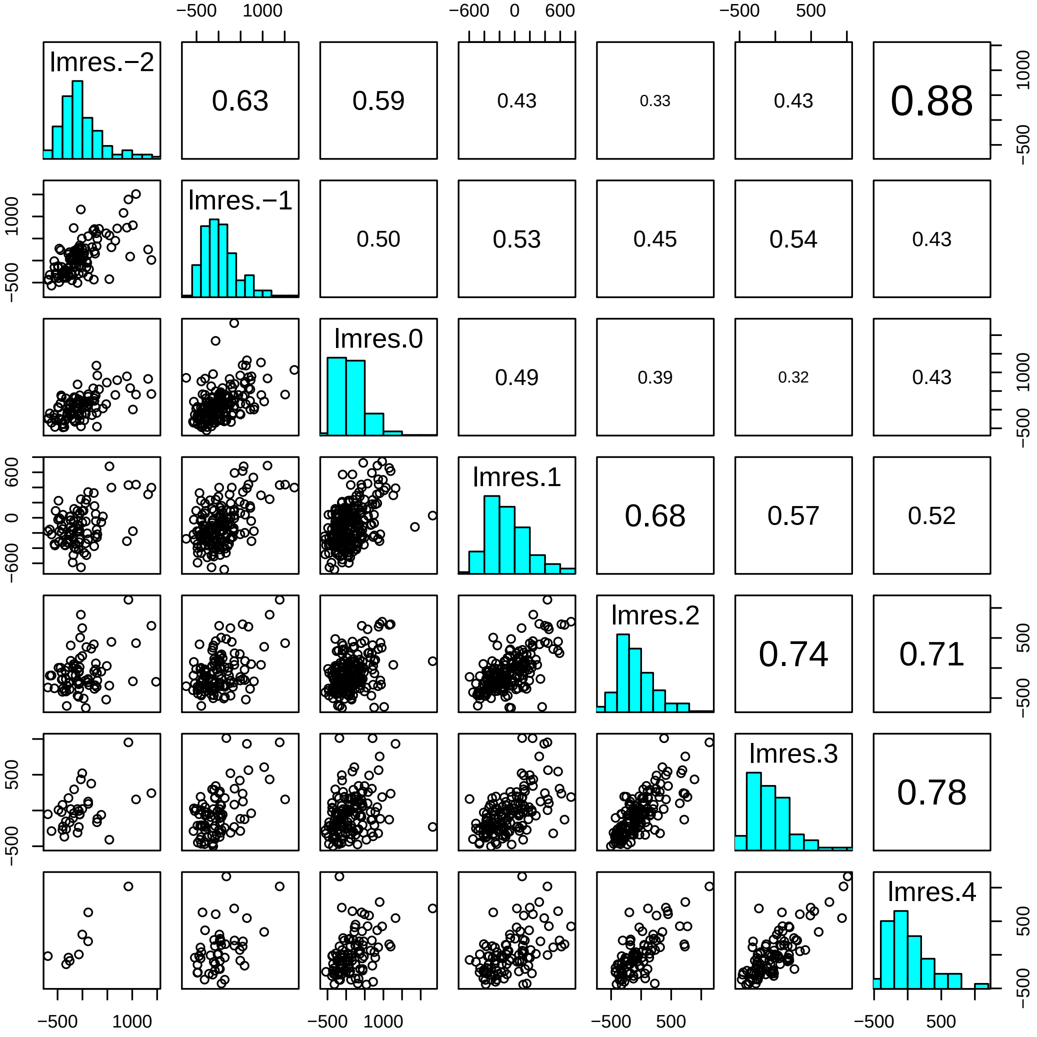
- To remove the effects of covariates, we first regress y_{ij} to the covariates $\boldsymbol{x}_{ij}$ using ordinary least-squares (OLS) and obtain the residuals:

$$r_{ij} = y_{ij} - \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}.$$

- We can then obtain the scatter-plot matrix using the residuals at different time points (or time intervals if the time points are not regular).

```
CD4.lm <- lm (CD4 ~ Time, data = CD4)
CD4$lmres <- resid (CD4.lm)
CD4$roundyr <- round (CD4$Time)
## Reshape the data to wide format
CD4w <- reshape (CD4[,c("ID", "lmres", "roundyr")],
                 direction = "wide",
                 v.names = "lmres", timevar = "roundyr",
                 idvar = "ID")
```

```
## Put histograms on the diagonal
panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks;
    nB <- length(breaks)
    y <- h$counts;
    y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)}
## Put (absolute) correlations on the upper panel, w/ size prop. to correlation.
panel.cor <- function(x, y, digits=2, prefix="", cex.cor) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs (cor(x, y, use = "pairwise.complete.obs"))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
         txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex * r)
}
pairs (CD4w[,c(5,2,3,6:9)], upper.panel = panel.cor,diag.panel = panel.hist)
```


lmres.−2
lmres.−1
lmres.0
lmres.1
lmres.2
lmres.3
lmres.4
0.63
0.59
0.43
0.33
0.43
0.88
0.50
0.53
0.45
0.54
0.43
0.49
0.39
0.32
0.43
0.68
0.57
0.52
0.74
0.71
0.78
−500
1000
−600
0
600
500

- The correlation is weaker for more distant observations.
- There is some hint that the correlation between observations at time t_i and t_j depends *primarily* on $|t_i - t_j|$.

Weakly Stationarity

Using the above CD4 data as an example, if the residuals r_{ij} have constant mean and variance for all j and if $\mathrm{Cor}(r_{ij}, r_{ik})$ depends only on the distance $|t_{ij} - t_{ik}|$, then this residual process is said to be *weakly stationary*.

How to check whether a process is weakly stationary?

- mean is constant w.r.t. j
- variance is constant w.r.t. j
- diagonals on correlation matrix are constant

Auto-correlation Function

- Assuming stationarity, a single correlation estimate can be obtained for each distinct values of the time separation or lag ($u = |t_{ij} - t_{ik}|$). This corresponds to pooling observation pairs along the diagonals of the scatterplot matrix.
- $\rho(u) = \text{Cor}(\epsilon_{ij}, \epsilon_{ij-u})$.
- If stationary seems appropriate, pool estimates from the same diagnal to increase precision.
- For example, for lag $u = 1$, we pool residuals from time points -2, -1, 0, 1, 2, 3 and pair them with residuals from time points -1, 0, 1, 2, 3, 4 and compute the correlation.
- Repeat this for each lag, u, and draw autocorrelation function $\hat{\rho}(u)$ vs u.
- Estimated autocorrelation function for CD4+ residuals

u:	1	2	3	4	5	6
$\hat{\rho}(u)$:	0.60	0.54	0.46	0.42	0.47	0.89

- The autocorrelation function is most effective for studying equally spaced data (note that we round observation times for the CD4+ data).
- Auto-correlations are more difficult to estimate with irregularly spaced data unless we round observation times as was done for the CD4+ data.

Variogram

An alternative function describing associations among repeated observations with irregular observation times is the *variogram*. For a stochastic process $Y(t)$, the variogram is defined as:

$$\gamma(u) = \frac{1}{2}\,\mathrm{E}\left[\{Y(t) - Y(t-u)\}^2\right], \quad u \geq 0. \tag{1}$$

If $Y(t)$ is stationary, the variogram is directly related to the autocorrelation function, $\rho(u)$, by

$$\gamma(u) = \sigma^2\{1 - \rho(u)\},$$

where σ^2 is the variance of $Y(t)$.

Proof.

$$\begin{aligned}
\gamma(u) &= \frac{1}{2}\mathrm{Var}[Y(t)] + \frac{1}{2}\mathrm{Var}[Y(t-u)] - \mathrm{Cov}[Y(t), Y(t-u)] \\
&= \frac{\sigma^2}{2} + \frac{\sigma^2}{2} - \sigma^2\rho(u) \\
&= \sigma^2(1 - \rho(u)).
\end{aligned}$$

Sample Variogram

- For each i and $j < k$, calculate

  $$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$$

  and the corresponding time-differences

  $$u_{ijk} = t_{ij} - t_{ik}.$$

- The sample variogram $\hat{\gamma}(u)$ is the average of all v_{ijk} corresponding to that particular value of u (we need more than one observations at each value of u).
- with highly irregular sample times, the variogram can be estimated from the data (u_{ijk}, v_{ijk}), by fitting a non-parametric curve (i.e., a loess line).
- the process variance σ^2 can be estimated by the variance of the residuals.
- Hence in the stationary case, the autocorrelation function at any lag u can be estimated from the sample Variogram by the formula

  $$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}. \tag{2}$$

- Do remember that in a typical longitudinal study, the correlation is usually not the main question of interest so it is not necessary to spend too much effort modeling the variogram.

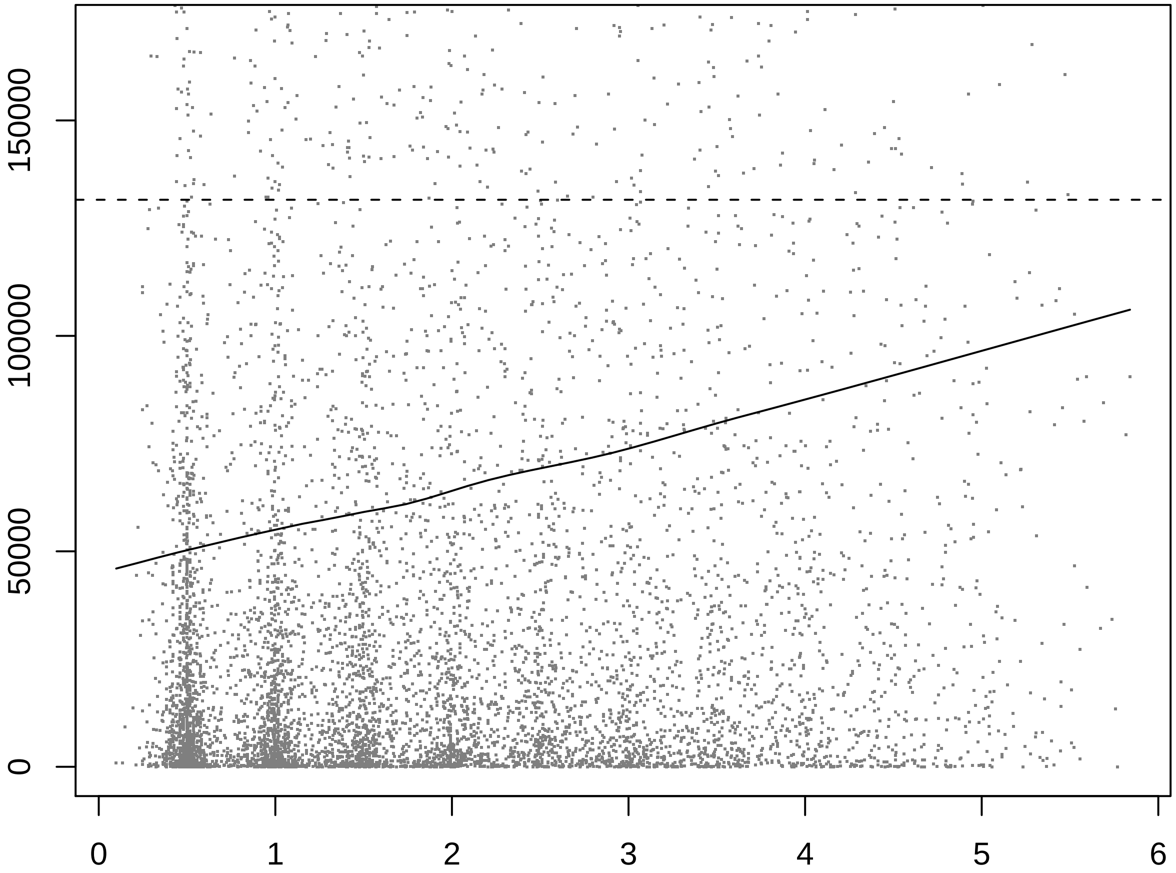
```
lda.vg <- function (id, res, time, plot = TRUE, ...) {
    vv <- tapply (res, id,
          function (x) outer (x, x, function (x, y) (x - y)^2/2))
    v <- unlist (lapply (vv, function (x) x[lower.tri (x)]))

    uu <- tapply (time, id,
          function (x) outer (x, x, function (x, y) (x - y)))
    u <- unlist (lapply (uu, function (x) x[lower.tri (x)]))

    if (plot) {
        vg.loess <- loess.smooth (u, v, family = "gaussian")

        plot (v ~ u, pch = ".", col = "gray50", ...)
        lines (vg.loess, lty = 1)
        abline (h = var (res), lty = 2)
    }
    invisible (data.frame (v = v, u = u))
}

lda.vg (CD4$ID, CD4$lmres, CD4$Time,
        ylim = c(0, 170000), ylab = "", xlab = "")
```



Sample variogram of CD4+ residuals.

A General Covariance Model

Diggle (1988) proposed the following model

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \tag{3}$$

where the error can be written as

$$\epsilon_{ij} = \epsilon_i(t_{ij}) = u_i + W_i(t_{ij}) + Z_{ij}. \tag{4}$$

In this decomposition, there are three sources of variation:

$$\begin{array}{ll} \text{random intercepts:} & u_i \\ \text{serial process:} & W_i(t_{ij}) \\ \text{measurement error:} & Z_{ij} \end{array}$$

If we further assume

$$\begin{aligned} \operatorname{Var}(u_i) &= \nu^2, \\ \operatorname{Cov}(W_i(s), W_i(t)) &= \sigma^2 \rho(|s-t|), \\ \operatorname{Var}(Z_{ij}) &= \tau^2, \end{aligned}$$

then,

$$\operatorname{Cov}(\epsilon_i(t), \epsilon_i(t-u)) = \begin{cases} \nu^2 + \sigma^2 \rho(u), & u > 0, \\ \nu^2 + \sigma^2 + \tau^2, & u = 0. \end{cases}$$

Variograms can be used to characterize these variance components (σ^2, τ^2, ν^2).

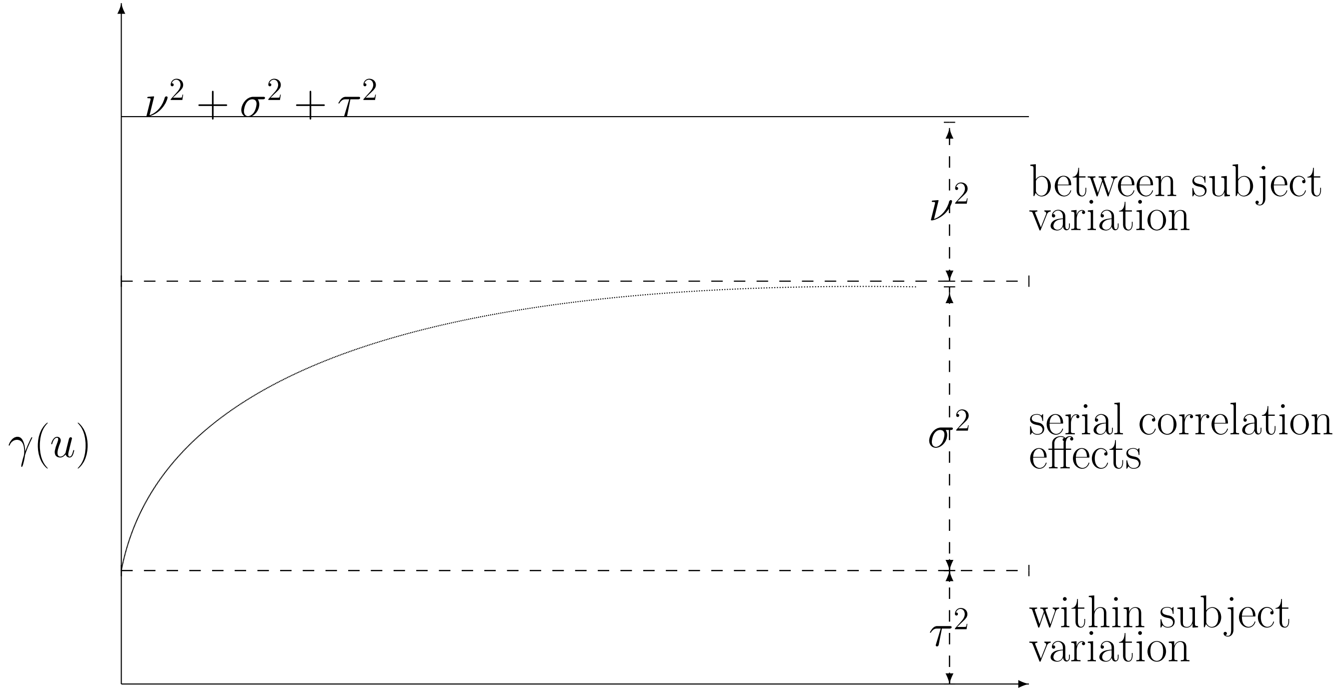
Characterizing Variance Components

Using Diggle's model, the variogram is defined as

$$\begin{aligned}\gamma(u) &= \frac{1}{2}\,\mathrm{E}[\epsilon_i(t) - \epsilon_i(t-u)]^2 \\ &= (\nu^2 + \sigma^2 + \tau^2) - (\nu^2 + \sigma^2\rho(u)) \\ &= \sigma^2(1-\rho(u)) + \tau^2.\end{aligned}$$

Assume $\rho(u) \longrightarrow 1$ as $u \longrightarrow 0$ and $\rho(u) \longrightarrow 0$ as $u \longrightarrow \infty$ so that

$$\gamma(u) \longrightarrow \begin{cases} \tau^2, & \text{as } u \longrightarrow 0, \\ \tau^2 + \sigma^2, & \text{as } u \longrightarrow \infty. \end{cases}$$


$\nu^2 + \sigma^2 + \tau^2$
$\gamma(u)$
ν^2
between subject variation
σ^2
serial correlation effects
τ^2
within subject variation

Exploring Association Amongst Categorical Response

- For a binary response, a correlation-based approach is less natural — the range of correlation between a pair of binary variables is constrained by their means.
- A more natural measure of association is log-odds ratio

  $$\mathrm{OR} = \gamma(Y_1, Y_2) = \frac{\Pr(Y_1 = 1, Y_2 = 1)\Pr(Y_1 = 0, Y_2 = 0)}{\Pr(Y_1 = 1, Y_2 = 0)\Pr(Y_1 = 0, Y_2 = 1)}$$

  Log-odds Ratio $= \log\gamma$.
- For a longitudinal sequence $Y_{i1}, \dots, Y_{in}$ at measurement times $t_1, \dots, t_n$, define *lorelogram* to be the function

  $$\mathrm{LOR}(t_j, t_k) = \log\gamma(Y_j, Y_k).$$
- For longitudinal data, y_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n_i$, replace theoretical probability with sample proportions across subjects.

Further Reading

- Chapters 3 and 5.2 of textbook (Diggle et al 2002).