eNote 12

# Repeated measures, part 2, advanced methods

# Indhold

## 12.1 Intro

This module describe a selection of models, with a covariance structure especially aimed at repeated measurements data. The simplest of these models is the random effects model known from the previous module, where all measurements on the same individual are equally correlated no matter how far apart. The models in this module elaborates and extends the random effects approach to models with fairly flexible covariance structures.

## 12.2 A different view on the random effects approach

Recall the random effects model presented in the last module, where the "individual" variable was added as a random effect. The covariance structure in this model turned out to be the structure where two measurements from the same individual are correlated, but equally correlated no matter how far apart the measurements were taken.

Remember from the first theory module that any mixed model can be expressed as:

$$\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{ZGZ}' + \mathbf{R}),$$

where $\mathbf{X}$ is the design matrix for the fixed effects part of the model, $\beta$ is the fixed effects parameters, and $\mathbf{Z}$ is the design matrix for the random effects. The two matrices $\mathbf{G}$ and $\mathbf{R}$ describe the covariance between the random effects in the model ($\mathbf{G}$), and the residual/remaining measurement errors ($\mathbf{R}$).

In the random effects approach for repeated measurements, the desired covariance structure for the observations $\mathbf{y}$ was obtained by adding the "individual" variable as a random effect. In terms of the general mixed model setup this corresponds to:

- $\mathbf{G}$ being a diagonal matrix with the variance between individuals in the diagonal and zeros everywhere else
- $\mathbf{Z}$ being the design matrix with one column for each individual with ones in the rows where the corresponding observation is from that individual and zeros everywhere else
- $\mathbf{R}$ being a diagonal matrix with the variance of the independent measurement error in the diagonal and zeros everywhere else

The desired covariance structure for the observations $\mathbf{y}$ (described in the previous module) is obtained by:

$$\mathbf{V} = \text{cov}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$$

This random effects approach corresponds well with the intuition behind the data, but in fact the exact same model could be obtained by leaving out the $\mathbf{ZGZ}'$ term and putting the desired variance structure directly into the $\mathbf{R}$ matrix.

The variance structure in the random effect model is:

$$\text{cov}(y_{i_1}, y_{i_2}) = \begin{cases} 0 & \text{, if } \texttt{individual}_{i_1} \neq \texttt{individual}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma^2_{\texttt{individual}} & \text{, if } \texttt{individual}_{i_1} = \texttt{individual}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma^2_{\texttt{individual}} + \sigma^2 & \text{, if } i_1 = i_2 \end{cases}$$

which simply states, that two measurements from the same individual are correlated, but equally correlated no matter how far apart the measurements were taken. This variance structure is known as *compound symmetry*.

### 12.2.1 Example: Activity of rats analyzed via compound symmetry model

The rats data set from the previous module is also used in this module. Recall the experiment:

- 3 treatments: 1, 2, 3 (concentration)
- 10 cages per treatment
- 10 contiguous months
- The response is activity ($\log(\text{count})$ of intersections of light beam during 57 hours).

In this setup the "individual" variable is `cage`.

The model is exactly the same as the random effects approach, but it will be written slightly different to better illustrate this new way of specifying it.

$$\begin{aligned} \mathbf{lnc} &\sim N(\mu, \mathbf{V}), \quad \text{where} \\ \mu_i &= \mu + \alpha(\texttt{treatm}_i) + \beta(\texttt{month}_i) + \gamma(\texttt{treatm}_i, \texttt{month}_i), \quad \text{and} \\ V_{i_1,i_2} &= \begin{cases} 0 & \text{, if } \texttt{cage}_{i_1} \neq \texttt{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma^2_d & \text{, if } \texttt{cage}_{i_1} = \texttt{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma^2_d + \sigma^2 & \text{, if } i_1 = i_2 \end{cases} \end{aligned}$$

This way of specifying the model is very direct. It is specified that the observations follow a multivariate normal distribution with a mean value depending on the fixed effects, and a covariance structure explicitly specified.

In the following we implement this model in R. In R this model without random effects cannot be specified using the function `lme`, nor with `lmer` but instead the function `gls` in the package `nlme` can be used

```
library(nlme)
rats <- read.table("rats.txt", header=T, sep=",", dec=".")
rats$monthQ <- rats$month # Make the quantitative version
rats$month <- factor(rats$month) # Make the factor version
rats$treatm <- factor(rats$treatm)
rats$cage <- factor(rats$cage)
model1<-gls(lnc~month+treatm+month:treatm,
            correlation = corCompSymm(form=~1|cage),data=rats)
```

The correlation structure is specified using the `correlation` argument. The value of this argument should an `corStruct` object. Typing `?corClasses` produces a list of predefined object classes. The given `corCompSymm` object corresponds to a compound symmetry structure.

Compare with the random effects approach for this data set (in the previous module) and notice that there is no random effect notation here - neither the `lme`-type nor the `lmer`-type. Instead it is replaced with the `correlation` argument.

Some of the relevant output is listed below:

```
summary(model1)
```

```
Linear mixed-effects model fit by REML
 Data: rats
       AIC      BIC    logLik
  72.61464 187.7641 -4.307319

Random effects:
 Formula: ~1 | cage
        (Intercept)  Residual
StdDev:   0.1657654 0.1946757
```

And the ANOVA:

```
xtable(anova(model1))
```

| | numDF | F-value | p-value |
|---|---|---|---|
| (Intercept) | 1 | 85524.70 | 0.00 |
| month | 9 | 46.11 | 0.00 |
| treatm | 2 | 3.22 | 0.04 |
| month:treatm | 18 | 2.12 | 0.01 |

Compare with the output from the random effects approach, from Module 11 and see that all estimates and tests are identical. This should be expected, as it is exactly the same model. Note that the default anova given here is the Type 1 (`type="sequential"`) anova. Use e.g. (`type="marginal"`) to get the type 3 table.

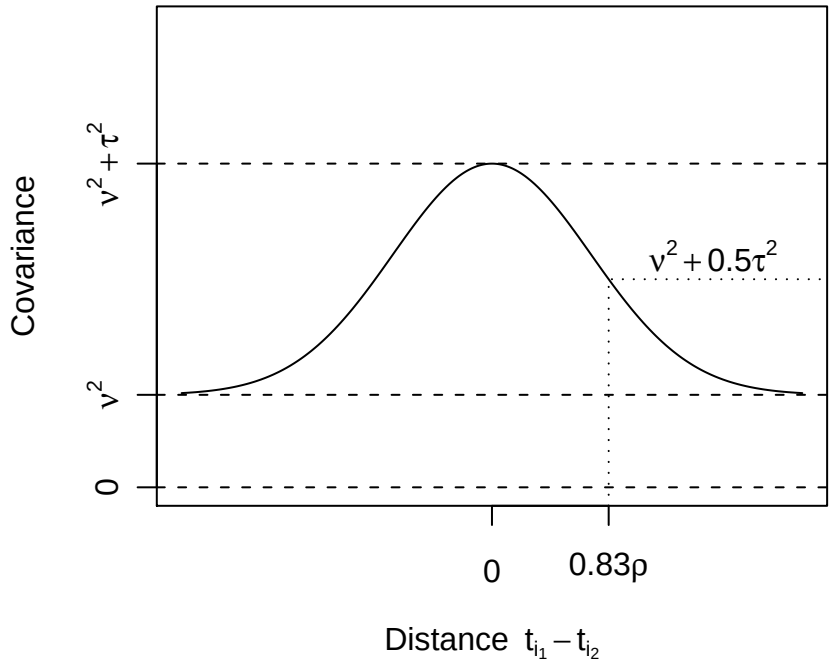
This section has described a different way to specify the random effects approach for repeated measurements. This is not very useful in itself, but this way of directly specifying the variance structure can be extended to include covariance structures that could not be specified by random effects alone. For instance time (or space) dependent correlation structures. Covariance structures depending on "how far" observations are apart are known as *spatial* covariance structures, also when the distance is time.

The two approaches can also be used in combination with each other. The `correlation` part only specifies the $\mathbf{R}$ matrix. Random effects can be added with the `lme` random effect notation.

## 12.3 Gaussian model of spatial correlation

The main problem with the random effects model (by now also known as the compound symmetry model) is that all measurements within the same individual are equally correlated. This is counterintuitive if some measurements are close (in time or space) and some are far apart. To fix this the following model has been proposed:

$$
\begin{aligned}
\mathbf{y} &\sim N(\mu, \mathbf{V}), \quad \text{where} \\
\mu_i &= \ldots \text{ (depends on fixed effects of the model), and} \\
V_{i_1,i_2} &= \begin{cases} 0 & \text{, if } \texttt{individual } i_1 \neq \texttt{individual } i_2 \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 \exp\left\{\frac{-(t_{i_1}-t_{i_2})^2}{\rho^2}\right\} & \text{, if } \texttt{individual } i_1 = \texttt{individual } i_2 \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 + \sigma^2 & \text{, if } i_1 = i_2 \end{cases}
\end{aligned}
$$




Figur 12.1: The Gaussian serial correlation illustrated. Two observations "very close" together have covariance $\nu^2 + \tau^2$ and two observations "very far" apart have covariance $\nu^2$. The curve indicate how the correlation drops as a function of the distance between the observations. How fast this decline from $\nu^2 + \tau^2$ to $\nu^2$ occur depends on the parameter $\rho$, as indicated on the graph ($0.83 \approx \sqrt{\log(2)}$)

The covariance structure of this model is an extension of the compound symmetry structure.

- Two observations from different individuals are independent
- Two observations "very close" together have covariance $\nu^2 + \tau^2$ and two observations "very far" apart have covariance $\nu^2$. How fast this decline in covariance from $\nu^2 + \tau^2$ to $\nu^2$ occur depends on the parameter $\rho$. The shape of this decline is illustrated in figure 12.1
- The total variance of a single observation is $\nu^2 + \tau^2 + \sigma^2$

This structure is know as *spatial Gaussian correlation*, because the covariance declines like the density of a normal/Gaussian distribution (see figure12.1).

### 12.3.1 Example: Activity of rats analyzed via spatial Gaussian correlation model

The natural model for the rats data set is a model where the correlation between two measurements on the same cage depends on how far apart the measurements are taken. One such model is the Gaussian serial correlation model.

$$
\begin{aligned}
\mathbf{lnc} &\sim N(\mu, \mathbf{V}), \quad \text{where} \\
\mu_i &= \mu + \alpha(\texttt{treatm}_i) + \beta(\texttt{month}_i) + \gamma(\texttt{treatm}_i, \texttt{month}_i), \quad \text{and} \\
V_{i_1,i_2} &= \begin{cases} 0 & , \text{ if } \texttt{cage}_{i_1} \neq \texttt{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 \exp\left\{ \frac{-(\texttt{month}_{i_1} - \texttt{month}_{i_2})^2}{\rho^2} \right\} & , \text{ if } \texttt{cage}_{i_1} = \texttt{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 + \sigma^2 & , \text{ if } i_1 = i_2 \end{cases}
\end{aligned}
$$

Notice that this is the same model as the random effects model, except for the added term in the covariance structure. The following lines implement this model in R:

```
model2<-lme(lnc~month+treatm+month:treatm,
            random=~1|cage,
            correlation=corGaus(form=~as.numeric(month)|cage,nugget=T),
            data=rats)
VarCorr(model2)


cage = pdLogChol(1)
            Variance   StdDev
(Intercept) 0.01971373 0.1404056
Residual    0.04715671 0.2171559


-2*logLik(model2)


'log Lik.' -105.3134 (df=34)


intervals(model2, which = "var-cov")


Approximate 95% confidence intervals

 Random Effects:
  Level: cage
```

```
                    lower      est.     upper
sd((Intercept)) 0.0880286 0.1404056 0.2239468

 Correlation structure:
           lower      est.     upper
range  1.8411387 2.3863954 3.0931310
nugget 0.1440834 0.2186743 0.3175538
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
    lower      est.     upper
0.1881918 0.2171559 0.2505779
```

And the anova table:

```
xtable(anova(model2), digits = 3)
```

| | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| (Intercept) | 1 | 243.000 | 79826.321 | 0.000 |
| month | 9 | 243.000 | 41.449 | 0.000 |
| treatm | 2 | 27.000 | 2.131 | 0.138 |
| month:treatm | 18 | 243.000 | 1.663 | 0.047 |

The spatial Gaussian correlation structure is specified in the `correlation` argument giving the object `corGaus` which has a first argument `form` specifying the time variable after $\sim$ and the grouping variable (independence between groups) after — and a second argument `nugget` taking a logical value deciding whether or not a fourth variance parameter should be added to the model. Notice the `as.numeric` around `month` in the specification of the correlation structure. The time variable (here `month`) should not be a factor, but a covariate.

The parameterisation is not exactly the same as the parameterisation used in the model expression. The square of the estimated residual standard deviation equals the sum $\hat{\sigma}^2 + \hat{\tau}^2$ of parameters in the model, the parameter estimate under `range` equals $\hat{\rho}^2$ and the parameter estimate under `nugget` equals $\hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\tau}^2)$. The estimated variance com-

ponent for cage is the same in the two parameterisations. Thus we have the equations

$$\begin{aligned}\hat{v}^2 &= 0.14040562^2,\\ \hat{\tau}^2 &= (1-0.2186744)\cdot 0.2171559^2 = 0.03684473,\\ \hat{\rho}^2 &= 2.3863954,\\ \hat{\sigma}^2 &= 0.2186744\cdot 0.2171559^2 = 0.01031196.\end{aligned}$$

From the ANOVA table it is seen that the interaction between treatment and month is not significant. The P–value is 0.047, which is just slightly below the the usual 5% level. This result is different from the random effects model where the same P–value was 0.0059. Judging from this, it made an important difference to extend the covariance structure. In the next section, a way to formally compare these two covariance structures, will be described.

We should note here that the ANOVA table produced for `lme`-results does NOT use the correction of degrees of freedom (Satterthwaithe and/or Kenward Roger) that we have otherwised used. (From a SAS analysis it has been seen that Satterthwaite corrected denominator degrees of freedom becomes 85.6 and the interaction $p$-value then becomes 0.0626 instead)

## 12.4 Test for model reduction

To test a reduction in the variance structure a *restricted/residual likelihood ratio test* can be used. A likelihood ratio test is used to compare two models $A$ and $B$, where $B$ is a submodel of $A$. Typically the model including some variance components (model $A$), and the model without these variance components (model $B$) is to be compared. To do this the negative restricted/residual log-likelihood values ($\ell_{re}^{(A)}$ and $\ell_{re}^{(B)}$) from both models must be computed. The test can now be computed as:

$$G_{A\to B} = 2\ell_{re}^{(B)} - 2\ell_{re}^{(A)}$$

The likelihood ratio test statistic follows a $\chi^2_{df}$–distribution, where $df$ is the difference between the number of parameters in $A$ and the number of parameters in $B$. In the case of comparing the spatial Gaussian correlation model ($A$) with the random effects model ($B$) $df = 2$.

The following table show this comparison for the rats data set:

```
xtable(as.matrix(anova(model1, model2))[,-1])
```

| | Model | df | AIC | BIC | logLik | Test | L.Ratio | p-value |
|---|---|---|---|---|---|---|---|---|
| model1 | 1 | 32 | 72.61464 | 187.76414 | -4.307319 | | | |
| model2 | 2 | 34 | -37.31339 | 85.03296 | 52.656695 | 1 vs 2 | 113.928 | 0 |

Or differently expressed:

| **Model** | $2\ell_{\mathbf{re}}$ | **G–value** | **df** | **P–value** |
|---|---|---|---|---|
| (A) Spatial Gaussian | -105.3 | $G_{A\to B} = 113.9$ | 2 | $P_{A\to B} < 0.0001$ |
| (B) Random effects | 8.6 | | | |

It follows that the spatial Gaussian correlation model cannot be reduced random effects model.

## 12.5 Other serial correlation structures

The spatial Gaussian correlation structure is only one among several possible covariance structures implemented in R. A few options are listed in the following table:

| **Write in R** | **Name** | **Correlation term** |
|---|---|---|
| `corGaus` | Gaussian | $\tau^2 \exp\{\frac{-(t_{i_1}-t_{i_2})^2}{\rho^2}\}$ |
| `corExp` | exponential | $\tau^2 \exp\{\frac{-\lvert t_{i_1}-t_{i_2}\rvert}{\rho}\}$ |
| `corAR1` | autoregressive(1) | $\tau^2\rho^{\lvert i_1-i_2\rvert}$ |
| `corSymm` | unstructured | $\tau^2_{i_1,i_2}$ |

This list only gives a brief idea about the different possible structures. A complete list can be found by writing `?corClasses`.

With all these possible covariance structures it would be nice with a bulletproof method to choose "the right one". Unfortunately such a method does not exist in general. The restricted/residual likelihood ratio test can only be used in those cases where the models are sub–models of each other, and even in some of those cases the $\chi^2$–approximation can be dubious.

Graphical methods can in some cases be used to aid the selection of covariance structure. These methods consists of estimating the correlation from the model residuals, and

then plotting these correlations as a function of the time difference. A variation of these plots is known as the *(semi)–variogram*. The semi–variogram can be used to get an impression of the shape of the spatial correlation, but it typically requires "many" observations on each individual to be able to distinguish between the different covariance structures[1].

R computes a few numerical *information criteria*, which can be used as a guideline when choosing between different covariance structures. Two of these are: Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Both are computed as the negative log–likelihood plus some penalty for the number of parameters used in the model. (BIC) gives a harder penalty for many parameters that (AIC). The best covariance structure (according to these criteria) is the one with the lowest criteria value.

For the rats data set the two criteria for the compound symmetry model is: AIC=12.6 and BIC=15.4, and for the spatial Gaussian model: AIC=-97.3 and BIC=-91.7. In this case there is little doubt that the spatial Gaussian is the better model (but this was already known from the likelihood ratio test).

Notice that even if the main interest is in the fixed effects it is important not to choose a wrong covariance structure. In the rats example the P–value for no interaction term in the compound symmetry model is 0.0059, and the same P–value in the spatial Gaussian model is 0.047, which could lead to different conclusions about the treatment effect.

## 12.6 Analysis strategy

Figure 12.2 shows a strategy that can adopted when analyzing repeated measurements via the mixed model. The first step is to identify the "individuals", within which the observations might be correlated. It need not be an animal or a person. Depending on the problem at hand it can be fields, test–tubes, meat–slices or something completely different. Once the "individuals" have been identified, the fixed effects of the model can be selected. The main and interaction terms of interest are included, just like in a pure fixed effects model.

The third step is to select a covariance structure. This is the tricky part. The choice can be aided by the different information criteria, but for short individual series this is really the modelers choice. The significance of the parameters of the covariance structure can be tested with a likelihood ratio test. This is indicated by the "change model" arrow on the left side of the diagram.

[1] In the field of geo-statistics the semi–variogram is a standard tool to investigate the covariance structure. In geo–statistics long data series are common.

Once the covariance structure is selected (and possibly reduced), the significance of the fixed effects part of the model can be tested. Whenever the fixed effects structure is reduced, the covariance structure should ideally be re–validated, which is indicated by the "change model" arrow on the right side of the diagram. This step is often omitted mainly for simplicity, but also by the argument that a non–significant change in the mean parameters should not change the covariance structure much. The final model can now be used draw inference (estimate parameters, setup confidence intervals and interpret results).

Given the nature and complexity of this type of models, it is recommended that main conclusions of a given study should be cross–validated with one of the simple methods from the last module whenever possible. For instance if a model with the spatial Gaussian covariance structure show a significant treatment effect, it might also be possible to validate this effect by analyzing a good summary measure.

## 12.7 The semi-variogram

This additional section briefly introduces the semi-variogram, mentioned above. Consider repeated measurements $Y_1, \ldots, Y_n$ taken over time at time points $t_1, \ldots, t_n$ for a single subject, and denote by $\lambda(|t_i - t_j|)$ the *serial* correlation between two measurements taken at time $t_i$ and $t_j$. For simplicity denote $u = |t_i - t_j|$ $(u \geq 0)$.

For the spatial Gaussian correlation model the correlation function is

$$\lambda(u) = \exp\{-u^2/\rho^2\}.$$

The parameter $\rho^2$ is sometimes called the *range*. In addition to the serial correlation, the spatial Gaussian correlation model (or any other spatial correlation model) usually also contains a random factor reflecting the variation in the subjects, resulting in the following covariance within a subject for measurements time $|t_i - t_j|$ apart

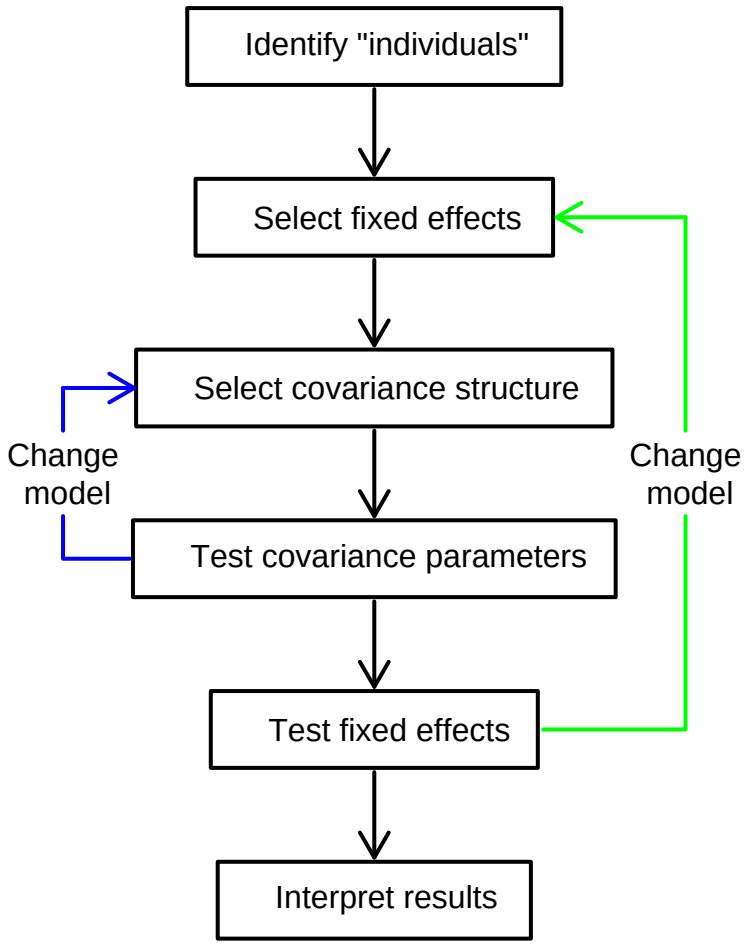
$$\nu^2 + \tau^2\lambda(u). \tag{12-1}$$

It follows from (12-1) that the variance is $\nu^2 + \tau^2$ (setting $u = 0$). Finally there is the residual variation which adds a component (sometimes called a *nugget* effect) to the variance

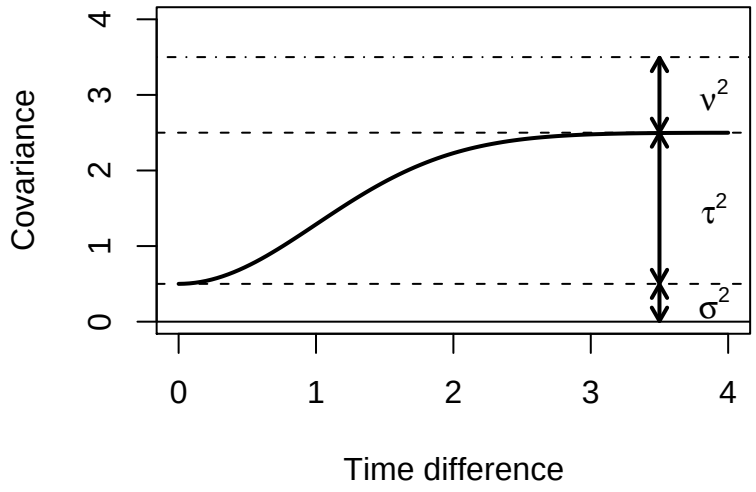
$$\text{cov}(Y_i, Y_j) = \sigma^2 + \nu^2 + \tau^2.$$

Having this decomposition of the variation in mind, we proceed to define the *semi-variogram* as the function

$$\gamma(u) = \tau^2\big(1 - \lambda(u)\big).$$




Figur 12.2: The strategy for analyzing repeated measurements via the mixed model




Figur 12.3: The structure of the (semi)variogram

The semi-variogram can be estimated from the data using residuals from a model without any spatial correlation structure. This estimate is called an empirical semi-variogram and, to be precise, it is an estimate of $\sigma^2 + \gamma(u)$. As $u$ approaches $\infty$ $\gamma(u)$ tends to $\tau^2$, implying that for large values of $u$ the value of the empirical semi-variogram is close to $\sigma^2 + \tau^2$. The difference between this term and the total variation is the contribution from the random factor $(\nu^2)$. This can be seen from the figure below for an idealised empirical semi-variogram.

Thus the dashed upper line indicates the total amount of variation in the data (remember that all measurements have the same variance in the spatial Gaussian correlation model), and in the spatial Gaussian correlation model this variation is then divided into three components: $\sigma^2$, $\tau^2$ and $\nu^2$. From the figure we get the following values

$$\sigma^2 = 0.5,$$
$$\tau^2 = 2,$$
$$\nu^2 = 1.$$

The points in the empirical semi-variogram will tend to be more variable as the time difference increases, because there are less points far apart than close. Therefore focus

should be on the section of the empirical semi-variogram corresponding to small time differences.

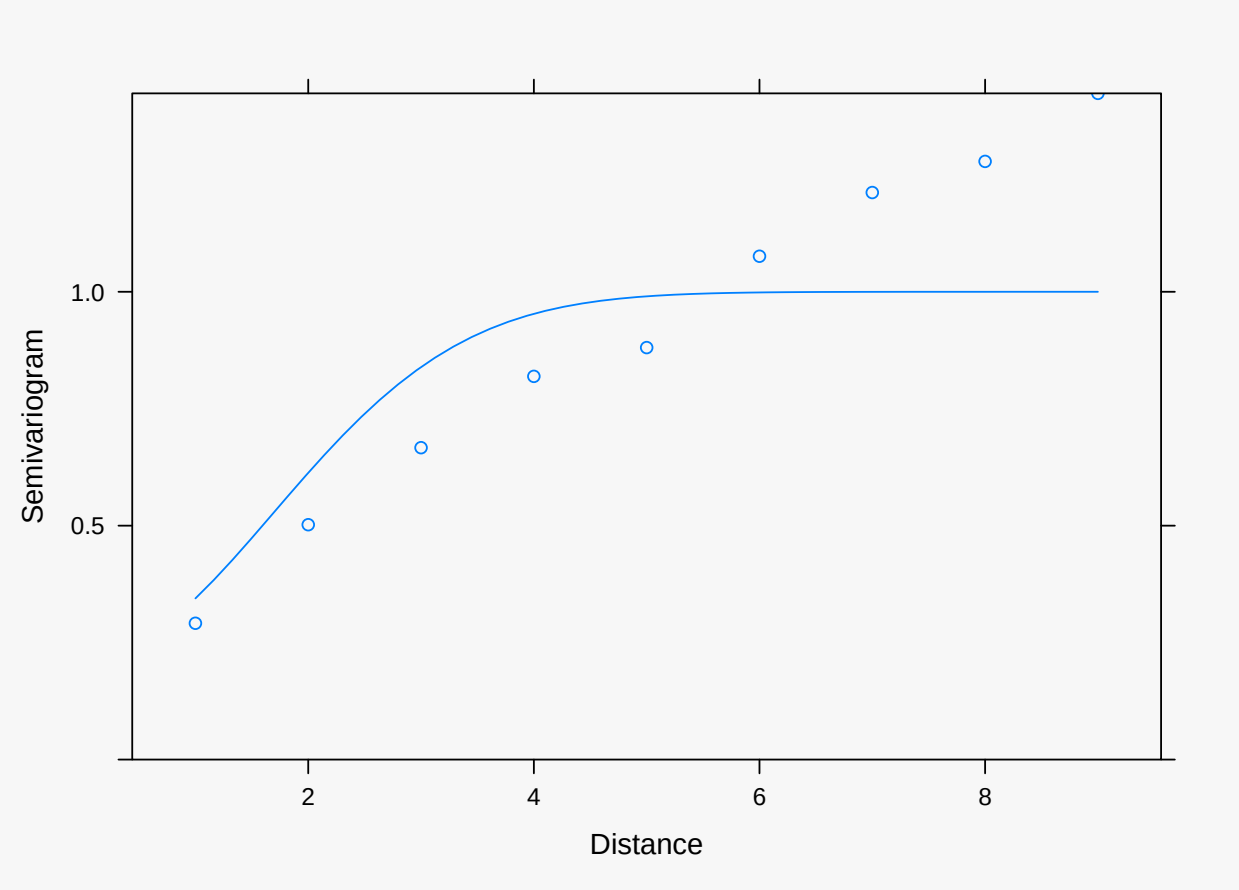
A model check is obtained by plotting both the empirical semi-variogram and the estimated (model-based) semi-variogram for the spatial correlation model.

The empirical semi-variogram can also be used to obtain initial values of the parameter estimates $\sigma_2$, $\tau^2$ and $\nu^2$ to facilitate the estimation of these parameters in the spatial Gaussian correlation model. These are simply read off the empirical semi-variogram (as described above).

### 12.7.1 Rats data example

To assess whether or not the correlation structure specified in the model is appropriate, the empirical semi-variogram and the estimated, model-based semi-variogram can be plotted using the function `Variogram`. This function also works for an `lme` object with no `correlation` argument, but in this case no model-based semi-variogram is supplied; instead the empirical semi-variogram is enhanced by adding a smoother to the plot. The arguments to `Variogram` are an `lme` object (the fitted model), a `form` argument indicating the time variable and the grouping variable for the spatial correlation model and a `data` argument containing the relevant data set (here `rats`).

```
print(plot(Variogram(model2, form = ~as.numeric(month) | cage, data = rats)))
```

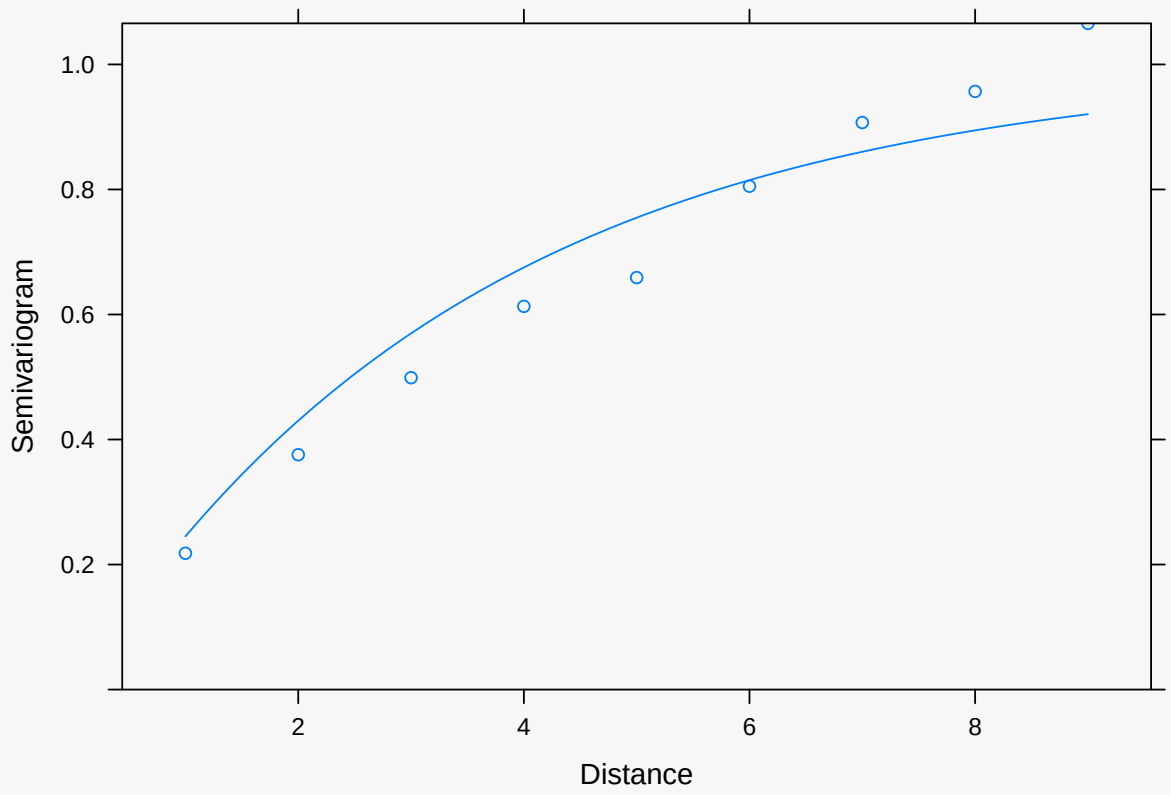



Notice that the y-axis is scaled to 1, meaning that the empirical semi-variogram in **R** cannot be used to to provide initial parameter estimates.

The agreement between the empirical semi-variogram and the spatial Gaussian correlation structure is not too good. Therefore two other correlation structures are fit: the spatial exponential correlation (`corExp`) and the autoregressive correlation (`corAR1`).

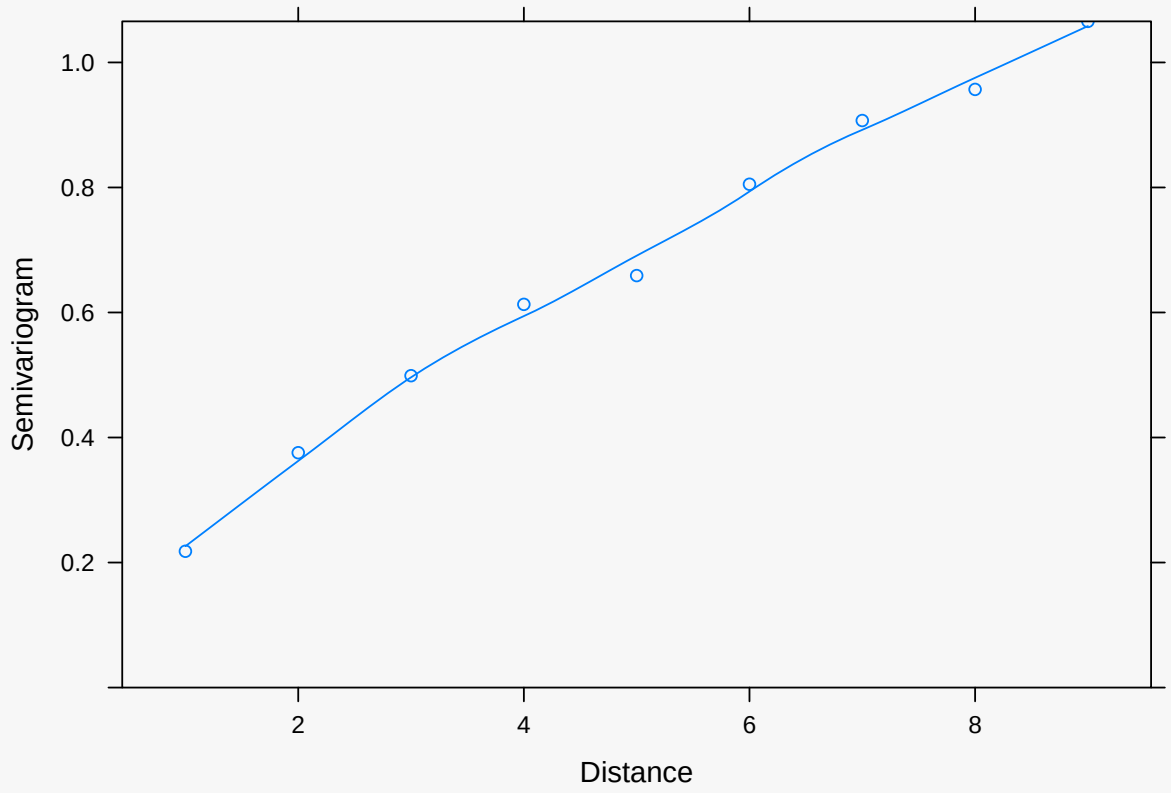
```
model3 <- lme(lnc ~ month + treatm + month:treatm, random = ~1 |
cage, correlation = corExp(form = ~as.numeric(month) |
cage, nugget = T), data = rats)

print(plot(Variogram(model3, form = ~as.numeric(month) | cage, data = rats)))
```




```
model4 <- lme(lnc ~ month + treatm + month:treatm, random = ~1 |
cage, correlation = corAR1(form = ~as.numeric(month) |
cage), data = rats)

print(plot(Variogram(model4, form = ~as.numeric(month) | cage, data = rats)))
```




It seems that both of them provide a somewhat better agreement with the empirical semi-variogram than the Gausian.

Comparison of `model2`, `model3` and `model4` by means of information criteria can be accomplished using the function `anova`:

```
xtable(as.matrix(anova(model2, model3, model4))[,-c(1,7:9)])
```

| | Model | df | AIC | BIC | logLik |
|---|---|---|---|---|---|
| model2 | 1 | 34 | -37.31339 | 85.03296 | 52.65669 |
| model3 | 2 | 34 | -42.71200 | 79.63434 | 55.35600 |
| model4 | 3 | 33 | -44.71200 | 74.03592 | 55.35600 |

The information criteria seem to favour the autoregressive correlation structure. We see that this is due to the fact that this model has one less variance parameter. If we look at the variance parameters of the spatial exponential correlation model:

```
summary(model3)
```

```
Correlation Structure: Exponential spatial correlation
 Formula: ~as.numeric(month) | cage
 Parameter estimate(s):
       range       nugget
3.556503e+00 3.341370e-08
```

we see that the nugget variance is basically zero, so let's try the Exponential spatial correlation but without the nugget:

```
model3b <- lme(lnc ~ month + treatm + month:treatm, random = ~1 |
cage, correlation = corExp(form = ~as.numeric(month) |
cage, nugget = FALSE), data = rats)
xtable(as.matrix(anova(model3b, model4))[,-1])
```

| | Model | df | AIC | BIC | logLik |
|---|---|---|---|---|---|
| model3b | 1 | 33 | -44.712 | 74.03592 | 55.356 |
| model4 | 2 | 33 | -44.712 | 74.03592 | 55.356 |

So these two models are giving exactly the same fit.

Some summary results:

```
xtable(anova(model3b))
```

| | numDF | denDF | F-value | p-value |
|---|---|---|---|---|
| (Intercept) | 1 | 243.00 | 78044.63 | 0.00 |
| month | 9 | 243.00 | 37.55 | 0.00 |
| treatm | 2 | 27.00 | 1.68 | 0.20 |
| month:treatm | 18 | 243.00 | 1.72 | 0.04 |

```
library(lsmeans)
lsmeans(model3b, "treatm", by="month")
```

```
month = 1:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.874280 0.08315533 29 9.704208 10.044352
 2      9.706260 0.08315533 27 9.535639  9.876881
 3      9.737014 0.08315533 27 9.566393  9.907635

month = 2:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.591410 0.08315533 29 9.421338  9.761482
 2      9.503630 0.08315533 27 9.333009  9.674251
 3      9.509809 0.08315533 27 9.339188  9.680430

month = 3:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.675270 0.08315533 29 9.505198  9.845342
 2      9.751012 0.08315533 27 9.580391  9.921633
 3      9.823261 0.08315533 27 9.652640  9.993882

month = 4:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.492995 0.08315533 29 9.322923  9.663067
 2      9.485335 0.08315533 27 9.314714  9.655956
 3      9.687759 0.08315533 27 9.517138  9.858380

month = 5:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.409991 0.08315533 29 9.239919  9.580063
 2      9.459219 0.08315533 27 9.288598  9.629840
 3      9.685679 0.08315533 27 9.515058  9.856300

month = 6:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.300140 0.08315533 29 9.130068  9.470212
 2      9.285450 0.08315533 27 9.114829  9.456071
 3      9.573260 0.08315533 27 9.402639  9.743881

month = 7:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.162416 0.08315533 29 8.992344  9.332488
 2      9.260123 0.08315533 27 9.089502  9.430744
```

```
 3      9.497330 0.08315533 27 9.326709  9.667951

month = 8:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.236198 0.08315533 29 9.066126  9.406270
 2      9.271053 0.08315533 27 9.100432  9.441674
 3      9.564056 0.08315533 27 9.393435  9.734677

month = 9:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      9.149882 0.08315533 29 8.979810  9.319954
 2      9.227982 0.08315533 27 9.057361  9.398603
 3      9.456433 0.08315533 27 9.285812  9.627054

month = 10:
 treatm   lsmean         SE df lower.CL  upper.CL
 1      8.972452 0.08315533 29 8.802380  9.142524
 2      8.854316 0.08315533 27 8.683695  9.024937
 3      9.037345 0.08315533 27 8.866724  9.207966

Confidence level used: 0.95
```

## 12.8 Analysing the time structure by polynomial regression

So far we have modelled the fixed effect time dependence with the time as a factor, hence a very general model of the patterns with no particular assumptions of the structure. This is often a good starting point for exactly that reason: It does not make any assumptions and the residual correlation structure is modelled for the ”pure” residuals, hence not running the risk of modelling a ”fixed time structure” as correlations in a fixed effect mis-specified model.

However, it may also sometimes provide a not so powerfull examination of time effects and/or time-treatment interaction effects, or differently put: it might, in some cases, be a perfectly reasonable model to express the time dependence either on average or by treatment as a function depending on the time. Linear regressions or more generally polynomial regressions are generic such functions that can be used for such models. And if nothing else, they could be used for a decomposition of the potential time dependence structures into linear, curvature etc. components.

Different analysis strategies could be used for this. What we suggest in the following is strongly influenced by what is easy for us to do using the two main linear mixed model functions in R: `lme` and `lmer`:

1. Do the factor based analysis as shown above.
2. Do some explorative plotting of individual and treatment average regression lines/curves.
3. Potentially make a "high degree" decomposition based on the simple "split-plot" repeated measures model using `lmer` and `lmerTest`.
4. Check if a linear or quadratic regression model could be used as an alternative to the factor based model: Fit the model by `lme` and compare by maximum likelihood. (Use the proper and chosen correlation structure)
5. IF a regression approach seems to capture what is going on, then try to fit the random coefficient model as an alternativ to the correlation structure used from above - chose the best one at the end.
6. A possibility is that a factor based model is needed for the main effect of time, whereas a quantitative model would fit the interaction effect. This model is not so easily fitted by `lme` due to some limitations of `lme` in handling over-specified fixed effects structures. Both `lm` and `lmer` handle those situations fine, so this combination is only (easily for us) available combined with either a random coefficient variance structure and/or the simple split-plot structure.

### 12.8.1 Example: Regression models for the rats data

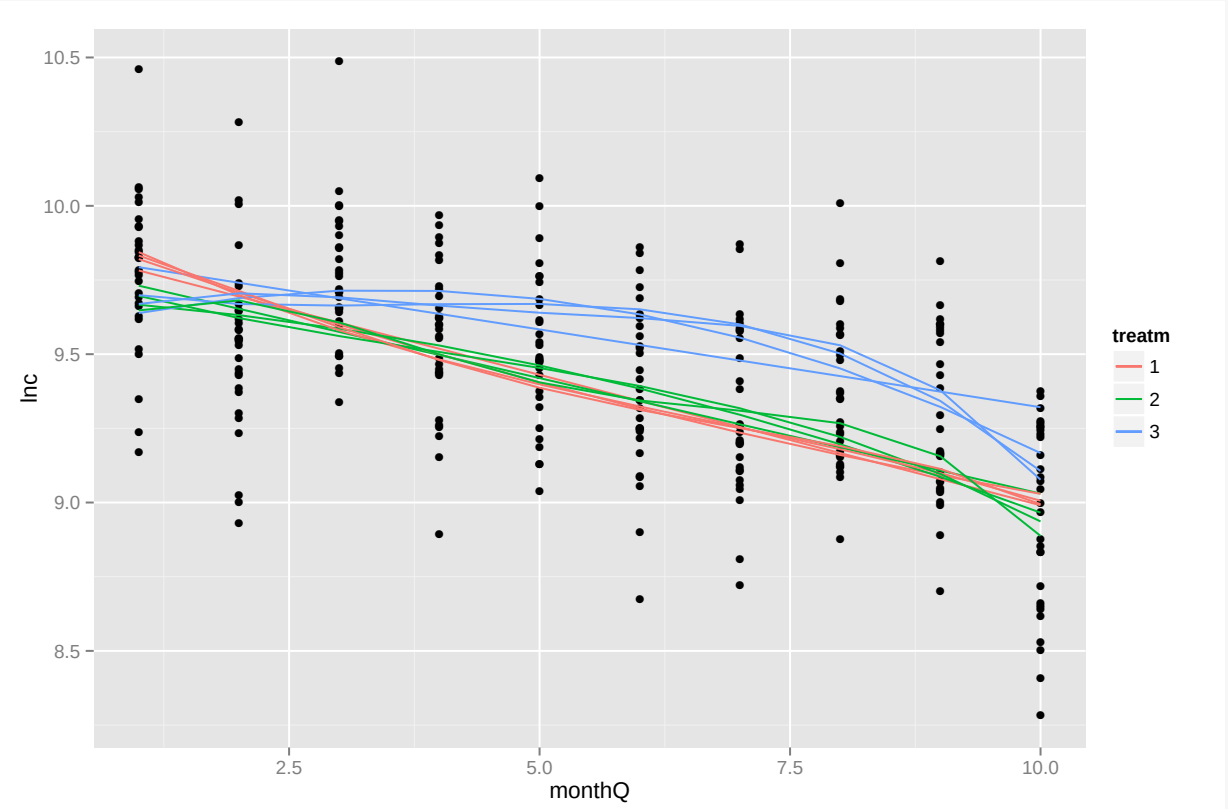
First let's do some plotting of individual and treatment average curves. We make the quantitative power versions of the time variable: (we re-scale them for numeric stability in the mixed model fitting)

```
rats$monthQ2 <- scale(rats$monthQ^2)
rats$monthQ3 <- scale(rats$monthQ^3)
rats$monthQ4 <- scale(rats$monthQ^4)
```

we do the plotting by fitting various linear models by `lm` and then plotting the fitted values from these, first average patterns, where we include the first four polynomials within each treatment group on a single plot:

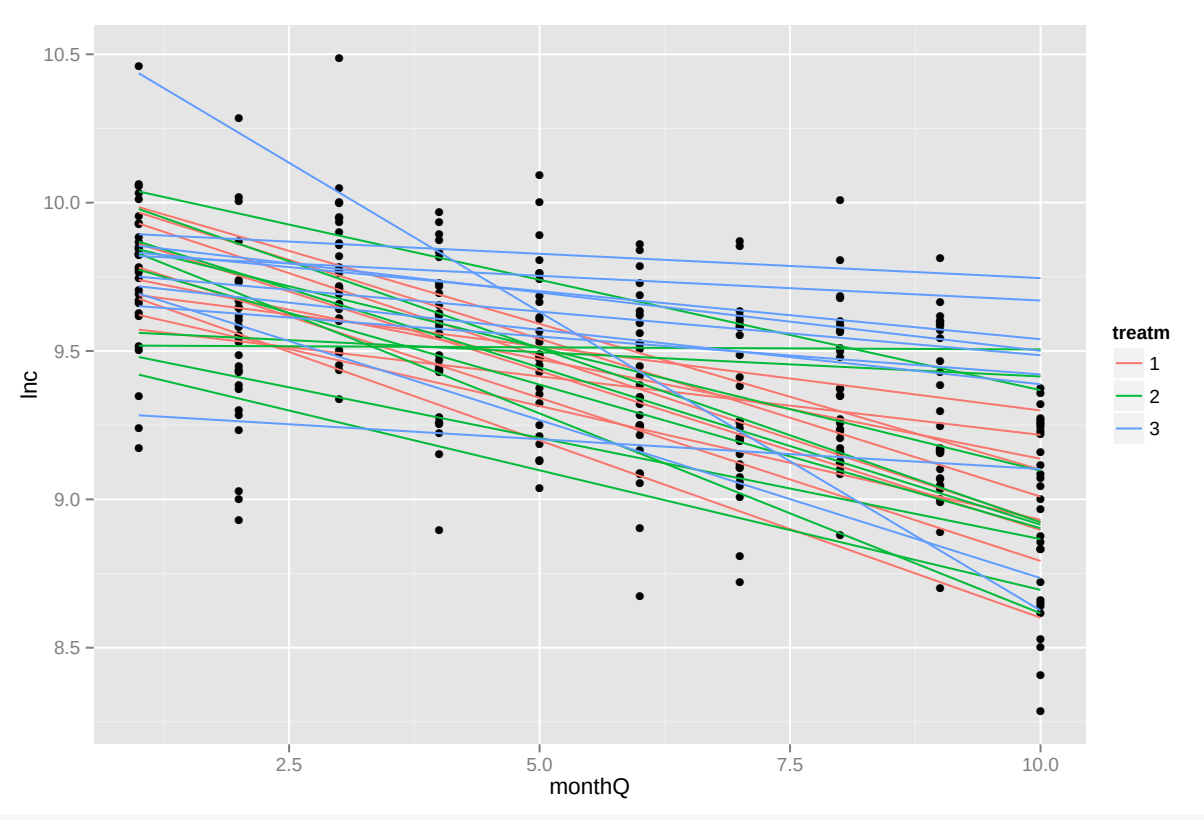
```
library(ggplot2)
p <- qplot(monthQ, lnc, data = rats)

lmQ <- lm(lnc ~ monthQ*treatm,data = rats)
lmQ2 <- lm(lnc ~ monthQ*treatm + monthQ2*treatm, data = rats)
lmQ3 <- lm(lnc ~ monthQ*treatm + monthQ2*treatm + monthQ3*treatm, data = rats)
lmQ4 <- lm(lnc ~ monthQ*treatm + monthQ2*treatm + monthQ3*treatm
           + monthQ4*treatm, data = rats)
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ), group=treatm, colour=treatm))
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ2), group=treatm, colour=treatm))
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ3), group=treatm, colour=treatm))
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ4), group=treatm, colour=treatm))
print(p)
```

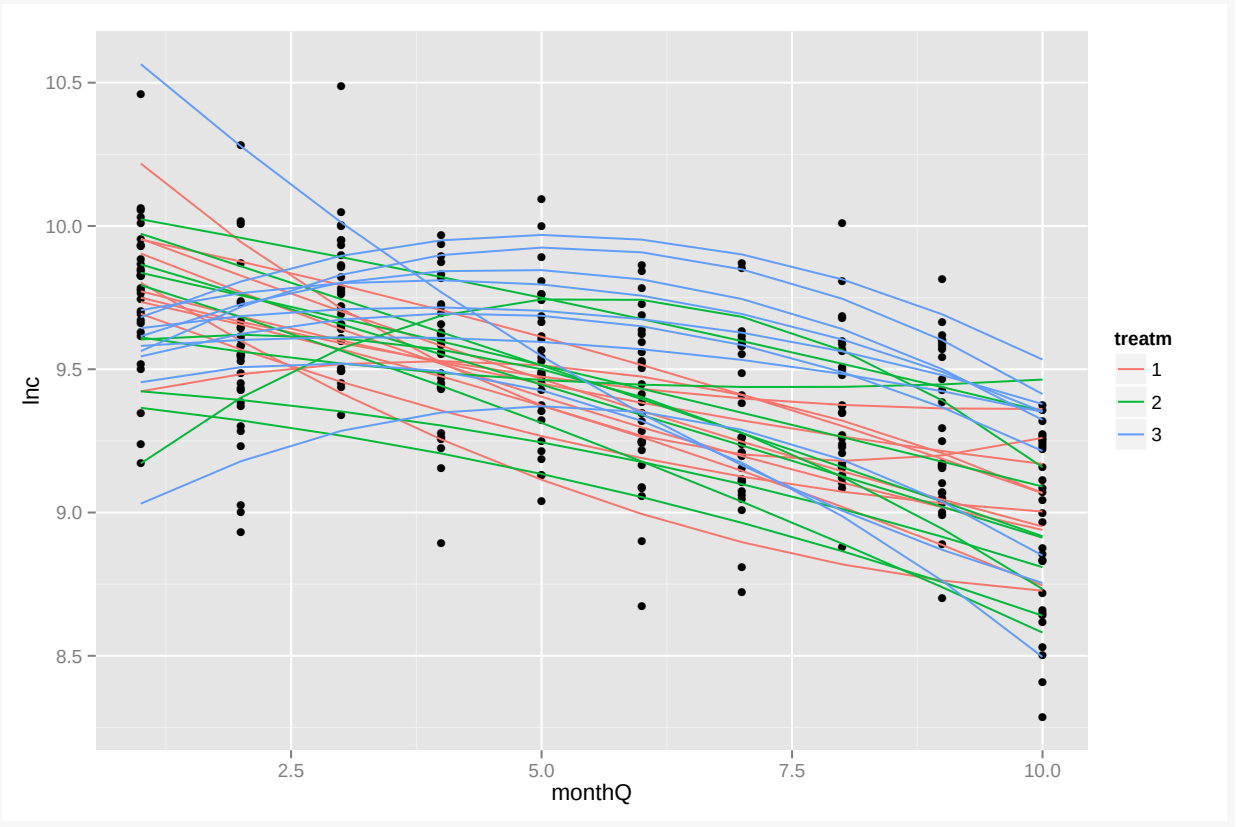



Next individual patterns where we make a plot for each degree of the polynomial:

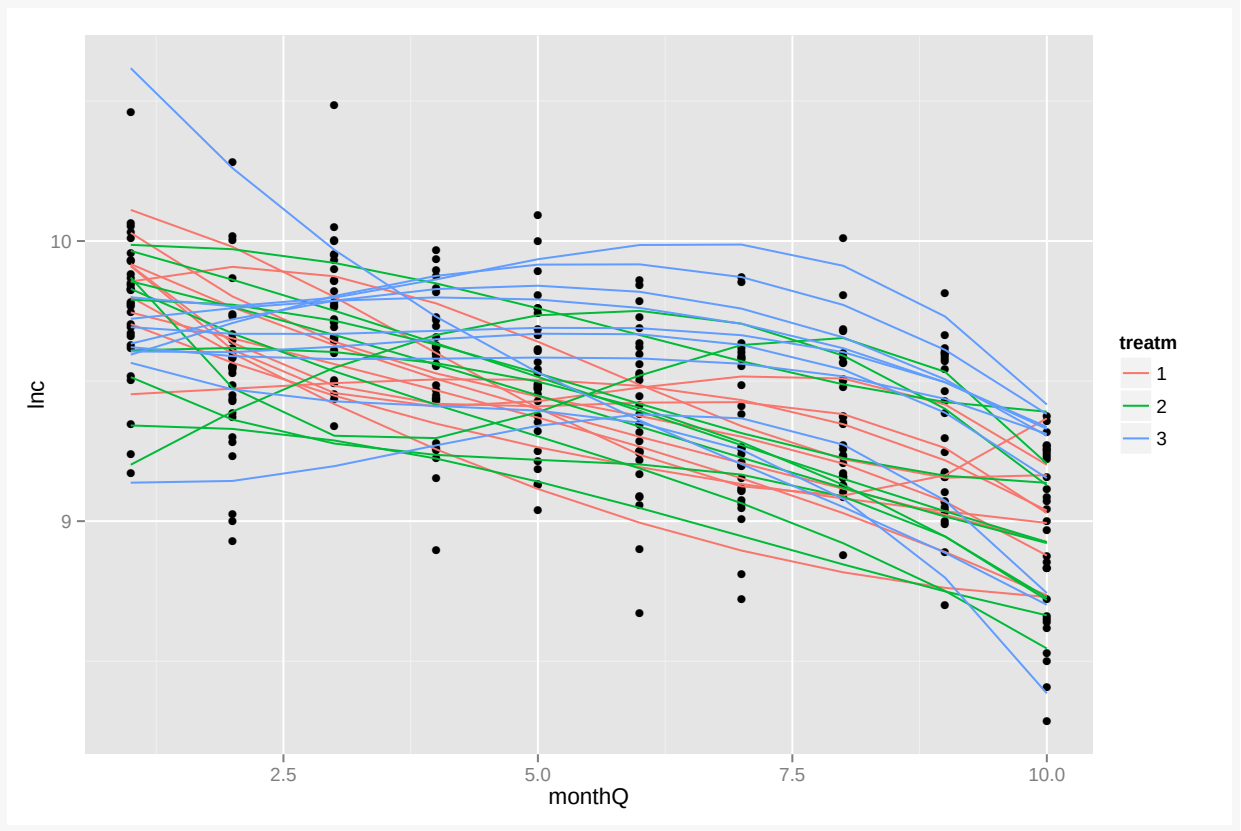
```
p <- qplot(monthQ, lnc, data = rats)
lmQ <- lm(lnc ~ monthQ*cage,data = rats)
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ), group=cage, colour=treatm))
print(p)
```



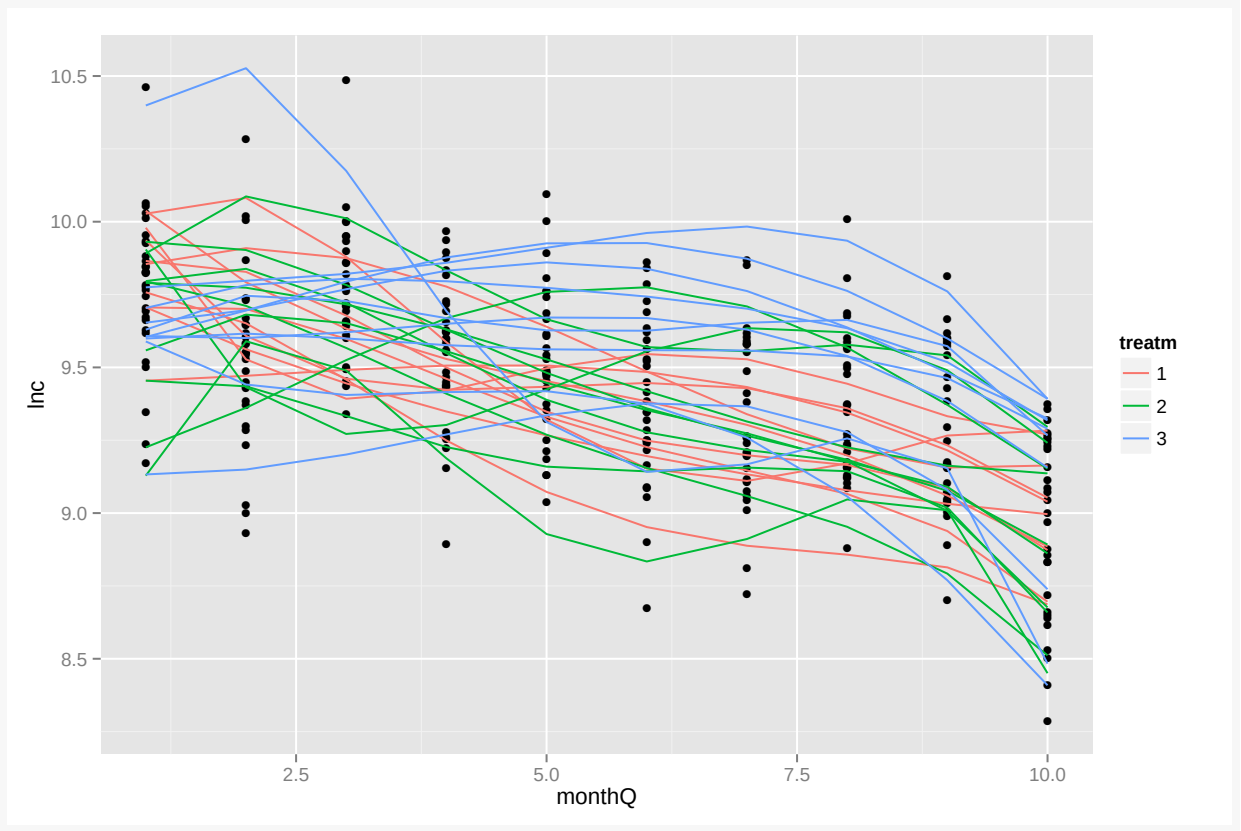

```
p <- qplot(monthQ, lnc, data = rats)
lmQ2 <- lm(lnc ~ monthQ*cage + monthQ2*cage,data = rats)
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ2), group=cage, colour=treatm))
print(p)
```




```
p <- qplot(monthQ, lnc, data = rats)
lmQ3 <- lm(lnc ~ monthQ*cage + monthQ2*cage+ monthQ3*cage,data = rats)
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ3), group=cage, colour=treatm))
print(p)
```




```
p <- qplot(monthQ, lnc, data = rats)
lmQ4 <- lm(lnc ~ monthQ*cage + monthQ2*cage+ monthQ3*cage+ monthQ4*cage,data = rats)
p<- p + geom_line(aes(x=monthQ, y=fitted(lmQ4), group=cage, colour=treatm))
print(p)
```




Let us try to do a "high degree" (4th order) decomposition: (this could easily be extended even higher)

```
lmerQ4 <- lmer(lnc ~ monthQ + monthQ2 + monthQ3 + monthQ4 + month +treatm
               + monthQ:treatm + monthQ2:treatm + monthQ3:treatm +
                 monthQ4:treatm + month:treatm +(1|cage),data = rats)
```

```
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
```

```
xtable(anova(lmerQ4, type=1))
```

```
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
```

```
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
fixed-effect model matrix is rank deficient so dropping 12 columns / coefficients
```

| | Sum Sq | Mean Sq | NumDF | DenDF | F.value | Pr(>F) |
|---|---|---|---|---|---|---|
| monthQ | 13.09 | 13.09 | 1.00 | 260.42 | 345.43 | 0.0000 |
| monthQ2 | 0.40 | 0.40 | 1.00 | 260.42 | 10.48 | 0.0014 |
| monthQ3 | 0.21 | 0.21 | 1.00 | 260.42 | 5.49 | 0.0199 |
| monthQ4 | 0.24 | 0.24 | 1.00 | 260.42 | 6.24 | 0.0131 |
| month | 1.80 | 0.36 | 5.00 | 260.43 | 9.48 | 0.0000 |
| treatm | 0.24 | 0.12 | 2.00 | 67.83 | 3.22 | 0.0461 |
| monthQ:treatm | 0.55 | 0.28 | 2.00 | 260.40 | 7.28 | 0.0008 |
| monthQ2:treatm | 0.68 | 0.34 | 2.00 | 260.42 | 9.01 | 0.0002 |
| monthQ3:treatm | 0.04 | 0.02 | 2.00 | 260.42 | 0.50 | 0.6061 |
| monthQ4:treatm | 0.06 | 0.03 | 2.00 | 260.42 | 0.81 | 0.4459 |
| month:treatm | 0.11 | 0.01 | 10.00 | 260.43 | 0.30 | 0.9818 |

Note: we are NOT using the correct correlation structure here. The concern would typically be that the tests from this analysis then would be "too siginificant". We see that the interaction effect seems to be nicely described by the first two components (linear and quadratic) (all other effects are non significant), whereas the main (average) time effect is not even closely described by a 4th order polynomium: The `month` effect as a factor is still clearly significant here.

We could still try (for the sake of the example) to test, using the correct, from above chosen, correlation structure, whether a polynomial approach "describes everything": (and let us try to go as high as a 6th degree poloynomial structure)

```
rats$monthQ5 <- scale(rats$monthQ^5)
rats$monthQ6 <- scale(rats$monthQ^6)

model4 <- lme(lnc ~ monthQ + monthQ2 + monthQ3 + monthQ4+ monthQ5 + monthQ6
              +treatm + monthQ:treatm + monthQ2:treatm + monthQ3:treatm +
                monthQ4:treatm + monthQ5:treatm + monthQ6:treatm,
              random = ~1 | cage, correlation =
                corExp(form = ~as.numeric(month) | cage, nugget = F), data = rats)
```

```
logLik(model4, REML=FALSE)

'log Lik.' 88.85816 (df=24)

logLik(model3b, REML=FALSE)

'log Lik.' 115.6872 (df=33)

1-pchisq(2*(logLik(model3b, REML=FALSE)-logLik(model4, REML=FALSE)), 9)

'log Lik.' 2.19253e-08 (df=33)
```

Note that the factor terms have been omitted here - both in the main effect as in the interaction part. We see that even the 6th degree polynomiaum does not fit the data in this case.

Finally, let us try the random coefficient structure on the model with factor structure on the main part and a 2nd order quantitative model for the interaction. We use maximum likelihood (not REML) to be able to compare the AIC values with the full fixed effect model using the proper correlation structure:

```
lmer3_ml <- lmer(lnc ~ month +treatm + monthQ:treatm + monthQ2:treatm
                +(1 + monthQ + monthQ2|cage), data = rats, REML = FALSE)

fixed-effect model matrix is rank deficient so dropping 2 columns / coefficients

model3b_ml <- lme(lnc ~ month + treatm + month:treatm, random = ~1 |
cage, correlation = corExp(form = ~as.numeric(month) |
cage, nugget = F), data = rats, method = "ML")

logLik(lmer3_ml)

'log Lik.' 101.4523 (df=23)

logLik(model3b_ml)
```

```
'log Lik.' 115.6872 (df=33)

AIC(lmer3_ml)

[1] -156.9046

AIC(model3b_ml)

[1] -165.3743
```

As expected, the analysis favors the model with full factor structure for this particular data.

Post hoc and summary of treatment-time interactions could potentially be nicer described by different slopes and/or curvatures than a by-time treatment story as given in the previous section.

**Remark 12.1**

A couple of model possibilitties not explicitly mentioned so far:

- It would be possible to combine random coefficient structures with residual correlation structures. This requires the handling of random coefficient models in `lme`, see e.g. the "old" course material: http://www.imm.dtu.dk/~perbb/st113/Module09/R.html
- It can be considered whether some simple transformation of either the observations OR the time scale could linearize the time profiles to make the story simple on a transformed scale - e.g. log(Y) as function of time and/or log(time).

## 12.9 Exercises

## |||| Exercise 1 PH in pigs

To investigate the effect of injection of Porcine Growth Hormone (PGH) on pH (among other things) a block experiment was carried out with two pigs from each of 6 litters (= blocks). There were two treatments:

1) control
2) pgh (daily injection with 0.08 mg Porcine Growth Hormone)

Apart from several other measurements the pH in the meat was measured 20 times from 30 minutes after until 24 hours after slaughter. There were 10 litters in the experiment but pH was measured for only 6 of these. The order of the data is: treatment, litter, pig number, followed by pH measurements at 30, 45, 60, 75, 90, 105, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 1440 minutes after slaughter.

The data file can be downloaded as: pgh.txt and is described also in eNote13 and has e following structure:

```
treatm   litter   pigno   min    ph
     1        2      21    30  6.45
     2        2      22    30  6.07
     1        4      41    30  6.77
     2        4      42    30  6.8
     .        .       .     .   .
     .        .       .     .   .(240 lines total)
     .        .       .     .   .
     2       10     102  1440  5.46
```

In this analysis the focus should be on the effect of the treatment over time.

a) Make one or more plots of the data. Comment on the plot(s).

b) Setup a suitable model for this data set, including both fixed and random effects, but no correlation structure. (Notice that besides the “pig” variable we also have information about litter, which could be included as an additional random effect.)

c) Reduce the initial model (if possible), both the random effects and fixed effects parts.

d) Extend the model by adding a correlation structure.

e) Use information criteria and/or semi-variograms to select an appropriate correlation structure.

f) Explain in words the correlation structure that was chosen.

g) Repeat the model reduction process.

h) What is the conclusion about the treatment?