Modeling the covariance structure

- The individuals represent a random sample from the population of interest.
- Observations from different individuals are independent, while repeated measurements of the same individual are not assumed to be independent.
- The elements of the vector of repeated measures $\mathbf{y}_i$, have a Multivariate Normal (MVN) distribution, with means $\mu_{i,j} = E(Y_{i,j} \mid Xi, j) = \beta_1 X_{i,j,1} + \beta_2 X_{i,j,2} + \ldots + \beta_p X_{i,j,p}$ and covariance matrix $\mathbf{\Sigma}_i$.

We say $\mathbf{\Sigma}_i$ as different subjects may have a different number of available observations or some individual-specific predictors may influence the variance. However, the observation times are fixed for the initial part of this note.

First, take a simpler case. Let $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$. The maximum likelihood-based methods for linear models usually assume the error $\boldsymbol{\epsilon}_i \sim \text{MVN}(0, \mathbf{\Sigma})$.

The basic implementations of `lme4` or `nlme` assume unstructured covariance. It has both pros and cons.

Pros:

- Appropriate when design is "balanced" and number of measurement occasions are relatively small.
- no assumptions made about the patterns of variances and covariances.

Cons:

- Number of covariance parameters grows rapidly with the number of measurement occasions. When there are n repeated measurements, there are $\frac{n(n+1)}{2}$ free parameters to be estimated.
- Unstructured covariance is problematic when there are mistimed measurements.

The choice of correlation would impact the estimate of the main effect $\boldsymbol{\beta}$.

- There is should be a balance in the imposed structure.
- With too little structure like in the unstructured case, there may be too many parameters to be estimated with a limited amount of data.
- With too much structure, potential risk of model misspecification and misleading inferences, concerning $\boldsymbol{\beta}$.

If $\boldsymbol{\epsilon}_i \sim \text{MVN}(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{n} \frac{1}{d_i} \mathbf{u}_i \mathbf{u}_i^T$, we can write $\boldsymbol{\epsilon}_i = \mathbf{U}\boldsymbol{\eta}_i$, such that $\boldsymbol{\eta}_i \sim \text{Normal}(0, \mathbf{D})$. (I use Normal(,) instead of MVN(,) even for multivariate data when the covariance is diagonal.)

[**Digression:** In principal component analysis, we visualize the data $\mathbf{y}_i$ assuming it follows $\text{MVN}(0, \boldsymbol{\Sigma})$ on the plan spanned by $\{\mathbf{u}_1, \mathbf{u}_2\}$ for 2-D visualization and $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ for 3-D visualization. On the plane spanned by $\{\mathbf{u}_1, \mathbf{u}_2\}$, the approximate representation of $\mathbf{y}_i$ is $(\mathbf{y}_i^T \mathbf{u}_1)\mathbf{u}_1 + (\mathbf{y}_i^T \mathbf{u}_2)\mathbf{u}_2$. $\mathbf{y}_i^T \mathbf{u}_1$ is the projection of $\mathbf{y}_i$ on $\mathbf{u}_1$. In 2-D PCA visualization we plot $(\mathbf{y}_i^T \mathbf{u}_1, \mathbf{y}_i^T \mathbf{u}_2)$]

Eigen-decomposition of $\boldsymbol{\Sigma}$ leads to $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{n} \frac{1}{d_i} \mathbf{u}_i \mathbf{u}_i^T$ with $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$. We often find a minimum K such that $\frac{\sum_{i=1}^{K} d_i}{\sum_{i=1}^{n} d_i} > C$, which often set as $C = 0.9$. If this happens, then $d_{K+1}, \ldots, d_n$ will be almost equal as they are very small relatively. Hence, an approximation for $\boldsymbol{\Sigma}$ is $\sum_{i=1}^{K} \frac{1}{d_i} \mathbf{u}_i \mathbf{u}_i^T + \sigma^2 \mathbf{I}$.

1 Latent factor model

The above approximation of $\boldsymbol{\Sigma}$ lets us $\boldsymbol{\epsilon}_i = \mathbf{U}_K \boldsymbol{\eta}_i + \boldsymbol{\delta}_i$, such that $\mathbf{U}_K$ is submatrix with fist K columns in $\mathbf{U}$ and $\boldsymbol{\eta}_i \sim \text{Normal}(0, \mathbf{D}_K)$ and $\boldsymbol{\delta}_i \sim \text{Normal}(0, \sigma^2 \mathbf{I})$. However, it is often the case that $\boldsymbol{\Sigma}$ is not known. Hence, we simplify the above model and write,

$$\boldsymbol{\epsilon}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\delta}_i, \quad \boldsymbol{\eta}_i \sim \text{Normal}(0, \mathbf{E}), \boldsymbol{\delta}_i \sim \text{Normal}(0, \sigma^2 \mathbf{I}),$$

where the columns of $\boldsymbol{\Lambda}$ are not required to be orthogonal. And our final model for the data becomes, $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\delta}_i$. Here $\boldsymbol{\eta}_i$ and $\boldsymbol{\delta}_i$ are both random. In the context of latent factor modeling, $\boldsymbol{\eta}_i$'s are called latent factors and $\boldsymbol{\Lambda}$ is the loading matrix of dimension $n \times K$ where $K < n$. The nature of the estimated loading matrix $\boldsymbol{\Lambda}$ is often of interest.

2 Random effect model

The latent factor model can be generalized further to allow $\boldsymbol{\Lambda}$ to vary with i based on some covariates that are expected to influence the variability of $\mathbf{y}_i$. Specifically, say blood biomarkers influence the variability in the lead-level, then an appropriate model will be

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\eta}_i + \boldsymbol{\delta}_i,$$

where $\mathbf{Z}_i$ will contain the blood biomarker numbers from different visits. Since, $\boldsymbol{\eta}_i \sim \text{Normal}(0, \mathbf{E}), \boldsymbol{\delta}_i \sim \text{Normal}(0, \sigma^2 \mathbf{I})$, we have $\text{V}(\mathbf{y}_i | \mathbf{X}_i) = \boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{E} \mathbf{Z}_i^T + \sigma^2 \mathbf{I}$.

These $\boldsymbol{\eta}_i$'s are called random 'effect's as they quantify the effects of some covariates on $\mathbf{y}_i$. For inference, $\boldsymbol{\eta}_i$'s are not of direct interest. But the incorporation of random effect may result in better estimation of the variability of $\hat{\beta}$ i.e. $\text{V}(\hat{\beta})$ for hypothesis testing and inference.

In all generality, we may model the outcome as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\eta}_i + \boldsymbol{\Lambda}\boldsymbol{\gamma}_i + \boldsymbol{\delta}_i, \quad \boldsymbol{\eta}_i \sim \text{Normal}(0, \mathbf{E}), \boldsymbol{\gamma}_i \sim \text{Normal}(0, \mathbf{F}), \boldsymbol{\delta}_i \sim \text{Normal}(0, \sigma^2 \mathbf{I})$$

but standard software will not allow you to have such additional latent factors on top of the random effects. But you can consider two-step estimation methods or REML-type methods to drop the first two terms $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\eta}_i$ from the model with the help of orthogonal projections based on $[\mathbf{X}, \mathbf{Z}]$. The best way will probably be a Bayesian approach in this case.

Major advantage: One of the major advantages of the random effect model is that one can include any covariate (even observation times) into $\mathbf{Z}_i$. This allows the model to have covariance structures dependent on time and thus can allow unbalanced longitudinal data as well. We will discuss more on this.

2.1 Linear mixed model

Based on the choices of $\mathbf{X}_i$ and $\mathbf{Z}_i$, the interpretation of linear mixed model changes. We may have $\mathbf{X}_i = \mathbf{Z}_i$, i.e. we may have the same set of predictors with both fixed and random effects.

Then the model reduces to:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\delta}_i,$$

where $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i$, where $\boldsymbol{\beta}$ is a 'population average estimate' which often makes sense if $\sum_i \boldsymbol{\eta}_i = 0$. Softwares often do a posthoc adjustment to ensure $\sum_i \boldsymbol{\eta}_i = 0$.

Example:

$$y_{i,j} = \beta_0 + t_{i,j}\beta_1 + \eta_{i,0} + t_{i,j}\eta_{i,1} + \delta_{i,j},$$

where $\beta_0 + \eta_{i,0}$ is an individual-specific intercept and $\beta_1 + \eta_{i,1}$ is an individual-specific slope with respect to time. If 'time' is not expected to influence the variance, we may also let $y_{i,j} = \beta_0 + t_{i,j}\boldsymbol{\beta}_1 + \eta_{i,0} + \delta_{i,j}$.

R implementation: We can fit above model using `lme4` or `nlme` packages. We need to just specify the 'fixed-effect' and 'random-effect' predictors.

```
library(lme4)
out <- lFormula(measurement ~ Week + (1+Week|V1), data=data_new)
X <- as.matrix(out$X)
Z <- t(as.matrix(out$reTrms$Zt))
```

```
fit <- lmer(measurement ~ Week + (1+Week|V1), data=data_new)
```

The big $\mathbf{X}$ is created by stacking $\mathbf{X}_i$ on top of each other as $\mathbf{X} = \texttt{rbind}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. But the big $\mathbf{Z}$ as $\mathbf{Z} = \texttt{bdiag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ in a block diagonal way so that by stacking $\mathbf{b} = \texttt{c}(\mathbf{b}_1, \ldots, \mathbf{b}_n)$. We can write the whole model in matrix-vector notation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\delta}$. This $\mathbf{X}$ and $\mathbf{Z}$, we have extracted in the above block of code.

Also, some covariates can only be included in $\mathbf{Z}$, but not in $\mathbf{X}$. However, in general, the variables that can control the variability of the data also control the mean. Hence, we usually have a subset of variables from $\mathbf{X}_i$ into $\mathbf{Z}_i$. In general, it is not a good idea to include factor variables with a lot of levels into $\mathbf{Z}_i$ as that makes the problem unstable and does not make sense.

```
library(lme4)
out <- lFormula(measurement ~ (1+Week|V1), data=data_new)
X <- as.matrix(out$X)
Z <- t(as.matrix(out$reTrms$Zt))


fit <- lmer(measurement ~ (1+Week|V1), data=data_new)
```

Using `nlme::lme`,

```
library(nlme)
res <- lme(measurement ~ Week, random = ~1+Week|V1, data = data_new)
```

Practical modeling for $\boldsymbol{\delta}_i$: We may again put another multivariate assumption of $\boldsymbol{\delta}_i$. But the main motivation to decompose $\boldsymbol{\epsilon}_i$ and take the $\boldsymbol{\delta}_i$ part out, assuming that this is a 'measurement-error'. This philosophical construction helps to justify a simple model like $\boldsymbol{\epsilon}_i \text{Normal}(0, \sigma^2\mathbf{I})$.

3 Time-series model

This will be particularly relevant for 'longitudinal' data analysis, as it includes the time component. The above approaches apply to any multivariate response data. In longitudinal case $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,n})$ are sequential errors and thus expected to exhibit among associations. 1) This also puts a restriction that the data has to be observed on a gridded observation times. Meaning the observation times should look like baseline, 3 months, 6 months, 9 months, etc. It CANNOT look like baseline, 1 month, 4 months, 5 months, 9 months, etc. In the latter case, one should consider the methods mentioned above or the one in the next section. However, it may often be reasonable to assume the following models even for non-gridded data after carefully checking the covariance structure. 2) These models will also be reasonable to apply when the observation times are all the same for all the subjects, i.e. the balanced data (there may be missing observations, but data collection or observation times need to be coming from the same set of possible times.).

3.1 Autoregressive (AR)

This model assumes that the present observation depends on the previous observation. Hence, the current value is regressed on the past values of itself (thus it's called autoregressive process. 'Auto' means on itself, and 'regressive' points to a regression model.). Specifically, if it is assumed that it depends on past p-values, the model (denoted as AR(p)) becomes,

$$\epsilon_{i,t} = \rho_1 \epsilon_{i,t-1} + \cdots \rho_p \epsilon_{i,t-p} + \delta_{i,t},$$

where $\delta_{i,t} \sim \text{Normal}(0, \sigma^2)$ for all t and independent. $\delta_{i,t}$'s are also called white noise.

In longitudinal data, there is always a limited number of time points. Thus, an AR(1) model is sufficient and the corresponding covariance matrix of $\boldsymbol{\epsilon}_i$ looks like,

$$\mathbf{\Sigma} = \frac{\sigma^2}{1-\rho_1^2}\begin{bmatrix} 1 & \rho_1 & \dots & \rho_1^{n-1} \\ \rho_1 & 1 & \dots & \rho_1^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1^{n-1} & \rho_1^{n-2} & \dots & 1 \end{bmatrix} \tag{1}$$

And precision matrix is banded and looks like,

$$\mathbf{\Omega} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -\rho_1 & 0 & \dots & 0 & 0 \\ -\rho_1 & 1+\rho_1^2 & -\rho_1 & \dots & 0 & 0 \\ 0 & -\rho_1 & 1+\rho_1^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 1+\rho_1^2 & -\rho_1 \\ 0 & 0 & 0 & \dots & -\rho_1 & 1 \end{bmatrix} \tag{2}$$

Determination of p: For AR(p), the bandwidth of the precision matrix is p. Precision and partial autocorrelation has a direct relation. Specifically, $\frac{\Omega_{\ell,k}}{\sqrt{\Omega_{\ell,\ell}}\sqrt{\Omega_{k,k}}}$ is the partial correlation between $\epsilon_{i,\ell}$ and $\epsilon_{i,k}$. Hence, to determine the order of an AR process, checking partial autocorrelation is an immediate way. Specifically, if $\epsilon_{i,t}$ follows AR(p), we must have partialcorr$(\epsilon_{i,t}, \epsilon_{i,t+h}) = 0$ for all $h > p$ and non-zero otherwise. R function `pacf` can be used.

3.2 Moving-average (MA)

Moving average assumes that there is a latent white noise process that is independent at each time point, and the observed process is a linear combination of some of those recent values. Specifically, if it is assumed that it depends on past p such white noise-values, the model (denoted as MA(q)) becomes,

$$\epsilon_{i,t} = \theta_1\delta_{i,t-1} + \cdots \theta_q\delta_{i,t-q} + \delta_{i,t},$$

where $\delta_{i,t} \sim \text{Normal}(0, \sigma^2)$.

Again due to the fewer number of time points in longitudinal data analysis, MA(1) model would be sufficient, and its covariance matrix is banded and looks like as,

$$\mathbf{\Sigma} = \sigma^2(1+\theta_1^2)\begin{bmatrix} 1 & \frac{\theta_1}{1+\theta_1^2} & 0 & \dots & 0 & 0 \\ \frac{\theta_1}{1+\theta_1^2} & 1 & \frac{\theta_1}{1+\theta_1^2} & \dots & 0 & 0 \\ 0 & \frac{\theta_1}{1+\theta_1^2} & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \dots & 1 & \frac{\theta_1}{1+\theta_1^2} \\ 0 & 0 & 0 & \dots & \frac{\theta_1}{1+\theta_1^2} & 1 \end{bmatrix} \tag{3}$$

Determination of q: For MA(q), the bandwidth of the correlation matrix is q. Hence, to determine the order of an MA process, checking autocorrelation is an immediate way. Specifically, if $\epsilon_{i,t}$ follows MA(q), we must have $\text{corr}(\epsilon_{i,t}, \epsilon_{i,t+h}) = 0$ for all $h > q$ and non-zero otherwise. R

function `acf` can be used.

3.3 Autoregressive Moving-average (ARMA)

If we combine the above two models, we get the ARMA model

$$\epsilon_{i,t} = \rho_1 \epsilon_{i,t-1} + \cdots \rho_p \epsilon_{i,t-p} + \theta_1 \delta_{i,t-1} + \cdots \theta_q \delta_{i,t-q} + \delta_{i,t}.$$

Checking orders of an ARMA process is hard. But, for an ARMA model, $p = 1, q = 1$ i.e. ARMA(1,1) is good enough in almost all applications due to a combination of theoretical and practical properties.

3.4 Symmetric Toeplitz

It is a generalization of some of the above covariance structures,

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \cdots & \cdots & \cdots \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \cdots & \cdots & \cdots \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \cdots & \cdots & \cdots \\ \rho_3 & \rho_2 & \rho_1 & 1 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Theoretically, it can be shown that there is a one-to-one relation between the Toeplitz covariance and some AR(p) process for some p.

When $\rho_1 = \rho_2 = \cdots = \rho_n$, the resulting covariance is called *compound symmetry*.

4 Generalization of Toeplitz for non-gridded data

We can define $\text{cov}(\epsilon_{i,t}, \epsilon_{i,s}) = \sigma^2 f(|t-s|)$, where σ is the marginal variance and f is a decreasing function.

4.1 Exponential kernel

In the case of the exponential kernel, we set $f(x) = \exp(-x/\rho)$, where ρ is the range parameter that controls the degree of dependence. As ρ increases, the correlation increases for a given pair (s, t)

4.2 Gaussian kernel

In the case of the Gaussian kernel, we set $f(x) = \exp(-x^2/\rho)$, where ρ is the range parameter that controls the degree of dependence. As ρ increases, the correlation increases for a given pair (s, t).

`nlme` has options for both of the above two kernels.

There are other kernels like Matérn kernel which is more popular for spatial data analysis. But there are more parameters involved, hence denser data is needed for considering a Matérn kernel.

5 Choosing the appropriate covariance structure

For the consideration of the above-mentioned models for errors in longitudinal data analysis, one may first get the least square estimate $\hat{\boldsymbol{\beta}}_{LS}$ of $\boldsymbol{\beta}$ and get the preliminary errors $\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{LS}$. Examine these errors carefully based on different diagnostic checks visually or some of the methods mentioned above, and then refit the model with the additional variability assumptions. Alternatively, you may check $\mathbf{y}_i$ directly as well. Here $\mathbf{X}_i$ should be formed adding as many relevant predictors as possible as since $\boldsymbol{\epsilon}_i$ is the part of $\mathbf{y}_i$, not explained by $\mathbf{X}_i\boldsymbol{\beta}$.

Here, we are always modeling the general nature of the variability of $\boldsymbol{\epsilon}_i$. From practical standpoint, we should choose a covariance model that encompasses all possible covariances in the data. In the case of nested covariance structure: You may consider the LRT test based on REML likelihoods. Like compound symetry is nested within the Toeplitz model, since if the former holds the latter must necessarily hold, with $\rho_1 = \rho_2 = \cdots = \rho_n$. All the above covariance structures are nested with unstructured covariance (the default covariance for `lmer`).

In the case of non-nested covariance structure: You may also compare the model fit for different structures using AIC or BIC.

6 Computational issues

To realize some of the errors in standard software, it is important to understand the problem both theoretically and from an implementation point of view.

From a theoretical perspective, we must have 'number of observations > number of random effects'. Otherwise, it is impossible to estimate the parameters. What it means is that $\mathbf{Z}_i$ should have fewer columns than rows in a balanced design.

Then unless we impose some variable selection approach, the fixed effects also should not be too large that it does not leave any degree of freedom for the estimation of variability. Remember, the unbiased estimate of error variance is $\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2$. Hence, we must have $n > p$ to leave any degree of freedom behind for estimating variability. For mixed model, it is a bit complicated.

7 Repeated measure ANOVA

We can take observations at each time point as a group and repeated measure ANOVA that tests the hypotheses that $H_0 : \mu_1 = \ldots = \mu_K$, where μ_k's are population means under different groups (or times in case of longitudinal design.)

The Repeated measure ANOVA.R code has all the steps. It also needs the normality assumption, and in addition, it assumes something called Sphericity which means equality of variance of the differences between each pair of group means as in $\mathrm{V}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$=$\mathrm{V}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_3)$=$\mathrm{V}(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)$ for any $i \neq j$. There are tests to verify this assumption. If violated, there is a way to adjust the degree of freedom of the sum of squares and run the analysis.

There are some issues in repeated measure ANOVA:

- Sphericity assumption and doing the required adjustments. But they are often inadequate.
- Some problems also require understanding the time effect on the response.
- There may be additional covariates to be adjusted for both in mean and variability.