

# Maximum likelihood estimation for linear mixed models

Rasmus Waagepetersen  
Department of Mathematics  
Aalborg University  
Denmark

February 8, 2023

# Outline for today

- ▶ linear mixed models
- ▶ the likelihood function
- ▶ maximum likelihood estimation
- ▶ restricted maximum likelihood estimation

# Linear mixed models

Consider mixed model:

$$Y_{ij} = \beta_1 + U_i + \beta_2 x_{ij} + \epsilon_{ij}$$

May be written in matrix vector form as

$$Y = X\beta + ZU + \epsilon$$

where  $\beta = (\beta_1, \beta_2)^T$ ,  $U = (U_1, \dots, U_k)^T$  and  $\epsilon = (\epsilon_{11}, \epsilon_{12}, \dots, \epsilon_{km})^T$ ,  $X$  is  $n \times 2$  and  $Z$  is  $n \times k$ .

# Linear mixed model: general form

Consider model

$$Y = X\beta + ZU + \epsilon$$

where  $U \sim N(0, \Psi)$  and  $\epsilon \sim N(0, \Sigma)$  are independent.

All previous models special cases of this.

Then  $Y$  has multivariate normal distribution

$$Y \sim N(X\beta, Z\Psi Z^T + \Sigma)$$

# Hierarchical version

1.  $U \sim N(0, \Psi)$
2.  $Y|U = u \sim N(X\beta + Zu, \Sigma)$

Useful for generalization to generalized linear mixed models.

**Ex:** Poisson log-normal:

Given  $U = u$ ,  $Y_i$  independent with  $Y_i \sim \text{Poisson}(\lambda_i)$  where  $\lambda_i = \exp(\eta_i)$  and  $\eta = X\beta + Zu$ .

Note likelihood (marginal density of  $Y$ ) typically not of simple form in case of generalized linear mixed models.

## Some useful matrix identities

Woodbury identity:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

$$(C^{-1} + DA^{-1}B)^{-1}DA^{-1} = CD(BCD + A)^{-1}$$

$$(C^{-1} + B^tA^{-1}B)^{-1}B^tA^{-1} = CB^t(BCB^t + A)^{-1}$$

# Inverse of covariance matrix

Assume  $\Sigma$  positive definite (e.g. scaled identity matrix).

Then  $Z\Psi Z^T + \Sigma$  guaranteed to be positive definite and

$$(Z\Psi Z^T + \Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1}Z(\Psi^{-1} + Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1}$$

Right hand side may be easier to evaluate if  $\Psi^{-1}$  and  $Z^T\Sigma^{-1}Z$  sparse (e.g. AR(1) random effects - next slide)

## Example AR(1) - covariance and inverse covariance

Consider  $U_1 = \nu_1$  and

$$U_i = aU_{i-1} + \nu_i, \quad i = 2, \dots, m$$

where  $\nu_i$  independent zero-mean normal with variances  $\text{Var}\nu_1 = \tau_1^2$  and  $\text{Var}\nu_i = \tau^2, i > 1$ .

Then  $U = B\nu$  for some  $B$  so  $U \sim N_n(0, BCB^T)$  where  $C = \text{diag}(\tau_1^2, \tau^2, \dots, \tau^2)$ . Hence  $\Psi = BCB^T$  and  $\Psi^{-1} = (B^{-1})^T C^{-1} B^{-1}$ .

**NB:**  $B^{-1}$  and  $C^{-1}$  are sparse (many zeros) and hence allows fast computations. So is  $\Psi^{-1}$  !

Expressions for covariances simplify in the stationary case  $|a| < 1$  and  $\tau_1^2 = \tau^2/(1 - a^2)$ .

Limiting case  $a \rightarrow 1$  is improper pairwise difference density (Lec 1, Excs 9).



# ANOVA models

ANOVA models arise when model specified using cross-combinations of factors/grouping variables or nested factors.

Example: one- and two-way analysis of variance.

Example: nested model for reflectance measurements.

E.g. one-way ANOVA:  $Z$  has entries  $Z_{(ij),q} = 1$   $i = q$  and 0 otherwise,  $i, q = 1, \dots, k$   $j = 1, \dots, m$ .

# Likelihood for linear mixed model

log likelihood for linear mixed model with covariance matrix  $V(\psi) = Z\Psi Z^T + \Sigma$ :

$$-\frac{1}{2} \log(|V(\psi)|) - \frac{1}{2} (y - X\beta)^T V(\psi)^{-1} (y - X\beta)$$

$\psi$ : parameters for covariance matrix (e.g. variance components)

# MLE and weighted least squares

Assume  $\psi$  known. MLE for  $\beta$  is weighted least squares estimate

$$\hat{\beta}(\psi) = \arg \min_{\beta} (y - X\beta)^T V(\psi)^{-1} (y - X\beta)$$

Differentiate and equate to zero:

$$X^T V(\psi)^{-1} (y - X\beta) = 0 \Leftrightarrow \hat{\beta}(\psi) = (X^T V(\psi)^{-1} X)^{-1} X^T V(\psi)^{-1} y$$

(provided relevant inverses exist)

Covariance parameters  $\psi$ : often numerical optimization is needed to maximize profile likelihood

$$-\frac{1}{2} \log(|V(\psi)|) - \frac{1}{2} (y - X\hat{\beta}(\psi))^T V(\psi)^{-1} (y - X\hat{\beta}(\psi))$$

## Estimation using orthogonal projections

Suppose  $Y \sim N_n(\mu, \sigma^2 I)$ ,  $\mu = X\beta$ . Let  $P$  be orthogonal projection on  $M = \text{span}(X)$  (assuming  $X$  full rank,  $P = X(X^\top X)^{-1}X^\top$ ).

Then by Pythagoras,  $\|Y - X\beta\|^2 = \|Y - PY\|^2 + \|PY - X\beta\|^2$ .  
Hence  $\hat{\mu} = PY$  and  $\hat{\beta} = (X^\top X)^{-1}X^\top y$ .

Moreover  $\hat{\sigma}^2 = \|Y - PY\|^2/n = \|Y - X\hat{\beta}\|^2/n$ .

Suppose now  $Y \sim N_n(\mu, \sigma^2 W)$  where  $W = LL^\top$  fixed. Then MLE based on  $Y$  and  $\tilde{Y} = L^{-1}Y$  equivalent. Note  $\text{Cov}(\tilde{Y}) = \sigma^2 I$  and  $\mathbb{E}\tilde{Y} = L^{-1}X\beta = \tilde{X}\beta$ . Hence by the above,

$$\hat{\beta} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{y} = (X^\top W^{-1} X)^{-1} X^\top W^{-1} y$$

and

$$\hat{\sigma}^2 = (y - X\hat{\beta})^\top W^{-1} (y - X\hat{\beta})/n$$

## Profile likelihood - uncorrelated noise

Suppose  $\text{Cov}\epsilon = \sigma^2 I$  ( $n \times n$ ) and  $\text{Cov}U = \Psi = \tau^2 L(\theta) L(\theta)^\top$  ( $k \times k$ )

Then  $(\psi = (\sigma^2, \theta, \phi))$

$$V(\psi) = \sigma^2 (I + \phi Z L(\theta) L(\theta)^\top Z^\top) = \sigma^2 W(\phi, \theta)$$

where  $\phi = \tau^2 / \sigma^2$  (signal to noise ratio).

Given  $\phi$  and  $\theta$ ,

$$\hat{\beta}(\phi, \theta) = (X^\top W^{-1}(\phi, \theta) X)^{-1} X^\top W(\phi, \theta)^{-1} y$$

and

$$\hat{\sigma}^2(\phi, \theta) = \frac{1}{n} (y - X \hat{\beta}(\phi, \theta))^\top W(\phi, \theta)^{-1} (y - X \hat{\beta}(\phi, \theta))$$

Then use matrix identity and result on next slide to get

$$W(\phi, \theta)^{-1} = (I + \phi Z L(\theta) L(\theta)^T Z^T)^{-1} = \\ I - Z(\phi^{-1}(L(\theta)^T)^{-1} L(\theta)^{-1} + Z^T Z)^{-1} Z^T$$

and

$$|I_n + \phi Z L(\theta) L(\theta)^T Z^T| = |I_k + \phi L(\theta) L(\theta)^T Z^T Z|$$

Note: now we just need to invert/compute determinant of  $k \times k$  and typically  $k < n$ .

Profile log likelihood for  $(\phi, \theta)$ :

$$l(\phi, \theta) = -\frac{1}{2} \log |\hat{\sigma}^2(\phi, \theta) W(\phi, \theta)| - \frac{n}{2} \equiv \\ -\frac{n}{2} \log \hat{\sigma}^2(\phi, \theta) - \frac{1}{2} \log |I_k + \phi L(\theta) L(\theta)^T Z^T Z|$$

## Some further useful matrix results

Consider

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Suppose  $A_{11}$  is invertible. Then

$$|A| = |A_{11}| |A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

Similarly, if  $A_{22}$  is invertible:  $|A| = |A_{22}| |A_{11} - A_{12}A_{22}^{-1}A_{21}|$

Proof: use that

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} = \begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} A$$

Moreover, if  $A : n \times k$  and  $B : k \times n$  then

$$|I_n + AB| = |I_k + BA|$$

Proof: use above result on

$$\begin{bmatrix} I_n & -A \\ B & I_k \end{bmatrix}$$

# MLE's of variances biased or inconsistent

For simple normal sample  $Y_i \sim N(\xi, \sigma^2)$ , MLE  $\hat{\sigma}^2$  is biased:

$$E\hat{\sigma}^2 = \sigma^2(n-1)/n$$

Bias arise from estimation of  $\xi$  ( $\sum_i (y_i - \xi)^2$  vs  $\sum_i (y_i - \bar{y})^2$ ).

Neyman-Scott example:  $y_{ij} = \xi_i + \epsilon_{ij}$ ,  $i = 1, \dots, k$  and  $j = 1, 2$ .  
MLE of  $\sigma^2$  not even consistent as  $k$  tends to infinity (exercise).



## REML (restricted/residual maximum likelihood)

Idea: linear transform of data which eliminates mean. Suppose design matrix  $X : n \times p$  and let  $A : n \times (n - p)$  have columns spanning the orthogonal complement  $M^\perp$  of  $M = \text{span}X$ . Then  $A^T X = 0$ .

Transformed data  $((n - p) \times 1)$

$$\tilde{Y} = A^T Y = A^T ZU + A^T \epsilon$$

has mean 0 and covariance matrix  $A^T V(\psi)A$ . Then proceed as for MLE.

NB: suppose  $A$  and  $B$  both span  $M^\perp$ . Then the same REML estimate of  $\psi$  is obtained (proof:  $B = AC$  for an invertible matrix  $C$ , write out likelihoods for  $\tilde{Y}$  using  $A$  and  $AC$ ).

NB:  $M^\perp$  is the null-space of  $X^T$ .

## REML continued

Given REML estimate  $\hat{\psi}$  we use weighted least squares estimate of  $\beta$ :

$$\hat{\beta} = (X^T V(\hat{\psi})^{-1} X)^{-1} X^T V^{-1}(\hat{\psi}) y$$

# REML examples

Simple normal sample:  $A$  has columns  $e_i - 1_n/n$ ,  $i = 1, \dots, n-1$  where  $1_n$  is the  $n$ -vector of 1's and  $e_i$  is the  $i$ th standard basis vector.

Alternative: use columns  $e_i - e_n$ ,  $i = 1, \dots, n-1$ .

Neyman-Scott problem:  $A^T$  has rows of the form  $(1, -1, 0, \dots, 0)$ ,  $(0, 0, 1, -1, 0, \dots, 0)$  etc.

## Implementation of REML - uncorrelated noise

Suppose  $\text{Cov}\epsilon = \sigma^2 I$  and  $\text{Cov}U = \Psi = \tau^2 L(\theta) L(\theta)^\top$

Then

$$V(\psi) = \sigma^2 (I + \phi Z L(\theta) L(\theta)^\top Z^\top) = \sigma^2 W(\phi, \theta)$$

where  $\phi = \tau^2 / \sigma^2$ .

Choose  $A$  so that columns form an orthogonal basis for  $M^\perp$  where  $M = \text{span}X$ . Then  $A^\top A = I$  and  $AA^\top = I - X(X^\top X)^{-1}X^\top$  (since  $AA^\top$  is a projection matrix).

$$\text{Cov}A^\top Y = A^\top V(\psi)A = \sigma^2 (I + \phi A^\top Z L(\theta) L(\theta)^\top Z^\top A) \quad (n-p) \times (n-p)$$

Hence given  $(\phi, \theta)$  estimate of  $\sigma^2$  is (using Woodbury)

$$\hat{\sigma}^2(\phi, \theta) = \frac{1}{n-p} [\tilde{Y}^\top \tilde{Y} - \tilde{Y}^\top A^\top Z [\phi^{-1} (L(\theta) L(\theta)^\top)^{-1} + Z^\top A A^\top Z]^{-1} Z^\top A \tilde{Y}]$$

Profile REML log likelihood for  $(\phi, \theta)$ :

$$l(\phi, \theta) = -\frac{n-p}{2} \log \hat{\sigma}^2(\phi, \theta) - \frac{1}{2} \log |I + \phi Z^T A A^T Z L(\theta) L(\theta)^T|$$

Note: depends only on  $A$  through  $A A^T = I - X(X^T X)^{-1} X^T$ . This again shows that specific choice of basis for  $M^\perp$  does not matter.

(if columns in  $A$  not orthogonal, we would have  $A(A^T A)^{-1} A^T = I - X(X^T X)^{-1} X^T$  and reach the same conclusion)

# MLE for balanced one-way ANOVA

Maximizing likelihood for balanced one-way (M&T Thm 5.4 and remarks 5.13-5.16)

$$\hat{\xi} = \bar{y}_{..}, \hat{\sigma}^2 = \frac{SSE}{k(m-1)}, \hat{\tau}^2 = \frac{SSB/k - \hat{\sigma}^2}{m}$$

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2 \quad \mathbb{E}SSB/k = \frac{k-1}{k}\sigma^2 + m\frac{k-1}{k}\tau^2$$

Hence  $\hat{\tau}^2$  biased. It is asymptotically unbiased as  $k$  tends to infinity.

In lecture 3 we derive the MLEs using orthogonal projections.

# REML for balanced one-way ANOVA

E.g.  $A$  as for simple normal sample, i.e.  $\tilde{y}_{ij} = y_{ij} - \bar{y}_{..}$ .

Then REML equations for estimating  $\tau^2$  and  $\sigma^2$  coincide with the moment equations from lecture 1.

# Maximization

**NB:** In general profile likelihoods (MLE or REML) must be maximized numerically (e.g. Newton-Raphson).

For one-way ANOVA we can do it by hand in closed form but tedious.

In special case of balanced ANOVA models orthogonal decomposition makes MLE very easy (later)



# Computational details

For the general linear mixed model computational complexity arises from the need to invert and compute determinant of  $V(\psi)$ .

Strategies covered here include using possible sparsity of  $\Psi$  or possible low dimension  $k \ll n$  of  $\Psi$

Usually we just need to specify  $X$  and  $Z$  and then general software (R or SAS) takes care of numerical details and maximization.

## Exercises

1. Slides 4-5 contain 3 specifications of the distribution of  $Y$ . Show that these are equivalent (hint: calculate characteristic function (lecture 1) in case of the hierarchical specification)
2. formulate random intercept and slope model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + U_i + V_i x_{ij} + \epsilon_{ij}$$

as general linear mixed model. What are the design matrices  $X$  and  $Z$  ?

3. (AR(1)-model) Consider the 'stationary' case  $|a| < 1$  and  $\tau_1^2 = \tau^2/(1 - a^2)$ .
  - 3.1 Show  $U_i \sim N(0, \tau^2/(1 - a^2)) \Rightarrow U_{i+1} \sim N(0, \tau^2/(1 - a^2))$ .
  - 3.2 Identify  $B^{-1}$  and  $B$  and compute  $\Psi$  and  $\Psi^{-1}$ .
  - 3.3 Formulate  $V^{-1} = (\Psi + \sigma^2 I)^{-1}$  in terms of sparse matrices where  $V$  is covariance matrix for the model  $Y_i = \xi + U_i + \epsilon_i$  (AR(1)+noise).
  - 3.4 Consider the limit as  $a \rightarrow 1$  of the density of an AR(1) with  $\tau_1^2 = \tau^2/(1 - a^2)$ . How is this related to the smoothing prior in Exercise 9 from lecture 1 ?

4. Show that the REML variance estimate for a simple normal sample coincides with  $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)$ .
5. Compute MLE and REML estimates for the Neyman-Scott example. Compute mean and variance for the estimates of  $\sigma^2$ .
6. Show that if  $A$  and  $B$  both span the orthogonal complement of  $\text{span} X$  then the same REML estimates are obtained from  $A^T Y$  and  $B^T Y$ .
7. Go carefully through the derivations leading to REML profile log likelihood (slide 20-21).
8. Suppose  $Y$  has a parametric density  $f_Y(\cdot; \theta)$  and  $\tilde{Y} = T(Y)$  for a differentiable and invertible transformation  $T$  that does not depend on  $\theta$ . Show that the MLE for  $\theta$  based on  $Y$  coincides with the MLE of  $\theta$  based on  $\tilde{Y}$ .

Further, assume  $\psi = g(\theta)$  for some invertible transformation  $g$ . Show that the MLE of  $\theta$  coincides with  $g^{-1}(\hat{\psi})$  where  $\hat{\psi}$  is the MLE of  $\psi$  ( $\hat{\psi} = \text{argmax}_{\psi} f(y; g^{-1}(\psi))$ ).

9. In case of simple normal sample compute variance of MLE  $\hat{\sigma}^2$  and REML estimate  $s^2$  given that  $\sum_{i=1}^n (Y_i - \bar{Y})^2$  is  $\sigma^2 \chi^2(n-1)$  (hint:  $\text{Var} \chi^2(f) = 2f$ ).

Which estimator has smallest mean square error ? (recall mean square error is sum of variance and squared bias)