Likelihood Methods for Mixed Models

Doctoral School in Statistics
Dept. Statistical Sciences, Univ. of Padua

Ruggero Bellio, University of Udine

http://www.dss.uniud.it/utenti/bellio/

Padua, July 2010

Summary

Linear Mixed Models (LMMs) slide 3

Linear Models (LMs): notation and set-up

- Normal linear regression model

$$y_j = \mathbf{x}_j^T \boldsymbol{\beta} + \varepsilon_j \,, \qquad\qquad j = 1, \dots, N \,,$$

with $\varepsilon_j \sim N(0, \sigma^2)$, i.i.d.

- Using matrix notation

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \mathbf{I}_N)\,. \end{aligned}$$

$\mathbf{y}$ random vector, $N \times 1$

$\mathbf{X}$ design matrix, $N \times p$

$\boldsymbol{\beta}$ parameter vector, $p \times 1$

$\boldsymbol{\varepsilon}$ random vector, $N \times 1$

Hierarchical structure

- Hierarchical data (M groups); let us include a factor B with M levels in the design matrix.

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_i + \varepsilon_{ij}\,, \qquad i = 1, \ldots, M, \qquad j = 1, \ldots, n_i\,.$$

γ_i: fixed effects (with a constraint, e.g. $\gamma_1 = 0$ or $\beta_0 = 0$).

M number of groups

n_i size of the i-th group, $N = \sum_i n_i$

- Matrix notation

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_{n_i} \gamma_i + \boldsymbol{\varepsilon}_i \\ \mathbf{y} &= \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \end{aligned}$$

with $\mathbf{Z} = \mathrm{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_M})$, of order $N \times M$.

Random effects

- Let us assume that the effects for B are random

$$b_i \sim \mathcal{N}(0, \sigma_b^2)\,, \text{ i.i.d.}$$

- The model becomes

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{1}_{n_i} b_i + \boldsymbol{\varepsilon}_i \\ \mathbf{y} &= \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{b} + \boldsymbol{\varepsilon} \end{aligned}$$

- This is a simple instance of the so-called Laird and Ware model (Laird and Ware, 1982, BMCS).

The Laird and Ware model

- The model has the general structure

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i\,, \qquad i = 1, \ldots, M.$$

$\mathbf{Z}_i$ design matrix, $n_i \times q$

$\mathbf{b}_i$ random vector, $q \times 1$

Usually, the columns of $\mathbf{Z}_i$ are a subset of those of $\mathbf{X}_i$

- Distributional assumptions

$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n_i})$

$\mathbf{b}_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Psi})$, $\boldsymbol{\Psi} > 0$

Independence assumption $\boldsymbol{\varepsilon}_i \perp\!\!\!\perp \mathbf{b}_i$

Example: random slopes

- Given a single covariate of interest x, the random slopes model is

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\, x_{ij} + \varepsilon_{ij}\,, \qquad i = 1, \ldots, M, \qquad j = 1, \ldots, n_i\,,$$

  with $(b_{0i}, b_{1i})^T \sim N\left((0,0)^T, \sigma^2 \boldsymbol{\Psi}\right)$, i.i.d. and $\boldsymbol{\Psi}$ is a 2×2 matrix ($q = 2$).

- For the i-th group

$$\mathbf{Z}_i = \left(\begin{array}{cc} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{array} \right) = (\mathbf{1}_{n_i} \quad \mathbf{x}_i)\,, \qquad \mathbf{b}_i = \left(\begin{array}{c} b_{0i} \\ b_{1i} \end{array} \right)$$

- In matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_0 + \mathbf{Z}_x \mathbf{b}_1 + \boldsymbol{\varepsilon}\,,$$

  where $\mathbf{Z} = \mathrm{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_M})$, $\mathbf{Z}_x = \mathrm{diag}(\mathbf{x}_1, \ldots, \mathbf{x}_M)$, $\mathbf{b}_0 = (b_{01}, \ldots, b_{0M})^T$, $\mathbf{b}_1 = (b_{11}, \ldots, b_{1M})^T$.

The Laird and Ware model

- If $q > 1$ (multiple random effects)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^{(1)}\mathbf{b}^{(1)} + \ldots + \mathbf{Z}^{(q)}\mathbf{b}^{(q)} + \boldsymbol{\varepsilon}\,.$$

  with $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(q)}$ are block-diagonal matrices of order $N \times M$ and $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(q)}$ are $M \times 1$ vectors.

- Let $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(q)})$ $(N \times Mq)$ and $\mathbf{b} = (\mathbf{b}^{(1)T}, \ldots, \mathbf{b}^{(q)T})^T$ $(Mq \times 1)$, we get

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$$

- $\mathbf{Y}$ has mean vector $\mathbf{X}\boldsymbol{\beta}$ and suitable block-diagonal covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ comprises all the parameters entering the covariance of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

- Inclusion of random effects introduces within-group correlation and heteroscedasticity.

Extensions

- Several possible extensions
  - Within-group variance function
    $$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Lambda}_i)\,,$$
    to model residual correlation structures (e.g. for longitudinal data) or structural heteroscedasticity.
  - More levels of nesting.
  - Crossed random effects.
  - Multivariate response.
- Non-trivial additional complications, but the logical structure is unchanged.

Maximum likelihood estimation

Back to the Laird and Ware model

- Model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha})$, with $\boldsymbol{\alpha}$ including the parameters of $\boldsymbol{\Psi}$, with $\dim(\boldsymbol{\alpha}) = p_a$.
- Parameter space
  $$\Theta = \Theta_{\boldsymbol{\beta}} \times (0, +\infty) \times \Theta_{\boldsymbol{\alpha}}\,,$$
  where $\Theta_{\boldsymbol{\beta}} = \mathbb{R}^p$, $\Theta_{\boldsymbol{\alpha}} = \{\boldsymbol{\alpha} \in \mathbb{R}^{p_a} : \boldsymbol{\Psi} > 0,\ \psi_{kk} > 0\}$.
- From the hypothesis of independent groups
  $$\Rightarrow L(\boldsymbol{\theta}) = \prod_{i}^{M} L_i(\boldsymbol{\theta})\,,$$
  with $L_i(\boldsymbol{\theta}) \propto p_{\mathbf{y}_i}(\mathbf{y}_i; \boldsymbol{\theta})$.

Maximum likelihood estimation

- Marginal distribution of $\mathbf{Y}_i$

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{V}_i(\boldsymbol{\alpha}))\,,$$

  with $\mathbf{V}_i(\boldsymbol{\alpha}) = \mathbf{Z}_i\,\boldsymbol{\Psi}\mathbf{Z}_i^T + \mathbf{I}_{n_i}$.

- Hence the likelihood function is

$$\begin{aligned} L_i(\boldsymbol{\theta}) \;\; &= \;\; c(\mathbf{y}_i)\,(\sigma^2)^{-n_i/2}\,|\mathbf{V}_i(\boldsymbol{\alpha})|^{-1/2} \\ & \quad\;\; \exp\left\{-\frac{1}{2\,\sigma^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\,\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}\,. \end{aligned}$$

- Maximization of $L(\boldsymbol{\theta})$ proceeds by separating the parameters.

Maximum Likelihood estimation

- When $\boldsymbol{\alpha}$ is known, ML estimates of $\boldsymbol{\beta}$ and σ^2 are

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} &= \left(\sum_i^M \mathbf{X}_i^T\,\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,\mathbf{X}_i\right)^{-1}\sum_i^M \mathbf{X}_i^T\,\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,\mathbf{y}_i\,, \\ \hat{\sigma}^2_{\boldsymbol{\alpha}} &= \frac{\sum_i^M(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}})^T\,\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}})}{N}\,. \end{aligned}$$

- When $\boldsymbol{\alpha}$ is unknown, we use the profile likelihood for $\boldsymbol{\alpha}$

$$L_P(\boldsymbol{\alpha}) = L(\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}, \hat{\sigma}^2_{\boldsymbol{\alpha}}, \boldsymbol{\alpha})$$

$$\Rightarrow \qquad \ell_P(\boldsymbol{\alpha}) = \log L_P(\boldsymbol{\alpha}) = c(\mathbf{y}) - \frac{N}{2}\log(\hat{\sigma}^2_{\boldsymbol{\alpha}}) - \textstyle\sum_i^M \displaystyle\frac{\log|\mathbf{V}_i(\boldsymbol{\alpha})|}{2}\,.$$

- $\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\Theta_\alpha}{\operatorname{argmax}}\,\ell_P(\boldsymbol{\alpha})$, hence $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}\,, \hat{\sigma}^2 = \hat{\sigma}^2_{\boldsymbol{\alpha}}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$.

Computational aspects

- Usually

  EM algorithm starting + Newton-Raphson refining

  ensures convergence, but there may be some local maxima.

- For maximizing $\ell_P(\boldsymbol{\alpha})$ is better to use a convenient reparametrization based on the relative precision factor $\boldsymbol{\Delta}$, giving

$$\boldsymbol{\Psi}^{-1} = \boldsymbol{\Delta}^T \boldsymbol{\Delta}\,.$$

  ($\boldsymbol{\Delta}$ from Cholesky decomposition of $\boldsymbol{\Psi}^{-1}$.)

- Example: with a single random effect

$$\Psi = \frac{\sigma_b^2}{\sigma^2} \quad \Rightarrow \quad \Delta = \sqrt{\frac{\sigma^2}{\sigma_b^2}}\,.$$

ML estimation: properties

- Under suitable conditions (Nie, 2007, JSPI), $\hat{\boldsymbol{\theta}}$ has the usual asymptotic properties of the MLE, for $N, M \to \infty$. In particular

- $\hat{\boldsymbol{\beta}}$ and $(\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2)$ asymptotically uncorrelated, as

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T}\right) = 0\,, \qquad E\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \sigma^2}\right) = 0\,.$$

- Asymptotic normality

$$\hat{\boldsymbol{\beta}} \,\dot{\sim}\, \mathcal{N}\left\{\boldsymbol{\beta}, \sigma^2 \left(\sum_i^M \mathbf{X}_i^T\, \mathbf{V}_i(\boldsymbol{\alpha})^{-1}\, \mathbf{X}_i\right)^{-1}\right\}$$

$$(\hat{\boldsymbol{\alpha}}, \hat{\sigma}^2)^T \,\dot{\sim}\, \mathcal{N}\left\{(\boldsymbol{\alpha}, \sigma^2)^T, [i^{-1}(\boldsymbol{\theta})]_{(\boldsymbol{\alpha},\sigma^2),(\boldsymbol{\alpha},\sigma^2)}\right\}\,,$$

  where i is the expected Fisher information matrix.

Asymptotic behaviour

- We are dealing with two different populations, hence there are two different sources of variability ($\rightarrow$ levels of data hierarchy).
- While the estimation of $\boldsymbol{\beta}$ and σ^2 is based on a data set of size N, estimation of $\boldsymbol{\alpha}$ is based on M groups. This implies, for example, that (typically)

$$E(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\alpha} + O\left(\frac{1}{M}\right),$$

  and if M is small (compared to p) the bias can be large.
- In short: with few groups, the between-group variability is poorly estimated.

Example: one-way ANOVA

- One-way ANOVA model

$$y_{ij} = \mu + b_i + \varepsilon_{ij} \qquad i = 1, \ldots, M, \quad j = 1, \ldots, n_i\,.$$

- Assume that $n_i = n$, $\forall i$ (balanced case), and $b_i \sim \mathcal{N}(0, \sigma_b^2)$.
- Let

$$\begin{aligned} \mathrm{MSE} &= \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{M\,(n-1)}\,, \\ \mathrm{MSB} &= \frac{\sum_i n\,(\bar{y}_i - \bar{y})^2}{M-1}\,, \\ \mathrm{SST} &= \sum_i \sum_j (y_{ij} - \bar{y})^2\,. \end{aligned}$$

- The ML estimate of μ is simply $\bar{y}$.

Example: one-way ANOVA

- ML estimates of variance components

$$\hat\sigma_b^2 = \max\left[\left\{\left(1-\frac{1}{M}\right)\mathrm{MSB}-\mathrm{MSE}\right\}/n, 0\right],$$

$$\hat\sigma^2 = \begin{cases}\mathrm{MSE} & \text{se} \quad \hat\sigma_b^2 > 0\,,\\ \dfrac{\mathrm{SST}}{M\,n} & \text{se} \quad \hat\sigma_b^2 = 0\,.\end{cases}$$

- Their expected values are complicated. However, if σ_b^2 is large so $P(\hat\sigma_b^2 = 0) \doteq 0$, we get

$$E(\hat\sigma_b^2) \doteq \sigma_b^2 - \frac{1}{M}\sigma_b^2 - \frac{\sigma^2}{M\,n}\,.$$

Restricted likelihood (REML)

- What seen in the example is generally true for LMMs: MLE of $(\boldsymbol{\alpha}, \sigma^2)$ is usually downward biased as it does not account for the d.f. lost for estimating $\boldsymbol{\beta}$.

- Resolution: remove the fixed effects by means of a marginal likelihood for $(\boldsymbol{\alpha}, \sigma^2)$, the restricted likelihood (Patterson and Thompson, 1971, BKA).

- The restricted likelihood can be obtained in several other ways, such as
  - as a conditional likelihood (Smyth and Verbyla, 1997, JRSS B);
  - as an integrated likelihood (Harville, 1974, BKA);
  - as a modified profile likelihood (Severini, 2000, book).

Restricted likelihood (REML)

- The original proposal is based on $\mathbf{U} = \mathbf{A}^T\mathbf{Y}$, with $\mathbf{A}$ matrix of order $N \times (N-p)$, with full rank and columns orthogonal to those of $\mathbf{X}$. We get

$$\mathbf{U} \sim \mathcal{N}(0, \sigma^2\,\mathbf{A}^T\,\mathbf{V}(\boldsymbol{\alpha})\,\mathbf{A})\,.$$

- The marginal likelihood based on $\mathbf{U}$ is

$$L_R(\sigma^2, \boldsymbol{\alpha}) = c(\mathbf{y})\,(\sigma^2)^{p/2}\,|\sum_i^M \mathbf{X}_i^T\,\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,\mathbf{X}_i|^{-1/2}\,L(\hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}, \sigma^2, \boldsymbol{\alpha})\,,$$

  from which we get $\hat\sigma_R^2$, $\hat{\boldsymbol{\alpha}}_R$.

- The REML estimate of $\boldsymbol{\beta}$ is then given by $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_R}$.

Example: one-way ANOVA

- □ REML estimate of σ_b^2

$$\hat{\sigma}_{bR}^2 = \max\left[\{\ \mathrm{MSB} - \mathrm{MSE}\}\,/n, 0\right]$$

- □ $\hat{\sigma}_R^2 = \hat{\sigma}^2$ (false in general), and $\hat{\mu}_R = \hat{\mu}$ (almost true in general).
- □ We get $\{\hat{\sigma}_{bR}^2 = 0\} \Rightarrow \{\hat{\sigma}_b^2 = 0\}$, while the opposite inclusion does not hold.
- □ If $P(\hat{\sigma}_{bR}^2 = 0) \doteq 0$, we get

$$E(\hat{\sigma}_{bR}^2) \doteq \sigma_b^2\,.$$

- □ Broadly speaking, the bias of $\hat{\boldsymbol{\alpha}}_R$ is always smaller than that of $\hat{\boldsymbol{\alpha}}$, and REML is less sensitive to outliers than ML (Verbyla, 1993, JRSS B).

Hypothesis testing

- □ Inference on fixed effects $\boldsymbol{\beta}$ is simpler than inference on variance parameters.
- □ Testing variance parameters may be delicate when the variance matrix $\boldsymbol{\Psi}$ is singular under H_0.

  For example, serious attention is needed if the hypothesis is about the absence of random effects.

Hypothesis testing: fixed effects

- □ Wald-type tests are very common (with ML or REML).
- □ Likelihood Ratio Tests $W = 2\,\{\ell(H_1) - \ell(H_0)\}$ are preferable. As usual $W \xrightarrow{d} \chi^2_{k_1-k_0}$, with k_1, k_0 number of model parameters under H_1, H_0.
- □ W for $\boldsymbol{\beta}$ is defined only with ML, not REML!
- □ For small samples, the χ^2 approximation may be inadequate. Possible remedies are
  - Use of simulation for computing P-values
  - Pretend that $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_R$, and use of exact tests for LMs (t tests, F tests)
  - Higher-order asympotics (Kenward and Roger, 1997, BMCS)

Hypothesis testing: variance parameters

- If the null hypothesis involves only elements of $(\boldsymbol{\alpha}, \sigma^2)$, we can use the REML likelihood $\Rightarrow$ good results in practice.
- Wald-type tests are not recommendable! At the very least, they should be performed after choosing a parameterization for which the normal approximation for the distribution of the MLE/REMLE is not too poor.

  Example: use $\eta = \log \sigma$ rather than $\eta = \sigma^2$.

Testing for no random effects

- Example: one-way ANOVA

$$H_0 : \sigma_b^2 = 0\,, \qquad\qquad H_1 : \sigma_b^2 > 0\,.$$

  The parameter under H_0 is on the boundary of $\Theta_{\boldsymbol{\alpha}}$ $\Rightarrow$ the usual asymptotic theory does not hold!

  One can show that

$$W \xrightarrow{d} Z^2\, I(Z > 0)\,, \qquad\qquad Z \sim N(0, 1)\,.$$

- There are some results for more complicated problems
- At any rate, they are just asymptotic results!
- A better method is to estimate the null distribution via simulation.

Model selection

- At a model-building stage, model selection procedures based on information criteria are often useful.
- The most common ones are
  - AIC$=-2\,\ell(\hat{\boldsymbol{\theta}}) + 2n_{par}$,
  - BIC$=-2\,\ell(\hat{\boldsymbol{\theta}}) + n_{par}\log(N)$.
- As usual, the smaller the better.
- To discriminate between model with the same fixed effects, we can use ℓ_R in place of ℓ, replacing N with $N - p$.

Inference on random effects

- In mixed models, the inferential interest is not only about the parameters $\boldsymbol{\theta}$.
- Often we are interested in the random effects, which are estimated using the observed data.
- We are interested in some $2°$-level residuals $\hat{\mathbf{b}}_i$, besides more usual $1°$-level residuals

$$\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\,\hat{\boldsymbol{\beta}} - \mathbf{Z}_i\,\hat{\mathbf{b}}_i\,.$$

- The random effects are $\mathbf{b}_i$ are latent quantities. To some extent, we can talk about both their estimation or their prediction (Robinson, 1991, Statist. Sci.)

BLUP

- $\mathbf{b}_i$ is typically estimated (predicted) from the conditional mean

$$E(\mathbf{b}_i|\mathbf{y}_i) = \int_{\mathbb{R}^q} \mathbf{b}_i\, p(\mathbf{b}_i|\mathbf{y}_i)\, d\mathbf{b}_i\,.$$

- This is the BLUP, Best Linear Unbiased Predictor.
- In LMMs, assuming that $\boldsymbol{\theta}$ is known, we get

$$\hat{\mathbf{b}}_i(\boldsymbol{\theta}) = \boldsymbol{\Psi}\,\mathbf{Z}_i^T\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\,\boldsymbol{\beta})$$

- The final estimate $\hat{\mathbf{b}}_i$ is obtained by replacing $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}_R$.
- "To a non-Bayesian, all things are BLUPs "(Speed, 1991, Statis. Sci.)

That BLUP is a good thing!

- The BLUP estimate has several good properties (Robinson, 1991, Stat.Sci.).
- In brief: $\hat{\mathbf{b}}_i(\boldsymbol{\theta})$ is the best (smallest variance) predictor of $\mathbf{b}_i$, linear function of $\mathbf{y}_i$, unbiased in the sense that

$$E(\hat{\mathbf{b}}_i(\boldsymbol{\theta})) = E(\mathbf{b}_i)\,.$$

- Inference on $\mathbf{b}_i$ can be made after computing $\mathrm{var}(\hat{\mathbf{b}}_i)$, usually obtained taking into account the variability of $\hat{\boldsymbol{\beta}}$ only.
- Adjusting for $\mathrm{var}(\hat{\boldsymbol{\alpha}})$ is difficult, and requires simulation methods (Booth and Hobert, 1998, JASA).

Shrinkage effect

- Typical of BLUPs, it is rather useful for applications (in epidemiology, veterinary, genetics, in social sciences...).
- For any linear combination λ of the random effects

$$\mathrm{var}(\boldsymbol{\lambda}^T \hat{\mathbf{b}}_i) \leq \mathrm{var}(\boldsymbol{\lambda}^T \mathbf{b}_i)$$

- A useful interpretation is

$$\begin{aligned} \hat{\mathbf{y}}_i &= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{b}}_i \\ &= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\boldsymbol{\Psi}} \mathbf{Z}_i^T \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \\ &= \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \hat{\mathbf{V}}_i^{-1}) \mathbf{y}_i \end{aligned}$$

  weighted average of $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ (estimated marginal mean) and $\mathbf{y}_i$ (observed response) (Verbeke and Molenberghs, 2000, book).

  Larger weight to the observed data if $\hat{\mathbf{V}}_i^{-1} \to \mathbf{0}_{n_i}$.

Example: one-way ANOVA

- Assume known $\psi = \sigma_b^2/\sigma^2$.
- We get

$$\hat{b}_i = \frac{n_i \psi}{1 + n_i \psi} (\bar{y}_i - \bar{y}) .$$

- as $n_i \psi/(1 + n_i \psi) < 1$, $\hat{b}_i$ is a weighted mean of $E(b_i) = 0$ and $\bar{y}_i - \bar{y}$.
- Notice that $\bar{r}_i = \bar{y}_i - \bar{y}$ is the estimate from the ANOVA model with fixed effects.
- $\hat{b}_i \to 0$ when $n_i \psi \to 0$, while $\hat{b}_i \to \bar{r}_i$ when $n_i \psi \to \infty$.

Henderson's mixed model equations

- BLUP estimates for known $\boldsymbol{\alpha}$ solve a particular system of equation (Henderson et al., 1959, BMCS).
- If $\boldsymbol{\Psi} = \mathrm{diag}(\sigma^2 \boldsymbol{\Psi})$, we get

$$\begin{aligned} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \mathbf{X}^T \mathbf{Z} \mathbf{b} &= \mathbf{X}^T \mathbf{y} \\ \mathbf{Z}^T \mathbf{X} \boldsymbol{\beta} + (\mathbf{Z}^T \mathbf{Z} + \boldsymbol{\Psi}^{-1}) \mathbf{b} &= \mathbf{Z}^T \mathbf{y} \end{aligned}$$

- The equations can be obtained from the joint distribution of $(\mathbf{Y}, \mathbf{b})$.
- They are the so-called Henderson's mixed model equations.

Diagnostics & robustness

- LMMs are an extension of the linear model.
- Analysis of residuals and diagnostic methods are as important as in LMs $\Rightarrow$ they have to be adapted to the more complex data structure.
- Active research area, like also the application of techniques which are robust to outliers or misspecified distributional assumptions.
- Some references: Lesaffre and Verbeke (1998, BMCS); Ghidey, Lesaffre and Eilers (2004, BMCS); Copt and Victoria-Feser (2006, JASA).

Generalized Linear Mixed Models (GLMMs) slide 33

Extending the linear model

- Possible extensions obtained by including random effects in nonlinear models, GLMs, survival models ....
- We will consider in particular the case of GLMs.
- Results for GLMs hold also for other classes of models, at least in principle.

Generalized Linear Models (GLMs)

- Distribution of the response

$$y_j \sim p_{Y_j}(y_j) \text{ , independent,}$$

$$p_{Y_j}(y_j) = \exp\left\{\frac{1}{\sigma^2}\left[y_j\,\theta(\eta_j) - h(\eta_j)\right]\right\}$$

  with θ, h suitable functions, σ^2 known.

- Linear predictor

$$\eta_j = \mathbf{x}_j^T \boldsymbol{\beta}\,, \qquad\qquad j = 1, \ldots, N\,.$$

- Link function

$$g(E[Y_j]) = \eta_j\,.$$

Generalized LMMs (GLMMs)

- Hierarchical data, same structure as in LMMs.
- the GLM specification still holds, conditional on the random effects.
- Linear predictor

$$\eta_{ij}^{\mathbf{b}} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i \,, \qquad i = 1, \ldots, M, \qquad j = 1, \ldots, n_i \,.$$

- Link function

$$g(E[Y_{ij}|\mathbf{b}_i]) = \eta_{ij}^{\mathbf{b}} \,.$$

- For the i-th group, we get

$$\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i \,, \qquad i = 1, \ldots, M.$$

GLMMs & LMMs

- Between LMMs and GLMMs there are close similarities but also some important differences.
- Common things
  - Suitable for hierarchical data.
  - Fixed and random effects.
- Differences
  - Marginal model $\neq$ conditional model in GLMMs.
  - For $n_i = 1$ LMMs=LMs, but GLMMs $\neq$ GLMs.
- While MLE for GLMs can be obtained by applying (iterated) algorithms for LMs, it is not true that MLE for GLMMs can be obtained from algorithms developed for LMMs.

Likelihood analysis

- Model parameters and parameter space are as in LMMs, but now σ^2 is known and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$.
- From the hypothesis of independent groups

$$L(\boldsymbol{\theta}) = \prod_{i}^{M} L_i(\boldsymbol{\theta}),$$

  with $L_i(\boldsymbol{\theta}) \propto p_{\mathbf{y}_i}(\mathbf{y}_i; \boldsymbol{\theta})$.
- Computation of $L_i(\boldsymbol{\theta})$ is now more involved, as the marginal distribution of $\mathbf{y}_i$ is not computable analytically $\Rightarrow$ numerical methods.
- Once computed $L(\boldsymbol{\theta})$, we can compute AIC and BIC.
- There are some REML-like solutions (Liao and Lipsitz, 2002, BKA).

Marginal distribution of $\mathbf{y}_i$

- Obtained after integrating out the random effects

$$p_{\mathbf{y}_i}(\mathbf{y}_i; \boldsymbol{\theta}) = \int_{\mathbb{R}^q} p_{\mathbf{y}_i|\mathbf{b}_i}(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\beta})\, p_{\mathbf{b}_i}(\mathbf{b}_i; \boldsymbol{\alpha})\, d\mathbf{b}_i .$$

- With few exceptions, the integral has to be computed numerically
  - Quadrature methods (Gaussian, adaptive);
  - Laplace approximation;
  - Simulation-based methods (Monte Carlo, MCMC).
- Similarly, we can obtain BLUP-type estimators for $\mathbf{b}$.

Quadrature methods

- Approximate the integral with a sum

$$\int_{\mathbb{R}} h^*(v)\, \exp\{-v^2\}\, dv \doteq \sum_{k=1}^{d} h^*(x_k)\, w_k\,,$$

  with x_k nodes, w_k weights.

- For Gaussian quadrature, x_k and w_k are fixed, at times many points $(> 40, 50)$ may be required.

- Adaptive Gaussian quadrature methods locate the maximum and the spread of the integrand function before computing the sum $\Rightarrow$ more reliable, slower.

- First-order Laplace approximation is a special case of adaptive Gaussian quadrature with $k = 1$. It often works well (Joe, 2008, CSDA).

Example (Lesaffre and Spiessens, 2001, JRSS C)

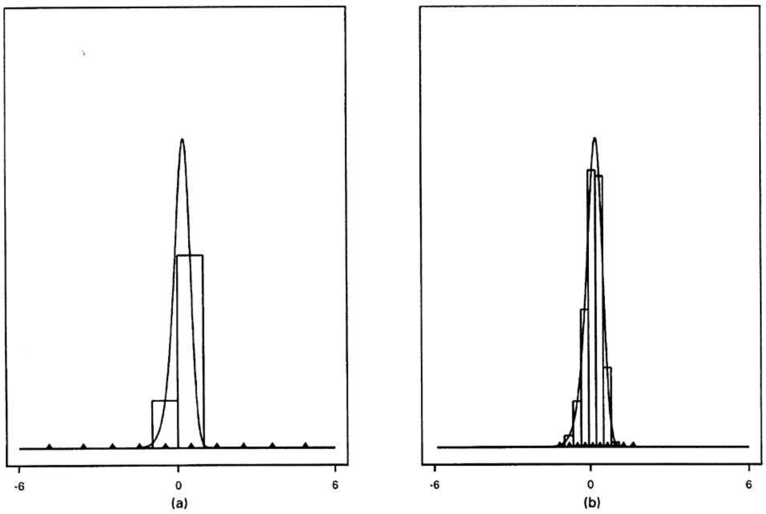

Fig. 2. Comparison of the positions of 10 quadrature points obtained from (a) an ordinary Gaussian quadrature and (b) an adaptive Gaussian quadrature for the same integrand: ▲, position of the quadrature points z_q; ▯, contribution of each point to the integral, i.e. $f(q)\omega_q$

Limitations of quadrature methods

- Quadrature methods are effective only for models with 2 or 3 levels.
- Problems arise with many random effects ($> 3, 4$).
- They are not suitable for crossed random effects, where the likelihood function does not factorize into factors from independent terms.
- Some references: Lesaffre and Spiessens (2001, JRSS C); Rodrigues and Goldman (2001, JRSS A); Clarkson and Zhan (2002, JCGS).

Simulation-based methods

- More complex and require some care, can virtually solve any kind of optimization problem.
- Main techniques for mixed models are Simulated Maximum Likelihood (SML) and Monte Carlo EM (MCEM).
- SML (Geyer and Thompson, 1992, JRSS B; Casella and Robert, 2004, book): Monte Carlo estimate of the entire likelihood

$$\int p_{\mathbf{y}|\mathbf{b}}(\mathbf{y}|\mathbf{b};\boldsymbol{\beta})\, p_{\mathbf{b}}(\mathbf{b};\boldsymbol{\alpha})d\mathbf{b} \doteq \frac{1}{K}\sum_{k=1}^{K} \frac{p_{\mathbf{y}|\mathbf{b}}(\mathbf{y}|\mathbf{b}^{(k)};\boldsymbol{\beta})\, p_{\mathbf{b}}(\mathbf{b}^{(k)};\boldsymbol{\alpha})}{h_{\mathbf{b}}(\mathbf{b}^{(k)})}$$

  with $\mathbf{b}^{(k)} \sim h_{\mathbf{b}}(\mathbf{b})$ and K rather large.

  The real issue is the choice of $h_{\mathbf{b}}(\mathbf{b})$; the optimal importance sampling distribution is given by $p_{\mathbf{b}|\mathbf{y}}(\mathbf{b}|\mathbf{y};\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the MLE. Several variations exist.

Simulation-based methods

- MCEM (McCulloch, 1997, JASA; Casella and Robert, 2004, book): Maximizes $\log L(\boldsymbol{\theta})$ by applying the EM algorithm.

  In particular, the E-step is performed by sampling from the distribution of the random effects given the data $p_{\mathbf{b}|\mathbf{y}}(\mathbf{b}|\mathbf{y};\tilde{\boldsymbol{\theta}})$, for the current value $\tilde{\boldsymbol{\theta}}$ of the parameter.

  Sampling from $p_{\mathbf{b}|\mathbf{y}}$ can be performed by MCMC.

  It is rather important to tune the accuracy of the Monte Carlo approximation (i.e. the number of simulations) as the algorithm evolves (Booth and Hobert, 1999, JRSS B).

Simulation-based methods

- Skaug (2002, JCGS) proposes a very efficient approach for computing the MLE
  - Use of simple methods (at least in principle) for computing $L(\boldsymbol{\theta})$, by Laplace approximation and importance sampling;
  - Fast and accurate evaluation of the required quantities using software for Automatic Differentiation and methods for the numerical resolution of sparse linear systems.
- For complex problems, Quasi-Monte Carlo methods can be also quite effective (Jank, 2006, Statist. & Comput.; Pan and Thompson, 2007, CSDA).

Alternative methods

- There are several methods based on estimating equations rather than the likelihood function.
- Sometimes they may represent a convenient alternative.
- Among others, they include
  - Methods that estimate $\boldsymbol{\theta}$ by using estimating equations based on approximate models: MQL, PQL.
  - Methods based on the concept of h-likelihood (Lee and Nelder, 1996; JRSS B, 2001, BKA).
  - Moment-type or quasi-likelihood estimating equation (McCullagh and Nelder, 1989, book; Jiang, 1998, JASA).
  - Methods based on some composite likelihood function (Varin, 2008, AStA).

MQL & PQL

Very used in empirical works, can be obtained in several ways

- as an extension of quasi-likelihood ideas (Breslow and Clayton, 1993, JASA);
- with Laplace-type approximations per $L(\boldsymbol{\theta})$ (Wolfinger, 1994, BKA);
- with some simple Taylor expansions (Goldstein, 1991, BKA; Schall, 1991, BKA), such as

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu}(\boldsymbol{\eta}) + \boldsymbol{\varepsilon} = \boldsymbol{\mu}(\mathbf{X}\,\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{b}) + \boldsymbol{\varepsilon} \\ &\doteq \boldsymbol{\mu}(\boldsymbol{\eta}_0) + \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}_0}\mathbf{X}\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}_0}\,\mathbf{Z}\,(\mathbf{b} - \mathbf{b}_0) + \boldsymbol{\varepsilon} \end{aligned}$$

  where $\mathbf{b}_0 = 0$ (MQL) or $\mathbf{b}_0 = \hat{\mathbf{b}}(\boldsymbol{\beta}_0)$ (PQL). Estimate the related LMM (by ML or REML), and iterate.

MQL & PQL

- The resulting estimating equations extend Henderson's mixed model equations (McGilchrist, 1994, JRSS B).
- They do not provide a valid $L(\hat{\boldsymbol{\theta}})$ $\Rightarrow$ no LRTs or AIC/BIC.
- The idea is simple and computationally efficient, but it does not always works: the estimators are consistent when $n_i \to \infty$, or when $\mathbf{Y}|\mathbf{b}$ is approximately normal. With small n_i and discrete data, such methods can be badly biased.
- There are methods that improve on the Taylor expansions (Goldstein and Rasbash, 1996, JRSS A), or make use of bootstrap corrections (Kuk, 1995, JRSS B).

Example: logistic regression

- Data from a multicenter clinical trial (Booth and Hobert, 1998, JASA)

Clinic	1	2	3	4	5	6	7	8
Trt 1	$11/36$	$16/20$	$14/19$	$2/16$	$6/17$	$1/11$	$1/5$	$4/6$
Trt 2	$10/37$	$22/32$	$7/19$	$1/17$	$0/12$	$0/10$	$1/9$	$6/7$

- Logistic regression with random intercepts, $M = 8$, $n_i = 13 - 73$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_b^2)$.

Example: logistic regression

- MLE based on quadrature methods

	GHQ 5	GHQ 50	AGH 1	AGQ 10
Int	$-1.36(.28)$	$-1.20(.56)$	$-1.20(.55)$	$-1.20(.56)$
Trt1	$0.77(.30)$	$0.74(.30)$	$0.74(.30)$	$0.74(.30)$
σ_b	$1.31(.22)$	$1.40(.43)$	$1.39(.38)$	$1.40(.43)$

- PQL-like methods

	PQL-ML	PQL-REML	PQL-REML (Bootstrap)
Int	$-1.14(.60)$	$-1.14(.56)$	$-1.21(.60)$
Trt1	$0.72(.30)$	$0.72(.31)$	$0.74(.30)$
σ_b	$1.32(.30)$	$1.42(.34)$	$1.54(.49)$

Semiparametric regression

- There are important connections between mixed models and semiparametric regression based on penalized splines (Robinson, 1991, Stat.Sci. + discussion).
- The basic idea can be grasped in the simple case of scatterplot smoothing.

Semiparametric regression: scatterplot smoothing

- Very simple model

$$y_i = f(x_i) + \varepsilon_i\,, \qquad i = 1, \ldots, n,$$

  where f is a smooth function and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, i.i.d.

- Mixed model formulation of penalized regression splines

$$f(x_i) = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \sum_{k=1}^{K} b_k s_k(x_i)\,,$$

  where $\{s_k(\cdot), k = 1, \ldots, K\}$ is a set of spline basis functions and K is the number of knots, whose choice depends on the sample size.

- We obtain a linear mixed model by assuming $b_k \sim \mathcal{N}(0, \sigma_b^2)$, i.i.d.

Semiparametric regression: scatterplot smoothing

- A very simple choice for $s_k(\cdot)$ is given by truncated lines, for which $p = 1$

$$s_k(x_i) = (x_i - x_k)_+ = \max(x_i - x_k, 0)\,.$$

- In matrix notation, we can write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}\,,$$

  where $\mathbf{X} = (\mathbf{1}_n \ \ \mathbf{x} \ \ \ldots \ \ \mathbf{x}^p)$ and $\mathbf{Z}$ corresponds to the basis function.

- For known variance components (σ_b^2, σ^2) (and with $\psi = \sigma_b^2/\sigma^2$), Henderson's mixed model equations minimize an objective function for $(\boldsymbol{\beta}, \mathbf{b})$ with a certain penalization for $\mathbf{b}$

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2 + \frac{1}{\psi}\|\mathbf{b}\|^2\,.$$

Semiparametric regression

- Mixed models can be used for extending in a suitable way basic models for semiparametric regression, providing a framework for including correlation structures, heteroscedasticity, clustering effects ...
- A detailed treatment of this fascinating (and very modern) approach is given in Ruppert, Wand and Carroll (2003, book).

  See also Ruppert, Wand, Carroll (2009, Elect. J. Stat.)

References slide 54

Bibliography

- Searle, Casella and McCulloch (1992, Wiley);
- Snijders and Bosker (1999, SAGE);
- McCulloch and Searle (2001, Wiley);
- Pinheiro and Bates (2000, Springer);
- Verbeke and Molenberghs (2000, Springer);
- Ruppert, Wand and Carroll (2003, Cambridge UP);
- Demidenko (2004, Wiley);
- Skrondal and Rabe-Hesketh (2004, Chapman & Hall/CRC);
- Molenberghs and Verbeke (2005, Springer);
- ...

Some software options

- LMMs
  - Statistical environments: R nlme, lme4; SAS MIXED; STATA gllamm, xt*.
  - Specialized packages: HLM, MLwIN.
- GLMMs - PQL/MQL
  - Statistical environments: R nlme, MASS, lme4; SAS GLIMMIX, NLINMIX.
  - Specialized packages: HLM; MLwIN.
- GLMMs - ML
  - Statistical environments: R nlme, MASS, lme4, glmmML; SAS NLMIXED; STATA gllamm, xt*.
  - Specialized packages: HLM; aML; AD Model Builder.