

MA/ST 810

Mathematical-Statistical Modeling and Analysis of Complex Systems

Hierarchical Statistical Models for Complex Data Structures

- Motivation
- Basic hierarchical model and assumptions
- Statistical inference
- Methods for implementation
- Model extensions
- Discussion

Motivation

Recall the theophylline data: 12 subjects, each observed over time following oral dose of theophylline

Objective 2: Learn about how absorption, distribution, elimination (ADME) differ from subject to subject \Rightarrow dosing recommendations for the *population* of likely subjects

- Are there certain types of subjects who need to be dosed *differently*?

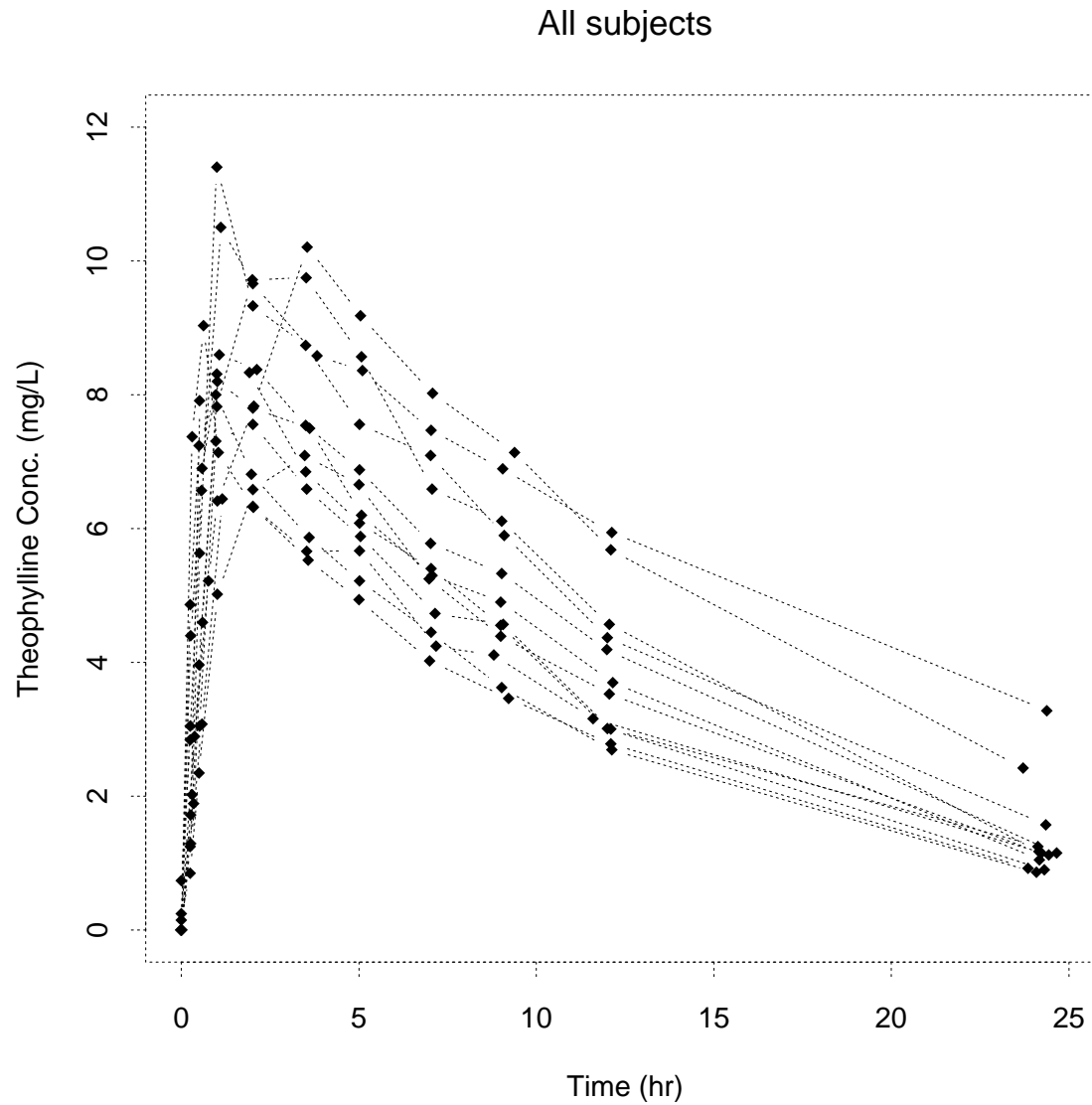
Mathematical (compartmental) model for any particular subject:

Theophylline concentration at time t

$$f(t, U, \theta) = \frac{k_a F D}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}, \quad \theta = (k_a, k_e, V)^T, \quad U = D$$

Motivation

All 12 subjects:



Motivation

Objective 2: ADME in the *population* of subjects

- *Similar pattern* for all subjects, but *different features* \Rightarrow each subject has his/her *own* θ
- *Subject-to-subject variation* \Rightarrow θ values *vary* in the *population*
- *Objective, restated* – learn about θ values in the (hypothetical) *population* of subjects like these
- But we have only seen a *sample* of 12 subjects from this population \Rightarrow *uncertainty* about *entire* population
- ...and we don't get to see θ values; we only can infer them *indirectly* through the *concentrations* on each subject
- ...which are subject to *uncertainty* due to *measurement error*, "*biological fluctuation*"

Motivation

Issue: How to *formalize* this objective and take into account *uncertainty* from all these *sources of variation*?

Required: A *statistical model* that reflects the *data-generating mechanism*

- *Sample* m subjects from the *population*
- For the i th subject, ascertain concentration at *each of* n_i *time points* (could be *different* for each subject i)
- A sort of “*two-stage*” data-generating process
- \Rightarrow The model must contain *components* that represent each stage. . .
- “*Hierarchical statistical model*”

Model and assumptions

Clearly: The *math model* $f(t, U, \theta)$ pertains to a *single subject*

- Representation of *pharmacokinetic processes* taking place *within* a given subject
- \Rightarrow An appropriate *statistical model* and *method for inference* must recognize this
- *In particular*, the *statistical model* we will now describe will lead to an *inferential approach* that respects this (more later...)

Model and assumptions

Formalization: “Learn about θ values in the population”

- *Conceptualize the* (infinitely-large) *population* as *all possible* θ (one for each subject) – how θ s would “*turn out*” if we sampled subjects
- \Rightarrow Represent by a (joint) *probability distribution* for θ (with *mean*, *covariance matrix*, etc)
- \Rightarrow “*Average value* of θ ” (*mean*), “*variability of* k_a, k_e, V ” (diagonal elements of *covariance matrix*), “*associations between PK processes*” (off-diagonal elements)

Model and assumptions

Thus: Think of potential θ values if we draw m subjects *independently* as *independent random vectors* θ_i , each with *this probability distribution*, e.g., a standard model is

$$\theta_i \sim \mathcal{N}_p(\beta, D), \quad i = 1, \dots, m \quad (1)$$

- Other distributional assumptions also possible (coming up...)
- When we *actually do this*, the i th subject in our sample has an associated *realization* of θ_i , $i = 1, \dots, m$
- *Equivalent representation*

$$\theta_i = \beta + b_i, \quad b_i \sim \mathcal{N}_p(0, D), \quad i = 1, \dots, m$$

- b_i is called a “*random effect*”

Model and assumptions

Objective, formalized: *Estimate* β and D

- β and D fully characterize the (assumed common for now) distribution of the b_i and hence of the θ_i
- And, *of course*, provide an assessment of the *uncertainty* involved

Complication: We *do not* observe a sample of θ_i directly

- This estimation must be based on the *concentration-time* data from all m subjects

First requirement: A model for *data-generating mechanism* from a given subject i ...

Model and assumptions

Data for subject i : *Given* θ_i (focus on subject i)

- Concentrations arise from an *assumed mechanism* for an *individual subject* as before, e.g.,

$$Y_{ij} = f(t_{ij}, U_i, \theta_i) + \epsilon_{ij}, \quad j = 1, \dots, n_i$$

$t_i = (t_{i1}, \dots, t_{in_i})^T$ are the time points at which i would be observed, $f_i(U_i, \theta_i) = \{f(t_{i1}, U_i, \theta_i), \dots, f(t_{in_i}, U_i, \theta_i)\}^T$

- $Y_i = (Y_{i1}, \dots, Y_{in_i})^T$ *random vector* of observations from i
- *Same considerations* for ϵ_{ij} as before, e.g., $\epsilon_{ij} = \epsilon_{1ij} + \epsilon_{2ij}$

Model and assumptions

Result: Specify a family of *conditional probability distributions* for $Y_i|U_i, \theta_i$, e.g.,

$$Y_i|U_i, \theta_i \sim \mathcal{N}_{n_i} \{f_i(U_i, \theta_i), \sigma^2 I_{n_i}\} \quad (2)$$

- $E(Y_i|U_i, \theta_i) = f_i(U_i, \theta_i),$

$$\text{var}(Y_i|U_i, \theta_i) = \sigma^2 I_{n_i} = \Lambda_i(U_i, \theta_i, \alpha), \quad \alpha = \sigma^2$$

- Taking σ^2 *the same* for all i reflects belief that *assay error* is *similar* for blood samples from *any* subject (as opposed to σ_i^2 ; could generalize)

Model and assumptions

Or a fancier model:

$$Y_i|U_i, \theta_i \sim \mathcal{N}_{n_i}\left\{f_i(U_i, \theta_i), \underbrace{\sigma_1^2 R_i(U_i, \theta_i, \gamma) + \sigma_2^2 \Gamma_i(\phi)}_{\Lambda_i(U_i, \theta_i, \alpha)}\right\} \quad (3)$$

$$R_i(U_i, \theta_i, \gamma) = \text{diag}\{f^{2\gamma}(t_{i1}, U_i, \theta_i), \dots, f^{2\gamma}(t_{in_i}, U_i, \theta_i)\}, \quad (n_i \times n_i)$$

$$\Gamma_i(\phi)_{jj} = 1, \quad \Gamma_i(\phi)_{jj'} = \exp\{-\phi(t_{ij} - t_{ij'})^2\}, \quad (n_i \times n_i)$$

- *Common* $\alpha = (\sigma_1^2, \sigma_2^2, \gamma, \phi)^T$ across subjects reflects *assumption* of similar pattern of *variation* due to each source, could be modified

- Here, $E(Y_{ij}|U_i, \theta_i) = f(t_{ij}, U_i, \theta_i)$

$$\text{var}(Y_{ij}|U_i, \theta_i) = \sigma_1^2 f^{2\gamma}(t_{ij}, U_i, \theta_i) + \sigma_2^2$$

- *Conditional covariance*

$$\text{cov}(Y_{ij}, Y_{ij'}|U_i, \theta_i) = \sigma_2^2 \exp\{-\phi(t_{ij} - t_{ij'})^2\}$$

Model and assumptions

Assumptions: $Y = (Y_1^T, \dots, Y_m^T)^T$, $\theta = (\theta_1^T, \dots, \theta_m^T)^T$,
 $U = (U_1, \dots, U_m)^T$

- Assume θ_i *independent across* $i \Rightarrow$ reasonable if individuals are *unrelated*, chosen *at random* from the population
- Assume $\theta_i \perp\!\!\!\perp U_{i'}$ for all $i, i' \Rightarrow$ e.g., dose received is unrelated to underlying PK
- Also assume Y_i are $\perp\!\!\!\perp$ of each other, and $Y_i \perp\!\!\!\perp \theta_{i'}$, $Y_i \perp\!\!\!\perp U_{i'}, i' \neq i \Rightarrow$ reasonable for *unrelated* individuals

Model and assumptions

Hierarchical statistical model: *Two-stage hierarchy*

1. *Assumption* for *family of conditional probability distributions* for $Y_i|U_i, \theta_i$ that governs data at the *individual level* \Rightarrow pmf/pdf

$$p(y_i|u_i, \theta_i, \alpha)$$

(n_i -variate, depending on t_i); e.g., like (2) or (3)

We highlight dependence on θ_i *separately* from other model parameters α (taken to be *the same* across individuals here; can generalize); e.g., $\alpha = (\sigma_1^2, \sigma_2^2, \gamma, \phi)^T$ in (3) on slide 13

2. *Assumption* for *probability distribution* for $\theta_i \Rightarrow$ pmf/pdf

$$p(\theta_i|\beta, \zeta)$$

p -variate, $p = \dim(b_i)$ [same for all i ; e.g., like (1) $\mathcal{N}_p(\beta, D)$]

Here, ζ represents *parameters* in addition to β required to describe this distribution, e.g., $\zeta = \text{vech}(D)^T$

Model and assumptions

Implications:

- With these assumptions, *joint pmf/pdf* of (Y^T, θ^T) is

$$p(y, \theta | u, \alpha, \beta, \zeta) = \prod_{i=1}^m p(y_i, \theta_i | u_i, \alpha, \beta, \zeta)$$

- Depends on *parameters*, e.g., $\psi = \{\sigma_1^2, \sigma_2^2, \gamma, \phi, \beta^T, \text{vech}(D)^T\}^T = (\alpha^T, \beta^T, \zeta^T)^T$ (to be *estimated*)
- $p(y_i, \theta_i | u_i, \alpha, \beta, \zeta) = p(y_i | u_i, \theta_i, \alpha) p(\theta_i | \beta, \zeta)$ for each $i \Rightarrow$

$$\prod_{i=1}^m p(y_i, \theta_i | u_i, \alpha, \zeta) = \prod_{i=1}^m p(y_i | u_i, \theta_i, \alpha) p(\theta_i | \beta, \zeta)$$

- The model contains *observable* and *unobservable* random components – $Y_i, i = 1, \dots, m$, are observed, θ_i are *not* – *but both* are required in the formulation to reflect all *sources of variation*
- Because θ_i are *not observed*, would like the *probability distribution* of $Y|U$ *alone*...

Model and assumptions

Illustration: What does the *hierarchy* (1) and (2) *together* say about $Y|U$? Consider just *mean* and *covariance matrix*

- From (2), $E(Y_i|U_i, \theta_i) = f_i(U_i, \theta_i)$, $\text{var}(Y_i|U_i, \theta_i) = \sigma^2 I_{n_i}$
- Thus, $E(Y_i|U_i) = E\{f_i(U_i, \theta_i)\} = \int f_i(U_i, \theta)p(\theta_i|\beta, \zeta) d\theta_i$, where $p(\theta_i|\beta, \zeta)$ is the *multivariate normal pdf* corresponding to (1)
- *Well-known*: $\text{var}(Y_i|U_i) = \sigma^2 I_{n_i} + \text{var}\{f_i(U_i, \theta_i)|U_i\}$
- The second term is *NOT* a *diagonal matrix* $\Rightarrow \text{var}(Y_i|U_i)$ implies that elements of Y_i , Y_{ij} and Y_{ij}' are *correlated* given U_i
- \Rightarrow Even if observations on *within* a single subject are *not serially correlated*, in the context of the *population of subjects*, observations on the *same* subject are *similar* (correlated) precisely because they are from the *same subject* and hence are “*more alike*” than observations from different subjects!
- \Rightarrow The *hierarchical model* takes expected “*similarity*” of observations on the same individual into account “*automatically*”

Model and assumptions

More generally: (*Marginal, given U_i*) *probability distribution* for *observable random vector Y* has pmf/pdf

$$\begin{aligned} p(y|u, \alpha, \beta, \zeta) &= \int p(y, \theta|u, \alpha, \beta, \zeta) d\theta \\ &= \int \prod_{i=1}^m p(y_i, \theta_i|u_i, \alpha, \beta, \zeta) d\theta_i \\ &= \prod_{i=1}^m \int p(y_i|u_i, \theta_i, \alpha) p(\theta_i|\beta, \zeta) d\theta_i \end{aligned}$$

Objective: Use to find *estimator* for $\psi = (\alpha^T, \beta^T, \zeta^T)^T$ via *maximum likelihood* (coming up...)

Model and assumptions

Model refinement: Recall the assumption

$$\theta_i \sim \mathcal{N}(\beta, D) \Leftrightarrow \theta_i = \beta + b_i, b_i \sim \mathcal{N}_p(0, D)$$

- *Implication* – all subjects are *exchangeable*
- *Practically speaking* – no attempt to make a more *refined* explanation of *variation* in the population
- *For example*, for theophylline, *elimination* characteristics may vary *systematically* with weight \Rightarrow e.g., k_{ei} are larger for smaller weights W_i (subjects *not exchangeable* this way)
- $\Rightarrow k_{ei} = \beta_{k_e,1} + \beta_{k_e,2}W_i + b_{k_e,i}$, $b_{k_e,i}$ is random effect with mean 0

Model and assumptions

More generally:

$$\theta_i = h(A_i, \beta) + b_i$$

for vector of (“*among individual*”) *covariates* A_i (e.g., weight, age, kidney function, disease status, etc.)

- *Parameter* β and *random effect* $b_i \sim \mathcal{N}_p(0, D)$

- Thus,

$$\theta_i | A_i \sim \mathcal{N}_p\{h(A_i, \beta), D\}$$

- Distribution of θ_i is *different* for each value of A_i

Model and assumptions

For example: Theophylline, $\beta = (\beta_1^T, \beta_2^T, \beta_3^T)^T$, $b_i = (b_{k_a,i}, b_{k_e,i}, b_{V,i})^T$

$$\left. \begin{aligned} \log k_{ai} &= A_i^T \beta_1 + b_{k_a,i} \\ \log k_{ei} &= A_i^T \beta_2 + b_{k_e,i} \\ \log V_i &= A_i^T \beta_3 + b_{V,i} \end{aligned} \right\} \theta_i = \underbrace{\mathcal{A}_i \beta}_{h(\mathcal{A}_i, \beta)} + b_i, \quad \mathcal{A}_i = \begin{pmatrix} A_i^T & 0 & 0 \\ 0 & A_i^T & 0 \\ 0 & 0 & A_i^T \end{pmatrix}$$

- $E(b_i) = 0$, $\text{var}(b_i) = D$; *often assumed* $b_i \sim \mathcal{N}_p(0, D)$
- *Subject matter* – population distributions of PK parameters are *skewed* rather than *symmetric*
- Might *parameterize* the PK model *directly* in terms of $\theta_i = (\log k_{ai}, \log k_{ei}, \log V)^T$
- In general $\Rightarrow E(\theta_i | A_i) = h(A_i, \beta)$ – determining β characterizes how PK parameters are *associated* with elements of A_i (important for dosing/usage recommendations, contraindications)
- ... and “*explains*” some of the variation in the *population*

Model and assumptions

Data collection: For each subject, measure *not only* Y_i under conditions U_i , but *also* a vector of *among-individual covariates* A_i

- $(Y_i^T, U_i^T, A_i^T)^T, i = 1, \dots, m$
- Some elements of A_i may be *fixed by design* (e.g., assignment to fed or fasting group); other elements may be *observed* (e.g., weight, creatinine clearance, smoking status)

Model and assumptions

Statistical model: The stages of the *hierarchy* become

1. *Individual model: Conditional probability distributions* $Y_i|U_i, A_i, \theta_i$
 \Rightarrow with *substitution of stage 2*, implies conditional pmf/pdf
 $p(y_i|u_i, a_i, b_i, \beta, \alpha)$
2. *Population model: Probability distribution for* $\theta_i|A_i \Rightarrow$ with
assumption $b_i \perp\!\!\!\perp A_i$, implies
 $p(a_i, b_i) = p(b_i|a_i, \zeta)p(a_i) = p(b_i|\zeta)p(a_i)$

Remark: The *assumption* $b_i \perp\!\!\!\perp A_i$ implies belief that for each possible value of A_i , variation among θ_i values is *the same* (although $E(\theta_i|A_i)$ changes with changing A_i) – this may be *relaxed*

Model and assumptions

Joint pdf: $b = (b_1^T, \dots, b_m^T)^T$, $A = (A_1^T, \dots, A_m^T)^T$

$$\begin{aligned} p(y, a, b|u, \alpha, \beta, \zeta) &= \prod_{i=1}^m p(y_i, a_i, b_i|u_i, \alpha, \beta, \zeta) \\ &= \prod_{i=1}^m p(y_i|u_i, a_i, b_i, \beta, \alpha) p(b_i|a_i, \zeta) p(a_i) \\ &= \prod_{i=1}^m p(y_i|u_i, a_i, b_i, \beta, \alpha) p(b_i|\zeta) p(a_i) \end{aligned}$$

and thus the joint pdf of the *observable data* $(Y^T, A^T)^T$ is

$$p(y, a|u, \alpha, \beta, \zeta) = \left\{ \prod_{i=1}^m \int p(y_i|u_i, a_i, b_i, \alpha, \beta) p(b_i|\zeta) db_i \right\} \left\{ \prod_{i=1}^m p(a_i) \right\}$$

- pdf/pmf of A_i *factors out* of the integral
- Will be important momentarily...

Model and assumptions

General statistical model: Recap in context of (3) on slide 13

$$1. Y_i = f_i(U_i, \theta_i) + \epsilon_i = f_i(U_i, A_i, \beta, b_i) + \underbrace{\Lambda_i^{1/2}(U_i, A_i, b_i, \beta, \alpha)\delta_i}_{\epsilon_i}$$

Note: Distribution of ϵ_i may depend on θ_i and hence b_i – write $\epsilon_i = \Lambda_i^{1/2}(U_i, \theta_i, \alpha)\delta_i = \Lambda_i^{1/2}(U_i, A_i, b_i, \beta, \alpha)\delta_i$ for $\delta_i \sim (0, I_{n_i})$

$$2. \theta_i = h(A_i, \beta) + b_i \Rightarrow E(\theta_i|A_i) = h(A_i, \beta)$$

Recall: Usual assumption $b_i \perp\!\!\!\perp A_i$ (More generally, could have model of form $\theta_i = h(A_i, \beta, b_i)$, *nonlinear* in b_i)

Together: Specifies $p(y_i|u_i, a_i, b_i, \beta, \alpha)$, $p(b_i|a_i, \zeta) = p(b_i|\zeta)p(a_i)$

- *Standard assumptions* – *normality* of δ_i and b_i

Model and assumptions

Terminology: In the statistical literature, such *hierarchical models* involving $f(t, U, \theta)$ *nonlinear* in θ are referred to as *nonlinear mixed effects models*

- The *first stage* model for $Y_i|U_i, A_i$ that leads to the specification of $p(y_i|u_i, a_i, b_i, \beta, \alpha)$ is referred to as the *individual-level model*
- The *second stage* model for θ_i that leads to the assumption on the *random effects* $b_i, p(b_i|\zeta)$, is referred to as the *population model*
- The formulation we have presented is for *scalar* observations, but *extends readily* to the case of *multivariate* observations where

$$Y_{ij} = (Y_{ij}^{(1)}, \dots, Y_{ij}^{(k)})^T$$

is observed at t_{ij} on subject i (or further with different components at different times)

- The methods for *inference* and *implementation* discussed next also extend

Inference

Natural approach: *Maximum likelihood* – e.g., in context of (3)

$$L(\psi|y, a) = \prod_{i=1}^m \int p(y_i|u_i, a_i, b_i; \beta, \alpha) p(b_i|\zeta) db_i$$

- $\psi = \{\alpha^T, \beta^T, \zeta^T\}^T = \{\sigma_1^2, \sigma_2^2, \gamma, \phi, \beta^T, \text{vech}(D)^T, \}^T$
- As long as $p(a_i)$ *does not depend on* elements of ψ , is *irrelevant*, so can *disregard*
- *MLE* maximizes $L(\psi|y, a)$ in ψ for y, a fixed
- *Complication* – integrals almost certainly *intractable* and possibly (probably) *high dimensional* (dimension p of b_i)
- *Integration required* on each internal iteration of optimization

Result: Methods for maximizing $L(\psi|y, a)$ in *popular software* rely on *approximations* to the integral

Inference

Sampling distribution of MLE $\hat{\psi}$:

- *Focus* is on *population* (as represented by β, D) \Rightarrow a good experiment would choose m *large*
- In many applications, n_i are *not too large* (limitations of sampling same individual many times; e.g., humans in a PK study)
- \Rightarrow *Usual approximation* to sampling distribution is under $m \rightarrow \infty$, n_i fixed
- *Alternatively*, in some applications, larger n_i are possible (e.g., rats in an exposure study)
- \Rightarrow Approximate under $m \rightarrow \infty$, $\min n_i \rightarrow \infty$
- Derivation of *approximate sampling distribution* for $\hat{\psi}$ is *standard* in either case – $\hat{\psi} \dot{\sim} \mathcal{N}_r(\psi_0, \Sigma_0)$, ψ ($r \times 1$)

Implementation

Key issues:

- Doing the *integral* for each subject at each internal iteration of optimization algorithm
- Computing the *forward solution* for each subject at each time within each internal iteration of optimization
- *Nasty objective function* (local maxima)

Implementation

Quadrature: A quadrature method *approximates* (deterministically) an integral by a weighted sum over predefined abscissæ

- For *normal* b_i , *Gauss-Hermite* quadrature
- The more abscissæ, the better the *accuracy* of the approximation
- If $\dim(b_i) = p > 1$ – need abscissæ in each dimension
- “*Curse of Dimensionality*” – as $p = \dim(b_i)$ grows, number of function evaluations gets large \Rightarrow takes *too much time*, *too complex*

Implementation

Modification: “*Adaptive Gaussian quadrature*”

- A “centered” and “scaled” version of Gauss-Hermite quadrature
- Requires much *fewer* abscissæ for *same accuracy*
- *Reference* – Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 4, 12–35.

Software: SAS proc nlmixed (quadrature and adaptive quadrature)

- Supports $Y_i|U_i, \theta_i \sim$ normal, binomial, Poisson, or allows user to write his/her own $p(y_i|u_i, a_i, b_i, \beta, \alpha)$
- Allows *variance models* for $\text{var}(Y_{ij}|U_i, \theta_i)$ but does *not support* fancy correlation structures (user must write)
- The b_i are *normal*
- *Experience* – often difficult to achieve convergence

Implementation

Stochastic approximations to integrals:

- *Obvious*: “Brute-force” *Monte Carlo integration* to do the integrals

$$\int p(y_i|u_i, a_i, b_i, \beta, \alpha)p(b_i|\zeta) db_i$$

⇒ *approximate* by

$$B^{-1} \sum_{b=1}^B p(y_i|u_i, a_i, b_i^{(b)}; \beta, \alpha)$$

where $b_i^{(b)}$, $b = 1, \dots, B$, are draws from the assumed $p(b_i|\zeta)$, e.g., $\mathcal{N}_p(0, D)$

- *Alternative*: “*Importance sampling*”
- *Problem*: Too *computationally intensive* to be practical; need *very large* B for adequate approximation to integrals

Implementation

Analytic approximations: Early, try to *avoid* “doing” the integral

“First-order” approximation: Approximate about $b_i = 0$

$$\begin{aligned} Y_i &= f_i(U_i, A_i, \beta, b_i) + \sigma^2 \Lambda_i^{1/2}(U_i, A_i, \beta, b_i, \alpha) \delta_i \\ &\approx f_i(U_i, A_i, \beta, 0) + Z_i(U_i, A_i, \beta, 0) b_i + \sigma^2 \Lambda_i^{1/2}(U_i, A_i, \beta, 0, \alpha) \delta_i, \\ Z_i(U_i, A_i, \beta, b_i) &= \partial / \partial b_i f_i(U_i, A_i, \beta, b_i) \end{aligned}$$

(*ignore* quadratic, crossproduct terms)

- \Rightarrow Using $b_i \sim \mathcal{N}_p(0, D)$ and $\delta_i \sim \mathcal{N}_{n_i}(0, I_{n_i}) \Rightarrow$ approximate

$$Y_i | U_i, A_i \sim \mathcal{N}_{n_i} \{ f_i(U_i, A_i, \beta, 0), Z_i(U_i, A_i, \beta, 0) D Z_i^T(U_i, A_i, \beta, 0) + \sigma^2 \Lambda_i^{1/2}(U_i, A_i, \beta, 0, \alpha) \}$$

- Under this *approximation*, may write $p(y_i | u_i, a_i, \alpha, \beta, \zeta)$ in a closed form

$$p(y, a | u, \alpha, \beta, \zeta) = \prod_{i=1}^m p(y_i, a_i | u_i, \alpha, \beta, \zeta) = \prod_{i=1}^m p(y_i | u_i, a_i, \alpha, \beta, \zeta) p(a_i)$$

(*ignore* $p(a_i)$ as before)

Implementation

Result: Implies *approximate likelihood function* $L_{\text{FOapprox}}(\psi|y, w)$

- Treat approximation as *exact* and *maximize* in ψ
- *Approximate sampling distribution* as if this were exact
- Is still a *nasty optimization* problem (still need forward solution for each subject at each internal iteration)
- If *interindividual variation* (variation across θ_i) is “*large*” (i.e., diagonal elements of D “*large*”), is a *lousy approximation* \Rightarrow leads to *biased estimator*

Implementation

Software:

- NONMEM
(<http://www.iconclinical.com/technology/products/nonmem/>)
Most popular way to implement among *pharmacokineticists*
- SAS proc nlmixed also does this
- A SAS macro, nlinmix
(<http://support.sas.com/kb/25/032.html>) does something related (analogous to the difference between *quadratic* and *linear* estimating equations discussed earlier)
- Other software in PK and statistical communities

Implementation

Better analytic approximation: Based on *Laplace's approximation* to integral of form $\int e^{-n_i \ell_i(b_i)} db_i$ for n_i "large"

$$\int e^{-n_i \ell_i(b_i)} db_i \approx (2\pi/n_i)^{p/2} |-\ell_{bb,i}(\hat{b}_i)|^{-1/2} e^{-n_i \ell_i(\hat{b}_i)}, \quad \hat{b}_i = \arg \min_{b_i} \ell_i(b_i)$$

- Under *normality* of both ϵ_i and b_i ,

$$\ell_i(b_i) = (1/2)\sigma^2 n_i^{-1} \left[\log |\Lambda_i(u_i, a_i, b_i, \beta, \alpha)| + B_i^T D^{-1} b_i + \sigma^{-2} \{y_i - f_i(u_i, a_i, \beta, b_i)\}^T \Lambda_i^{-1}(u_i, a_i, b_i, \beta, \alpha) \{y_i - f_i(u_i, a_i, \beta, b_i)\} \right]$$

- Can use this with *further approximation* to show that $Y_i|U_i, A_i$ is approximately n_i -variate normal with

$$E(Y_i|U_i, A_i) \approx f_i(U_i, A_i, \beta, \hat{b}_i) - Z_i(U_i, A_i, \beta, \hat{b}_i) \hat{b}_i$$

$$\text{var}(Y_i|U_i, A_i) \approx Z_i(U_i, A_i, \beta, \hat{b}_i) D Z_i^T(U_i, A_i, \beta, \hat{b}_i) + \sigma^2 \Lambda_i(U_i, A_i, \hat{b}_i, \beta, \alpha)$$

- Such approximations often referred to as *conditional* because \hat{b}_i maximizes the *conditional pdf* $p(b_i|y_i, u_i, a_i, \psi)$ (the *posterior pdf* of b_i)

Implementation

Result: Implies *iterative scheme*

1. *Maximize* the implied *approximate likelihood function*

$L_{C_{\text{approx}}}(\psi|y, a)$ to obtain $\hat{\psi}$ holding \hat{b}_i fixed

2. *Update* \hat{b}_i by maximizing the *posterior pdf* with $\hat{\psi}$ held fixed

Continue until “*convergence*,” approximate *sampling distribution* for final $\hat{\psi}$ treating as exact with final \hat{b}_i fixed

Remarks: Often *better approximation* than FO but computationally *more complex*

Implementation

Software:

- NONMEM
(<http://www.iconclinical.com/technology/products/nonmem/>)
does a fancier version of this
- SAS proc nlmixed does the same fancier version
- R/Splus nlme() function and SAS macro nlinmix
(<http://support.sas.com/kb/25/032.html>) do something
similar (*linear* rather than *quadratic* estimating equations)

Implementation

Different approximation: When *sufficient* number of observations are available on each subject to estimate θ_i *individually*

- Obtain *individual-specific* estimates $\hat{\theta}_i$, $i = 1, \dots, m$ using whatever method is most appropriate (OLS, GLS, etc)
- Use as “*data*” for inference on β and ζ that characterize the *population distribution* of the θ_i
- Must take into account that the $\hat{\theta}_i$ *are not* the true θ_i but are only *estimators* of them

To illustrate: Assume *Stage 2* model

$$\theta_i = \mathcal{A}_i\beta + b_i$$

as on slide 21, with the assumption $b_i \sim \mathcal{N}_p(0, D)$

Implementation

Idea: Use the *large sample approximation* to the distribution of $\hat{\theta}_i$

- For “*large*” n_i

$$\hat{\theta}_i | U_i, \theta_i \sim \mathcal{N}_p(\theta_i, C_i)$$

for a matrix C_i depending on θ_i and α

- Estimate C_i by \hat{C}_i obtained by *substituting* $\hat{\theta}_i$ (if α unknown substitute an estimate)

- Further approximation $\hat{\theta}_i | U_i, \theta_i \sim \mathcal{N}_p(\theta_i, \hat{C}_i)$

- Can write *equivalently* as

$$\hat{\theta}_i \approx \theta_i + e_i, \quad e_i \sim \mathcal{N}_p(0, \hat{C}_i)$$

- *Substitute* the *population model* $\theta_i = \mathcal{A}_i\beta + b_i$

$$\hat{\theta}_i \approx \mathcal{A}_i\beta + b_i + e_i, \quad b_i \sim \mathcal{N}_p(0, D), \quad e_i \sim \mathcal{N}_p(0, \hat{C}_i)$$

Implementation

$$\hat{\theta}_i \approx A_i \beta + b_i + e_i, \quad b_i \sim N_p(0, D), \quad e_i \sim N_p(0, \hat{C}_i)$$

Result: This is in the form of a so-called *linear mixed effects model* for which *standard software* is available for fitting

- Can estimate β and $\zeta = \text{vech}(D)$ using such software
- Alternatively, can fit this “model” using an *EM algorithm* referred to in the PK literature as the *global two stage* (GTS) algorithm

Implementation

GTS algorithm: At iteration $c + 1$, for $i = 1, \dots, m$

E-step: Produce “*refined*” estimates

$$\tilde{\theta}_i^{(c+1)} = (\hat{C}_i^{-1} + \hat{D}_{(c)}^{-1})^{-1} \hat{C}_i^{-1} \hat{\theta}_i + \hat{D}_{(c)}^{-1} \mathcal{A}_i \hat{\theta}^{(c)}$$

M-step: Obtain updated estimates

$$\hat{\theta}^{(c+1)} = \sum_{i=1}^m W_i^{(c)} \tilde{\theta}_i^{(c+1)}, \quad W_i^{(c)} = \left(\sum_{i=1}^m \mathcal{A}_i^T \hat{D}_{(c)}^{-1} \mathcal{A}_i \right)^{-1} \mathcal{A}_i^T \hat{D}_{(c)}^{-1}$$

$$\hat{D}_{(c+1)}^{-1} = m^{-1} \sum_{i=1}^m (\hat{C}_i^{-1} + \hat{D}_{(c)}^{-1})^{-1}$$

$$+ m^{-1} \sum_{i=1}^m (\tilde{\theta}_i^{(c+1)} - \mathcal{A}_i \hat{\theta}^{(c+1)}) (\tilde{\theta}_i^{(c+1)} - \mathcal{A}_i \hat{\theta}^{(c+1)})^T$$

Implementation

Remarks:

- This method may be shown to be just a different way to achieve an *analytical approximation* to the integrals
- If solving the inverse problem is “do-able” for each individual, can be easier to implement than the *linearization* methods
- Can be easier to explain to *non-quantitative* collaborators
- If the n_i are not too *small*, generally gives results similar to those from the *first order conditional* approximation
- *Issue:* The *approximate sampling covariance matrices* \hat{C}_i need to all be *non-singular* for $i = 1, \dots, m$

Model extensions

Leap of faith: The θ_i (and equivalently, the b_i) are *unobservable*, *latent* model components

- Accordingly, there are no “*direct*” data that can inform the assumption on the *population model*
- Obtaining *individual-specific estimates* $\hat{\theta}_i$ and using this *sample* to gain insight into the nature of this distribution can be *misleading* (e.g., *histograms* for each component)
- The $\hat{\theta}_i$ are subject to *uncertainty*, and if $p = \dim(\theta_i)$ is “*large*,” this is very difficult (*covariance structure*?)
- Why should the distribution of such latent quantities be *normal*?

Model extensions

Instead: Make *no* or only *mild* assumption on distribution of θ_i or b_i

- *Least restrictive*: Make *no assumption* on the *joint distribution* of (θ_i, A_i)
 - *any* probability distribution is a viable candidate
- If there are no *individual covariates*, this boils down to allowing the possibility that the distribution of θ_i could be *anything* – including *discrete*, exhibit *pathological behavior* etc.
- Would take θ_i to have *unspecified cdf* $F(\theta_i)$, say
- The marginal pdf of interest on slide 18 becomes

$$p(y|u, \alpha) = \prod_{i=1}^m \int p(y_i|u_i, \theta_i, \alpha) dF(\theta_i)$$

- $\beta = E(\theta_i)$ and $D = \text{var}(\theta_i)$ under $F(\cdot)$
- *Implementation*: Estimate $F(\theta_i)$ *jointly* with α by ML \Rightarrow *discrete estimator*

Model extensions

Result:

- $\beta = E(\theta_i)$ and $D = \text{var}(\theta_i)$ under $F(\cdot)$
- *Implementation*: Estimate $F(\theta_i)$ *jointly* with $\alpha \Rightarrow$ *discrete estimator* $\hat{F}(\theta_i)$
- $\hat{\beta}$ and \hat{D} are moments of $\hat{F}(\theta_i)$
- *Analogously*: Given an assumed *population model* $\theta_i = \mathcal{A}_i\beta + b_i$, can assume only that b_i have *unspecified cdf* $F(b_i)$ with $E(b_i) = 0$

Model extensions

Alternatively: The phenomena represented by components of θ_i likely have *continuous distributions*

- By placing *no restrictions* on $F(\theta_i)$ or Fb_i , are including *infeasible*
- \Rightarrow Assume distributions of θ_i or b_i have a “*smooth*” *pdf* (exclude pdfs with “*weird*” behavior)
- Illustrate with $b_i \dots$

Model extensions

Assume: b_i have a “*smooth*” *density* $p(b_i)$ that can be represented by a *flexible form* involving *parameters* δ ; $p(b_i|\delta)$

- The marginal pdf of interest on slide 27 becomes

$$L(\psi|y, a) = \prod_{i=1}^m \int p(y_i|u_i, a_i, b_i; \beta, \alpha) p(b_i|\delta) db_i$$

- Popular representations for $p(b_i)$ include *mixtures of normals* and the “*semiparametric*” (SNP) series expansion used by Davidian and Gallant (1993, *Biometrika*)
- *Degree of flexibility* required, e.g., number of mixture components, terms in series expansion is a “*tuning parameter*” whose choice is *data-driven*

Model extensions

For example: The SNP approach assumes $p(b_i)$ is in a *class* of densities that can be represented by an *infinite Hermite series expansion*

$$p(b) = \{P_\infty(L^{-1}b)\}^2 n_p(b|0, LL^T)$$

where $n_p(b|0, D)$ is p -variate normal density, L ($p \times p$) is upper triangular, and $P_\infty(z)$ is infinite dimensional polynomial

- E.g., $p = 2$ so $z = (z_1, z_2)^T$

$$P_\infty(z) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{11}z_1z_2 + a_{20}z_1^2 + a_{02}z_2^2 \\ + a_{30}z_1^3 + a_{103}z_2^3 + a_{12}z_1z_2^2 + a_{21}z_1^2z_2 + \dots$$

- *Truncate* at degree K and represent as $P_K(z)$; K is the *tuning parameter* to be determined from the data
- Must constrain $\int \{P_K(L^{-1}b)\}^2 n_p(b|0, LL^T) db = 1$
- Require $K \rightarrow \infty$ with $m \rightarrow \infty$ for properties

Model extensions

Implementation: ML with δ estimated along with β, α

Usefulness: In addition to providing a more *faithful characterization*, can use for *model selection*

- If the estimated $p(b_i)$ exhibits *multimodality* may indicate an important *individual covariate* has been *excluded* from the *population model*
- \Rightarrow Obtain *estimated posterior modes* \hat{b}_i and *plot* against covariates

Model extensions

Other extensions: Numerous other modifications of the basic model we have discussed here have been proposed and studied

- *Censored observations*: If some observations of responses are *censored* by a *limit of quantification*, the *observed data* for subject i become, e.g., using the previous notation when there is a *lower limit*

$$(Z_i, \Delta_i) = \{(Z_{i1}, \Delta_{i1}), \dots, (Z_{in_i}, \Delta_{i,n_i})\}$$

where $Z_{ij} = \max(Y_{ij}, Q_{ij})$, $\Delta_{ij} = I(Y_{ij} > Q_{ij})$, and Q_{ij} is the lower limit for subject i at t_{ij} .

- Deduce the *observed data pdf* $p(z_i, \delta_i | u_i, a_i, b_i, \beta, \alpha)$ from the *full data pdf* $p(y_i | u_i, a_i, b_i, \beta, \alpha)$
- Likelihood based on observed data

$$L(\psi | z, \delta, a) = \prod_{i=1}^m \int p(z_i, \delta_i | u_i, a_i, b_i, \beta, \alpha) p(b_i | \zeta) db_i, \quad \psi = (\alpha^T, \beta^T, \zeta^T)^T$$

Model extensions

Other extensions: *Functional valued parameters*

- For some models $\dot{x}(t) = g\{t, x(t), \theta\}$, some components of θ may in fact depend on t , e.g.,

$$\theta(t) = \{\theta_1^T, \theta_2(t)\}$$

for $\theta_2(t)$ an unspecified real-valued (or vector-valued) function of time

- Represent $\theta_2(t)$ by some approximation (e.g., via some chosen *spline basis*)
- Incorporate in likelihood

Model extensions

Other extensions: *Multi-level models*

- Observations may have a *nested* structure; e.g., in *forestry applications*, observations are made on each *tree* within a *plot* or *stand*, and the math model describes the evolution of *growth* on an *individual tree*
- Trees in the same plot/stand may tend to be *more alike* than those from disparate plots/stands (*share* common features)
- Modified *population model*: Plots/stands $i = 1, \dots, m$, *Trees within plots* $k = 1, \dots, s_i$, Observations on tree (i, k) , $j = 1, \dots, n_{ik}$

$$\theta_{ik} = \mathcal{A}_i\beta + b_i + b_{ik}$$

- b_i is a mean zero random effect pertaining to plot i
- b_{ik} is a mean zero random effect pertaining to tree (i, k) – how tree k deviates from the *conditional mean* $\mathcal{A}_i\beta + b_i$ for plot i
- \mathcal{A}_i contains elements of *plot-specific* characteristics \mathcal{A}_i

Model extensions

Other extensions: “*Inter-occasion variation*” (relevant in *pharmacokinetics*)

- Subjects are observed over more than one “*occasion*;” e.g., dosing intervals, weeks, etc.
- Subject *characteristics* are ascertained during each “occasion;” e.g., enzyme levels, weight, renal function, etc.
- *Perspective*: PK parameters θ_i may themselves *fluctuate* within an individual over occasions, which may be in part associated with *changing subject characteristics*
- Assume q “*occasion intervals*” \mathcal{I}_h , $h = 1, \dots, q$, where characteristics = A_{ih}

Model extensions

Other extensions: “*Inter-occasion variation*”

- Modified *population model*:

$$\theta_{ij} = \mathcal{A}_{ih}\beta + b_i,$$

where θ_{ij} is the value of θ_i at time $t_{ij} \in \mathcal{I}_h$

⇒ “*inter-occasion*” variation of θ_i *entirely attributable* to changes in \mathcal{A}_{ih}

- Or

$$\theta_{ij} = \mathcal{A}_{ih}\beta + b_i + b_{ih}$$

Discussion

- Fitting hierarchical nonlinear models using *approximate methods* is *commonplace*, e.g., routine in PK (“*population pharmacokinetics*”)
- Can involve *computational challenges*, e.g. *starting values*, *local maxima*, etc.
- With “*high quality*” data, is often fruitful
- *Bayesian formulation* also common, implemented via *Markov chain Monte Carlo* techniques
- How to *design studies* recognizing that different individuals have different *parameters*?
- Hierarchical nonlinear models are the foundation for *modeling and simulation* efforts to be discussed shortly