Published in final edited form as:
Comput Stat Data Anal. 2007 August 15; 51(12): 6614–6623. doi:10.1016/j.csda.2007.03.008.

Nonlinear Random Effects Mixture Models: Maximum Likelihood Estimation via the EM Algorithm

Xiaoning Wang[a], **Alan Schumitzky**[b], and **David Z. D'Argenio**[a,*]

[a] Department of Biomedical Engineering, University of Southern California, Los Angeles, CA 90089, USA

[b] Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA

[*] Department of Biomedical Engineering, University of Southern California, 1042 Downey Way, DRB 140, Los Angeles, CA 90089, voice, (213) 740-0341, fax, (213) 740-0343, email: dargenio@bmsr.usc.edu

Abstract

Nonlinear random effects models with finite mixture structures are used to identify polymorphism in pharmacokinetic/pharmacodynamic phenotypes. An EM algorithm for maximum likelihood estimation approach is developed and uses sampling-based methods to implement the expectation step, that results in an analytically tractable maximization step. A benefit of the approach is that no model linearization is performed and the estimation precision can be arbitrarily controlled by the sampling process. A detailed simulation study illustrates the feasibility of the estimation approach and evaluates its performance. Applications of the proposed nonlinear random effects mixture model approach to other population pharmacokinetic/pharmacodynamic problems will be of interest for future investigation.

Keywords

Finite mixture models; Mixed effects models; Pharmacokinetics/pharmacodynamics

1. Introduction

There is substantial variability in the way individuals respond to medications, in both treatment efficacy and toxicity. The sources of a drug's underlying pharmacokinetic and pharmacodynamic variability can include demographic factors (such as age, sex, weight), physiological status (such as renal, liver, cardiovascular function), disease states, genetic differences, interactions with other drugs and environmental factors. In their seminal work, Sheiner, Rosenberg and Melmon (1972) proposed a parametric nonlinear mixed-effects modeling framework for quantifying both within and between subject variability in a drug's pharmacokinetics, and developed an approximate maximum likelihood solution to the problem. Since the introduction by Beal and Sheiner (1979) of the general purpose software package NONMEM implementing this approach, other approximate maximum likelihood algorithms have been introduced to solve the nonlinear random and mixed effects modeling problem (see Davidian and Giltinan (1995) for an extensive review). An exact maximum

likelihood (i.e., no linearization) solution to the parametric population modeling problem based on the EM algorithm has also been proposed by Schumitzky (1995) and fully developed and implemented by Walker (1996). The population modeling framework has had a significant impact on how pharmacokinetic (and pharmacodynamic) variability is quantified and studied during drug development, and on the identification of important covariates associated with a drug's inter-individual kinetic/dynamic variability.

While population models incorporating measured covariates have proven to be useful in drug development, it is recognized that genetic polymorphisms in drug metabolism and in the molecular targets of drug therapy, for example, can also have a significance influence on the efficacy and toxicity of medications (Evans and Relling, 1999). There is, therefore, a need for population modeling approaches that can extract and model important subpopulations using pharmacokinetic/pharmacodynamic data collected in the course of drug development trials and other clinical studies, in order to help identify otherwise unknown genetic determinants of observed pharmacokinetic/pharmacodynamic phenotypes. The nonparametric maximum likelihood approach for nonlinear random effects modeling developed by Mallet (1986), as well as the nonparametric Bayesian approaches of Wakefield and Walker (1997) and Rosner and Mueller (1997), and the smoothed nonparametric maximum likelihood method of Davidian and Gallant (1993) all address this important problem. In this paper we propose a parametric approach using finite mixture models to identify subpopulations with distinct pharmacokinetic/pharmacodynamic properties.

An EM algorithm for exact maximum likelihood estimation of nonlinear random effects finite mixture models is introduced, extending the previous work of Schumitzky (1995) and Walker (1996). The EM algorithm has been used extensively for linear mixture model applications (see McLachlan and Peel (2000) for a review). The algorithm for nonlinear mixture models presented below has an analytically tractable M step, and uses sampling-based methods to implement the E step. Section 2 of this paper describes the finite mixture model within a nonlinear random effects modeling framework. Section 3 gives the EM algorithm for the maximum likelihood estimation of the model. Section 4 addresses individual subject classification, while an error analysis is presented in section 5. A detailed simulation study of a pharmacokinetic model is presented in section 6. Section 7 contains a discussion.

2. Nonlinear Random Effects Finite Mixture Models

A two-stage nonlinear random effects model that incorporates a finite mixture model is given by

$$\boldsymbol{Y}_i|\boldsymbol{\theta}_i,\boldsymbol{\beta} \sim N(\mathbf{h}_i(\boldsymbol{\theta}_i), \mathbf{G}_i(\boldsymbol{\theta}_i,\boldsymbol{\beta})), \quad i=1,\ldots,n \tag{1}$$

and

$$\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_n \sim_{i.i.d} \sum_{k=1}^{K} w_k N(\boldsymbol{\mu}_k, \sum\nolimits_k), \tag{2}$$

where $i=1,\ldots,n$ indexes the individuals and $k=1,\ldots,K$ indexes the mixing components.

At the first stage represented by (1), $\mathbf{Y}_i = (y_{1i},\ldots,y_{m_i i})^T$ is the observation vector for the ith individual ($\mathbf{Y}_i \in R^{m_i}$); $\mathbf{h}_i(\boldsymbol{\theta}_i)$ is the function defining the pharmacokinetic/pharmacodynamic (PK/PD) model, including subject specific variables (e.g., drug doses), and $\boldsymbol{\theta}_i$ is the vector of

model parameters (random effects) ($\boldsymbol{\theta}_i \in R^p$). In (1) $\mathbf{G}_i$ ($\boldsymbol{\theta}_i$, $\boldsymbol{\beta}$) is a positive definite covariance matrix ($\mathbf{G}_i \in R^{m^i \times m^i}$) that may depend upon $\boldsymbol{\theta}_i$ as well as on other parameters $\boldsymbol{\beta}$ (fixed effects) ($\boldsymbol{\beta} \in R^q$).

At the second stage given by (2), a finite mixture model with K multivariate normal components is used to describe the population distribution. The weights $\{w_k\}$are nonnegative numbers summing to one, denoting the relative size of each mixing component (subpopulation), for which $\boldsymbol{\mu}_k$ ($\boldsymbol{\mu}_k \in R^p$) is the mean vector and $\boldsymbol{\Sigma}_k$ ($\boldsymbol{\Sigma}_k \in R^{p\times p}$) is the positive definite covariance matrix.

Letting φ represent the collection of parameters, $\{\boldsymbol{\beta},(w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),k = 1,\ldots,K\}$, the population problem involves estimating φ given the observation data$\{\mathbf{Y}_1,\ldots,\mathbf{Y}_n\}$. The maximum likelihood estimate (MLE) can be obtained by maximizing the overall data likelihood L with respect to φ. Under the *i.i.d.* assumption of the individual parameters $\{\boldsymbol{\theta}_i\}$, L is given by the expression

$$L(\varphi)=\prod_{i=1}^{n}\int p(\boldsymbol{Y}_i|\boldsymbol{\theta}_i,\boldsymbol{\beta})\sum_{k=1}^{K} w_k p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k, \sum_k)d\boldsymbol{\theta}_i. \tag{3}$$

The MLE of φ is defined as φ_{ML} with $L(\varphi_{ML}) \geq L(\varphi)$ for all φ in the parameter space.

3. Solution via the EM Algorithm

The EM algorithm, originally introduced by Dempster, Laird and Rubin (1977), is a widely applicable approach to the iterative computation of MLEs. It was used by Schumitzky (1995) and Walker (1996) to solve the nonlinear random effects maximum likelihood problem for a second stage model consisting of a single normal distribution. The EM algorithm is typically formulated in terms of “complete” versus “missing” data structure. Consider the model given by (1) and (2) for the important case

$$\mathbf{G}_i(\boldsymbol{\theta}_i,\boldsymbol{\beta})=\sigma^2\mathbf{H}_i(\boldsymbol{\theta}_i) \tag{4}$$

where $\mathbf{H}_i$ ($\boldsymbol{\theta}_i$) is a known function and $\boldsymbol{\beta} = \sigma^2$. The component label vector $\mathbf{z}_i$ is introduced as a K dimensional indicator such that z_i (k) is one or zero depending on whether or not the parameter $\boldsymbol{\theta}_i$ arises from the kth mixing component. The individual subject parameters ($\boldsymbol{\theta}_1$, $\ldots,\boldsymbol{\theta}_n$) are regarded as unobserved random variables. The “complete” data is then represented by $\mathbf{Y}_c = \{(\mathbf{Y}_i, \boldsymbol{\theta}_i, \mathbf{z}_i),i = 1,\ldots,n\}$ with $\{\boldsymbol{\theta}_i, \mathbf{z}_i\}$ representing the “missing” data.

The algorithm starts with $\varphi^{(0)}$ and moves from $\varphi^{(r)}$ to $\varphi^{(r+1)}$ at the rth iteration. At the E-step, define

$$Q(\varphi, \varphi^{(r)})=E\{\log L_c(\varphi)|\boldsymbol{Y}, \varphi^{(r)}\},$$

where log L_c (φ) is the complete data likelihood given by

$$\log L_c(\varphi)=\sum_{i=1}^{n}\sum_{k=1}^{K} z_i(k)\log p(\boldsymbol{Y}_i,\boldsymbol{\theta}_i|\sigma^2,\boldsymbol{\mu}_k,\textstyle\sum_k). \tag{5}$$

Now

$$E\{z_i(k)|\boldsymbol{Y},\varphi\}=pr\{z_i(k)=1|\boldsymbol{Y},\varphi\}=\tau_i(k),$$

and by Bayes' Theorem,

$$\tau_i(k)=\frac{w_k p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\mu}_k,\textstyle\sum_k)}{\sum_{k=1}^{K} w_k p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\mu}_k,\textstyle\sum_k)}=\frac{w_k\int p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\textstyle\sum_k)d\boldsymbol{\theta}_i}{\sum_{k=1}^{K} w_k\int p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\textstyle\sum_k)d\boldsymbol{\theta}_i}.$$

Introducing the notation

$$g_{ik}(\boldsymbol{\theta}_i,\varphi)=\frac{w_k p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\textstyle\sum_k)}{\sum_{k=1}^{K} w_k\int p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\textstyle\sum_k)d\boldsymbol{\theta}_i},$$

then

$$\tau_i(k)=\int g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i.$$

The expected value of (5) is given by

$$Q(\varphi,\varphi^{(r)})=\sum_{i=1}^{n}\int g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})\log p(\boldsymbol{Y}_i,\boldsymbol{\theta}_i|\sigma^2,\boldsymbol{\mu}_k,\textstyle\sum_k)d\boldsymbol{\theta}_i,$$

where

$$\log p(\boldsymbol{Y}_i,\boldsymbol{\theta}_i|\sigma^2,\boldsymbol{\mu}_k,\textstyle\sum_k)=C-\frac{m_i}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{Y}_i-\mathbf{h}_i(\boldsymbol{\theta}_i))^T\mathbf{H}_i(\boldsymbol{\theta}_i)^{-1}(\boldsymbol{Y}_i-\mathbf{h}_i(\boldsymbol{\theta}_i))-\frac{1}{2}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)^T\textstyle\sum_k^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)-\frac{1}{2}\log|\sum_k|$$

for some constant C.

The M-step takes $\varphi^{(r)}\rightarrow\varphi^{(r+1)}$ where $\varphi^{(r+1)}$ is the unique optimizer of $Q(\varphi,\varphi^{(r)})$ such that $\varphi^{(r+1)}=\arg\max_{\varphi} Q(\varphi,\varphi^{(r)})$. Let $\varphi'=\{\boldsymbol{\beta},(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),k=1,\ldots,K\}$, then the optimizer of $Q(\varphi,\varphi^{(r)})$

relative to φ' occurs at interior points, and the corresponding components of $\varphi^{(r+1)}$ are the unique solution to

$$\frac{\partial}{\partial \varphi'} Q(\varphi, \varphi^{(r)})|_{\varphi^{(r+1)}} = 0. \tag{6}$$

From the expression of $\log p(\mathbf{Y}_i, \boldsymbol{\theta}_i \mid \sigma^2, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \log L_c(\varphi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i(k) \{ (\sum\nolimits_k)^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_k) \},$$

so

$$E\{ \frac{\partial}{\partial \boldsymbol{\mu}_k} \log L_c(\varphi) | \boldsymbol{Y}, \varphi^{(r)} \} = \sum_{i=1}^{n} \int g_{ik}(\boldsymbol{\theta}_i, \varphi^{(r)}) \{ \sum\nolimits_k{}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_k) \} d\boldsymbol{\theta}_i.$$

Also,

$$\frac{\partial}{\partial \sum_k} \log L_c(\varphi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i(k) \{ \frac{1}{2} \sum\nolimits_k{}^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_i - \boldsymbol{\mu}_k)^T \sum\nolimits_k{}^{-1} - \frac{1}{2} \sum\nolimits_k{}^{-1} \}$$

and

$$\frac{\partial}{\partial (\sigma^2)} \log L_c(\varphi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i(k) \{ -\frac{m_i}{2} \frac{1}{\sigma^2} + \frac{1}{2} \frac{(\boldsymbol{Y}_i - \mathbf{h}_i(\boldsymbol{\theta}_i))^T \mathbf{H}_i(\boldsymbol{\theta}_i)^{-1} (\boldsymbol{Y}_i - \mathbf{h}_i(\boldsymbol{\theta}_i))}{\sigma^4} \}.$$

The unique solution of (6) is thus given by

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n} \int \boldsymbol{\theta}_i g_{ik}(\boldsymbol{\theta}_i, \varphi^{(r)}) d\boldsymbol{\theta}_i}{\sum_{i=1}^{n} \int g_{ik}(\boldsymbol{\theta}_i, \varphi^{(r)}) d\boldsymbol{\theta}_i}, \tag{7}$$

$$\sum\nolimits_k^{(r+1)} = \frac{\sum_{i=1}^{n} \int (\boldsymbol{\theta}_i - \boldsymbol{\mu}_k{}^{(r+1)})(\boldsymbol{\theta}_i - \boldsymbol{\mu}_k{}^{(r+1)})^T g_{ik}(\boldsymbol{\theta}_i, \varphi^{(r)}) d\boldsymbol{\theta}_i}{\sum_{i=1}^{n} \int g_{ik}(\boldsymbol{\theta}_i, \varphi^{(r)}) d\boldsymbol{\theta}_i}, \tag{8}$$

and

$$(\sigma^2)^{(r+1)}=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\int(\boldsymbol{Y}_i-\mathbf{h}_i(\boldsymbol{\theta}_i))^T\mathbf{H}_i(\boldsymbol{\theta}_i)^{-1}(\boldsymbol{Y}_i-\mathbf{h}_i(\boldsymbol{\theta}_i))g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i}{\sum_{i=1}^{n}m_i}. \tag{9}$$

The updated estimates $\{w_k^{(r+1)}\}$ are calculated independently. If $\mathbf{z}_i$ were observable, then the MLE of w_k would be $\widehat{w}_k=\frac{1}{n}\sum_{i=1}^{n}z_i(k)$. By replacing each $\mathbf{z}_i$ by its conditional expectation from the E step, the updating for w_k is given by (see McLachlan and Peel, 2000):

$$w_k{}^{(r+1)}=\frac{1}{n}\sum_{i=1}^{n}\tau_i(k)=\frac{1}{n}\int g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i. \tag{10}$$

Dempster et al. (1977) showed that the resulting sequence $\{\varphi^{(r+1)}\}$ has the likelihood improving property $L(\varphi^{(r+1)})\geq L(\varphi^{(r)})$. It can be shown that the above updates are well-defined, that is for all $1\leq k\leq K$, if $w_k^{(0)}>0$ then $w_k^{(r+1)}>0$ so that $\sum_{i=1}^{n}\int g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i>0$ and $\sum_k^{(r+1)}$ are positive definite. Wu (1983) and Tseng (2005) gave the sufficient conditions for the convergence of $\varphi^{(r)}$ to a stationary point of the likelihood function $L(\varphi)$. A number of starting positions are suggested, however, in an effort to ensure convergence to a global maximum.

In order to implement the algorithm all the integrals in (7)–(10) must be evaluated at each iterative step. For the non-mixture problem involving a relatively simple pharmacokinetic model, Walker (1996) proposed Monte Carlo integration to evaluate the required integrals. We and others (Ng et al., 2005) have found that importance sampling is preferable to the Monte Carlo integration for approximating the integrals in the EM algorithm for a number of representative models of interest in PK/PD. We have also applied importance sampling to the current mixture model.

We note that all the integrals above have the following form

$$\int f(\boldsymbol{\theta}_i)g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i=\frac{\int f(\boldsymbol{\theta}_i)w_k p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\sum_k)d\boldsymbol{\theta}_i}{\sum_{k=1}^{K}\int w_k p(\boldsymbol{Y}_i|\sigma^2,\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_k,\sum_k)d\boldsymbol{\theta}_i}.$$

For each mixing component, a numbers of samples are taken from an envelope distribution, $\boldsymbol{\theta}_{i(k)}^{(1)},\ldots,\boldsymbol{\theta}_{i(k)}^{(T)}\sim_{i.i.d.}p_{e(k)}(\boldsymbol{\theta}_i)$, and used to approximate the integrals as follows

$$\int f(\boldsymbol{\theta}_i) g_{ik}(\boldsymbol{\theta}_i, \varphi) d\boldsymbol{\theta}_i \cong \frac{\sum_{l=1}^{T} f\left(\boldsymbol{\theta}_{i(k)}^{(l)}\right) w_k p(\mathbf{Y}_i|\sigma^2, \boldsymbol{\theta}_{i(k)}^{(l)}) p(\boldsymbol{\theta}_{i(k)}^{(l)}|\boldsymbol{\mu}_k, \Sigma_k) / p_{e(k)}(\boldsymbol{\theta}_{i(k)}^{(l)})}{\sum_{k=1}^{K}\sum_{l=1}^{T} w_k p(\mathbf{Y}_i|\sigma^2, \boldsymbol{\theta}_{i(k)}^{(l)}) p(\boldsymbol{\theta}_{i(k)}^{(l)}|\boldsymbol{\mu}_k, \Sigma_k) / p_{e(k)}(\boldsymbol{\theta}_{i(k)}^{(l)})}. \tag{11}$$

For each mixing component for each subject, the envelope distribution is taken to be a multivariate normal density using the subject's previously estimated conditional mean and conditional covariance as its mean and covariance. Therefore all the random samples are independent and individual specific. For details of the importance sampling approach in general, see Geweke (1989). The number of independent samples, *T*, will depend on the complexity of the model and the required accuracy in the integral approximations.

4. Classification of Subjects

It is of interest to assign each individual subject to a subpopulation. Such a classification will allow further investigation into the genetic basis of any identified PK/PD polymorphism. The quantity $\tau_i(k) = E\{z_i(k) \mid \mathbf{Y}, \varphi\}$ in the E step is the posterior probability that the *i*th individual belongs to the *k*th mixing component. The classification involves assigning an individual to the subpopulation associated with the highest posterior probability of membership. For example, for each *i*, ($i = 1,\ldots,n$), set

$$\widehat{z_i}(k) = 1 \text{ if } k = \arg\max_c \tau_i(c),$$

or to zero otherwise. No additional computation is required since all the $\tau_i(k)$ are evaluated during each EM step.

5. Standard Errors

Assuming the regularity conditions from Philppou and Roussas (1975), it can be shown that asymptotically as $n \to \infty$,

$$Cov(\varphi_{ML}) \approx \left(\sum_{i=1}^{n} \mathbf{V}_i(\varphi_{ML})\right)^{-1},$$

where $\mathbf{V}_i(\varphi) = \left(\frac{\partial}{\partial\varphi}\log p(\mathbf{Y}_i|\varphi)\right)\left(\frac{\partial}{\partial\varphi}\log p(\mathbf{Y}_i|\varphi)\right)^T$.

Now

$$\frac{\partial}{\partial\varphi}\log p(\mathbf{Y}_i|\varphi) = \sum_{k=1}^{K}\int\left\{\frac{\partial}{\partial\varphi}\log p(\mathbf{Y}_i, \boldsymbol{\theta}_i|\varphi)\right\} g_{ik}(\boldsymbol{\theta}_i, \varphi) d\boldsymbol{\theta}_i,$$

and the gradient components are calculated for $k = 1,\ldots,K$ as

$$\mathbf{s}_{\boldsymbol{\mu}_k}=\frac{\partial}{\partial \boldsymbol{\mu}_k}\log p(\boldsymbol{Y}_i|\varphi)=\int g_{ik}(\boldsymbol{\theta}_i,\varphi)\{\textstyle\sum_k{}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)\}d\boldsymbol{\theta}_i,$$
$$\mathbf{s}_{\sum_k}=\frac{\partial}{\partial \sum_k}\log p(\boldsymbol{Y}_i|\varphi)=\int g_{ik}(\boldsymbol{\theta}_i,\varphi)\{(-\tfrac{1}{2})\textstyle\sum_k{}^{-1}+\frac{1}{2}\sum_k{}^{-1}(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)^T\sum_k{}^{-1}\}d\boldsymbol{\theta}_i,$$
$$\mathbf{s}_{\sigma^2}=\frac{\partial}{\partial \sigma^2}\log p(\boldsymbol{Y}_i|\varphi)=\sum_{k=1}^{K}\int g_{ik}(\boldsymbol{\theta}_i,\varphi)\{-\tfrac{1}{2}\frac{m_i}{\sigma^2}+\tfrac{1}{2}\frac{(\boldsymbol{Y}_i-\mathbf{h}_i(\theta_i))^T\mathbf{H}_i^{-1}(\theta_i)(\boldsymbol{Y}_i-\mathbf{h}_i(\theta_i))}{\sigma^4}\}d\boldsymbol{\theta}_i,$$

and for $k=1,\ldots,K-1$,

$$s_{w_k}=\frac{\partial}{\partial w_k}\log p(\boldsymbol{Y}_i|\varphi)=\frac{1}{w_k}\int g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i-\frac{1}{w_K}\int g_{iK}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i.$$

Introduce the notation $\mathbf{s}_{\boldsymbol{\varpi}k}=((\mathbf{s}_{\boldsymbol{\Sigma}k})_{1,1},(\mathbf{s}_{\boldsymbol{\Sigma}k})_{2,1}\ldots,(\mathbf{s}_{\boldsymbol{\Sigma}k})_{p,p})$, where $(\mathbf{s}_{\boldsymbol{\Sigma}k})_{i,j}$ is the component of the lower triangular part of $\mathbf{s}_{\boldsymbol{\Sigma}k}$ in the (i,j) position. Put these results together to produce the vector

$$\mathbf{s}_i=(\mathbf{s}_{\boldsymbol{\mu}_1},\ldots,\mathbf{s}_{\boldsymbol{\mu}_K},\mathbf{s}_{\boldsymbol{\varpi}_1},\ldots,\mathbf{s}_{\boldsymbol{\varpi}_K},s_{w_1},\ldots,s_{w_{K-1}},s_{\sigma^2}),$$

so

$$Cov(\varphi_{ML})=\Big(\sum_{i=1}^{n}\mathbf{s}_i\mathbf{s}_i{}^T\Big)^{-1}.$$

All the computations can be performed during the importance sampler calculation at the final iteration of the EM algorithm.

6. Example

In this section a simulation study is conducted to evaluate the proposed algorithm for calculating the exact MLEs for a population finite mixture model. A one compartment pharmacokinetic model is used, with the observations of plasma concentration given by

$$y_{ji}=\frac{D}{V_i}\exp(-k_i*t_j)(1+\varepsilon_{ji}),$$

where D is a bolus drug administration with 100 units of dose, V represents the volume of distribution and k is the elimination rate constant. For all the individuals, $m_i=5$ with $t_1=1.5$, $t_2=2$, $t_3=3$, $t_4=4$ and $t_5=5.5$. The within-individual error is assumed to be *i.i.d.* with variance 0.01. Data sets were simulated from this model for each of 100 subjects sampled from the following population model:

$$V_i\sim_{i.i.d}N(20,2^2),$$
$$k_i\sim_{i.i.d}0.8N(0.3,0.06^2)+0.2N(0.6,0.06^2),$$

where V_i and k_i are assumed to be independent. A total of 200 such population data sets were generated. This model represents the pharmacokinetics of a drug with an elimination that can be characterized by two distinct subpopulations.

The formulation of Section 2 has been modified to accommodate the important case where a subset of parameters are modeled by a multivariate normal distribution and the remaining parameters follow a mixture of normals (see modified updating formulas in the Appendix). The MLEs were obtained for each of the 200 population sets using the EM algorithm with importance sampling described above. For each of the estimated parameters φ, its percent prediction error was calculated for each population data set as:

$$\mathrm{pe}_j = 100 \times (\varphi_j^{\mathrm{ML}} - \varphi_j)/\varphi_j, \quad j=1,\ldots,200.$$

These percent prediction errors were used to calculate the mean prediction error and root mean square prediction error for each parameter. In addition, for each population data set, the calculated standard errors were used to construct 95% confidence intervals for all estimated parameters. The percent coverage of these confidence intervals was then tabulated over the 200 population data sets. Finally, the individual subject classification accuracy was evaluated for each population data set.

Figure 1 provides a graphical illustration of the results showing the true population distributions of V and k along with the estimated distribution obtained from the 200 simulated population data sets. Quantitative results are presented in Table 1, which gives mean and root mean square prediction errors (RMSE) as well as the percent coverage of the calculated confidence intervals for each of the estimated parameters. The parameter estimates, overall, match the population values and the percent coverage of confidence intervals is reasonable. The estimates of population variance have relatively greater biases and RMSE. Over the 200 population data sets, on average 1.54 out of 100 subjects were classified in the wrong subpopulation. The largest number of subjects misclassified was 4, while all the subjects were correctly classified in 83 of the 200 population data sets.

Central to the calculation of the MLEs is the computation of the integrals in (7)–(9), as approximated by importance sampling in our implementation. Using one of the 200 population data sets we examined the influence of the number of samples (T) used in the importance sampler, as well as the number of EM iterations required to achieve two digits of accuracy for each of the estimated parameters. Table 2 presents the parameter estimates from 50 EM iterations by using 1000, 2000 and 3000 samples in the important sampling. Accuracy to two digits was obtained with 1000 samples. Based on this experience, T was taken to be 1000 in this simulations study and 50 EM iterations were run on each data set.

Error! Reference source not found

demonstrates the convergence of log-likelihood values for a particular data set by starting the 50 EM iterations from 9 different positions. The log-likelihood values were approximated via Monte Carlo integration

$$\log\{L(\varphi)\} \cong \sum_{i=1}^{n} \log\{\frac{1}{T}\sum_{k=1}^{K}\sum_{l=1}^{T} w_k p(\boldsymbol{Y}_i|\sigma^2, \boldsymbol{\theta}_{i(k)}^{(l)}) p(\boldsymbol{\theta}_{i(k)}^{(l)}|\boldsymbol{\mu}_k, \textstyle\sum_k) / p_{e(k)}(\boldsymbol{\theta}_{i(k)}^{(l)})\},$$

where $\boldsymbol{\theta}_{i(k)}^{(1)}, \ldots, \boldsymbol{\theta}_{i(k)}^{(T)} \sim_{i.i.d.} p_{e(k)}(\boldsymbol{\theta}_i)$. In any particular example, of course, the number of EM iterations, the number of samples T used to approximate integrals, as well as the use of different starting guess will depend on the experiment design and complexity of the model.

7. Discussion

In this paper, an EM algorithm for maximum likelihood estimation of nonlinear random effects mixture models is presented that has application in pharmacokinetic/pharmacodynamic population modeling studies. It extends the previous work on the use of the EM algorithm for MLE of nonlinear random effects models, to the case of finite mixture models, and reinforces the practicability of using exact (no linearizing approximation) MLE estimation in PK/PD modeling studies (see also, Kuhn and Lavielle (2005) for a stochastic EM variation). The parametric mixture model MLE approach presented also complements previous work on nonparametric Bayesian and smoothed nonparametric MLE, in addressing the increasingly important problem of identifying subpopulations with distinct PK/PD properties in drug development trials. We note that approximate maximum likelihood methods using mixture models are also available in NONMEM.

The EM algorithm has been used extensively for fitting linear mixture models in numerous applications in diverse fields of study. Even for linear problems involving mixture of normal components, a number of challenges attend the use of the EM algorithm for maximum likelihood estimation (McLachlan and Peel (2000)) that are also relevant to the nonlinear random effects problem. These include: potential unboundness of the likelihood for heteroscedastic covariance components; local maxima of the likelihood function; and choice of the number of mixing components. Application of the algorithm for nonlinear random effects mixture models presented here can be guided by the extensive work related to these issues for linear mixture modeling.

We have investigated the possible unboundness of the likelihood for the example considered in this paper. If, for example, in the first component of the mixture, $\boldsymbol{\mu}_1$ satisfies $\mathbf{h}_i(\boldsymbol{\mu}_1) = \mathbf{Y}_i$ for any i and $\boldsymbol{\Sigma}_1 \to \mathbf{0}$, then the likelihood will tend to infinity, and the global maximizer will not exist. By restricting the covariance matrices $\boldsymbol{\Sigma}_k$, $k = 1,\ldots, K$ to be equal (homoscedastic components), as is often done in mixture modeling, the unboundness of the likelihood will be eliminated. In our example with heteroscedastic variance components, each individual has five error-associated observations, while the parameter space is of dimension two. The condition for likelihood singularity is therefore very unlikely to be satisfied.

Future work is also needed to extend the algorithm to include important practical cases involving more general error variance models and random effects covariates.

Acknowledgments

This work was supported in part by National Institute of Health grants P41-EB001978 and R01-GM068968.

References

Beal, SL.; Sheiner, LB. NONMEM User's Guide, Part I. San Francisco: Division of Clinical Pharmaology, University of California; 1979.

Davidian M, Gallant AR. The non-linear mixed effects model with a smooth random effects density. Biometrika 1993;80:475–488.

Davidian, M.; Giltinan, M. Nonlinear Models for Repeated Measurement Data. Chapman and Hall; New York: 1995.

Dempster AP, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 1977;39:1–38.

Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 1999;286:487–491. [PubMed: 10521338]

Geweke J. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 1989;57:1317–1340.

Kuhn E, Lavielle M. Maximum likelihood estimation in nonlinear mixed effects models. Comput Stat Data Anal 2005;49:1020–1038.

Mallet A. A maximum likelihood estimation method for random coefficient regression models. Biometrika 1986;73:645–656.

McLachlan, GJ.; Peel, D. Finite Mixture Models. Wiley; New York: 2000.

Ng CM, Joshi A, Dedrick RL, Garovoy MR, Bauer RJ. Pharmacokinetic-pharmacodynamic-efficacy analysis of Efalizumab in patients with moderate to severe psoriasis. Pharm Res 2005;77:1088–1100. [PubMed: 16028009]

Philppou A, Roussas G. Asymptotic normality of the maximum likelihood estimate in the independent but not identically distributed case. Ann Inst Stat Math 1975;27:45–55.

Rosner GL, Muller P. Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. J Pharmacokinet Biopharm 1997;25:209–234. [PubMed: 9408860]

Schumitzky, A. EM Algorithms and two stage methods in pharmacokinetic population analysis. In: D'Argenio, DZ., editor. Advanced Methods of Pharmacokinetic and Pharmacodynamic Systems Analysis. Vol. II. Plenum Press; New York: 1995. p. 145-160.

Sheiner LB, Rosenberg B, Melmon KL. Modeling of individual pharmacokinetics for computer-aided drug dosing. Comput Biomed Res 1972;5:441–459.

Tierney L. Markov chains for exploring posterior distributions (with discussion). Ann Stat 1994;22:1701–1762.

Tseng P. An analysis of the EM algorithm and entropy-like proximal point methods. Math Oper Res 2005;29:27–44.

Wakefiled JC, Walker SG. Bayesian nonparametric population models: formulation and comparison with likelihood approaches. J Pharmacokinet Biopharm 1997;25:235–253. [PubMed: 9408861]

Walker S. An EM algorithm for nonlinear random effects models. Biometrics 1996;52:934–944.

Wu CF. On the convergence properties of the EM algorithm. Ann Stat 1983;11:95–103.

Appendix

Appendix

For the example in Section 6, as in other PK/PD problem, it is often reasonable to assume that the mechanism of genetic polymorphism applies to only part of the system, for example, drug metabolism or drug target. It is therefore desirable to partition the parameter $\boldsymbol{\theta}_i$ into two components, one ($\boldsymbol{\alpha}_i$) that follows a mixture of multivariate normals and the second ($\boldsymbol{\beta}_i$) defined by a single multivariate normal distribution: $\boldsymbol{\theta}_i = \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\}$, where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are independent. The EM updates for this special case are given by:

$$(\boldsymbol{\mu}_{-}\boldsymbol{\alpha})_k^{(r+1)}=\frac{\sum_{i=1}^{n}\int \boldsymbol{\alpha}_i g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i}{\sum_{i=1}^{n}\int g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i},$$

$$(\sum\nolimits_{-}\boldsymbol{\alpha})_k^{(r+1)}=\frac{\sum_{i=1}^{n}\int(\boldsymbol{\alpha}_i-(\boldsymbol{\mu}_{-}\boldsymbol{\alpha})_k^{(r+1)})(\boldsymbol{\alpha}_i-(\boldsymbol{\mu}_{-}\boldsymbol{\alpha})_k^{(r+1)})^T g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i}{\sum_{i=1}^{n}\int g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i},\ \text{for } 1\le k\le K;$$

$$(\boldsymbol{\mu}_{-}\boldsymbol{\beta})^{(r+1)}=\frac{\sum_{i=1}^{n}\{\sum_{k=1}^{K}\int \boldsymbol{\beta}_i g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i\}}{n},$$

and

$$(\sum\nolimits_{-}\boldsymbol{\beta})^{(r+1)}=\frac{\sum_{i=1}^{n}\{\sum_{k=1}^{K}\int(\boldsymbol{\beta}_i-(\boldsymbol{\mu}_{-}\boldsymbol{\beta})^{(r+1)})(\boldsymbol{\beta}_i-(\boldsymbol{\mu}_{-}\boldsymbol{\beta})^{(r+1)})^T g_{ik}(\boldsymbol{\theta}_i,\varphi^{(r)})d\boldsymbol{\theta}_i\}}{n}.$$

The updates for $\{w_k, 1\le k\le K\}$ and σ^2 are the same as in Section 3.

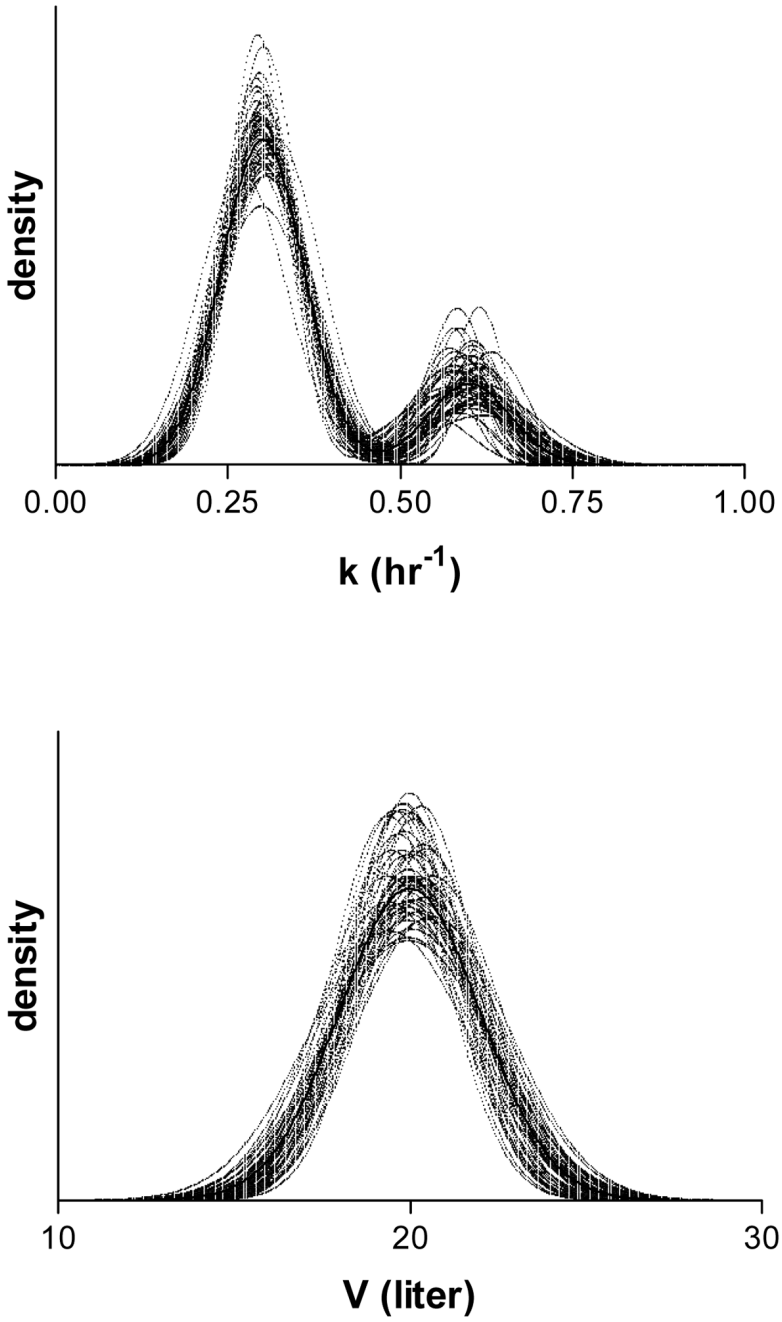

Fig. 1.
True (solid line) and estimated (dotted line) population densities of *k* (upper panel) and *V* (lower panel) from the population simulation analysis.

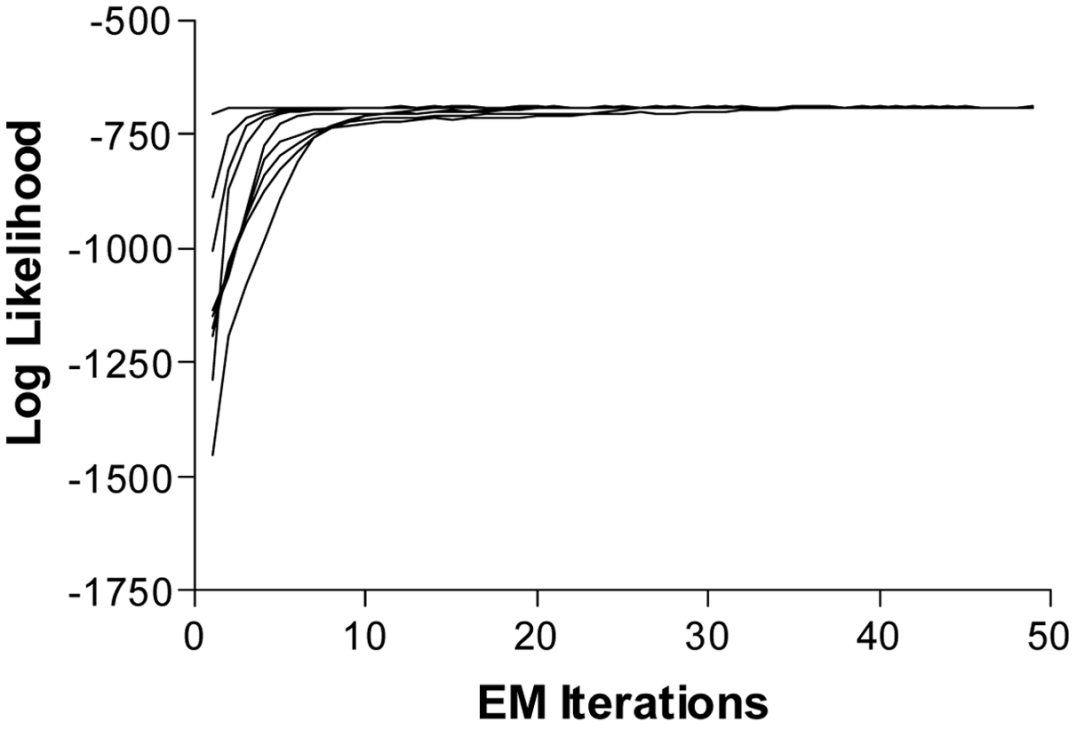

Fig. 2.
Convergence of log-likelihood by starting the EM algorithm from 9 positions.

Table 1

Mean of parameter estimates (over 200 simulations); Mean percent prediction error (PE) and root mean square percent prediction error (RMSE); Percent coverage of 95% confidence interval

Parameter	Population Values	Mean of Estimates	Mean PE (%)	RMSE (%)	Coverage of 95% CI (%)
μ_V	20	19.978	0.043586	1.0399	94.5
μ_{k1}	0.3	0.29982	−0.09045	1.6491	96.5
μ_{k2}	0.6	0.60029	0.10042	2.6455	90.5
w_1	0.8	0.80448	0.55991	5.4248	94.5*
σ_V^2	4	3.7605	−5.0867	23.822	94.5
σ_{k1}^2	0.0036	0.003573	−1.0539	14.88	91.0
σ_{k2}^2	0.0036	0.003195	−10.857	40.236	83.5
σ	0.1	0.099931	−0.06857	4.0618	95.5

*The coverage of the transformed variable $\frac{w_1}{1-w_1}$ is shown.

Table 2

Parameter estimates by importance sampling with 1000, 2000 and 3000 samples

Parameter	T=1000	T=2000	T=3000
μ_V	19.85039	19.84843	19.85201
μ_{k1}	0.30840	0.30837	0.30853
μ_{k2}	0.60038	0.60029	0.60079
w_1	0.75164	0.75157	0.75269
σ_V^2	5.73544	5.72227	5.70365
σ_{k1}^2	0.00327	0.00323	0.00328
σ_{k2}^2	0.00202	0.00203	0.00200
σ	0.09706	0.09687	0.09692