Published in final edited form as:
Comput Stat Data Anal. 2009 October 1; 53(12): 3907–3915. doi:10.1016/j.csda.2009.04.017.

Population Pharmacokinetic/Pharmacodyanamic Mixture Models via Maximum *a Posteriori* Estimation

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

[a] Clinical Discovery, Strategic Modeling & Simulation Group, Bristol-Myers Squibb Co., Princeton, NJ 08543, USA

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

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

Abstract

Pharmacokinetic/pharmacodynamic phenotypes are identified using nonlinear random effects models with finite mixture structures. A maximum *a posteriori* probability estimation approach is presented using an EM algorithm with importance sampling. Parameters for the conjugate prior densities can be based on prior studies or set to represent vague knowledge about the model parameters. A detailed simulation study illustrates the feasibility of the approach and evaluates its performance, including selecting the number of mixture components and proper subject classification.

Keywords

Nonlinear random effects; EM algorithm; Importance sampling; Bayesian Analysis

1. Introduction

The use of mathematical modeling is central to the study of the absorption, distribution and elimination of therapeutic drugs and to understanding how drugs produce their effects. From its inception the field of pharmacokinetics and pharmacodynamics has incorporated methods of mathematical modeling, simulation and computation in an effort to better understand and quantify the processes of uptake, disposition and action of therapeutic drugs. These methods for pharmacokinetic/pharmacodynamic (PK/PD) systems analysis impact all aspects of drug development including *in vitro*, animal and human testing, as well as drug therapy. Modeling methodologies developed for studying pharmacokinetic/pharmacodynamic processes confront many challenges related in part to the severe restrictions on the number and type of measurements that are available from laboratory experiments and clinical trials, as well as the variability in the experiments and the uncertainty associated with the processes themselves.

*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.

Since their initial application to pharmacokinetics in the 1970's, Bayesian methods have provided a framework for PK/PD modeling in drug development that can address some of the above-mentioned challenges. Sheiner *et al.* (1975) applied Bayesian estimation (maximum *a posteriori* probability estimation, MAP) to combine population information with drug concentration measurements obtained in an individual, in order to determine a patient-specific dosage regimen. Katz, Azen and Schumitzky (1981) reported the first efforts to calculate the complete posterior distribution of model parameters in a nonlinear pharmacokinetic individual estimation problem, for which they used numerical quadrature methods to perform the needed multi-dimensional integrations.

The population PK/PD problem has also been cast in a Bayesian framework, initially by Racine-Poon (1985) using a two-stage approach, and more generally by Wakefield *et al.* (1994). Solution to this computationally demanding problem is accomplished through the application of Markov chain Monte Carlo methods pioneered by Gelfand and Smith (1990) and now available in general purpose software (see Lunn *et al.*, 2002 for discuss relevant to PK/PD population modeling).

Population PK/PD modeling, including Bayesian approaches, are used in drug development to identify the influence of measured covariates (e.g., demographic factors and disease status) on drug kinetics and response. It is now recognized, however, that genetic polymorphisms in drug metabolism and in the molecular targets of drug therapy can also influence the efficacy and toxicity of medications (Evans and Relling, 1999). Population modeling approaches that can identify and model distinct subpopulations not related to available measured covariates may, therefore, help determine otherwise unknown genetic and other determinants of observed pharmacokinetic/pharmacodynamic phenotypes.

We have previously reported on a maximum likelihood approach using finite mixture models to identify subpopulations with distinct pharmacokinetic/pharmacodynamic properties (Wang *et al.*, 2007). Wakefield and Walker (1997) and Rosner and Mueller (1997) have introduced Bayesian approaches to address this problem within a nonparametric mixture model framework. In this paper a computationally practical maximum *a posteriori* probability estimation (MAP) approach is proposed using finite mixture models. Section 2 of this paper describes the finite mixture model within a nonlinear random effects framework and defines the MAP estimation problem. A solution via the EM algorithm is presented in Section 3. Subject classification and model selection issues are discussed in Section 4. An example and detailed simulation study are presented in Section 5 and Section 6 contains a discussion. The Appendix discusses important asymptotic properties of the MAP estimator, further motivating its use, and derives the formulas for the asymptotic covariance.

2. Nonlinear Random Effects Finite Mixture Model and MAP Estimation Problem

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

$$Y_i|\theta_i,\beta \sim N(h_i(\theta_i), G_i(\theta_i,\beta)),\ \ i=1,\ldots,n \tag{1}$$

and

$$\theta_1, \ldots, \theta_n \sim_{i.i.d} \sum_{k=1}^{K} w_k N(\mu_k, \sum_k), \tag{2}$$

where $i = 1, \ldots, n$ indexes the individuals and $k = 1, \ldots, K$ indexes the mixing components. The alternate problem formulation presented in Cruz-Mesía *et al.* (2008) and in Pauler and Laird (2000), also results in the solution outlined below.

At the first stage represented by (1), $Y_i = (y_{1i}, \ldots, y_{m_i i})^T$ is the observation vector for the ith individual ($Y_i \in R^{m_i}$); $h_i(\theta_i)$ is the function defining the pharmacokinetic/pharmacodynamic (PK/PD) model, including subject specific variables (e.g., drug doses); θ_i is the vector of model parameters (random effects) ($\theta_i \in R^p$); and $G_i(\theta_i, \beta)$ is a positive definite covariance matrix ($G_i \in R^{m_i \times m_i}$) that may depend upon θ_i as well as on other parameters β (fixed effects) ($\beta \in R^q$).

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 μ_k ($\mu_k \in R^p$) is the mean vector and Σ_k ($\Sigma_k \in R^{p \times p}$) is the positive definite covariance matrix.

Letting φ represent the collection of parameters $\{\beta, (w_k, \mu_k, \Sigma_k), k = 1, \ldots, K\}$, the population problem involves estimating φ given the observed data $\{Y_1, \ldots, Y_n\}$. If φ is regarded as a random variable with a known prior distribution $p(\varphi)$, the maximum *a posteriori* probability (MAP) estimate (φ^{MAP}) is the mode of the posterior distribution $p(\varphi|Y^n)$:

$$p(\varphi|Y^n) = \frac{p(Y^n|\varphi)p(\varphi)}{p(Y^n)} = \frac{\prod_{i=1}^{n} p(Y_i|\varphi)p(\varphi)}{p(Y^n)}.$$

where $Y^n = \{Y_1, \ldots, Y_n\}$. The MAP estimator enjoys the same large sample properties as the ML estimator, namely consistency and asymptotic normality. In addition, the MAP objective function is a regularization of the likelihood function and as such avoids the well-documented singularities and degeneracies of mixture models (see, Fraley and Raftery (2005) and Ormoneit and Tresp(1998)).

For mixture of normal distributions, a multivariate normal prior on the mean for each mixing component is given by:

$$\mu_k | \sum_k \sim N(\lambda_k, \sum_k / \tau_k), k = 1, \ldots, K, \tag{3}$$

where λ_k and τ_k can be viewed as the mean and shrinkage respectively. An inverse Wishart prior with degrees of freedom q_k and scale matrix Ψ_k is assigned to each covariance component:

$$(\sum_k)^{-1} \sim Wishart(q_k, \Psi_k), \; k = 1, \ldots, K. \tag{4}$$

The mixing weights have a Dirichlet distribution as the prior:

$$(w_1, \ldots, w_K) \sim Dirichlet(a_1, \ldots, a_K). \quad (5)$$

We consider the model given by (1) and (2) for the important case $G_i(\theta_i, \beta) = \sigma^2 H_i(\theta_i)$, where $H_i(\theta_i)$ is a known function and $\beta = \sigma^2$. The prior for σ^2 is an inverse Gamma distribution:

$$(\sigma^2)^{-1} \sim Gamma(a, b). \quad (6)$$

These densities are conjugate priors to the multivariate normal mixtures (see appendix for the parameterizations used).

The hyper-parameters $\{a, b, (a_k, \lambda_k, \tau_k, q_k, \Psi_k), k = 1, \ldots, K\}$ can be based on prior studies or set to represent vague knowledge about the parameters (see below).

3. Solution via the EM Algorithm

The EM algorithm, originally introduced by Dempster, Laird and Rubin (1977), is used to perform iterative computation of maximum likelihood (ML) estimates. It is applied below to solve the MAP estimation problem defined in the previous section.

The component label vector 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 θ_i arises from the kth mixing component. Letting $\varphi^{(r)} = \{\beta, (w_k^{(r)}, \mu_k^{(r)}, \Sigma_k^{(r)}, k = 1, \ldots, K)\}$ represent the parameters at the rth iteration of the EM algorithm, the E step computes a conditional posterior expectation, given by

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

where $\log L_c(\varphi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_i(k) \log \left(w_k p \left(Y_i, \theta_i | \sigma^2, \mu_k, \sum_k \right) \right)$. In the M step, the posterior mode $\varphi^{(r+1)}$ is estimated as the optimizer of $Q(\varphi, \varphi^{(r)})$ such that $\varphi^{(r+1)} = \arg\max_{\varphi} Q(\varphi, \varphi^{(r)})$. Under the prior defined above, the updating process of the M step is:

$$w_k^{(r+1)} = \frac{\sum_{i=1}^{n} \int g_{ik}(\theta_i, \varphi^{(r)}) d\theta_i + (a_k - 1)}{n - K + \sum_{k=1}^{K} a_k},$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} \int \theta_i g_{ik}(\theta_i, \varphi^{(r)}) d\theta_i + \tau_k \lambda_k}{\sum_{i=1}^{n} \int g_{ik}(\theta_i, \varphi^{(r)}) d\theta_i + \tau_k},$$

$$\sum_k^{(r+1)} = \frac{\sum_{i=1}^{n} \int (\theta_i - \mu_k^{(r+1)})(\theta_i - \mu_k^{(r+1)}) g_{ik}(\theta_i, \varphi^{(r)}) d\theta_i + \tau_k (\lambda_k - \mu_k^{(r+1)})(\lambda_k - \mu_k^{(r+1)})^T + \boldsymbol{\psi}_k}{\sum_{i=1}^{n} \int g_{ik}(\theta_i, \varphi^{(r)}) d\theta_i + q_k - d},$$

and

$$(\sigma^2)^{(r+1)}=\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\int(Y_i-h_i(\theta_i))^T H_i(\theta_i)^{-1}(Y_i-h_i(\theta_i))g_{ik}(\theta_i,\varphi^{(r)})d\theta_i+2b}{\sum_{i=1}^{n}m_i+2(a-1)},$$

$$g_{ik}(\theta_i,\varphi)=\frac{w_k p(Y_i|\sigma^2,\theta_i)p(\theta_i|\mu_k,\sum_k)}{\sum_{k=1}^{K}w_k\int p(Y_i|\sigma^2,\theta_i)p(\theta_i|\mu_k,\sum_k)d\theta_i}\quad\text{for } k=1,\ldots,K$$

where and $d = \dim(\theta_i)$.

The E and M steps are repeated until convergence. Discussions on the sufficient conditions for the convergence can be found at Dempster et al. (1977), Wu (1983) and Tseng (2005). In the appendix, an error analysis for φ^{MAP} is presented.

In order to implement the algorithm, all the integrals in the updating equations must be evaluated at each iterative step. For the maximum likelihood mixture model problem we have successfully used importance sampling to calculate the corresponding integrals in the EM algorithm (see Wang *et al.* 2007). The same method can be applied here and was used in the examples presented below. In brief, an envelope function $p_{e(k)}$ is selected for each mixing component and for each subject, so that a number of individual specific random samples are taken from $p_{e(k)}$: $\boldsymbol{\theta}_{i(k)}^{(1)},\ldots,\boldsymbol{\theta}_{i(k)}^{(T)}\sim_{i.i.d.}p_{e(k)}(\boldsymbol{\theta}_i)$. As all the integrals in the algorithm share the form $\int f(\boldsymbol{\theta}_i)g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i$, they are approximated as follows:

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

The envelope distribution is taken to be a multivariate normal density using the subject's previously estimated mixing-component specific conditional mean (i.e., $\int\boldsymbol{\theta}_i g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i$) and conditional covariance (i.e., $\int(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)(\boldsymbol{\theta}_i-\boldsymbol{\mu}_k)^T g_{ik}(\boldsymbol{\theta}_i,\varphi)d\boldsymbol{\theta}_i$) as its mean and covariance. The number of independent samples (T) depends on the complexity of the model and the required accuracy in the integral approximations.

4. Subpopulation Classification and Model Selection

Assigning each individual subject to a subpopulation follows the same method as presented in Wang *et al*. (2007) for ML estimation. The E-step computes the posterior probability that subject *i* belongs to the *k*th subgroup:

$$\tau_i(k)=\frac{w_k p(Y_i|\sigma^2,\mu_k,\sum_k)}{\sum_{k=1}^{K}w_k p(Y_i|\sigma^2,\mu_k,\sum_k)}=\frac{w_k\int p(Y_i|\sigma^2,\theta_i)p(\theta_i|\mu_k,\sum_k)d\theta_i}{\sum_{k=1}^{K}w_k\int p(Y_i|\sigma^2,\theta_i)p(\theta_i|\mu_k,\sum_k)d\theta_i},$$

and each individual is classified to the subpopulation associated with the highest posterior probability of membership. For example, for each *i*, (*i* =1, …, *n*), set $\hat{z}_i$ (*k*) = 1 if $k = \arg\max_c \tau_i(c)$ or to zero otherwise. No additional computation is required since all the τ_i (*k*) are evaluated at each EM step.

Determining the number of mixing components *K* in a finite mixture model is an important but difficult problem. To compare two non-nested models, in contrast to likelihood ratio procedures for comparing nested models, several criteria have been proposed. For example, Lemenuel-diot *et al.* (2006) based such model selection on the Kullback-Leibler test, with the null hypothesis of p_0 components versus the alternative hypothesis of p_1 components ($p_1 > p_0$). Other criteria are based on a penalized objective function. For two models A and B with different values of *K*, say K_A and K_B, one would compare objective functions Q_A and Q_B, and prefer the model with the larger value. The Akaike Information Criterion (AIC) takes the form *AIC* = *L* − *P*, where *L* is the maximized log-likelihood for the model and *P* is the total number of estimated parameters. The Bayesian information criterion (BIC) is defined as *BIC* = *L* − 1/2*P* log *N*, where *N* is the number of observations in the data set. Fraley and Raftery (2005) reported using a BIC criterion for mixture model selection. Based on its simplicity and good behavior, their version of BIC criterion is also adopted here, with *L* replaced by the log-likelihood evaluated at φ^{MAP}.

5. Example

A one-compartment model with first-order elimination and first-order absorption is used as the model of the drug's plasma concentrations y_{ji}, where

$$y_{ji} = \frac{DKa_i}{V_i Ka_i - CL_i}\left(e^{-\frac{CL_i}{V_i}t_j} - e^{-Ka_i t_j}\right)\left(1+\varepsilon_{ji}\right)$$

The model parameters *V* (*L*) and K_a (hr^{-1}) were assumed to be independent with $V_i \sim_{i.i.d.} LN(\mu_V = 50, \sigma_V^2 = 10^2)$ and $Ka_i \sim_{i.i.d.} LN(\mu_{Ka} = 1, \sigma_{Ka}^2 = 0.2^2)$. The drug's clearance *CL* (*L*/*hr*) was assumed to be a mixture of two lognormal densities:

$$CL_i \sim_{i.i.d.} w_1 LN(\mu_{CL_1} = 5, \sigma_{CL_1}^2 = (0.5\mu_{CL_1})^2) + w_2 LN(\mu_{CL_2} = 10, \sigma_{CL_2}^2 = (0.5\mu_{CL_2})^2)$$

with mixing weights w_1 =0.3 and w_2 =0.7. A sparse sampling schedule was used (m=4) with t_1 =2, t_2 =8, t_3 =12 and t_{24} = 24 hrs. This is a difficult mixture problem, as is evident from inspection of the density for *CL* shown in Fig. 1, and was taken from a study of the kinetics of the drug metoprolol used by Kaila et al. (2006). The within individual error ε_{ji} is assumed to be i.i.d with $\sigma^2 = 0.1^2$. The number of samples was reduced from the six sample times used by Kaila *et al*. (1, 2, 4, 8, 12, 24 hrs) to the four sample times used here in order to provide an even more challenging test. The number and timing of samples can influence estimation results and formal approaches to sample schedule design for population studies are available (Mentré *et al*., (1997)). A total of 300 population data sets were simulated from this model each consisting of *n* =50 subjects.

The MAP estimates were obtained for each of the 300 population data sets using the EM algorithm with importance sampling as described above (the lognormal kinetic parameters were transformed). Very dispersed priors were assumed for the transformed parameters as follows:

$$\mu_1|\sum_1 \sim N(\log(5),\ \sum_1/0.01)\ \ \mu_2|\sum_2 \sim N(\log(10),\ \sum_2/0.01)$$
$$(\sum_1)^{-1}=(\sum_2)^{-1} \sim Wishart(1, 0.25)$$
$$(w_1, w_2) \sim Dirichlet(1, 1)$$
$$(\sigma^2)^{-1} \sim Gamma(1, 0)$$

For each of the estimated parametersφ, its percent prediction error was calculated for each population data set as:

$$\mathrm{pe}_j = 100 \times (\varphi_j^{\mathrm{MAP}} - \varphi_j)/\varphi_j,\ j=1, \ldots, 300$$

These percent prediction errors were used to calculate the mean prediction error and root mean square prediction error (RMSE) for each parameter. In addition, the individual subject classification accuracy was evaluated for each population data set.

Figure 2 plots the estimated parameter population distributions of the 300 simulated data sets, as well as the true distributions used in the simulation, while Table 1 shows the mean estimates, estimation biases and root mean square errors (RMSE). The histogram of the percent correct classification is presented in Figure 3.

The data were also analyzed assuming a single component model for *CL* (lognormal) as well as for *V* and K_a as above. Model selection was based on the Bayesian information criterion (BIC) to select between a one or two component model for *CL* in each of the 300 population data sets. The estimated densities of clearance are displayed in Figure 4, while the detailed estimation results are given in Table 2. Based on BIC values, the two-component lognormal model for *CL* was correctly selected in 198 out of the 300 population trials.

6. Discussion

In this paper, a maximum *a posteriori* probability estimation approach is presented for nonlinear random effects finite mixtures models that has application to identifying hidden subpopulations in pharmacokinetic/pharmacodynamic studies. Previously reported nonparametric Bayesian approaches to this problem (Wakefield and Walker (1997), Rosner and Mueller (1997)) have advantages over the MAP estimation approach presented herein, including calculation of the complete posterior distribution of model parameters including the number of mixing components. However, the computational challenges associated with the proper solution to the nonparametric Bayesian mixture problem are considerable, whereas the MAP estimation approach using the EM algorithm with importance sampling presented here is straightforward in comparison.

In calculating the MAP estimator, the values of the parameters defining the conjugate prior densities may be available from previous studies. When no prior information is available, these parameters can be set to reflect very disperse priors (e.g, $\tau_k \to 0$ and other parameters set as illustrated in the example presented). For linear mixture models with diffuse priors, as noted by Fraley and Raftery (2005), it is expected that the MAP results will be similar to the MLE results when they latter can be calculated. The advantage of the MAP estimator in such cases is that it avoids the unboundedness that can be associated with maximum likelihood mixture model problems (Fraley and Raftery (2005), Ormoneit and Tresp (1998)).

For determining the number of components in mixture models, several measures have been suggested, including the BIC criterion used in the example presented in this paper. In addition, *a priori* knowledge or assumptions about the biological mechanism for the modeled PK/PD

polymorphism can also facilitate the model selection procedure, when combined with a model selection measure. For example, for drugs primarily metabolized by the liver, the information on hepatic cytochrome P450 family can help to decide a reasonable range for the number of clearance subgroups (e.g., K=3 accounting for extensive, intermediate, and poor metabolizer subpopulations) thus limiting the number of competing models to be tested.

Acknowledgments

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

References

Bernardo, JM.; Smith, AFM. Bayesian Theory, Section 5.3.2. Wiley; New York: 2001.

Cruz-Mesía R, Quintana FA, Marshall G. Model-based clustering for longitudinal data. Computational Statistics & Data Analysis 2008;52:1441–1457.

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]

Fraley, C.; Raftery, AE. Technical Report no. 486. Department of Statistics, University of Washington; 2005. Bayesian regularization for Normal mixture estimation and model-based clustering. http://handle.dtic.mil/100.2/ADA454825

Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J of the Am Stat Assoc 1990;85:398–409.

Katz D, Azen SP, Schumitzky A. Bayesian approach to the analysis of nonlinear models: Implementation and evaluation. Biometrics 1981;37:137–142.

Kaila N, Straka RJ, Brundage RC. Mixture models and subpopulation classification: a pharmacokinetic simulation study and application to metoprolol CYP2D6 phenotype. J Pharmacokinet Pharmacodyn 2007;34:141–156. [PubMed: 17053980]

Lemenuel-diot A, Laveille C, Frey N, Jochemsen R, Mallet A. Mixture modeling for the detection of subpopulations in a pharmacokinetic/pharmacodynamic analysis. J Pharmacokinet Pharmacodyn 2006;34:157–181. [PubMed: 17151938]

Lunn DJ, Best N, Thomas A, Wakefield J, Spiegelhalter D. Bayesian analysis of population PK/PD models: general concepts and software. J of Pharmacokin and Pharmacodyn 2002;29:217–307.

Mentré F, Mallet A, Baccar D. Optimal design in random effect regression models. Biometrika 1997;84:429–442.

Minka, T. Old and new matrix algebra useful for statistics. MIT Media Lab. 2000. http://research.microsoft.com/minka/papers/matrix

Ormoneit D, Tresp V. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. IEEE Transactions on Neural Networks 1998;9:639–650. [PubMed: 18252487]

Pauler DK, Laird NM. A mixture model for longitudinal data with application to assessment of noncompliance. Biometrics 2000;56:464–472. [PubMed: 10877305]

Racine-Poon A. A Bayesian approach to non-linear random effects models. Biometrics 1985;41:1015–1024. [PubMed: 4096913]

Rosner GL, Muller P. Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. J Pharmacokinet Biopharm 1997;2:209–234. [PubMed: 9408860]

Sheiner LB, Halkin H, Peck C, Rosenberg B, Melmon KL. Improved computer assisted digoxin therapy: a method using feedback of measured serum digoxin concentrations. Ann Intern Med 1975;82:619–627. [PubMed: 1137256]

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

Wakefield JC, Smith AFM, Racine-Poon A, Gelfand AE. Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. Applied Statistics 1994;43:201–221.

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

Wang X, Schumitzky A, D'Argenio DZ. Nonlinear random effects mixture models: Maximum likelihood estimation via the EM algorithm. Comput Stat Data Anal 2007;51:6614–6623. [PubMed: 19756256]

White, H. Estimation, Inference and Specification Analysis. Cambridge University Press; 1994. Chapter 6

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

Appendix: Asymptotic Properites

Let φ_n^{MAP} be the posterior mode of φ given Y^n. Assuming that there is a "true" parameter φ_0 which lies in the interior of the support of $p(\varphi)$, it can be shown under suitable hypotheses that φ_n^{MAP} is consistent and asymptotically normal (White, 1994). However, from a Bayesian perspective, it is more natural to investigate the asymptotic behavior of the conditional density $p(\varphi|Y^n)$. Under the assumptions stated below, it is found that $p(\varphi|Y^n)$ is asymptotically normal with mean φ_n^{MAP}. This gives another justification for calculating the MAP estimate.

This result is now stated more precisely. Assume that the Hessian matrix $A_n(\varphi)=\frac{\partial^2 \log p(\varphi|Y^n)}{\partial\varphi\partial\varphi}$ is invertible at $\varphi=\varphi_n^{MAP}$. Then given the regularity conditions of Philppou and Roussas (1975) and Bernardo and Smith (2001), it can be shown that $\Gamma_n^{-1/2}(\varphi_n - \varphi_n^{ML})$ converges in distribution to a standard multivariate normal random variable as $n \to \infty$, where $\varphi_n \sim p(\varphi|Y^n)$ and $\Gamma_n^{-1} = -A_n(\varphi_n^{MAP})$. It follows that asymptotically, φ_n^{MAP} and Γ_n are the posterior mean and posterior covariance of $p(\varphi|Y^n)$. For any component of φ_n^{MAP}, the posterior standard error is the corresponding diagonal element of Γ_n.

The computation of Γ_n proceeds as follows:

$$\frac{\partial^2 \log p(\varphi|Y^n)}{\partial\varphi\partial\varphi}=\left(\sum_{i=1}^{n}\frac{\partial^2 \log p(Y_i|\varphi)}{\partial\varphi\partial\varphi}\right)+\frac{\partial^2 \log p(\varphi)}{\partial\varphi\partial\varphi}$$

and

$$-\sum_{i=1}^{n}\frac{\partial \log p(Y_i|\varphi_n^{MAP})}{\partial\varphi\partial\varphi} \approx \sum_{i=1}^{n} V_i(\varphi_n^{MAP})$$

where

$$V_i(\varphi)=\frac{\partial \log p(Y_i|\varphi)}{\partial\varphi}\left(\frac{\partial \log p(Y_i|\varphi)}{\partial\varphi}\right)^T$$

It follows that

$$\Gamma_n \approx \left(\sum_{i=1}^{n} V_i(\varphi_n^{MAP}) - \frac{\partial^2 \log p(\varphi_n^{MAP})}{\partial\varphi\partial\varphi} \right)^{-1}.$$

Note that in the maximum likelihood setting of Wang et al (2007), $Cov(\varphi_n^{ML}) \approx \left(\sum_{i=1}^{n} V_i(\varphi_n^{MAP}) \right)^{-1}$. So the difference between MAP and ML is the term $\frac{\partial^2 \log p(\varphi_n^{MAP})}{\partial\varphi\partial\varphi}$. As $n \to \infty$, the contribution of this term diminishes.

In this asymptotic analysis the precisions $(\sigma^2)^{-1}$ and $\sum_k^{-1}$ are considered as primary variables instead of the variances σ^2 and Σ_k. This is standard practice in the normal, gamma, Wishart model. Further using $\sum_k^{-1}$ instead of Σ_k greatly simplifies the second derivative calculations.

We first calculate $V_i(\varphi)$, $i = 1, \ldots, n$. The formulas below are similar to those given in given in Wang, et al (2007). The gradient components are calculated using the relation:

$$\frac{\partial}{\partial\varphi} \log p(Y_i|\varphi) = \sum_{k=1}^{K} \int \{ \tfrac{\partial}{\partial\varphi} \log p(Y_i|\varphi, \gamma) w_k \eta(\theta|\mu_k, \textstyle\sum_k) \} g_{ik}(\theta_i, \varphi) d\theta,$$

We have:

$$s_{(\sigma^2)^{-1}} = \frac{\partial}{\partial(\sigma^2)^{-1}} \log p(Y_i|\varphi) = \sum_{k=1}^{K} \int \{ (1/2) m_i \sigma^2 - (1/2)(Y_i - h_i(\theta))^T H_i(\theta)^{-1} (Y_i - h_i(\theta)) \} g_{ik}(\theta, \varphi) d\theta$$

Next using the constraint: $w_K = 1 - w_1 - \ldots - w_{K-1}$, we have for $k = 1, \ldots, K-1$

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

And for $k = 1, \ldots, K$

$$s_{\mu_k} = \tfrac{\partial}{\partial\mu_k} \log p(Y_i|\varphi) = \int g_{ik}(\theta, \varphi) \{ \textstyle\sum_k^{-1} (\theta - \mu_k) \} d\theta,$$
$$\tfrac{\partial}{\partial \sum_k^{-1}} \log p(Y_i|\varphi) = \int \{ (1/2) \textstyle\sum_k - (1/2)(\theta - \mu_k)(\theta - \mu_k)^T \} g_{ik}(\theta, \varphi) d\theta,$$

so that

$$s_{vech(\sum_k^{-1})} = \frac{\partial}{\partial vech(\sum_k^{-1})} \log p(Y_i|\varphi) = vech[\, 2 \frac{\partial}{\partial \sum_k^{-1}} \log p(Y_i|\varphi) - Diagonal(\frac{\partial}{\partial \sum_k^{-1}} \log p(Y_i|\varphi))]$$

where *vech*(*X*) is the vector of the lower triangular components of a symmetric matrix *X*. Putting these results together produces the vector

$$s_i=(s_{(\sigma^2)^{-1}}, s_{w_1}, \ldots, s_{w_{K-1}}, s_{\mu_1}, \ldots, s_{\mu_K}, s_{vech(\sum_1^{-1})}, \ldots, s_{vech(\sum_K^{-1})}),$$

and the formula

$$V_i(\varphi)=(\sum_{i=1}^{n} s_i s_i^T)^{-1}.$$

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

It remains to calculate $\frac{\partial^2 \log p(\varphi)}{\partial\varphi\partial\varphi}$. For this calculation we represent the parameter φ as $\varphi=\{(\sigma^2)^{-1}, w_1, \ldots, w_{K-1}, (\mu_k, \sum_k^{-1}), k=1, \ldots, K\}$. Now suppose$\theta$ has p components. Then each μ_k has p components and each $\sum_k^{-1}$ has $p(p+1)/2$ unique components. Therefore $\frac{\partial^2 \log p(\varphi)}{\partial\varphi\partial\varphi}$ is an $R\times R$ matrix, where $R = 1 + (p - 1) + Kp + Kp(p+1)/2$. Now

$$\log p(\varphi)=\log p(\sigma^{-2})+\log p(w_1, \ldots w_{K-1})+\sum_{k=1}^{K}\log p(\mu_k|\sum_k^{-1})+\sum_{k=1}^{K}\log p(\sum_k^{-1})$$

so that $B=\frac{\partial^2 \log p(\varphi)}{\partial\varphi\partial\varphi}$ is a block diagonal matrix of the form $B = Diag\ (B_{\sigma^{-2}}; B_w; B(\mu_k v_k), k = 1, \ldots, K)$, $v_k=vech(\sum_k^{-1})$.

To calculate B the following parameterizations for the prior $p(\varphi)$ are assumed:

$$p(\kappa)=c\kappa^{(a-1)}\exp(-b\kappa),\quad \kappa=\sigma^{-2}$$
$$p(w_1, \ldots, w_{K-1})=c(w_1)^{a_1-1}\cdot\cdots\cdot(w_K)^{a_K-1},\quad w_K=1-(w_1+\ldots+w_{K-1})$$
$$p(\mu_k|\sum_k^{-1})=C(\det\sum_k^{-1})^{1/2}\exp((-\tau_k/2)(\mu_k-\lambda_k)^T\sum_k^{-1}(\mu_k-\lambda_k)$$
$$p(\sum_k^{-1})=C(\det\sum_k^{-1})^{\beta}\exp(-tr(\Psi_k\sum_k^{-1})/2),\quad \beta=(q_k-d-1)/2$$

It follows:

$$
\begin{aligned}
&B_{\sigma^{-2}}=\frac{\partial^2}{\partial\sigma^{-2}\partial\sigma^{-2}}\log p(\sigma^{-2})=-(a-1)\sigma^4\\
&B_w=\frac{\partial^2}{\partial w\partial w}\log p(w),\ w=(w_1,\ldots,w_{K-1})\\
&[B_w]_{rr}=-(a_r-1)/(w_r)^2+(a_K-1)/(w_K)^2,\ r=1,\ldots,p-1\\
&[B_w]_{rs}=(a_K-1)/(w_K)^2,\ 1\le r,s\le p-1,\ r\ne s\\
&B_{(\mu_k,v_k)}=\begin{bmatrix} G_{p\times p} & H_{p\times p(p+1)/2}\\ H^T & J_{p(p+1)/2\times p(p+1)/2}\end{bmatrix}\\
&G=\frac{\partial^2}{\partial\mu_k\partial\mu_k}\log p(\mu_k|\textstyle\sum_k^{-1})=-\tau_k\sum_k^{-1}\\
&H=\frac{\partial^2}{\partial\mu_k\partial vech(\sum_k^{-1})}\log p(\mu_k|\textstyle\sum_k)=-\tau_k D^T((\mu_k-\lambda_k)\otimes I)D,\\
&J=J_1+J_2\\
&J_1=\frac{\partial^2}{\partial vech(\sum_k^{-1})\partial vech(\sum_k^{-1})}\log p(\mu_k|\textstyle\sum_k^{-1})=-(1/2)D^T\textstyle\sum_k\otimes\sum_k D\\
&J_2=\frac{\partial^2}{\partial vech(\sum_k^{-1})\partial vech(\sum_k^{-1})}\log p(\textstyle\sum_k)=-\beta D^T\textstyle\sum_k\otimes\sum_k D
\end{aligned}
$$

In the above equations, D is the permutation matrix satisfying $D\,vech\,(X)=vec\,(X)$, where $vec\,(X)$ is the vector of the stacked columns of the symmetric matrix X, and $\otimes$ is the Kronecker product (see Minka (2000)).

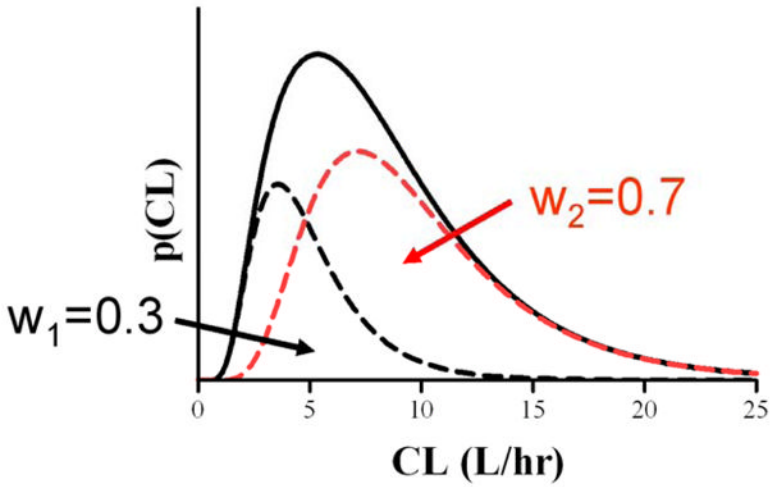

Fig. 1.
Simulated population density of clearance CL

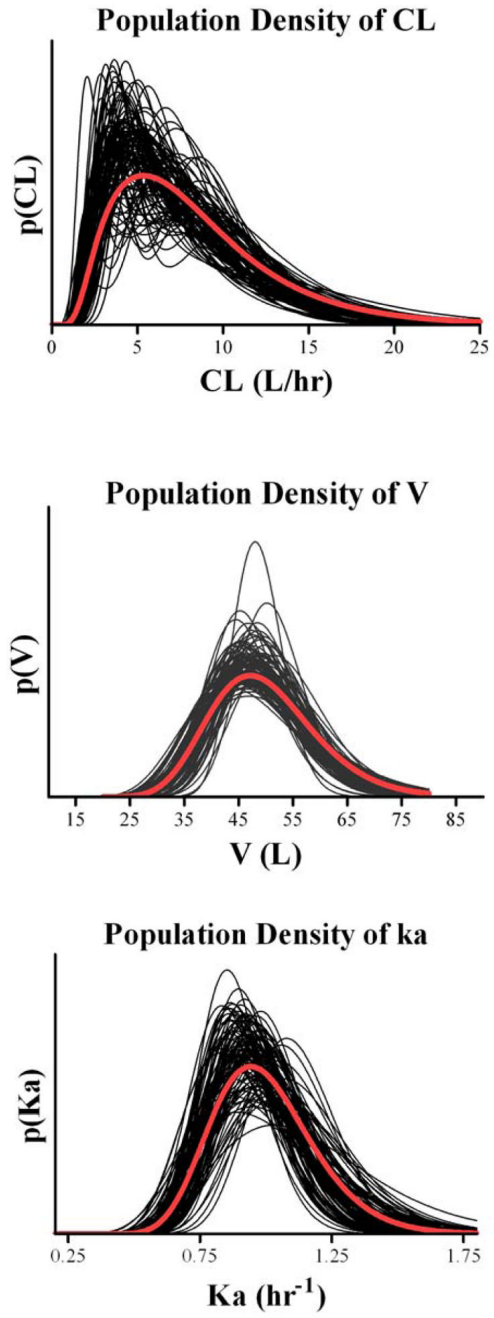

Fig. 2.
True (bold lines) and estimated (thin lines) population densities of CL (upper panel), V (middle panel) and Ka (lower panel).

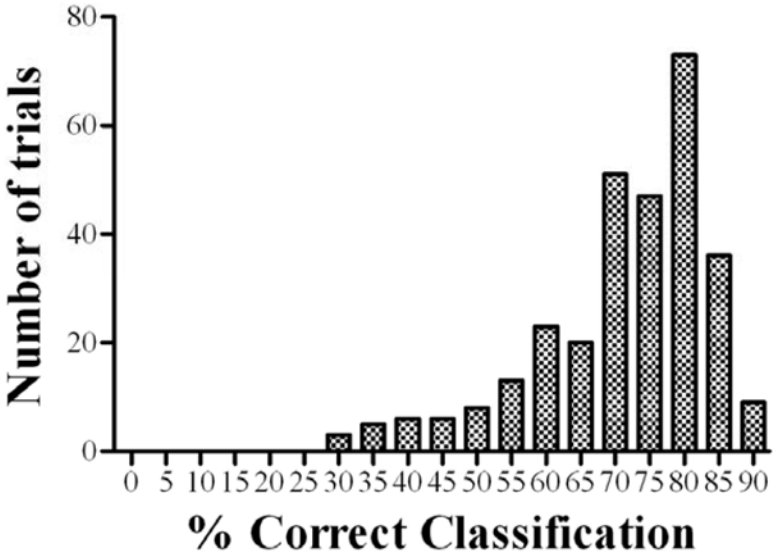

Fig. 3.
Histogram of percent correct classification.

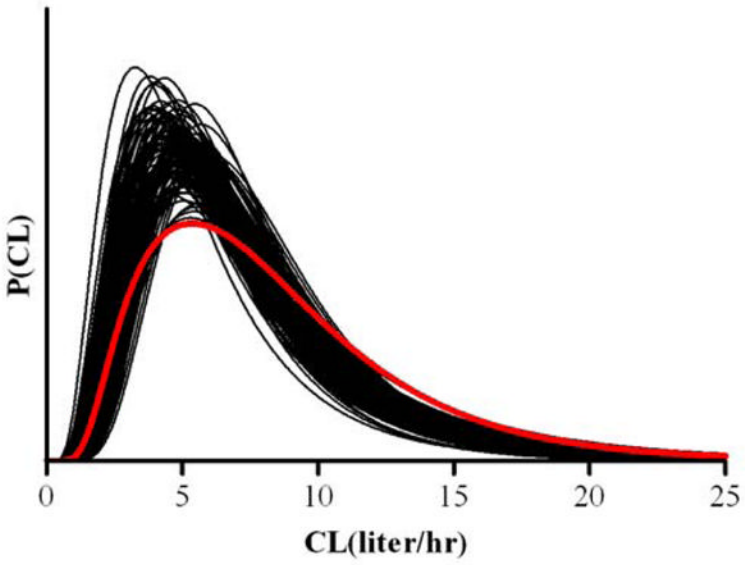

Fig. 4.
True (bold lines) and estimated (thin lines) population densities of CL for the single component analysis

Table 1

True population values and mean of parameter estimates (300 trials), mean percent prediction error (PE) and root mean square percent prediction error (RMSE)

Parameter	Population Values	Mean of Estimates	Mean PE (%)	RMSE (%)
μ_{CL1}	5	5.177	3.541	30.50
μ_{CL2}	10	9.709	−2.910	20.55
μ_V	50	49.09	−1.816	3.374
μ_{ka}	1	0.986	−1.401	6.287
σ_{CL}^2 (cv%)	50	45.21	−18.25	36.79
σ_V^2 (cv%)	20	19.31	−6.796	22.64
σ_{ka}^2 (cv%)	20	19.77	−2.270	42.25
w_1	0.3	0.4486	49.53	68.44
σ	0.1	0.1094	9.421	15.46

Table 2

True population values and mean of parameter estimates (300 trials) for the single component model

Parameter	Population Values	Mean of Estimates
μ_{CL1}	5	7.271
μ_{CL2}	10	X
μ_V	50	49.16
μ_{ka}	1	0.9891
σ_{CL}^2 (cv%)	50	56.97
σ_V^2 (cv%)	20	18.84
σ_{ka}^2 (cv%)	20	18.30
σ	0.1	0.1163