Technical Guide on Various Methods in NONMEM 7

Robert J. Bauer

Contents

Basic Theory of Nonlinear Mixed Effects

Individual parameters phi ($\boldsymbol{\phi}$) to a PK/PD model are assumed to have a random distribution in a population of subjects, typically a normal distribution with mean MU ($\boldsymbol{\mu}$) and variance OMEGA ($\boldsymbol{\Omega}$). The mean $\boldsymbol{\mu}$ may in turn be modeled as a function of a set of unknown but to be estimated fixed effects parameters THETA ($\boldsymbol{\theta}$), and a set of covariates, or information about individual i, $\mathbf{x}_i$. The deviation of the individual parameter $\boldsymbol{\phi}$ from its mean is designated $\boldsymbol{\eta}$, so that the following relation holds:

$$\boldsymbol{\phi} = \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu, \mathbf{x}_i) + \boldsymbol{\eta} \tag{1.1}$$

where $\boldsymbol{\theta}_\mu$ are those thetas that are related to etas through a mu function, of the above format. Thus, the distribution of $\boldsymbol{\phi}$ can be described as

$$h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \propto \frac{1}{\det(\boldsymbol{\Omega})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i) \right) \tag{1.2}$$

The population parameter density $h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})$ is the probability that the particular $\boldsymbol{\phi}$ would occur for an individual, given mean population parameters $\boldsymbol{\mu}_i$ and its inter-individual covariance $\boldsymbol{\Omega}$. The distribution of $\boldsymbol{\eta}$ is therefore centered about zero (**0**), and can be described as

$$h(\boldsymbol{\eta} \mid \mathbf{0}, \boldsymbol{\Omega})) \propto \frac{1}{\det(\boldsymbol{\Omega})^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \right) \tag{1.3}$$

Not all fixed effects theta are involved in an eta ($\boldsymbol{\eta}$) relationship as shown above. For those theta that are not exclusively expressed in the PK/PD model or error model via mu (**μ**), these are considered not mu modeled. We shall designate these thetas as $\boldsymbol{\theta}_{\neq}$. The entire vector of thetas is then

$$\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_{\neq}\} \tag{1.4}$$

The parameters as designated in NONMEM are THETA for $\boldsymbol{\theta}$, ETA for $\boldsymbol{\eta}$, and OMEGA for $\boldsymbol{\Omega}$. There are also a set of parameters designated as SIGMA in NONMEM, which are never mu modeled, and because in our discussion they will be treated in exactly the same way as non mu modeled theta, we shall include them in $\boldsymbol{\theta}_{\neq}$ to reduce the complexity of the

nomenclature. There may also be some etas that are not related to a MU model, in which case

$$\boldsymbol{\phi} = \boldsymbol{\eta} \tag{1.5}$$

Thus the phi vector includes all etas, whether or not they are involved in a mu function.

For observed data that are modeled as normally distributed, a predictive function may be evaluated using the individual PK/PD parameters phi, and/or may be modeled directly from fixed effects parameters not be phi/mu modeled, $\mathbf{f}_i(\boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})$. In addition, a residual variance matrix **V** describes the uncertainty of the observed values, and may be directly a function of the predicted value $\mathbf{f}_i$, sigma parameters and other non-mu modeled thetas, and rarely, individual parameters $\boldsymbol{\phi} : \mathbf{V}_i(\mathbf{f}_i, \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})$. The normal data density can be expressed as

$$l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}) \propto \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{f}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{f}_i)\right]}{\sqrt{\det\left(\mathbf{V}_i\right)}} \tag{1.6}$$

where $l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})$ is the individual data density, the probability of data $\mathbf{y}_i$ occurring for individual *i*, given individual PK/PD parameters $\boldsymbol{\phi}$, and fixed effect parameters $\boldsymbol{\theta}_{\neq}$ that are not mu modeled.

The joint density of data $\mathbf{y}_i$ and $\boldsymbol{\phi}$ for an individual is then

$$p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i(\boldsymbol{\theta}_{\mu}), \boldsymbol{\Omega}) = l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i(\boldsymbol{\theta}_{\mu}), \boldsymbol{\Omega}) \tag{1.7}$$

The $l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})$ is the joint likelihood density of $\boldsymbol{\phi}$ and $\mathbf{y}_i$ for a given individual. It is integrated over all possible values of $\boldsymbol{\phi}$ for each individual, so that the "best" population parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ are determined by taking into account the joint probability to an individual's data over the entire parameter space of $\boldsymbol{\phi}$, rather than at just one particular location, such as at the individual's best fit. We are therefore interested in evaluating the marginal density of $\mathbf{y}_i$ for any given $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ (or $\boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}$):

$$p(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) = \int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} = \int_{-\infty}^{\infty} l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.8}$$

for each subject *i*. The total marginal density for all *m* subjects is then

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) = \prod_{i=1}^{m} \int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.9}$$

It is convenient at this stage to use the negative logarithm of the density, and refer to this as the objective function, for each individual:

$$L_i = -\log(\int_{-\infty}^{+\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}) \tag{1.10}$$

and for the total data set:

$$L = -\log(\, p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\Omega})) = \sum_{i=1}^{m} L_i \tag{1.11}$$

Thus, the negative logarithm of the parameter density is

$$-\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i) \tag{1.12}$$

And the negative logarithm of the data density is:

$$-\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})) = \frac{1}{2}(\mathbf{y}_i - \mathbf{f}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{f}_i) + \frac{1}{2}\log\left(\det\left(\mathbf{V}_i\right)\right) \tag{1.13}$$

To fit a model with mean population parameters $\boldsymbol{\theta}$ and population variance $\boldsymbol{\Omega}$ to data $\mathbf{y}$, the marginal density (1.9) is to be maximized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$. These parameter values are then considered the most likely for the observed data $\mathbf{y}$. Therefore, the maximization of the marginal density with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ is called the maximum likelihood method. In practice, as an equivalent process, the negative logarithm of the marginal density (1.11) is minimized. This is the goal of the first order (FO), first order conditional estimation (FOCE/FOCEI), Laplace, iterative two stage (ITS), and expectation maximization (EM) methods.

FOCE and Laplace Methods

Generally the integral of the joint density (1.10) is very difficult to evaluate deterministically, but it may be approximated for classical methods FOCE and Laplace using a method described by Beal (part VII of NONMEM manuals [1]). The derivation is given in [2], while we will merely report the results. Classical NONMEM (FO, FOCE, and Laplace) does not require the use of $\boldsymbol{\mu}$ modeling, so for this section, we will use the

parameterization of $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, for all parameters, rather than distinguishing between $\boldsymbol{\mu}$ modeled and non-$\boldsymbol{\mu}$ modeled $\boldsymbol{\theta}$. For example, the individual's joint density may be alternately expressed as

$$p(\mathbf{y}_i,\boldsymbol{\phi}\mid\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})=l(\mathbf{y}_i\mid\boldsymbol{\phi},\boldsymbol{\theta}_{\neq})h(\boldsymbol{\phi}\mid\boldsymbol{\mu}_i,\boldsymbol{\Omega})=l(\mathbf{y}_i\mid\boldsymbol{\theta},\boldsymbol{\eta})h(\boldsymbol{\eta}\mid\mathbf{0},\boldsymbol{\Omega})=p(\mathbf{y}_i,\boldsymbol{\eta}\mid\boldsymbol{\theta},\boldsymbol{\Omega}) \quad (1.14)$$

and integration over all values of $\boldsymbol{\eta}$ is equivalent to integrating over all values of $\boldsymbol{\phi}$:

$$L_i=-\log(\textstyle\int_{-\infty}^{+\infty}p(\mathbf{y}_i,\boldsymbol{\eta}\mid\boldsymbol{\theta},\boldsymbol{\Omega})d\boldsymbol{\eta}) \quad (1.15)$$

In order to integrate the individual's joint density over all $\boldsymbol{\eta}$ using the approximation suggested by Beal, we wish first to determine the set of $\boldsymbol{\eta}$ at the maximum of this joint density, or equivalently, at the minimum of the negative logarithm of the joint density:

$$-\log(l(\mathbf{y}_i\mid\boldsymbol{\theta},\boldsymbol{\eta})h(\boldsymbol{\eta}\mid\mathbf{0},\boldsymbol{\Omega}))=-\log(l(\mathbf{y}_i\mid\boldsymbol{\theta},\boldsymbol{\eta}))+\frac{1}{2}\log(\det(\boldsymbol{\Omega}))+\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\eta} \quad (1.16)$$

We minimize with respect to eta by evaluating

$$\frac{\partial-\log(l(\mathbf{y}_i\mid\boldsymbol{\eta},\boldsymbol{\theta})h(\boldsymbol{\eta}\mid\mathbf{0},\boldsymbol{\Omega}))}{\partial\boldsymbol{\eta}}=\frac{\partial-\log(l(\mathbf{y}_i\mid\boldsymbol{\eta},\boldsymbol{\theta}))}{\partial\boldsymbol{\eta}}+\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}=\mathbf{0} \quad (1.17)$$

using typical search strategies. The $\boldsymbol{\eta}$ at which equation (1.17) is satisfied is called the mode of the joint density for subject *i*, and shall be designated $\hat{\boldsymbol{\eta}}_i$ (the hat over the parameter shall refer to a mode or point estimate, whereas the line over a parameter refers to a mean). Finding the $\hat{\boldsymbol{\eta}}_i$ parameters that provide the minimum of the individual's joint density is called *mode a posteriori* (MAP) estimation. This is used to then evaluate an approximation of the negative logarithm of the individual's integral of his joint density as follows:

$$\begin{aligned}&L_i=-\log(\textstyle\int_{-\infty}^{+\infty}l(\mathbf{y}_i\mid\boldsymbol{\eta},\boldsymbol{\theta})h(\boldsymbol{\eta}\mid\mathbf{0},\boldsymbol{\Omega})d\boldsymbol{\eta})\approx\\&-\log(l(\mathbf{y}_i\mid\hat{\boldsymbol{\eta}}_i,\boldsymbol{\theta}))+\frac{1}{2}\log(\det(\boldsymbol{\Omega}))+\frac{1}{2}\hat{\boldsymbol{\eta}}_i'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\eta}}_i+\\&\frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1}+\mathbf{S}^{-1}(\mathbf{y}_i\mid\hat{\boldsymbol{\eta}}_i,\boldsymbol{\theta})))=L_{Ni}\end{aligned} \quad (1.18)$$

where $\mathbf{S}^{-1}(\mathbf{y}_i\mid\hat{\boldsymbol{\eta}}_i,\boldsymbol{\theta})$ is the hessian or information matrix to the data density $l(\mathbf{y}_i\mid\hat{\boldsymbol{\eta}}_i,\boldsymbol{\theta})$. The total approximate objective function is in turn, for *m* subjects:

$$L_N=\sum_{i=1}^{m}L_{Ni} \quad (1.19)$$

(the subscript *N* refers to classical NONMEM) where the first three terms of (1.18) are simply the negative logarithm of the joint density evaluated at the mode $\hat{\boldsymbol{\eta}}_i$, and the last term is ½ of the negative logarithm of the determinant of the variance of $\boldsymbol{\eta}$ under the joint density, $l(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) h(\boldsymbol{\eta} \mid \mathbf{0}, \boldsymbol{\Omega})$. One can therefore think of the joint density evaluated at the mode (that is, the first three terms of equation (1.18) as the "height" of the joint density:

$$-\log(H_i) = -\log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta})) + \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}\hat{\boldsymbol{\eta}}_i' \boldsymbol{\Omega}^{-1} \hat{\boldsymbol{\eta}}_i \tag{1.20}$$

where H_i is the "height" of individual *i*'s joint density. Similarly, one may think of one-half the determinant of the variance under the joint density (the last term of the equation (1.18)) as its "width":

$$-\log(W_i) = \frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta})) \tag{1.21}$$

Thus equation (1.18) represents the negative logarithm of the height multiplied by the width, resulting in the "area" of the joint density. The $\mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta})$ may be evaluated several ways. One method is to evaluate the second derivative, usually by finite difference methods:

$$\mathbf{S}^{-1}(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \left\{ \frac{\partial^2 -\log(l(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\eta}_{k_1} \partial \boldsymbol{\eta}_{k_2}}, k_1 = 1 \text{to } n, k_2 = 1 \text{to } n, \right\} \tag{1.22}$$

where {} means "matrix containing elements". This evaluation is used in the Laplace method in NONMEM. This evaluation guarantees positive definiteness (assuming no numerical difficulties arise) when evaluated at the mode (see appendix B). Another method is by the cross product of the first derivatives of the individual data point densities:

$$\mathbf{S}^{-1}(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \left\{ \sum_{j=1}^{m_i} \frac{\partial -\log(l(y_{ij} \mid \boldsymbol{\eta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\eta}_{k_1}} \frac{\partial -\log(l(y_{ij} \mid \boldsymbol{\eta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\eta}_{k_2}} \right\} \tag{1.23}$$

where m_i is the number of data points for patient *i* (assuming all data are independent). Based on its structure, positive definiteness is guaranteed even when evaluated not at the mode (see appendix B). A third method for evaluating $\mathbf{S}^{-1}(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta})$ is by the expected value of the second derivative:

$$\mathbf{S}^{-1}(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) = \left\{ E\left[\frac{\partial^2 - \log(l(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\eta}_{k_1} \partial \boldsymbol{\eta}_{k_2}} \right] \right\} =$$
$$\left\{ \frac{\partial \mathbf{f}(t_{ij}, \boldsymbol{\eta}, \boldsymbol{\theta})}{\partial \boldsymbol{\eta}_{k_1}} \mathbf{V}_i^{-1} \frac{\partial \mathbf{f}(t_{ij}, \boldsymbol{\eta}, \boldsymbol{\theta})}{\partial \boldsymbol{\eta}_{k_2}} + \frac{1}{2} tr\left(\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\eta}_{k_1}} \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\eta}_{k_2}} \right) \right\} \tag{1.24}$$

which is also positive definite even when not evaluated at the mode. Equation (1.24) is used as the non-Laplace (CONDITIONAL) method in NONMEM. The $\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1}$ is in turn the Hessian (information) matrix of the joint density:

$$\left\{ \begin{array}{l} E\left[\dfrac{\partial^2 - \log(l(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) h(\boldsymbol{\eta} \mid \mathbf{0}, \boldsymbol{\Omega}))}{\partial \boldsymbol{\eta}_{k_1} \partial \boldsymbol{\eta}_{k_2}} \right] = \\ E\left[\dfrac{\partial^2 - \log(l(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}))}{\partial \boldsymbol{\eta}_{k_1} \partial \boldsymbol{\eta}_{k_2}} \right] + E\left[\dfrac{\partial^2 - \log(h(\boldsymbol{\eta} \mid \mathbf{0}, \boldsymbol{\Omega}))}{\partial \boldsymbol{\eta}_{k_1} \partial \boldsymbol{\eta}_{k_2}} \right] \end{array} \right\} = \tag{1.25}$$
$$\mathbf{S}^{-1}(\mathbf{y}_i \mid \boldsymbol{\eta}, \boldsymbol{\theta}) + \boldsymbol{\Omega}^{-1}$$

and hence its inverse is the variance matrix of $\boldsymbol{\eta}$ under the joint density, as mentioned earlier. Because the sum of two positive definite matrices is itself positive definite, the variance of the joint density as evaluated above is positive definite. For joint densities that are exactly multivariate normally distributed with respect to $\boldsymbol{\eta}$, equation (1.18) evaluates the joint area exactly. We shall also refer to $\mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta})$ as $\hat{\mathbf{S}}_i^{-1}$. The $\hat{\mathbf{S}}_i^{-1}$ must be evaluated at the individual *i*'s mode of his joint density, at $\hat{\boldsymbol{\eta}}_i$, and not at the mean population position of $\boldsymbol{\eta}$=**0**, so the INTERACTION option in NONMEM must be used.

Keep in mind that while the $\hat{\boldsymbol{\eta}}_i$ represents the *i*th individual's "best fit" parameters for its data, based on its joint density, it is only needed here to evaluate the area under his joint density using the above approximation. In other words, we really don't need an individual's best fit parameter set theoretically, but we need it practically, in order to evaluate the "height" of the density, and thus approximate his joint density area. There are alternative methods of finding the area without needing to know the individual's "best fit" parameters, which we will explore later.

Following the evaluation of each individual's objective function in the manner described above, these are summed to form the total approximate objective function L_N. NONMEM optimizes L_N with respect to THETAS, OMEGAS, and SIGMAS using a variable metric method, in which L_N is evaluated at a series of values of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, to provide a directional search to find the set of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ that optimizes L_N. The description of the variable metric method is beyond the scope of this document, but a good reference is [3].

Expectation Maximization (EM) Principles

Maximization-expectation methods separate the process of expectation (integration) and maximization. To find improved estimates for $\boldsymbol{\mu}$ modeled $\boldsymbol{\theta}_{\mu}$, it is convenient to first minimize the negative logarithm of $p(\mathbf{y} \mid \boldsymbol{\theta}_{\neq\mu}, \boldsymbol{\mu}, \boldsymbol{\Omega})$ with respect to $\boldsymbol{\mu}$, which is equivalent to maximizing $p(\mathbf{y} \mid \boldsymbol{\theta}_{\neq\mu}, \boldsymbol{\mu}, \boldsymbol{\Omega})$. We can do this as follows:

$$\frac{\partial L_i}{\partial \boldsymbol{\mu}_i} = \tag{1.26}$$

$$\frac{\partial -\log\left(\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}\right)}{\partial \boldsymbol{\mu}_i} = \tag{1.27}$$

$$\frac{-\partial\left(\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}\right) / \partial \boldsymbol{\mu}}{\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}} = \frac{\int_{-\infty}^{\infty}\left[-\partial p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) / \partial \boldsymbol{\mu}_i\right] d\boldsymbol{\phi}}{\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}} \tag{1.28}$$

$$\frac{\int_{-\infty}^{\infty}\left[\partial -\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega})) / \partial \boldsymbol{\mu}\right] p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}}{\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}} = \tag{1.29}$$

$$\int_{-\infty}^{\infty}\left[\frac{\partial -\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\partial \boldsymbol{\mu}_i}\right] z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} = \tag{1.30}$$

$$E_{\boldsymbol{\phi},i}\left(\frac{\partial -\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\partial \boldsymbol{\mu}_i}\right) = \mathbf{g}_{\boldsymbol{\mu}_i} \tag{1.31}$$

where $\mathbf{g}_{\boldsymbol{\mu}i}$ is the gradient with respect to $\boldsymbol{\mu}_i$, and $E_{\boldsymbol{\phi},i}(\ )$ represents the expected value when integrating over all $\boldsymbol{\phi}$, and

$$z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) = \frac{p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega})}{\int_{-\infty}^{\infty} p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi}} \tag{1.32}$$

is called the conditional density of $\boldsymbol{\phi}$ for individual *i*. The conditional density integrated over all possible $\boldsymbol{\phi}$ evaluates to 1:

$$z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}) = \int_{-\infty}^{\infty} \frac{p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\mu}_i,\boldsymbol{\Omega})}{\int_{-\infty}^{\infty} p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi}}\,d\boldsymbol{\phi} = 1 \tag{1.33}$$

The relationship

$$\frac{\partial -\log(p(\mathbf{y}_i\,|\,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\mu}_i} = E_{\boldsymbol{\phi},i}\left(\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\mu}_i}\right) \tag{1.34}$$

holds for any joint density $p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\theta}_{\neq}\boldsymbol{\mu}_i,\boldsymbol{\Omega})$. Now, to evaluate specifically for a parameter density *h* that is multivariate normal:

$$\int_{-\infty}^{\infty}\left[-\frac{\partial\log(p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\mu}_i}\right] z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} = \tag{1.35}$$

$$\int_{-\infty}^{\infty}\left[-\frac{\partial\log(l(\mathbf{y}_i\,|\,\boldsymbol{\phi},\boldsymbol{\theta}_{\neq})h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\mu}_i}\right] z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} = \tag{1.36}$$

$$\int_{-\infty}^{\infty} -\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}_i)z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} = \tag{1.37}$$

$$-\boldsymbol{\Omega}^{-1}\int_{-\infty}^{\infty}(\boldsymbol{\phi}-\boldsymbol{\mu}_i)z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} = \mathbf{g}_{\boldsymbol{\mu}_i} \tag{1.38}$$

We can perform the above algebraic manipulation because $\boldsymbol{\mu}$ (and therefore $\boldsymbol{\theta}_{\mu}$) appears only in the parameter density *h*, but does not appear in the data density *l*. We define

$$\bar{\boldsymbol{\phi}}_i = \int_{-\infty}^{\infty}\boldsymbol{\phi}z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} \tag{1.39}$$

as the conditional mean $\boldsymbol{\phi}$ vector for individual *i*, so that

$$\frac{\partial L_i}{\partial\boldsymbol{\mu}_i} = -\boldsymbol{\Omega}^{-1}\int_{-\infty}^{\infty}(\boldsymbol{\phi}-\boldsymbol{\mu}_i)z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} =$$

$$-\boldsymbol{\Omega}^{-1}[\int_{-\infty}^{\infty}\boldsymbol{\phi}z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} - \boldsymbol{\mu}_i\int_{-\infty}^{\infty}z(\boldsymbol{\phi}\,|\,\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi}] =$$

$$-\boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) = \mathbf{g}_{\boldsymbol{\mu}_i} \tag{1.40}$$

There are several ways of determining $\bar{\boldsymbol{\phi}}_i$ which are described later, and are called the expectation (=integrating or averaging) step.

Maximization

To determine the $\boldsymbol{\mu}$ modeled theta that reduces the objective function, we must solve:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} = -\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} \boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) = \mathbf{g}_{\boldsymbol{\theta}_\mu} \tag{1.41}$$

So that

$$\mathbf{g}_{\boldsymbol{\theta}_\mu} = \mathbf{0} \tag{1.42}$$

To evaluate (1.41) fully, an optimization algorithm is necessary which varies $\boldsymbol{\theta}_\mu$, and evaluating *L* at each $\boldsymbol{\theta}_\mu$. Keep in mind that in addition to $\boldsymbol{\mu}_i$ varying with $\boldsymbol{\theta}_\mu$, $\bar{\boldsymbol{\phi}}_i$ also varies with $\boldsymbol{\theta}_\mu$ through the conditional density *z*, so this minimization process can be computationally expensive. Alternatively, we can perform a limited maximization step in which $\bar{\boldsymbol{\phi}}_i$ is kept constant, while only $\boldsymbol{\mu}_i$ is varied with changes in $\boldsymbol{\theta}_\mu$. This separation of the expectation step from the maximization step is characteristic of the EM algorithm.

Evaluating (1.41) by this limited optimization is equivalent to minimizing the following surrogate objective function (keeping $\bar{\boldsymbol{\phi}}_i$ constant):

$$E_{\boldsymbol{\phi},i}\left(-\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega}))\right) =$$

$$\frac{1}{2} E_{\boldsymbol{\phi},i}\left(\sum_{i=1}^{m} (\boldsymbol{\phi} - \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu))' \boldsymbol{\Omega}^{-1} (\boldsymbol{\phi} - \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu))\right) + \frac{1}{2} E_{\boldsymbol{\phi},i}\left(m \log(\det(\boldsymbol{\Omega}))\right) =$$

$$= \frac{1}{2} \sum_{i=1}^{m} \left[\overline{\boldsymbol{\phi}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{\phi}_i} - 2\boldsymbol{\mu}_i' \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{\phi}}_i + \boldsymbol{\mu}' \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}_i \right] + \frac{1}{2} m \log(\det(\boldsymbol{\Omega}) =$$

$$= \frac{1}{2} \sum_{i=1}^{m} \left[\bar{\boldsymbol{\phi}}_i' \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{\phi}}_i - 2\boldsymbol{\mu}_i' \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{\phi}}_i + \boldsymbol{\mu}' \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}_i \right] + \frac{1}{2} m \log(\det(\boldsymbol{\Omega}) + \frac{1}{2} \sum_{i=1}^{m} \left[\overline{\boldsymbol{\phi}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{\phi}_i} - \bar{\boldsymbol{\phi}}_i' \boldsymbol{\Omega}^{-1} \bar{\boldsymbol{\phi}}_i \right] \propto$$

$$\frac{1}{2} \sum_{i=1}^{m} (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}}_\mu))' \boldsymbol{\Omega}^{-1} (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}}_\mu)) = L_c \tag{1.43}$$

The L_c is called the (negative) complete data log likelihood, and it can be shown (see [4]) that any $\boldsymbol{\theta}_\mu$ that reduces L_c, will reduce *L* by an at least equivalent amount, or:

$$L(\boldsymbol{\theta}_\mu) - L(\hat{\boldsymbol{\theta}}_\mu) \geq L_c(\boldsymbol{\theta}_\mu) - L_c(\hat{\boldsymbol{\theta}}_\mu) \tag{1.44}$$

where $\hat{\boldsymbol{\theta}}_\mu$ is an improved value over the present value $\boldsymbol{\theta}_\mu$. That is, any improvement value $\hat{\boldsymbol{\theta}}_\mu$ that reduces L_c (where $\bar{\boldsymbol{\phi}}_i$ was kept constant), will also reduce *L* (in which $\bar{\boldsymbol{\phi}}_i$ varies with $\boldsymbol{\theta}_\mu$), by at least the same amount as it reduced L_c.

The easiest way to minimize L_c is to perform a least squares analysis, by producing the following positive definite Hessian matrix:

$$\mathbf{H}_{\boldsymbol{\theta}_\mu} = E\left(\frac{\partial^2 L_c}{\partial \boldsymbol{\theta}_\mu^{\,2}}\right) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} E\left(\frac{\partial^2 L_{ci}}{\partial \boldsymbol{\mu}_i^2}\right) \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \tag{1.45}$$

And performing the following update with a variable step size $\alpha \leq 1$:

$$\hat{\boldsymbol{\theta}}_\mu = \boldsymbol{\theta}_\mu + \alpha \mathbf{H}_{\boldsymbol{\theta}_\mu}^{-1} \mathbf{g}_{\boldsymbol{\theta}_\mu} \tag{1.46}$$

This is the maximization step of the EM algorithm. If all of the mu's have linear relationships with respect to theta, then the step size that minimizes L_c with respect to the mu's is α=1. However, if the mu's are not linearly related to thetas, then α must be adjusted to minimize L_c with respect to mu. This can be done by selecting a value α, evaluate $\hat{L}_c$ using the proposed $\hat{\boldsymbol{\theta}}_\mu$, and if $\hat{L}_c$ is not smaller than the present L evaluated at the present $\boldsymbol{\theta}_\mu$, try another value of α, etc. In NONMEM, α=1 is first selected, tested, and if necessary, α is reduced by geometrical decrements of square root of 2 until an $\hat{L}_c$ is found that is less than L_c. More elaborate search algorithms (such as conjugate gradient or variable metric methods) for thetas not linearly modeled with respect to mu could be used for the expectation-maximization methods, but no real time savings occurs in doing so for population analysis problems.

In the next iteration, the updated $\boldsymbol{\theta}_\mu$ are used to evaluate a new set of condional means $\bar{\boldsymbol{\phi}}_i$ in the expectation step, followed by a limited maximization step to update $\boldsymbol{\theta}_\mu$ again. By repeatedly performing the expectation step (1.39), and evaluating the maximization step as expressed in equations (1.40) through (1.46), the gradient $\mathbf{g}_{\boldsymbol{\theta}_\mu}$ becomes smaller, and estimates $\hat{\boldsymbol{\theta}}_\mu$ that maximize the marginal density (satisfy equation (1.42)) are eventually obtained [4].

Again, because $\boldsymbol{\mu}$ appeared only in the parameter density *h* as the mean to this multivariate normal density, and does not appear in the data density *l*, and the parameters $\boldsymbol{\theta}_\mu$ to be

estimated appear in the objective function only through $\boldsymbol{\mu}$, this allowed us to obtain a gradient evaluation with a simple construction as given in (1.41) For those $\boldsymbol{\theta}$ that are not expressed in the model through $\boldsymbol{\mu}$ the $\boldsymbol{\theta}$ may appear anywhere in the joint density. No shortcut evaluation can be made by maximizing just the parameter density portion. Thus, to optimize the population objective function in these $\boldsymbol{\theta}$ as well, we need to differentiate the entire joint density. Through a similar process as we showed in differentiating with respect to $\boldsymbol{\mu}$,

$$\frac{\partial L_i}{\partial \boldsymbol{\theta}_{\neq}} = \tag{1.47}$$

$$\int_{-\infty}^{\infty}\left[\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\mid\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\theta}_{\neq}}\right]z(\boldsymbol{\phi}\mid\mathbf{y}_i,\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi} = \tag{1.48}$$

$$E_{\boldsymbol{\phi},i}\left(\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\mid\boldsymbol{\theta}_{\neq},\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\boldsymbol{\theta}_{\neq}}\right)=\mathbf{g}_{\boldsymbol{\theta}_{\neq}i} \tag{1.49}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{\neq}}=\sum_{i=1}^{m}\frac{\partial L_i}{\partial\boldsymbol{\theta}_{\neq}}=\sum_{i=1}^{m}\mathbf{g}_{\boldsymbol{\theta}_{\neq}i}=\mathbf{g}_{\boldsymbol{\theta}_{\neq}} \tag{1.50}$$

A Hessian matrix may be constructed as follows:

$$\mathbf{H}_{\boldsymbol{\theta}_{\neq}}=\sum_{i=1}^{m}\mathbf{g}_{\boldsymbol{\theta}_{\neq}i}\mathbf{g}'_{\boldsymbol{\theta}_{\neq}i} \tag{1.51}$$

$$\hat{\boldsymbol{\theta}}_{\neq}=\boldsymbol{\theta}_{\neq}+\mathbf{H}^{-1}_{\boldsymbol{\theta}_{\neq}}\mathbf{g}_{\boldsymbol{\theta}_{\neq}} \tag{1.52}$$

To minimize the objective function with respect to the inter-subject variance parameters, we recognize that $\boldsymbol{\Omega}$ is symmetrical, and we must vary only the lower triangular portion of the matrix. Defining **A** as the lower triangular matrix of $\boldsymbol{\Omega}$, and minimizing with respect to **A**, we have

$$\frac{\partial -\log(p(\mathbf{y}\mid\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\mathbf{A}}=\sum_{i=1}^{m}E_{\boldsymbol{\phi},i}\left(\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\mid\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\mathbf{A}}\right)= \tag{1.53}$$

$$\sum_{i=1}^{m}\int_{-\infty}^{\infty}\left[-\frac{\partial\log(l(\mathbf{y}_i\mid\boldsymbol{\phi})h(\boldsymbol{\phi}\mid\boldsymbol{\mu}_i,\boldsymbol{\Omega}))}{\partial\mathbf{A}}\right]z(\boldsymbol{\phi}\mid\mathbf{y}_i,\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi}= \tag{1.54}$$

$$\sum_{i=1}^{m}\int_{-\infty}^{\infty}\text{Lower}\left[\mathbf{R}_i-\frac{1}{2}\text{diag}\left(\mathbf{R}_i\right)\right]z(\boldsymbol{\phi}\mid\mathbf{y}_i,\boldsymbol{\mu}_i,\boldsymbol{\Omega})d\boldsymbol{\phi}=\mathbf{g}_{\mathbf{A}} \tag{1.55}$$

where

$$\mathbf{R}_i = \boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega} - (\boldsymbol{\phi} - \boldsymbol{\mu}_i)(\boldsymbol{\phi} - \boldsymbol{\mu}_i)'\right)\boldsymbol{\Omega}^{-1} \tag{1.56}$$

and $\mathbf{g}_\mathbf{A}$ is the gradient with respect to **A**. The derivation from equation (1.54) to (1.55) requires evaluating partial derivatives of matrix components, the tools for which are derived in appendix **A**.

We define

$$\bar{\boldsymbol{\Omega}}_i = \int_{-\infty}^{\infty} (\boldsymbol{\phi} - \boldsymbol{\mu}_i)(\boldsymbol{\phi} - \boldsymbol{\mu}_i)' z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} = \mathbf{E}_{\boldsymbol{\phi},i}\left[(\boldsymbol{\phi} - \boldsymbol{\mu}_i)(\boldsymbol{\phi} - \boldsymbol{\mu}_i)'\right] \tag{1.57}$$

as the contribution to the evaluated population variance from each individual *i*. Then,

$$E_{\boldsymbol{\phi},i}(\mathbf{R}_i) = \boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega} - \bar{\boldsymbol{\Omega}}_i\right)\boldsymbol{\Omega}^{-1} \tag{1.58}$$

and

$$\text{Lower}\left[\sum_{i=1}^{m} E_{\boldsymbol{\phi},i}(\mathbf{R}_i) - \frac{1}{2}\text{diag}\left(\sum_{i=1}^{m} E_{\boldsymbol{\phi},i}(\mathbf{R}_i)\right)\right] = \mathbf{g}_\mathbf{A} = \mathbf{0} \tag{1.59}$$

is equivalent to solving for

$$\sum_{i=1}^{m} E_{\boldsymbol{\phi},i}(\mathbf{R}_i) = \boldsymbol{\Omega}^{-1}\left(m\boldsymbol{\Omega} - \sum_{i=1}^{m} \bar{\boldsymbol{\Omega}}_i\right)\boldsymbol{\Omega}^{-1} = \mathbf{0} \tag{1.60}$$

which suggests the following update for $\boldsymbol{\Omega}$:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m}\sum_{i=1}^{m} \bar{\boldsymbol{\Omega}}_i \tag{1.61}$$

Note for any given $\boldsymbol{\Omega}$

$$\mathbf{g}_\mathbf{A} = \text{Lower}\left[m\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}})\boldsymbol{\Omega}^{-1} - \frac{m}{2}\text{diag}\left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}})\boldsymbol{\Omega}^{-1}\right)\right] \tag{1.62}$$

Thus, with repeatedly evaluating the expectation step (1.57), and utilizing the result to evaluate the next estimate of the intersubject variance (maximization step (1.61), when the "output" $\hat{\boldsymbol{\Omega}}$ equals the "input" $\boldsymbol{\Omega}$, then the gradient $\mathbf{g}_\mathbf{A}$ is equal to **0**.

Note that equation (1.57) may be rearranged as follows, which will be useful later:

$$\bar{\boldsymbol{\Omega}}_i = (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' + \int_{-\infty}^{\infty} (\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_i)(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_i)' z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.63}$$

Defining

$$\bar{\mathbf{B}}_i = \int_{-\infty}^{\infty} (\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_i)(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_i)' z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.64}$$

as the conditional variance of θ for individual *i*, then

$$\bar{\boldsymbol{\Omega}}_i = (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)' + \bar{\mathbf{B}}_i \tag{1.65}$$

so that

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m}\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)' + \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{B}}_i \tag{1.66}$$

Thus, the update variance inter-subject variance is evaluated as the sum of the sample variance of the conditional means and the average conditional variance. To summarize, the EM algorithm consists of an expectation step evaluating conditional means $\bar{\boldsymbol{\phi}}_i$ and conditional variances $\bar{\mathbf{B}}_i$, keeping θ_μ and $\boldsymbol{\Omega}$ constant, followed by a limited maximization step to obtain updated θ_μ and $\boldsymbol{\Omega}$, keeping $\bar{\boldsymbol{\phi}}_i$ and $\bar{\mathbf{B}}_i$ constant.

Evaluating the Expectation step: Importance Sampling

One can evaluate the area under the joint density and the other integrals by Monte Carlo techniques. The advantage to these methods is that the actual mathematical expression of the integral is not necessary for its computation, and the precision to which the integral is evaluated depends on the number of random samples generated to evaluate the integral. One Monte Carlo method is to use a sampling function that approximates the joint density, from which one obtains sample values of η or ϕ.

One possible sampling function is the multivariate normal density that has mean at $\hat{\boldsymbol{\phi}}_i$ and variance of $(\boldsymbol{\Omega}^{-1} + \hat{\mathbf{S}}_i^{-1})^{-1}$ as described in the previous section. To get these values, therefore, one first maximizes the joint density with respect to ϕ (or η) as one would for a MAP estimation. The negative logarithm of the area of this sampling function is exactly L_{Ni} of equation (1.18). Thus, the purpose of the randomization method is to modify L_{Ni} to the extent that the joint density deviates from this sampling density. In practice, one may start with a sampling function that is somewhat larger, by multiplying the variance by a value γ>1: $\gamma(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1})^{-1}$. The area of this sampling function is then

$$E_i = -\log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i, \boldsymbol{\theta})) + \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) + \frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1})) - \frac{1}{2}n\log(\gamma) \quad (1.67)$$

where n is the number of $\boldsymbol{\phi}$ parameters to be integrated, since $\boldsymbol{\phi}$ is integrated to form L_i.

For the kth random sample selected from this sampling density, the parameter vectors $\boldsymbol{\phi}_{(k)}$ are used to evaluate the logarithm of the joint density at that position:

$$\log(\pi(\boldsymbol{\phi}_{(k)})) = \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}_{(k)}, \boldsymbol{\theta}_{\neq})h(\boldsymbol{\phi}_{(k)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \quad (1.68)$$

To evaluate the normalized log of the joint density, we subtract

$$\log(\pi(\hat{\boldsymbol{\phi}}_i)) = \log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i, \boldsymbol{\theta}_{\neq})h(\hat{\boldsymbol{\phi}}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \quad (1.69)$$

To obtain

$$\log(\pi(\boldsymbol{\phi}_{(k)})) - \log(\pi(\hat{\boldsymbol{\phi}}_i)) \quad (1.70)$$

so that this normalized log joint density is 0 at the mode $\hat{\boldsymbol{\phi}}_i$. We also evaluate the logarithm of the normalized sampling function (which is also equal to 0 at $\hat{\boldsymbol{\phi}}_i$),

$$\log(e_i(\boldsymbol{\phi}_{(k)})) = -\frac{1}{2}(\boldsymbol{\phi}_{(k)} - \hat{\boldsymbol{\phi}}_i)'(\boldsymbol{\Omega}^{-1} + \hat{\mathbf{S}}_i^{-1})(\boldsymbol{\phi}_{(k)} - \hat{\boldsymbol{\phi}}_i)/\gamma \quad (1.71)$$

The logarithm of the ratio between joint density and sampling density is then:

$$q_{(k)i} = \log(\pi(\boldsymbol{\phi}_{(k)})) - \log(\pi(\hat{\boldsymbol{\phi}}_i)) - \log(e_i(\boldsymbol{\phi}_{(k)})) \quad (1.72)$$

which evaluates to

$$q_{(k)i} = \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}_{(k)}, \boldsymbol{\theta}_{\neq})) - \log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i, \boldsymbol{\theta}_{\neq})) + \frac{1}{2}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) - \frac{1}{2}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i) + \frac{1}{2}(\boldsymbol{\phi}_{(k)} - \hat{\boldsymbol{\phi}}_i)'(\boldsymbol{\Omega}^{-1} + \hat{\mathbf{S}}_i^{-1})(\boldsymbol{\phi}_{(k)} - \hat{\boldsymbol{\phi}}_i)/\gamma \quad (1.73)$$

Its exponent is the probability of accepting this position by the joint density, relative to the sampling density, which we may consider as a weight:

$$u_{(k)i} = \exp(q_{(k)i}) \quad (1.74)$$

Thus the following fraction,

$$\psi_i = \frac{1}{r}\sum_{k=1}^{r} u_{(k)i} \tag{1.75}$$

represents the ratio of the area of the conditional density to the area of the sampling density. The r=ISAMPLE is the number of random samples selected for each individual. This fraction is now used to adjust the area of the sampling density E_i, which is known, to obtain the true area of the conditional density, which is unknown:

$$\begin{aligned} &L_i = E_i - \log(\psi_i) = \\ &-\log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i, \boldsymbol{\theta}_{\#})) + \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) + \frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1})) \\ &-\log(\psi_i \gamma^{n/2}) \end{aligned} \tag{1.76}$$

so that $-\log(\psi_i \gamma^{n/2})$ is the "correction factor" that the randomization method adds to our original area equation to improve its accuracy. In NONMEM, γ is continually adjusted so that ψ_i approximates IACCEPT, up to the limit of the boundaries of γ being between ISCALE_MIN and ISCALE_MAX (available in NONMEM 7.2).

The above derivation of sample weights and likelihood evaluation for importance sampling resulting in equations (1.73) and (1.76) was developed to demonstrate that they are based on general principles of obtaining integrals by Monte Carlo methods. These relationships can be simplified by moving all of the elements from E_i to $q_{(k)i}$, given that the components in E_i are constant for all random samples k, so that the use of $\exp(q_{(k)i})$ as a weight factor will not be affected. Furthermore, we may generalize for all sampling densities $\boldsymbol{\phi}_{(k)} \sim N(\boldsymbol{\mu}_{si}, \gamma_i \boldsymbol{\Omega}_{si})$, by substituting a general sampling density mean $\boldsymbol{\mu}_{si}$ in place of $\hat{\boldsymbol{\phi}}_i$, and a general sampling density variance $\gamma_i \boldsymbol{\Omega}_{si}$ in place of $\gamma\left(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1}\right)^{-1}$, so that we obtain:

$$\begin{aligned} &q_{(k)i} = \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}_{(k)}, \boldsymbol{\theta}_{\#})) - \frac{1}{2}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i) - \frac{1}{2}\log(\det(\boldsymbol{\Omega})) \\ &+ \frac{1}{2}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_{si})'(\gamma_i \boldsymbol{\Omega}_{si})^{-1}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_{si}) + \frac{1}{2}\log(\det(\gamma_i \boldsymbol{\Omega}_{si})) \end{aligned} \tag{1.77}$$

$$u_{(k)i} = \exp(q_{(k)i}) \tag{1.78}$$

$$\psi_i = \frac{1}{r}\sum_{k=1}^{r} u_{(k)i} \tag{1.79}$$

$$L_i = -\log(\psi_i) \tag{1.80}$$

The new and improved objective function is then

$$L = \sum_{i=1}^{m} L_i \tag{1.81}$$

With this technique, we can also evaluate an improved mean $\boldsymbol{\theta}$ and improved variance-covariance matrix. Letting

$$z_{(k)i} = \frac{u_{(k)i}}{\sum_{k=1}^{r} u_{(k)i}} \tag{1.82}$$

so that

$$\sum_{k=1}^{r} z_{(k)i} = 1 \tag{1.83}$$

then

$$\bar{\boldsymbol{\phi}}_i = \sum_{k=1}^{r} z_{(k)i} \boldsymbol{\phi}_{(k)i} \tag{1.84}$$

$$\bar{\mathbf{B}}_i = \sum_{k=1}^{r} z_{(k)i} (\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_i)(\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_i)' \tag{1.85}$$

(note also that these are means and variance about the means, as indicated by the line above the parameter). The update equations yield the following:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} = -\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} \boldsymbol{\Omega}^{-1} (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) = \mathbf{g}_{\boldsymbol{\theta}_\mu} = \mathbf{0} \tag{1.86}$$

$$\mathbf{H}_{\boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \tag{1.87}$$

The easiest way to maximize is to perform the following updates:

$$\hat{\boldsymbol{\theta}}_\mu = \boldsymbol{\theta}_\mu + \alpha \mathbf{H}_{\boldsymbol{\theta}_\mu}^{-1} \mathbf{g}_{\boldsymbol{\theta}_\mu} \tag{1.88}$$

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}}_\mu)$$

And, according to equations (1.66), (1.64) and (1.57),

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m}\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)(\bar{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)' + \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{B}}_i =$$
$$\frac{1}{m}\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)(\bar{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)' + \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{r_i} z_{(k)i}(\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_i)(\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_i)' = \tag{1.89}$$
$$\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{r_i} z_{(k)i}(\boldsymbol{\phi}_{(k)i} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{\phi}_{(k)i} - \hat{\boldsymbol{\mu}}_i)'$$

This is equivalent to performing summary statistics on all of the random samples among all of the individuals. Note that the normalized weights $z_{(k)i}$ defined in equation (1.82) are obtained from sampled evaluations under the joint density $l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_\mu)h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})$, and are therefore empirical evaluations of the conditional density of equation (1.32). As the number of samples approaches infinity ($r \to \infty$), equations (1.88) and (1.89) approach the exact evaluation of updates that are required, as expressed in equations (1.39) and (1.66).

For subsequent iterations, the Monte Carlo evaluated conditional mean and variances of the previous iteration for that subject may be used as the parameters to the sampling density. This multivariate density has mean at $\bar{\boldsymbol{\phi}}_{pi}$ and, and variance of $\bar{\mathbf{B}}_{pi}$, so we sample from $\boldsymbol{\phi}_{(k)} \sim N(\bar{\boldsymbol{\phi}}_{pi}, \gamma_i\bar{\mathbf{B}}_{pi})$ where subscript *p* refers to previous iteration, so the pertinent weighting function is:

$$q_{(k)i} = \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}_{(k)}, \boldsymbol{\theta}_{\#})) - \frac{1}{2}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}_{(k)} - \boldsymbol{\mu}_i) - \frac{1}{2}\log(\det(\boldsymbol{\Omega}))$$
$$+\frac{1}{2}(\boldsymbol{\phi}_{(k)} - \bar{\boldsymbol{\phi}}_{pi})'(\gamma_i\bar{\mathbf{B}}_{si})^{-1}(\boldsymbol{\phi}_{(k)} - \bar{\boldsymbol{\phi}}_{pi}) + \frac{1}{2}\log(\det(\gamma_i\bar{\mathbf{B}}_{si})) \tag{1.90}$$

Followed by

$$u_{(k)i} = \exp(q_{(k)i}) \tag{1.91}$$

$$\psi_i = \frac{1}{r}\sum_{k=1}^{r} u_{(k)i} \tag{1.92}$$

$$L_i = -\log(\psi_i) \tag{1.93}$$

and the additional computations are carried out as before. Whether the parameters to the proposal density are obtained from a MAP estimation, or from conditional means and variances determined from a previous iteration, depends on whether METHOD=IMP or

METHOD=IMPMAP is used, and the settings of MAPITER and MAPINTER (available in NONMEM 7.2).

For those $\boldsymbol{\theta}$ that are not expressed in the model through $\boldsymbol{\mu}$, the $\boldsymbol{\theta}$ may appear anywhere in the likelihood. To optimize the population objective function in these $\boldsymbol{\theta}$ as well, we need to perform a finite difference on the entire likelihood for each non-mu modeled $\theta_{\neq j}$ of $\boldsymbol{\theta}_{\neq}$

$$\frac{\partial L_i}{\partial \theta_{\neq j}} \approx \frac{L_i(\boldsymbol{\theta}_{\neq} + \Delta\theta_{\neq j}) - L_i(\boldsymbol{\theta}_{\neq})}{\Delta\theta_{\neq j}} = \tag{1.94}$$

$$\int_{-\infty}^{\infty} \left[\frac{-\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid (\boldsymbol{\theta}_{\neq} + \Delta\theta_{\neq j}), \boldsymbol{\mu}_i, \boldsymbol{\Omega})) + \log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\Delta\theta_{\neq j}} \right] z(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \approx \tag{1.95}$$

$$\sum_{k=1}^{r_i} z_{(k)i} \left[\frac{-\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid (\boldsymbol{\theta}_{\neq} + \Delta\theta_{\neq j}), \boldsymbol{\mu}_i, \boldsymbol{\Omega})) + \log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\Delta\theta_{\neq j}} \right] \approx \tag{1.96}$$

$$E_{\boldsymbol{\phi},i} \left(\frac{\partial - \log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\partial \theta_{\neq j}} \right) = g_{\theta_{\neq j} i} \tag{1.97}$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\neq i}} = \mathbf{g}_{\boldsymbol{\theta}_{\neq}} = \mathbf{0} \tag{1.98}$$

where $\mathbf{g}_{\boldsymbol{\theta}_{\neq i}}$ is the vector of all $g_{\theta_{\neq j} i}$ for which $\theta_{\neq j} \in \boldsymbol{\theta}_{\neq}$. A Hessian matrix may be constructed as follows:

$$\mathbf{H}_{\boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\neq i}} \mathbf{g}'_{\boldsymbol{\theta}_{\neq i}} \tag{1.99}$$

$$\hat{\boldsymbol{\theta}}_{\neq} = \boldsymbol{\theta}_{\neq} + \alpha \mathbf{H}^{-1}_{\boldsymbol{\theta}_{\neq}} \mathbf{g}_{\boldsymbol{\theta}_{\neq}} \tag{1.100}$$

We now consider the computational expense for importance sampling required to update mu modeled theta parameters versus non-mu modeled parameters. For complex PK/PD problems that use the numerical integration ($DES), the greatest computational expense is in evaluating the predictive function $\mathbf{f}_i(\mathbf{t}, \boldsymbol{\phi}_i, \boldsymbol{\theta}_{\neq})$. The evaluation of the individual objective function $-2\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}, \boldsymbol{\Omega})) = -2\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})))$, in particular the data likelihood portion $l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})$ requires evaluation of $\mathbf{f}_i(\mathbf{t}, \boldsymbol{\phi}_i, \boldsymbol{\theta}_{\neq})$ for every observed value

of subject *i*. In importance sampling, the individual likelihood is evaluated *r* times in the evaluation of the conditional means and variances, per subject per iteration, regardless of how many mu modeled parameters are to be evaluated, according to equation (1.84). Once the conditional means and variances are determined, the individual objective function is no longer needed to evaluate the update for these thetas, according to equations (1.86) and (1.88).

For non-MU modeled parameters, however, equations (1.96), (1.97),(1.98), (1.99), and (1.100) suggest that $n_{\neq} r$ individual objective function calls are required, where $n_{\neq}$ is the number of non-mu modeled parameters, one for each $\log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid (\boldsymbol{\theta}_{\neq} + \Delta\theta_{\neq j}), \boldsymbol{\mu}_i, \boldsymbol{\Omega}))$ evaluation.

There is a sub-class of non-mu modeled parameters for which some computation efficiency can be made, and these are the SIGMA parameters, or Sigma-like, theta parameters. Such parameters are not used in evaluating the predictive function $\mathbf{f}_i(\mathbf{t}, \boldsymbol{\phi}_i, \boldsymbol{\theta}_{\neq})$, the most computationally expensive component, but only in evaluating the residual variance $\mathbf{V}_i(\mathbf{f}_i, \boldsymbol{\theta}_{\neq})$, so NONMEM uses $\mathbf{f}_i(\mathbf{t}, \boldsymbol{\phi}_i, \boldsymbol{\theta}_{\neq})$ in evaluating $\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq}))$ as well as $\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq} + \Delta\theta_{\neq j}))$ during the finite difference step, and will not re-evaluate **f**:

$$-2\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq})) = [\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\phi})]' \mathbf{V}_i^{-1}(\mathbf{f}_i, \boldsymbol{\theta}_{\neq})[\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\phi})] + \log\left(\det\left(\mathbf{V}_i(\mathbf{f}_i, \boldsymbol{\theta}_{\neq})\right)\right) \quad (1.101)$$

$$-2\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\neq} + \Delta\theta_{j\neq})) = [\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\phi})]' \mathbf{V}_i^{-1}(\mathbf{f}_i, \boldsymbol{\theta}_{\neq} + \Delta\theta_{j\neq})[\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\phi})] + \log\left(\det\left(\mathbf{V}_i(\mathbf{f}_i, \boldsymbol{\theta}_{\neq} + \Delta\theta_{j\neq})\right)\right) \quad (1.102)$$

Note that for these parameters only the $\mathbf{V}_i(\mathbf{f}_i, \boldsymbol{\theta}_{\neq})$ has to be re-evaluated (as the Y value in the NONMEM control stream file), which is usually a simple algebraic relation. SIGMA parameters are automatically recognized by NONMEM as those for which it can make this short-cut. THETA parameters that are used only in evaluating the residual variance (in the evaluation of Y in the control stream file) but not, directly or indirectly, in evaluating the predictive function (in the evaluation of F in the control stream file), may be given an S

designation in the GRD setting of $EST, and only then will NONMEM utilize the short cut for evaluating its partial derivative.

Sigma parameters (but not Sigma-like THETA parameters) can be additionally updated efficiently by evaluating their partial derivative gradient contributions analytically, as given in Appendix F (available in NONMEM 7.2). However, if the user specifies that Sigma's GRD value with an N, then their partial derivatives are evaluated numerically by finite difference method.

In general therefore, it is best to mu model THETA parameters whenever possible, to take advantage of the efficiency available for EM methods, and to specify when thetas may be considered Sigma-like, or to take advantage of modeling residual variances via SIGMA parameters, as much as possible.

Direct Sampling (available in NONMEM 7.2)

Direct sampling is much less efficient than importance sampling, and can require 10000 to 300000 random samples per subject to properly obtain conditional means and variances. Direct sampling does not use an "importance" region sampling density, but creates samples $\boldsymbol{\phi}_{(k)}$ directly from the normal distribution population parameter density: $\boldsymbol{\phi}_{(k)} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Omega})$ (see [5]). The following weight is then associated with the sample, based on the appropriate substitutions into equation (1.77):

$$u_{(k)i} = l(\mathbf{y}_i \mid \boldsymbol{\phi}_{(k)}, \boldsymbol{\theta}_{\#}) \tag{1.103}$$

Conditional means and variances are obtained as shown earlier:

$$z_{(k)i} = \frac{u_{(k)i}}{\sum_{k=1}^{r_i} u_{(k)i}} \tag{1.104}$$

$$\bar{\boldsymbol{\phi}}_i = \sum_{k=1}^{r_i} z_{(k)i} \boldsymbol{\phi}_{(k)i} \tag{1.105}$$

$$\bar{\mathbf{B}}_i = \sum_{k=1}^{r_i} z_{(k)i} (\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_{Ri})(\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_{Ri})' \tag{1.106}$$

As with importance sampling, an average of weights is obtained,

$$\psi_i = \frac{1}{r_i}\sum_{k=1}^{r_i} u_{(k)i} \tag{1.107}$$

From which the integrated objective function is obtained

$$L_i = -\log(\psi_i) \tag{1.108}$$

Iterative Two Stage

Iterative two stage approximates the expectation step by using the conditional modes and approximate conditional variances that are evaluated during the MAP estimation method that is also used in FOCE or LAPLACE methods, as described earlier. We can consider an approximate update for mu modeled thetas that is applied in iterative two stage, by evaluating:

$$\frac{\partial L_i}{\partial \boldsymbol{\mu}_i} = -\boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) \approx (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) = -\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\eta}}_i = \mathbf{g}_{\boldsymbol{\mu}_i} = \mathbf{0} = \frac{\partial L_{Ai}}{\partial \boldsymbol{\mu}_i} \tag{1.109}$$

Where subscript A refers to "approximate". This is an approximation to that extent that the mode $\hat{\boldsymbol{\phi}}_i$

$$\hat{\boldsymbol{\phi}}_i = \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu) + \hat{\boldsymbol{\eta}}_i$$

approximates the mean $\bar{\boldsymbol{\phi}}_i$. Then as before,

$$\frac{\partial L_A}{\partial \boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial L_{Ai}}{\partial \boldsymbol{\mu}_i}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} = -\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu}\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\eta}}_i = \mathbf{g}_{\boldsymbol{\theta}_\mu} = \mathbf{0} \tag{1.110}$$

We then perform a Gauss-Newton update:

$$\mathbf{H}_{\boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\mu i}}\mathbf{g}'_{\boldsymbol{\theta}_{\mu i}} \tag{1.111}$$

$$\hat{\boldsymbol{\theta}}_\mu = \boldsymbol{\theta}_\mu + \alpha\mathbf{H}^{-1}_{\boldsymbol{\theta}_\mu}\mathbf{g}_{\boldsymbol{\theta}\mu} \tag{1.112}$$

Then, updating the mus:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}}_\mu)$$

Similarly, to update Omega, iterative two stage approximates update (1.66):

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m}\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)' + \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{B}}_i \approx \tag{1.113}$$

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)(\hat{\boldsymbol{\phi}}_i - \hat{\boldsymbol{\mu}}_i)' + \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{B}}_i \tag{1.114}$$

where

$$\hat{\mathbf{B}}_i = \left(\boldsymbol{\Omega}^{-1} + \hat{\mathbf{S}}_i^{-1}\right)^{-1} \tag{1.115}$$

is the approximate conditional variance evaluated during the FOCE or LAPLACE integration step.

The approximate optimization of the iterative two stage method is related to an approximate optimization of FOCE's or Laplace's L_N. To consider optimizing L_N for mu modeled thetas, we can conveniently rephrase equation (1.18) as

$$L_{Ni} = -\log(l(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i, \boldsymbol{\theta}_{\boldsymbol{\mu}})) + \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) + \frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i))) \tag{1.116}$$

Since $\hat{\boldsymbol{\phi}}_i$ is at the mode of the posterior density, then

$$\left(\frac{\partial -\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\boldsymbol{\mu}}))}{\partial \boldsymbol{\phi}}\right)_{\hat{\boldsymbol{\phi}}_i} + \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{1.117}$$

It follows that:

$$\frac{\partial L_N}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} = \sum_{i=1}^{m}\frac{\partial L_{Ni}}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} = \sum_{i=1}^{m}\left(\left(\frac{\partial -\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\boldsymbol{\mu}}))}{\partial \boldsymbol{\phi}}\right)_{\hat{\boldsymbol{\phi}}_i} + \frac{1}{2}\frac{\partial(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)}{\partial \hat{\boldsymbol{\phi}}_i}\right)\frac{\partial \hat{\boldsymbol{\phi}}_i}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} + -\sum_{i=1}^{m}\frac{\partial \boldsymbol{\mu}_i'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i)))}{\partial \hat{\boldsymbol{\phi}}_i}\frac{\partial \hat{\boldsymbol{\phi}}_i}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} = \tag{1.118}$$

$$-\sum_{i=1}^{m}\frac{\partial \boldsymbol{\mu}_i'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} + \frac{1}{2}\sum_{i=1}^{m}\frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\phi}}_i)))}{\partial \hat{\boldsymbol{\phi}}_i}\frac{\partial \hat{\boldsymbol{\phi}}_i}{\partial \boldsymbol{\theta}_{\boldsymbol{\mu}}} = \mathbf{0} \tag{1.119}$$

where the term in parentheses cancels because of equation (1.117). Comparing equation (1.119) with that of (1.110) shows that iterative two stage only approximates the optimization of FOCE's L_N because it does not include the contribution of change in the information matrix of the joint density with respect to theta (the log(det) term in equation (1.119)).

Similarly, we consider differentiating FOCE's objective function with respect to OMEGA:

$$\frac{\partial L_N}{\partial \boldsymbol{\Omega}^{-1}} = \sum_{i=1}^{m} \frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}^{-1}} = \sum_{i=1}^{m} \left(\left(\frac{\partial - \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_\mu))}{\partial \boldsymbol{\phi}} \right)_{\hat{\boldsymbol{\phi}}_i} + \frac{1}{2} \frac{\partial (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Omega}^{-1} (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)}{\partial \hat{\boldsymbol{\phi}}_i} \right) \frac{\partial \hat{\boldsymbol{\phi}}_i}{\partial \boldsymbol{\Omega}^{-1}}$$

$$- \frac{1}{2} \sum_{i=1}^{m} \frac{\partial \log(\det(\boldsymbol{\Omega}^{-1})}{\partial \boldsymbol{\Omega}^{-1}} + \frac{1}{2} \sum_{i=1}^{m} \frac{\partial (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Omega}^{-1} (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\Omega}^{-1}} + \frac{1}{2} \sum_{i=1}^{m} \frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta}))}{\partial \boldsymbol{\Omega}^{-1}} + \quad (1.120)$$

$$\frac{1}{2} \sum_{i=1}^{m} \frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta}))}{\partial \hat{\boldsymbol{\eta}}_i} \frac{\partial \hat{\boldsymbol{\eta}}_i}{\partial \boldsymbol{\Omega}^{-1}} =$$

$$\sum_{i=1}^{m} \left[-\frac{1}{2} \boldsymbol{\Omega} + \frac{1}{2} \hat{\boldsymbol{\eta}}_i \hat{\boldsymbol{\eta}}_i' + \frac{1}{2} \left(\boldsymbol{\Omega}^{-1} + \hat{\mathbf{S}}_i^{-1} \right)^{-1} \right] + \frac{1}{2} \sum_{i=1}^{m} \frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1})}{\partial \hat{\boldsymbol{\eta}}_i} \frac{\partial \hat{\boldsymbol{\eta}}_i}{\partial \boldsymbol{\Omega}^{-1}} = \mathbf{0} \quad (1.121)$$

Here we differentiate the objective function with respect to $\boldsymbol{\Omega}^{-1}$ for convenience. When the gradient with respect to $\boldsymbol{\Omega}^{-1}$ equals $\mathbf{0}$, then the gradient with respect to $\boldsymbol{\Omega}$ also equals $\mathbf{0}$. The details of the linear algebra manipulations leading to the last part of (1.121) are given in appendix A. Then we can express the exact minimization of L_N with respect to Omega as:

$$\frac{\partial L_N}{\partial \boldsymbol{\Omega}^{-1}} = \sum_{i=1}^{m} \left[-\frac{1}{2} \boldsymbol{\Omega} + \frac{1}{2} \hat{\boldsymbol{\eta}}_i \hat{\boldsymbol{\eta}}_i' + \frac{1}{2} \hat{\mathbf{B}}_i \right] + \frac{1}{2} \sum_{i=1}^{m} \frac{\partial \log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1})}{\partial \hat{\boldsymbol{\eta}}_i} \frac{\partial \hat{\boldsymbol{\eta}}_i}{\partial \boldsymbol{\Omega}^{-1}} = \mathbf{0} \quad (1.122)$$

Note that $\hat{\mathbf{B}}_i$ represents a linearized approximation to the true conditional variance $\bar{\mathbf{B}}_i$. We may consider an approximate partial gradient to L_N with respect to Omega as:

$$\frac{\partial L_N}{\partial \boldsymbol{\Omega}^{-1}} = \sum_{i=1}^{m} \frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}^{-1}} \approx \sum_{i=1}^{m} \left[-\frac{1}{2} \boldsymbol{\Omega} + \frac{1}{2} \hat{\boldsymbol{\eta}}_i \hat{\boldsymbol{\eta}}_i' + \frac{1}{2} \hat{\mathbf{B}}_i \right] = \mathbf{0} \quad (1.123)$$

Solving for the next estimate of Omega from equation (1.123):

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\eta} \boldsymbol{\eta}' + \frac{1}{m} \sum_{i=1}^{m} \hat{\mathbf{B}}_i \quad (1.124)$$

which is similar to the update of Omega in the iterative two stage algorithm (1.114). Again, the ITS method updates L_N only approximately, as it does not take into account the log(det) term in equation (1.122).

To summarize, in iterative two stage, $\boldsymbol{\theta}$ is updated here using the average of the modes of the individual joint densities, which serves only as an approximation to the more precise update of the average of the means of the $\boldsymbol{\phi}$ under the joint density, as dictated by the exact equation (1.39). If there is a skewness to each individual's joint density, such that the means tend to differ systematically from the modes, then the iterative two stage update may yield biased results.

For non-mu modeled theta, iterative two stage in NONMEM falls back on a forward difference evaluation of the full likelihood:

$$\frac{\partial L_N}{\partial \boldsymbol{\theta}_{\neq}} \approx \frac{L_N(\boldsymbol{\theta}_{\neq} + \Delta\boldsymbol{\theta}_{\neq}) - L_N(\boldsymbol{\theta}_{\neq})}{\Delta\boldsymbol{\theta}_{\neq}} = \mathbf{g}_{\boldsymbol{\theta}_{\neq}} \tag{1.125}$$

Followed by a single step Gauss-Newton update:

$$\mathbf{H}_{\boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\neq i}} \mathbf{g}'_{\boldsymbol{\theta}_{\neq i}} \tag{1.126}$$

$$\hat{\boldsymbol{\theta}}_{\neq} = \boldsymbol{\theta}_{\neq} + \mathbf{H}^{-1}_{\boldsymbol{\theta}_{\neq}} \mathbf{g}_{\boldsymbol{\theta}_{\neq}} \tag{1.127}$$

To summarize, the NONMEM FOCE method optimizes L_N , which is an approximation to the true objective function L, and iterative two stage further approximates the optimization of L_N.

All iterative update methods that rely on updating the population parameters $\boldsymbol{\theta}$ using the average to the individual estimates guarantee "centeredness" of the population parameters about the individual parameters by definition. However, because the FOCE NONMEM method uses a search algorithm on an approximate objective function, it does not guarantee centeredness. One can impose the "CENTERING" option to the estimation process in NONMEM, which then optimizes a modified objective function of equation (1.18):

$$L'_{Ni} \approx$$
$$-\log(l(\mathbf{y}_i \mid (\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}), \boldsymbol{\theta})) + \frac{1}{2}\log(\det(\boldsymbol{\Omega})) + \frac{1}{2}(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}})'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\eta}}_i - \hat{\boldsymbol{\eta}}) \tag{1.128}$$
$$+\frac{1}{2}\log(\det(\boldsymbol{\Omega}^{-1} + \mathbf{S}^{-1}(\mathbf{y}_i \mid \hat{\boldsymbol{\eta}}_i, \boldsymbol{\theta})))$$

where

$$\hat{\boldsymbol{\eta}} = \frac{1}{m}\sum_{i=1}^{m} \hat{\boldsymbol{\eta}}_i \tag{1.129}$$

to ensure statistical centering, although not exact centering.

The MCMC method of Expectation in SAEM

In Markov Monte Carlo sampling, used in the SAEM and BAYES methods, samples are generated from a larger variety of proposal densities than in importance sampling. As

implemented in NONMEM, for a given set of population parameters mu and omega, proposed parameters phi for each individual are generated by a three mode process. The following is based on references [6] and [7].

During mode 1, a vector of model parameters is generated from the following proposal density or kernel:

$$\log(k_1(\boldsymbol{\phi})) = \log(N(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = \log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = -\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i)' - \frac{1}{2}\log|\boldsymbol{\Omega}| \quad (1.130)$$

For the acceptance test, we need to evaluate the above density along with the following backward density, at the present $\boldsymbol{\phi}_i$ for subject *i*:

$$\log(k_1(\boldsymbol{\phi}_i)) = \log(N(\boldsymbol{\phi}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = \log(h(\boldsymbol{\phi}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = -\frac{1}{2}(\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}_i - \boldsymbol{\mu}_i)' - \frac{1}{2}\log|\boldsymbol{\Omega}| \quad (1.131)$$

Also, the joint density is evaluated at the present $\boldsymbol{\phi}_i$:

$$\log(\pi(\boldsymbol{\phi}_i)) = \log(p(\mathbf{y}_i, \boldsymbol{\phi}_i \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = \log(l(\mathbf{y}_i, \boldsymbol{\phi}_i \mid \boldsymbol{\theta}_{\neq})) + \log(h(\boldsymbol{\phi}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \quad (1.132)$$

And at the proposed $\boldsymbol{\phi}$

$$\log(\pi(\boldsymbol{\phi})) = \log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega})) = \log(l(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq})) + \log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \quad (1.133)$$

Then the test statistic is created:

$$t_1 = \log(\pi(\boldsymbol{\phi})) - \log(\pi(\boldsymbol{\phi}_i)) + \log(k_1(\boldsymbol{\phi}_i)) - \log(k_1(\boldsymbol{\phi})) = \log(l(\mathbf{y}_i, \boldsymbol{\phi}_i \mid \boldsymbol{\theta}_{\neq})) - \log(l(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq})) \quad (1.134)$$

A unform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_1 \quad (1.135)$$

then the proposed sample set $\boldsymbol{\phi}$ of parameters is accepted and becomes the new $\boldsymbol{\phi}_i$ for subject *i*.

Following mode 1 sampling, proposal kernel mode 1A sampling and testing is performed, in which a sample from one of the other subjects is randomly selected. It is assumed that the set of parameters among subjects is normally distributed with mean and variance of the present $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Thus, the statistic t_1 of equation (1.134) is used as the acceptance test. This method has limited use to assist certain subjects to find good parameter values by borrowing from their neighbors, in case the neighbors had obtained good values. This

mode should generally not be used, and can be inaccurate if not all subjects share the same $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, such as in covariate modeling. Alternatively, use mode 1A sampling at the beginning of an SAEM analysis for a few burn in iterations, then continue with a complete SAEM analysis with mode 1A sampling turned off, with more burn in and accumulated sampling iterations.

Following mode 1A sampling, proposal kernel mode 2 sampling and testing is performed, using the proposal density:

$$\log(k_2(\boldsymbol{\phi} \mid \boldsymbol{\phi}_i)) = \log(N(\boldsymbol{\phi} \mid \boldsymbol{\phi}_i, \mathbf{Z})) = -\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}_i)'(\mathbf{Z})^{-1}(\boldsymbol{\phi} - \boldsymbol{\phi}_i)' - \frac{1}{2}\log|\mathbf{Z}| \quad (1.136)$$

where $\boldsymbol{\phi}_i$ is the present set of parameters for individual *i* (it could have been the one accepted from the just completed mode 1 sampling), and where

$$\mathbf{Z} = \kappa\boldsymbol{\Omega}$$

which includes a scaling factor κ. This scaling factor is adjusted for each subject such that samples are accepted at a fractional rate ρ_M =IACCEPT. This scaling factor κ is similar to the scaling factor γ in importance sampling, and is also subject to the boundary values of ISCALE_MIN and ISCALE_MAX (available in NONMEM 7.2). The backward density is

$$\log(k_2(\boldsymbol{\phi}_i \mid \boldsymbol{\phi})) = \log(N(\boldsymbol{\phi}_i \mid \boldsymbol{\phi}, \mathbf{Z})) = -\frac{1}{2}(\boldsymbol{\phi}_i - \boldsymbol{\phi})'(\mathbf{Z})^{-1}(\boldsymbol{\phi}_i - \boldsymbol{\phi})' - \frac{1}{2}\log|\mathbf{Z}| = k_2(\boldsymbol{\phi} \mid \boldsymbol{\phi}_i) \quad (1.137)$$

so the test statistic is calculated as:

$$t_2 = \log(\pi(\boldsymbol{\phi})) - \log(\pi(\boldsymbol{\phi}_i)) + \log(k_2(\boldsymbol{\phi}_i \mid \boldsymbol{\phi})) - \log(k_2(\boldsymbol{\phi} \mid \boldsymbol{\phi}_i)) = \log(\pi(\boldsymbol{\phi})) - \log(\pi(\boldsymbol{\phi}_i)) \quad (1.138)$$

A uniform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_2 \quad (1.139)$$

then the proposed sample set $\boldsymbol{\phi}$ of parameters is accepted and becomes the new $\boldsymbol{\phi}_i$ for subject *i*.

For proposal kernel mode 3, each parameter of the vector $\boldsymbol{\phi}$ is sequentially sampled using the univariate density:

$$k_3(\phi_l \mid \phi_{li}) = \log(N(\phi_l \mid \phi_{li}, z_{ll}^{-1})) = -\frac{1}{2}(\phi_l - \phi_{li})'(z_{ll}^{-1})(\phi_l - \phi_{li})' + \frac{1}{2}\log|z_{ll}^{-1}| \quad (1.140)$$

where subscript *l* refers to the *l*th parameter, and z_{ll}^{-1} is the *l*th diagonal element of $\mathbf{Z}^{-1}$. The backward density is

$$k_3(\phi_l \mid \phi_{li}) = k_3(\phi_{li} \mid \phi_l) \tag{1.141}$$

so the test statistic is:

$$t_3 = \log(\pi(\boldsymbol{\phi}_l)) - \log(\pi(\boldsymbol{\phi}_i)) \tag{1.142}$$

Where $\boldsymbol{\phi}_l$ equals $\boldsymbol{\phi}_i$ but with the *l*th element replaced with ϕ_l:

$$\boldsymbol{\phi}_l = \{\phi_l, \boldsymbol{\phi}_{i,+}\}$$

Once a parameter is tested, the result contributes to the new $\boldsymbol{\phi}_i$ for the next parameter in the vector to be sampled.

The mode 1B kernel obtains samples using the individual conditional mean and individual conditional variance collected from previous iterations as proposal density (a type of importance sampling kernel for SAEM).

During the MCMC sampling process, the IACCEPT sets ρ_M, ISAMPLE_M1 determines the number of mode 1 samplings, followed by ISAMPLE_M1B samplings, followed by ISAMPLE_M1A samplings, followed by ISAMPLE_M2 mode 2 samplings, followed by ISAMPLE_M3 mode 3 samplings. The final parameter set $\boldsymbol{\phi}_i$ after the cycle of ISAMPLE_M1+ISAMPLE_M2+*p*ISAMPLE_M3 samplings (where p=number of elements in vector $\boldsymbol{\phi}_i$) serves as the results of one chain for subject *i*. During each iteration, *r*=ISAMPLE separate chains of vectors $\boldsymbol{\phi}_i$ may be collected. Then, as described with importance sampling, the following may be calculated:

$$\bar{\boldsymbol{\phi}}_l = \sum_{k=1}^{r} \boldsymbol{\phi}_{(k)i} \tag{1.143}$$

$$\bar{\mathbf{B}}_i = \sum_{k=1}^{r} (\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_l)(\boldsymbol{\phi}_{(k)i} - \bar{\boldsymbol{\phi}}_l)' \tag{1.144}$$

Note that the acceptance/rejection process assured that the collection of $\boldsymbol{\phi}_{(k)i}$ reflect the distribution of the desired conditional density, and weights *z* are not needed.

During the stochastic mode, the updates to the population parameters (both mu and non-mu modeled) are then performed as described in importance sampling. During the accumulation mode, update results from previous *k*-1 iterations are averaged into the updates of the present *k*th iteration.

For mu modeled theta, and Omegas, the conditional means and variances are accumulatively updated and saved as follows:

$$\bar{\boldsymbol{\phi}}_{iS_k} = \frac{k-1}{k}\bar{\boldsymbol{\phi}}_{iS_{k-1}} + \frac{1}{k}\bar{\boldsymbol{\phi}}_{ik}$$

$$\bar{\mathbf{S}}_{iS_k} = \frac{(k-1)}{k}\left(\bar{\mathbf{B}}_{iS_{k-1}} + \bar{\boldsymbol{\phi}}^2_{iS_{k-1}}\right) + \frac{1}{k}\left(\bar{\mathbf{B}}_{ik} + \bar{\boldsymbol{\phi}}^2_{ik}\right)$$

$$\bar{\mathbf{B}}_{iS_k} = \bar{\mathbf{S}}_{iS_k} - \bar{\boldsymbol{\phi}}^2_{iS_k}$$

followed by update of the main population parameters in the usual manner:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_\mu} = -\sum_{i=1}^{m}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu}\boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_{iS_k} - \boldsymbol{\mu}_i) = \mathbf{g}_{\boldsymbol{\theta}_\mu} = \mathbf{0} \tag{1.145}$$

$$\mathbf{H}_{\boldsymbol{\theta}_\mu} = \sum_{i=1}^{m}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}}\boldsymbol{\Omega}^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \tag{1.146}$$

$$\hat{\boldsymbol{\theta}}_\mu = \boldsymbol{\theta}_\mu + \mathbf{H}^{-1}_{\boldsymbol{\theta}_\mu}\mathbf{g}_{\boldsymbol{\theta}_\mu} \tag{1.147}$$

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\theta}}_\mu)$$

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m}\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_{iS_k} - \hat{\boldsymbol{\mu}}_i)(\bar{\boldsymbol{\phi}}_{iS_k} - \hat{\boldsymbol{\mu}}_i)' + \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{B}}_{iS_k} \tag{1.148}$$

For non-mu modeled theta, the thetas $\hat{\boldsymbol{\theta}}_{\neq k}$ of the present *k*th iteration are updated using equations (1.47)-(1.52) using the present iteration's sampling process, followed by:

$$\hat{\boldsymbol{\theta}}_{\neq S_k} = ((k-1)\hat{\boldsymbol{\theta}}_{\neq S_{k-1}} + \hat{\boldsymbol{\theta}}_{\neq k})/k \tag{1.149}$$

First derivative gradients of non-mu modeled theta are also accumulated (for use in first order approximations of standard errors, see Appendix C):

$$\mathbf{g}_{iS_k} = \frac{k-1}{k}\mathbf{g}_{iS_{k-1}} + \frac{1}{k}\mathbf{g}_{ik} \tag{1.150}$$

In general, the order of accuracy for the various methods is
Monte Carlo EM (IMP, SAEM, DIRECT) >Laplace>FOCEI>ITS.

Three Stage EM Analysis

There are times when one desires to use information from a previous analysis and incorporate it into the present analysis. This would be in the form of prior information for thetas and/or omegas. The principle on which this is based is as follows. Let $\boldsymbol{\theta}_{\boldsymbol{\theta}}$ be the priors to the thetas (theta priors, which could be estimates of theta from a previous analysis). Let the matrix $\boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}$ be the information matrix (which could be the theta portion of the inverse of the standard error variance matrix of a previous analysis) of the theta priors. Then $\boldsymbol{\Omega}_{\boldsymbol{\theta}}$ may be called variance to theta priors, or theta variance priors. Let $\boldsymbol{\Omega}_{\boldsymbol{\Omega}}$ be the prior to the omegas of the population inter-subject variance-covariance matrix, of dimension p (Omega priors, could be Omega estimates of a previous analysis), let ρ be the degrees of freedom of $\boldsymbol{\Omega}_{\boldsymbol{\Omega}}$ (degrees of freedom priors, could be the number of subjects in the previous analysis). The contribution to the objective function that incorporates this prior information is

$$L_p = -\log\left(N(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_{\boldsymbol{\theta}},\boldsymbol{\Omega}_{\boldsymbol{\theta}})\right) - \log\left(W^{-1}(\rho\boldsymbol{\Omega}\,|\,\boldsymbol{\Omega}_{\boldsymbol{\Omega}},\rho+p+1)\right) \tag{1.151}$$

And is then added to equation (1.11):

$$L = -\sum_{i=1}^{m}\log\left(\int_{-\infty}^{+\infty} p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega})d\boldsymbol{\phi}\right) + L_p \tag{1.152}$$

where

$$-\log(N(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_{\boldsymbol{\theta}},\boldsymbol{\Omega}_{\boldsymbol{\theta}})) = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\boldsymbol{\theta}})'\boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\boldsymbol{\theta}}) + \frac{1}{2}\log(\det(\boldsymbol{\Omega}_{\boldsymbol{\theta}})) \tag{1.153}$$

$$-\log(W^{-1}(\boldsymbol{\Omega}\,|\,\rho\boldsymbol{\Omega}_{\boldsymbol{\Omega}},d_W)) = \frac{1}{2}\left(\rho\,\mathrm{tr}(\boldsymbol{\Omega}_{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}) + (d_W-n-1)\ln\left(|\boldsymbol{\Omega}|\right) - d_W\left[\ln\left(|\boldsymbol{\Omega}_{\boldsymbol{\Omega}}|\right) + n\ln(\rho)\right]\right) \tag{1.154}$$

(not including constants) where n is the dimension of $\boldsymbol{\Omega}$. The degrees of freedom d_w will be described later. It follows that the partial derivatives to L contributed by this prior information are:

$$\frac{\partial L_P}{\partial\boldsymbol{\theta}} = -\boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\boldsymbol{\theta}}) \tag{1.155}$$

$$\frac{\partial L_P}{\partial\omega_{j_1 j_2}} = \rho c(j_1,j_2)\mathbf{I}'_{j_1}\boldsymbol{\Omega}^{-1}(z_W\boldsymbol{\Omega}-\boldsymbol{\Omega}_{\boldsymbol{\Omega}})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2} \tag{1.156}$$

where

$$c(j_1, j_2) = 1 \text{ for } j_1 \neq j_2 \\ = 1/2 \text{ for } j_1 = j_2 \quad (1.157)$$

and

$$z_W = (d_W - n - 1)/\rho \quad (1.158)$$

Which suggest the following updates. To determine the $\boldsymbol{\mu}$ modeled theta that minimize the objective function, we must solve adding the contribution from the prior:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_\mu} = -\sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_\mu} \boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) - \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}) = \mathbf{g}_{\boldsymbol{\theta}_\mu} \quad (1.159)$$

and

$$\mathbf{H}_{\boldsymbol{\theta}_\mu} = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} + \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1} \quad (1.160)$$

And performing the following update As before:

$$\hat{\boldsymbol{\theta}}_\mu = \boldsymbol{\theta}_\mu + \alpha \mathbf{H}_{\boldsymbol{\theta}_\mu}^{-1} \mathbf{g}_{\boldsymbol{\theta}_\mu} \quad (1.161)$$

For non-mu modeled thetas,

$$\frac{\partial L_i}{\partial \boldsymbol{\theta}_{\neq}} = E_{\boldsymbol{\phi},i}\left(\frac{\partial - \log(p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\neq}, \boldsymbol{\mu}_i, \boldsymbol{\Omega}))}{\partial \boldsymbol{\theta}_{\neq}} \right) = \mathbf{g}_{\boldsymbol{\theta}_{\neq} i} \quad (1.162)$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \boldsymbol{\theta}_{\neq}} - \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\neq i}} - \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\boldsymbol{\theta}}) = \mathbf{g}_{\boldsymbol{\theta}_{\neq}} = \mathbf{0} \quad (1.163)$$

$$\mathbf{H}_{\boldsymbol{\theta}_{\neq}} = \sum_{i=1}^{m} \mathbf{g}_{\boldsymbol{\theta}_{\neq i}} \mathbf{g}'_{\boldsymbol{\theta}_{\neq i}} + \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1} \quad (1.164)$$

$$\hat{\boldsymbol{\theta}}_{\neq} = \boldsymbol{\theta}_{\neq} + \mathbf{H}_{\boldsymbol{\theta}_{\neq}}^{-1} \mathbf{g}_{\boldsymbol{\theta}_{\neq}} \quad (1.165)$$

The inter-subject variances are updated as

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m + d_W - n - 1}\left[\sum_{i=1}^{m} (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' + \sum_{i=1}^{m} \bar{\mathbf{B}}_i + \rho \boldsymbol{\Omega}_H \right] \quad (1.166)$$

These were derived from setting partial derivatives of the objective function to 0 and using the appropriate "inverse" densities and particular degrees of freedom in the objective function.

For maximization methods (all methods except BAYES), the degrees of freedom to the inverse Wishart is selected as

$$d_W = \rho + n + 1 \tag{1.167}$$

so that the maximization of these densities leads to a centering about the prior inter-subject variances, weighted according to the number of subjects from that previous analysis, and with a denominator term of $m+\rho$, yielding an intuitive update. That is, the density whose mode is at $\boldsymbol{\Omega}_{\boldsymbol{\Omega}}^{-1}$ is $W^{-1}(\boldsymbol{\Omega} \mid \rho\boldsymbol{\Omega}_{\boldsymbol{\Omega}}, \rho + n + 1)$. We shall call this the "modal" or maximization version of adding the prior information. The density whose mean is at $\boldsymbol{\Omega}_{\boldsymbol{\Omega}}^{-1}$ is $W^{-1}(\boldsymbol{\Omega} \mid \rho\boldsymbol{\Omega}_{\boldsymbol{\Omega}}, \rho))$. A BAYES analysis is concerned with obtaining average population parameters rather than best fit or modal population parameters, so it utilizes the degrees of freedom

$$d_W = \rho \tag{1.168}$$

which we shall call the "mean" version of adding the prior information.

The above equations are also suggested by the sample distribution equations listed on page 341 of [8].

The priors to $\boldsymbol{\Sigma}$ are also inverse Wishart distributed with prior parameters ($\boldsymbol{\Sigma}_{\boldsymbol{\Sigma}}$, ρ_Σ) so similar relationships hold, as for $\boldsymbol{\Omega}$ priors. However, $\boldsymbol{\Sigma}$ parameters are embedded in the data likelihood portion of the total likelihood in a non-linear manner, so updates need to be performed by extending the first and second derivatives of the total likelihood with respect to $\boldsymbol{\Sigma}$, and then using them in the Gauss-Newton update process. With this in mind, we need to find the partial derivative of the prior portion of the objective function with respect to each of the cholesky elements to $\boldsymbol{\Sigma}$, since this is how we vary the parameters in $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Lambda}$ be the cholesky matrix to $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' \tag{1.169}$$

So,

$$L_{\boldsymbol{\Sigma}_\Sigma} = \frac{1}{2}\left(\rho_\Sigma \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (d_\Sigma - n_\Sigma - 1)\ln\left(|\boldsymbol{\Sigma}|\right) - d_\Sigma\left[\ln\left(|\boldsymbol{\Sigma}_{\boldsymbol{\Sigma}}|\right) + n_\Sigma \ln\left(\rho_\Sigma\right)\right]\right) \tag{1.170}$$

$$\frac{1}{2}\left(\rho_\Sigma \text{ tr}(\boldsymbol{\Sigma}_\Sigma \boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}) + (d_\Sigma - n_\Sigma - 1)\ln\left(\left|\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\right|\right) - d_\Sigma\left[\ln\left(\left|\boldsymbol{\Sigma}_\Sigma\right|\right) + n_\Sigma \ln\left(\rho_\Sigma\right)\right]\right) \tag{1.171}$$

It is the cholesky elements in $\boldsymbol{\Lambda}$ that are varied to optimize the likelihood, so,

$$\frac{\partial L_{\Sigma_\Sigma}}{\partial \boldsymbol{\Lambda}} = -\rho_\Sigma \boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_\Sigma \boldsymbol{\Lambda}'^{-1} + (d_\Sigma - n_\Sigma - 1)\boldsymbol{\Lambda}'^{-1} \tag{1.172}$$

$$\frac{\partial L_{\Sigma_\Sigma}}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_2}} = \rho_\Sigma \mathbf{I}'_{j_1}\boldsymbol{\Lambda}'^{-1}\mathbf{I}_{k_2j_2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_\Sigma\boldsymbol{\Lambda}'^{-1}\mathbf{I}_{k_1} + \rho_\Sigma \mathbf{I}'_{j_1}\boldsymbol{\Sigma}^{-1}\mathbf{I}_{j_2k_2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_\Sigma\boldsymbol{\Lambda}'^{-1}\mathbf{I}_{k_1} - \mathbf{I}'_{j_1}\frac{\partial L_{\Sigma_\Sigma}}{\partial \boldsymbol{\Lambda}}\mathbf{I}_{k_2j_2}\boldsymbol{\Lambda}'^{-1}\mathbf{I}_{k_1} \tag{1.173}$$

Where

$\mathbf{I}_j = 1$ for vector element j

$= 0$ otherwise

$\mathbf{I}_{jk} = 1$ for matrix element j,k

$= 0$ otherwise

Population Mixture Modeling

Sometimes the data may be derived from two or more sub-populations, as evidenced by a distribution of a parameter among the subjects that appears to be bi-modal, or skewed. For example, suppose the data is first fit with a simple one compartment model, with volume Vc and rate constant of elimination k10. A histogram analysis of the individual k10's suggests a bimodal or skewed distribution. However, none of the known binary covariates (gender, for example) explains this bimodality. Under these circumstances, one can specify the probability of an individual belonging to a sub-group, without insisting on the certainty of belonging to that particular sub-group.

Consider that we have N_m sub-populations. Then for each subject i, and for each sub-population j we have the probability

$$p(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) = \int_{-\infty}^{+\infty} \sum_{j=1}^{N_m} a_j \, p_j(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.174}$$

where

$$p_j(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) \tag{1.175}$$

is the density for sub-population model *j*, for subject *i*, and a_j is the probability of belonging to sub-population *j*. Then define

$$L_{ji} = -\log(\int_{-\infty}^{+\infty} p_j(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) d\boldsymbol{\phi}) \tag{1.176}$$

so the negative log-likelihood of an individual is:

$$L_i = -\log(\int_{-\infty}^{+\infty} \sum_{j=1}^{N_m} a_j p_j(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) d\boldsymbol{\phi}) = \tag{1.177}$$

$$-\log(\sum_{j=1}^{N_m} a_j \int_{-\infty}^{+\infty} p_j(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) d\boldsymbol{\phi}) = \tag{1.178}$$

$$-\log(\sum_{j=1}^{N_m} a_j \exp(-L_{ji})) \tag{1.179}$$

Consider that equations for updating the non-proportion (that is, non-*a*) population parameters $\mathbf{q} = \{\boldsymbol{\theta}, \boldsymbol{\Omega}\}$ are derived from obtaining the partial derivatives of the objective function *L*:

$$\frac{\partial L_i}{\partial \mathbf{q}} = \sum_{j=1}^{N_m} \frac{a_j \exp(-L_{ji}) \frac{\partial L_{ji}}{\partial \mathbf{q}}}{\sum_{k=1}^{N_m} a_k \exp(-L_{ki})} \tag{1.180}$$

or

$$\frac{\partial L_i}{\partial \mathbf{q}} = \sum_{j=1}^{N_m} \frac{a_j r_{ji} \frac{\partial L_{ji}}{\partial \mathbf{q}}}{\sum_{k=1}^{N_m} a_k r_{ki}} = \sum_{j=1}^{N_m} a_{ji} \frac{\partial L_{ji}}{\partial \mathbf{q}} \tag{1.181}$$

where

$$r_{ji} = \exp(-L_{ji}) \tag{1.182}$$

and

$$a_{ji} = \frac{a_j r_{ji}}{\sum_{k=1}^{N_m} a_k r_{ki}} \tag{1.183}$$

is the probability or weight for individual *i*, sub-population model *j*. As an example,

$$\frac{\partial L_{ji}}{\partial \boldsymbol{\mu}_i} = -\boldsymbol{\Omega}^{-1} \sum_{i=1}^{m} \int_{-\infty}^{\infty} (\boldsymbol{\phi} - \boldsymbol{\mu}_i) z_j(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} = \mathbf{g}_{\boldsymbol{\mu}_{ji}} \tag{1.184}$$

where

$$z_j(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) \tag{1.185}$$

is the conditional density for subject *i* modeled under sub-model *j*, then the appropriate conditional mean for subject *i* would be

$$\bar{\boldsymbol{\phi}}_i = \sum_{j=1}^{N_m} a_{ji} \bar{\boldsymbol{\phi}}_{ji} \tag{1.186}$$

where

$$\bar{\boldsymbol{\phi}}_{ji} = \int_{-\infty}^{\infty} \boldsymbol{\phi} z_j(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.187}$$

which are then used in the usual way to update the thetas.

Similarly:

$$\bar{\mathbf{B}}_i = \sum_{j=1}^{N_m} a_{ji} (\bar{\mathbf{B}}_{ji} + \bar{\boldsymbol{\phi}}_{ji} \bar{\boldsymbol{\phi}}_{ji}') \quad - \quad \bar{\boldsymbol{\phi}}_i \bar{\boldsymbol{\phi}}_i' \tag{1.188}$$

where

$$\bar{\mathbf{B}}_{ji} = \int_{-\infty}^{\infty} (\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_{ji})(\boldsymbol{\phi} - \bar{\boldsymbol{\phi}}_{ji})' z_j(\boldsymbol{\phi} \mid \mathbf{y}_i, \boldsymbol{\mu}, \boldsymbol{\Omega}) d\boldsymbol{\phi} \tag{1.189}$$

is the conditional variance for individual *i*, sub-model *j*, whereupon the update is the usual:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m} \sum_{i=1}^{m} (\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' + \frac{1}{m} \sum_{i=1}^{m} \bar{\mathbf{B}}_i \tag{1.190}$$

The weighted average of the other expectation results are also performed, using the same weightings. The L_{ji}, and therefore a_{ji}, is readily obtained during the expectation step as the objective function to subject *i*, under sub-population model *j*. In practice therefore, the expectation step is done N_m times for each individual, collecting the resulting conditional means, variances, and objective function values to each sub-model, and then performing the weighted average, as shown above.

A method in keeping with minimizing the total objective function would be to construct partial derivatives and partial second derivatives, where for each subject *i*:

$$\frac{\partial L_i}{\partial a_j} = -\frac{\exp(-L_{ji}) - \exp(-L_{N_m i})}{\sum_{k=1}^{N_m} a_k \exp(-L_{ki})} = -\frac{r_{ji} - r_{N_m i}}{\sum_{k=1}^{N_m} a_k r_{ki}} \tag{1.191}$$

$$-\frac{\partial^2 L_i}{\partial a_{j1}\partial a_{j2}} = \frac{\exp(-L_{j1i}) - \exp(-L_{N_m i})}{\sum_{k=1}^{N_m} a_k \exp(-L_{ki})} \frac{\exp(-L_{j2i}) - \exp(-L_{N_m i})}{\sum_{k=1}^{N_m} a_k \exp(-L_{ki})} = \frac{\partial L_i}{\partial a_{j1}} \frac{\partial L_i}{\partial a_{j2}} \tag{1.192}$$

since

$$a_{N_m} = 1 - \sum_{j=1}^{N_m - 1} a_j \tag{1.193}$$

Then, perform the usual Gauss-Newton update, where $\boldsymbol{\theta}_a$ are all thetas that model the sub-population proportions in the $MIX module:

$$\mathbf{g}_{\boldsymbol{\theta}_a} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \boldsymbol{\theta}_a} \tag{1.194}$$

$$\mathbf{H}_{\boldsymbol{\theta}\mathbf{a}} = \sum_{i=1}^{m} \frac{\partial L_i}{\partial \boldsymbol{\theta}_a} \frac{\partial L_i}{\partial \boldsymbol{\theta}_a}' = \sum_{i=1}^{m} g_{i\boldsymbol{\theta}_a} g'_{i\boldsymbol{\theta}_a} \tag{1.195}$$

$$\boldsymbol{\theta}_{anew} = \boldsymbol{\theta}_{aold} - \left(\mathbf{H}_{\boldsymbol{\theta}_a}\right)^{-1} \left(\mathbf{g}_{\boldsymbol{\theta}_a}\right) \tag{1.196}$$

MCMC Bayesian Analysis for Evaluating a Distribution of Population Parameters

The Markov chain Monte Carlo (MCMC) Bayesian analysis can be used to obtain many thousands of population parameter and variance parameters that represent the distribution according to their ability to fit the data. This information is similar to what is obtained by boot strap methods, and MCMC Bayesian analysis can be used in their place. The Bayesian analysis may be performed with or without including prior information, but it is recommended that there at least be prior information for OMEGA.

There are two main types of Bayesian analysis available in NONMEM. The most efficient is the Gibbs sampling method, and is used to create samples of thetas that are linearly modeled with respect to their mu's, and the inter-subject variances. This is performed in the manner of page 341 of [8]. Updating linearly modeled thetas (designated as $\boldsymbol{\theta}_{\mu_L}$) is done as follows. Use the EM update method to obtain estimates $\hat{\boldsymbol{\theta}}_{\mu_L}$:

$$\mathbf{H}_{\boldsymbol{\theta}_{\mu_L}} = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}_{\mu_L}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\theta}_{\mu_L}} + \boldsymbol{\Omega}_{\boldsymbol{\theta}_{\mu_L}}{}^{-1} \tag{1.197}$$

Followed by

$$\hat{\boldsymbol{\theta}}_{\mu_L} = \boldsymbol{\theta}_{\mu_L} + a\mathbf{H}^{-1}_{\boldsymbol{\theta}_{\mu_L}} \mathbf{g}_{\boldsymbol{\theta}_{\mu_L}} \tag{1.198}$$

Next, sample from the following conditional density:

$$[\boldsymbol{\theta}_{\mu_L} | .] \sim N\left(\hat{\boldsymbol{\theta}}_{\mu_L}, \mathbf{H}^{-1}_{\boldsymbol{\theta}_{\mu_L}}\right) \tag{1.199}$$

For the Omegas:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{m+\rho}\left[\sum_{i=1}^{m}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' + \sum_{i=1}^{m}\bar{\mathbf{B}}_i + \rho\boldsymbol{\Omega}_{\boldsymbol{\Omega}}\right] \tag{1.200}$$

Followed by sampling from an inverse Wishart density:

$$[\boldsymbol{\Omega}^{-1} | .] \sim W^{-1}\left(\left(\hat{\boldsymbol{\Omega}}\right)^{-1}, m+\rho\right) \tag{1.201}$$

A matrix with an inverse Wishart distribution of $m+\rho$ degrees of freedom could be constructed as follows. Create *k* vectors of normally distributed random samples:

$$\mathbf{x}_k \sim N(0,1) \tag{1.202}$$

Then construct

$$\mathbf{S}_{m+\rho} = \sum_{k=1}^{m+\rho}\mathbf{x}_k\mathbf{x}'_k \tag{1.203}$$

$$\boldsymbol{\Omega} = \mathbf{L}_{\hat{\boldsymbol{\Omega}}}\mathbf{S}^{-1}\mathbf{L}'_{\hat{\boldsymbol{\Omega}}} \tag{1.204}$$

Where $\mathbf{L}_{\hat{\boldsymbol{\Omega}}}$ is the cholesky of $\hat{\boldsymbol{\Omega}}$. More efficient methods of creating an inverse Wishart matrix sample are available. Because these sample densities are also the conditional densities for the respective parameters, the samples are always accepted, and no acceptance/rejection analysis needs to be performed.

Sigma parameters (but not Sigma-like THETA parameters) that are isolated residual variance coefficients are updated as follows:

$$\hat{\sigma}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m_i}\frac{(y_{ij} - f_{ij})^2}{(\partial y_{ij} / \partial \varepsilon)^2} \tag{1.205}$$

Followed by sampling from an inverse chi-square:

$$\sigma^{-2} = \chi^{-2}(N, \hat{\sigma}^{-2})$$

Where N is the total number of data points involved in evaluation of that particular sigma.

Metropolis-Hastings sampling must be performed on all other types of theta parameters, as follows. For the first mode, for thetas not linearly mu modeled and cholesky decomposed sigma elements, designated collectively as θ, proposed sample parameters for the *k+1*th iteration are created using

$$\log(N(\theta \mid \theta_0, \mathbf{Z})) = -\frac{1}{2}(\theta - \theta_0)'(\mathbf{Z})^{-1}(\theta - \theta_0) - \frac{1}{2}\log|\mathbf{Z}| \quad (1.206)$$

θ_0 and $\mathbf{Z}$ vary according to how many samples have so far been created. During the first several hundred iterations of burn-in, θ_0 are the initial thetas at the start of the analysis, and $\mathbf{Z}$ is a diagonal matrix with diagonal elements that are equal to $(0.5*\theta_0)^2$. During the subsequent iterations of burn-in, θ_0 and $\mathbf{Z}$ are the sample means and variances of θ collected during the previous several hundred iterations. During the stationary phase, θ_0 and $\mathbf{Z}$ are the sample means and variances of all θ collected so far since the beginning of the stationary phase.

To reflect the probability of choosing these values, the following log density values are therefore calculated, based on the respective proposal densities, for mode 1

$$\log(k_1(\theta \mid \theta_0)) = \log(N(\theta \mid \theta_0, \mathbf{Z}_0)) \quad (1.207)$$

The log likelihood of the *k*th set of population parameters with respect to the data, and with respect to positions of the *k*th set of individual parameters ϕ_{ik} is evaluated also:

$$\log(\pi(\theta_k)) = \sum_{i=1}^{m} \log(p(\mathbf{y}_i, \phi_{ik} \mid \theta_k, \mu_{ik}, \Omega_k)) \quad (1.208)$$

The log likelihood of the proposed sample set of population parameters with respect to the data, and with respect to positions of the present *k*th set of individual thetas is evaluated also:

$$\log(\pi(\theta)) = \sum_{i=1}^{m} \log(p(\mathbf{y}_i, \phi_{ik} \mid \theta, \mu_i, \Omega_k)) \quad (1.209)$$

The following test statistic is created:

$$t_1 = -\log(k_1(\theta \mid \theta_0)) + \log(k_1(\theta_k \mid \theta_0)) - \log(\pi(\theta_k)) + \log(\pi(\theta)) \quad (1.210)$$

A unform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_1 \quad (1.211)$$

then the proposed sample set of population parameters is accepted. If rejected, then the *k*th sample set is used as the *k*+1th sample set. This is done PSAMPLE_M1 (an option in NONMEM) times.

Next, during the second kernel density mode, the population parameters of the present position *k* may be used to create a sample for the next iteration:

$$\log(N(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_k,\mathbf{Z})) = -\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_k)'\left(w\mathbf{Z}\right)^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_k) - \frac{1}{2}\log\left|w\mathbf{Z}\right| \tag{1.212}$$

Where $\boldsymbol{\theta}_k$ is the accepted theta of the *k*th iteration, *w* is a scaling parameter, which is adjusted throughout the analysis so that a fraction PACCEPT (option) of random sample sets are accepted. The PACCEPT (option) parameter is set by the user.

To reflect the probability of choosing these values, the following log density values are therefore calculated, based on the respective proposal densities:

$$\log(k_2(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_k)) = \log(N(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_k, w\mathbf{Z})) \tag{1.213}$$

as well as their backward density of mode 2:

$$\log(k_2(\boldsymbol{\theta}_k\,|\,\boldsymbol{\theta})) = \log(N(\boldsymbol{\theta}_k\,|\,\boldsymbol{\theta}, w\mathbf{Z})) = \log(k_2(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_k)) \tag{1.214}$$

The test statistic is created:

$$t_2 = -\log(k_2(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_k)) + \log(k_1(\boldsymbol{\theta}_k\,|\,\boldsymbol{\theta}_k)) - \log(\pi(\boldsymbol{\theta}_k)) + \log(\pi(\boldsymbol{\theta})) \tag{1.215}$$

A uniform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_2 \tag{1.216}$$

Then the sample is accepted. This is done PSAMPLE_M2 times.

As a third kernel sampling mode, samples on each parameter separately and sequentially may be made using the univariate distribution

$$\log(N(\theta_l\,|\,\theta_{kl}, z_{ll}^{-1})) = -\frac{1}{2}(\theta_l-\theta_{kl})'z_{ll}^{-1}(\theta_l-\theta_{kl}) + \frac{1}{2}\log\left|z_{ll}^{-1}\right| \tag{1.217}$$

where z_{ll}^{-1} is the *ll*th diagonal element to $\mathbf{Z}^{-1}$, for parameter *l*. The other parameters are not moved when in this mode.

To reflect the probability of choosing these values, the following log density values are therefore calculated, based on the respective proposal densities:

$$\log(k_{3l}(\boldsymbol{\theta}_l \mid \boldsymbol{\theta}_{kl})) = \log(N(\boldsymbol{\theta}_l \mid \boldsymbol{\theta}_{kl}, w\mathbf{Z})) \tag{1.218}$$

and backward density of mode 3:

$$\log(k_{3l}(\boldsymbol{\theta}_{kl} \mid \boldsymbol{\theta}_l)) = \log(N(\boldsymbol{\theta}_{kl} \mid \boldsymbol{\theta}_l, w\mathbf{Z})) \tag{1.219}$$

The test statistic is created for each parameter *l*:

$$t_3 = -\log(k_{3l}(\boldsymbol{\theta} \mid \theta_{kl})) + \log(k_1(\theta_{kl} \mid \boldsymbol{\theta}_k)) - \log(\pi(\boldsymbol{\theta}_k)) + \log(\pi(\theta_{kl})) \tag{1.220}$$

A uniform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_3 \tag{1.221}$$

then the proposed sample set of population parameters is accepted as the *k*+1th sample set. If rejected, then the *k*th sample set is used as the *k*+1th sample set.

The third mode is done for each parameter PSAMPLE_M3 times , for *n**PSAMPLE_M3 times in a given iteration, where *n* is the number of population parameters in the vector $\boldsymbol{\theta}$.

If the user has selected to perform Metropolis-Hastings samplings for Omega elements, then for each time that samples of population mean parameters and covariates are created, samples of population variances are also created using the inverse Wishart distribution.

For mode 1, using the starting position values (*k*=0) (OSAMPLE_M1 times):

$$\log(W^{-1}(\boldsymbol{\Omega} \mid (\rho+m)\boldsymbol{\Omega}_0, (\rho+m))) = \\ -\frac{1}{2}\left((\rho+m)tr(\boldsymbol{\Omega}_0\boldsymbol{\Omega}^{-1}) + (\rho+m-n-1)\ln\left(|\boldsymbol{\Omega}|\right) - (\rho+m)\left[\ln\left(|\boldsymbol{\Omega}_0|\right) + n\ln\left(\rho+m\right)\right]\right) \tag{1.222}$$

To reflect the probability of choosing these values, the following log density values are therefore calculated, based on the respective proposal densities:

$$\log(k_1(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}_0)) = \log(W^{-1}(\boldsymbol{\Omega} \mid (\rho+m)\boldsymbol{\Omega}_0, (\rho+m))) \tag{1.223}$$

The log likelihood of the *k* set of population parameters with respect to the data, and with respect to positions of the *k* set of individual parameters $\boldsymbol{\phi}_{ik}$ is evaluated also:

$$\log(\pi(\boldsymbol{\Omega}_k)) = \sum_{i=1}^{m} \log(p(\mathbf{y}_i, \boldsymbol{\phi}_{ik} \mid \boldsymbol{\theta}_k, \boldsymbol{\mu}_{ik}, \boldsymbol{\Omega}_k)) \tag{1.224}$$

The log likelihood of the proposed sample set of population variances with respect to the data, and with respect to positions of the present *k*th set of individual thetas is evaluated also:

$$\log(\pi(\boldsymbol{\Omega})) = \sum_{i=1}^{m} \log(p(\mathbf{y}_i, \boldsymbol{\phi}_{ik} \mid \boldsymbol{\theta}_k, \boldsymbol{\mu}_i, \boldsymbol{\Omega})) \tag{1.225}$$

During mode 1, the following test statistic is created:

$$t_1 = -\log(k_1(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}_0)) + \log(k_1(\boldsymbol{\Omega}_k \mid \boldsymbol{\Omega}_0)) - \log(\pi(\boldsymbol{\Omega}_k)) + \log(\pi(\boldsymbol{\Omega})) \tag{1.226}$$

A uniform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_1 \tag{1.227}$$

then the proposed sample set of variances is accepted as the *k*+1th sample set. If rejected, then the *k*th sample set is used as the *k*+1th sample set.

For mode 2, the present position *k* is used (OSAMPLE_M2 times):

$$\log(W^{-1}(\boldsymbol{\Omega} \mid w(\rho+m)\boldsymbol{\Omega}_k, w(\rho+m))) =$$
$$-\frac{1}{2}\left(w(\rho+m)tr(\boldsymbol{\Omega}_k\boldsymbol{\Omega}^{-1}) + (w(\rho+m)-n-1)\ln\left(|\boldsymbol{\Omega}|\right) - (w(\rho+m))\left[\ln\left(|\boldsymbol{\Omega}_k|\right) - n\ln\left(w(\rho+m)\right)\right]\right) \tag{1.228}$$

where *w* is the scaling parameter (separate from that used for the normal distribution proposal density for the theta parameters) to allow OACCEPT acceptance rate.

To reflect the probability of choosing these values, the following log density values are therefore calculated, based on the respective proposal densities:

$$\log(k_2(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}_k)) = \log(W^{-1}(\boldsymbol{\Omega} \mid w(\rho+m)\boldsymbol{\Omega}_k, w(\rho+m))) \tag{1.229}$$

as well as their backward density of mode 2:

$$\log(k_2(\boldsymbol{\Omega}_k \mid \boldsymbol{\Omega})) = \log(W^{-1}(\boldsymbol{\Omega}_k \mid w(\rho+m)\boldsymbol{\Omega}, w(\rho+m))) \tag{1.230}$$

The test statistic is created:

$$t_2 = -\log(k_2(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}_k)) + \log(k_1(\boldsymbol{\Omega}_k \mid \boldsymbol{\Omega}_k)) - \log(\pi(\boldsymbol{\Omega}_k)) + \log(\pi(\boldsymbol{\Omega})) \tag{1.231}$$

A uniform random deviate *u* is then generated, log transformed, and if

$$\log(u) < t_2 \tag{1.232}$$

then the proposed sample set of variances is accepted, and serves as the *k*+1th sample set. If rejected, then the *k*th sample set is used as the *k*+1th sample set. This is done OSAMPLE_M2 times.

A single iteration consists of: Gibbs sampling of THETAS, SIGMAS and OMEGAS, followed by Metropolis-Hastings sampling of other THETAS, PSAMPLE_M1 times for mode 1, then PSAMPLE_M2 times for mode 2, followed by OMEGAS sampled OSAMPLE_M1 times for mode 1, then OSAMPLE_M2 times for mode 2. The final sample set of THETAS, OMEGAS and SIGMAS after going through this process is then stored in the raw output file as the results to that particular iteration.

Conditional Weighted Residuals

We consider the following linear-epsilon residual error model:

$$y_i = f_i(\boldsymbol{\eta}) + \sum_{m=1}^{M} q'_{im}(\boldsymbol{\eta})\varepsilon_m \tag{2.1}$$

for data points i=1 to N of a particular subject, which takes into account intra-subject error components within a subject (such as homoscedastic eps(1) mixed with heteroscedastic eps(2)) plus possible intra-individual error interaction with other data points, as well as the possibility that q_{im} depends on $\boldsymbol{\eta}$. We define the Mx1 normally distributed random vector:

$$\boldsymbol{\varepsilon} = \left\{\varepsilon_m,\ m = 1 \text{ to } M\right\} \tag{2.2}$$

with the properties:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \tag{2.3}$$

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{I}$$

$$E(\varepsilon_i \eta_k) = E(\varepsilon_i)E(\eta_k) = 0 \tag{2.4}$$

and the NxN matrix

$$\mathbf{Q}' = \left\{q'_{im}, i = 1 \text{ to } N, m = 1 \text{ to } M\right\} \tag{2.5}$$

with 1xM row vectors

$$\mathbf{q}'_i = \left\{q'_{im}, m = 1 \text{ to } M\right\} \tag{2.6}$$

Then, for a given $\boldsymbol{\eta}$, the expected value over all epsilon is

$$E_\varepsilon(y_i) = E(f_i(\boldsymbol{\eta})) + E(\sum_{m=1}^{M} q'_{im}(\boldsymbol{\eta})\varepsilon_m) = f_i(\boldsymbol{\eta}) + \sum_{m=1}^{M} q'_{im}(\boldsymbol{\eta})E(\varepsilon_m) = f_i(\boldsymbol{\eta}) \tag{2.7}$$

and

$$\text{var}(y_i y_j) = E_\varepsilon((y_i - f_i(\boldsymbol{\eta})(y_j - f_j(\boldsymbol{\eta})) =$$

$$E_\varepsilon\left(\left(\sum_{m=1}^{M} q'_{im}(\boldsymbol{\eta})\varepsilon_m\right)\left(\sum_{k=1}^{M} q'_{jk}(\boldsymbol{\eta})\varepsilon_k\right)\right) = E_\varepsilon\left(\left(\sum_{m=1}^{M}\sum_{k=1}^{M} q'_{im}(\boldsymbol{\eta})\varepsilon_m q'_{jk}(\boldsymbol{\eta})\varepsilon_k\right)\right) = \tag{2.8}$$

$$\sum_{m=1}^{M}\sum_{k=1}^{M} q'_{im}(\boldsymbol{\eta})q'_{jk}(\boldsymbol{\eta})E(\varepsilon_m\varepsilon_k) = \sum_{m=1}^{M} q'_{im}q'_{jm} = \sum_{m=1}^{M} q'_{im}q_{mj}$$

or

$$\text{var}(\mathbf{y}) = \mathbf{Q}'\mathbf{Q} = \mathbf{V} \tag{2.9}$$

To integrate over all $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ and have an analytical solution, first define

$$w_{i(mk)} = \frac{\partial q_{mi}}{\partial \eta_k} \tag{2.10}$$

as the series of *M*x*n* matrices

$$\mathbf{W}_i = \left\{ w_{i(mk)}, m = 1 \text{ to } M, k = 1 \text{ to } n \right\} \tag{2.11}$$

and the *N*x*n* matrix

$$\mathbf{G} = \left\{ g_{ik} = \frac{\partial f_i(\boldsymbol{\eta})}{\partial \eta_k}, i = 1 \text{ to } N, k = 1 \text{ to } n \right\} \tag{2.12}$$

with 1x*n* row vectors

$$\mathbf{g}_i' = \left\{ g_{ik} = \frac{\partial f_i(\boldsymbol{\eta})}{\partial \eta_k}, k = 1 \text{ to } n \right\} \tag{2.13}$$

We now linearize by Taylor series expansion about $\boldsymbol{\eta}'$ as follows:

$$y_i = f_i(\hat{\boldsymbol{\eta}}) + \mathbf{g}_i'\boldsymbol{\eta} - \mathbf{g}_i'\hat{\boldsymbol{\eta}} + (\mathbf{q}_i' - \hat{\boldsymbol{\eta}}'\mathbf{W}_i')\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\mathbf{W}_i'\boldsymbol{\varepsilon} \tag{2.14}$$

If we now integrate over all $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$, we have the marginal density of **y** with mean

$$E_{\eta,\varepsilon}(y_i) = f_i(\hat{\boldsymbol{\eta}}) - \mathbf{g}_i'\hat{\boldsymbol{\eta}} \tag{2.15}$$

and variance

$$Var(y_i y_j) = E((y_i - f_i(\hat{\boldsymbol{\eta}}) + \mathbf{g}_i'\hat{\boldsymbol{\eta}})(y_j - f_j(\hat{\boldsymbol{\eta}}) + \mathbf{g}_j'\hat{\boldsymbol{\eta}})) = \tag{2.16}$$

$$\begin{aligned} &E_{\eta,\varepsilon}((\mathbf{g}_i'\boldsymbol{\eta} + (\mathbf{q}_i' - \hat{\boldsymbol{\eta}}'\mathbf{W}_i')\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\mathbf{W}_i'\boldsymbol{\varepsilon})(\mathbf{g}_j'\boldsymbol{\eta} + (\mathbf{q}_j' - \hat{\boldsymbol{\eta}}'\mathbf{W}_j')\boldsymbol{\varepsilon} + \boldsymbol{\eta}'\mathbf{W}_j'\boldsymbol{\varepsilon})') = \\ &\mathbf{g}_i' E_\eta(\boldsymbol{\eta}\boldsymbol{\eta}')\mathbf{g}_j + (\mathbf{q}_i' - \hat{\boldsymbol{\eta}}'\mathbf{W}_i')E_\varepsilon(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')(\mathbf{q}_j - \mathbf{W}_j\hat{\boldsymbol{\eta}}) + E_\eta(\boldsymbol{\eta}'\mathbf{W}_i' E_\varepsilon(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{W}_j\boldsymbol{\eta}) \end{aligned} \tag{2.17}$$

But

$$\begin{aligned} &E(\boldsymbol{\eta}'\mathbf{W}_i' E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{W}_j\boldsymbol{\eta}) = E(\boldsymbol{\eta}'\mathbf{W}_i'\mathbf{W}_j\boldsymbol{\eta}) = \\ &E(\sum_{m=1}^{n}\sum_{l=1}^{M}\sum_{k=1}^{n}\eta_k w_{i(kl)}' w_{j(lm)}\eta_m) = E(\sum_{m=1}^{n}\sum_{l=1}^{M}\sum_{k=1}^{n}\eta_k w_{i(lk)} w_{j(lm)}\eta_m) = \\ &\sum_{m=1}^{n}\sum_{l=1}^{M}\sum_{k=1}^{n} w_{i(lk)} E(\eta_k\eta_m) w_{j(lm)} = \sum_{l=1}^{M}\sum_{k=1}^{n}\sum_{m=1}^{n} w_{i(lk)}\omega_{km} w_{j(lm)} = \\ &\text{tr}(\mathbf{W}_i\boldsymbol{\Omega}\mathbf{W}_j') = \text{tr}(\boldsymbol{\Omega}\mathbf{W}_i'\mathbf{W}_j) \end{aligned} \tag{2.18}$$

And similarly,

$$\hat{\boldsymbol{\eta}}'\mathbf{W}_i'\mathbf{W}_j\hat{\boldsymbol{\eta}} = \text{tr}(\mathbf{W}_i\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}'\mathbf{W}_j') = \text{tr}(\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}'\mathbf{W}_i'\mathbf{W}_j) \tag{2.19}$$

So (CWRESI)

$$Var(y_i y_j) = \mathbf{g}_i'\boldsymbol{\Omega}\mathbf{g}_j + \mathbf{q}_i'\mathbf{q}_j - (\mathbf{q}_i'\mathbf{W}_j + \mathbf{q}_j'\mathbf{W}_i)\hat{\boldsymbol{\eta}} + \text{tr}\left[(\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}' + \boldsymbol{\Omega})\mathbf{W}_i'\mathbf{W}_j\right] \tag{2.20}$$

If $\hat{\boldsymbol{\eta}} = 0$ (WRESI) then

$$Var(y_i y_j) = \mathbf{g}_i' \boldsymbol{\Omega} \mathbf{g}_j + \mathbf{q}_i' \mathbf{q}_j + \mathrm{tr}\left[\boldsymbol{\Omega} \mathbf{W}_i' \mathbf{W}_j \right] \tag{2.21}$$

If $\mathbf{W}_i$=0, that is, interactive component is not taken into account, then (CWRES, [9])

$$Var(y_i y_j) = \mathbf{g}_i' \boldsymbol{\Omega} \mathbf{g}_j + \mathbf{q}_i' \mathbf{q}_j \tag{2.22}$$

In NONMEM 7.2, if $EST INTERACTION was specified prior to requesting $TABLE CWRES, then **g** and **q** are evaluated at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ in equation (2.22). If INTERACTION was not specified prior to requesting $TABLE CWRES, then **g** and **q** are evaluated at $\boldsymbol{\eta} = 0$ in equation (2.22). In NONMEM 7.1.0 and 7.1.2, regardless of INTERACTION setting in a previous $EST statement, **g** and **q** are evaluated at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ in equation (2.22).

In NONMEM, the residual error is modeled as follows:

$$y_i = f_i + h_{i1}' e_1 + h_{i2}' e_2 \ldots \tag{2.23}$$

for data point *i* of a particular subject, where e_k refers to the *k*th residual error component, that is in turn modeled to be normally distributed with variance

$$E(e_{k_1} e_{k_2}) = \Sigma_{k_1 k_2}$$

And

$$h_{ik}' = \frac{\partial y_i}{\partial e_k}$$

Consider a problem where PK data are modeled with mixed homoscedastic error and heteroscedastic error, as is PD data, and there is a correlation between certain PK and PD data that are sampleed at the same time. Such a correlation is indicated by the L2 variable listed in the data set. For such a problem, we could have:

$$y_i = f_i + (2 - C_i) f_i e_1 + (2 - C_i) e_2 + (C_i - 1) f_i e_1 + (C_i - 1) e_2 \tag{2.24}$$

Where C=1 if the datum is PK, and C=2 if datum is PD. The Sigma matrix would be modeled as:

$$\begin{bmatrix} \sigma_{11} & 0 & \sigma_{13} & 0 \\ 0 & \sigma_{22} & 0 & \sigma_{24} \\ \sigma_{31} & 0 & \sigma_{33} & 0 \\ 0 & \sigma_{42} & 0 & \sigma_{44} \end{bmatrix} \tag{2.25}$$

With correlation between certain paired PK,PD data, between their homoscedastic errors and heteroscedastic errors.

A grand matrix **H'** is produced among all data points for a subject. Suppose a particular subject has four data points:

1: PK datum at time 1 hour
2: PD datum at time 1 hour
3: PK datum at time 2 hours
4: PD datum at time 3 horus

Data points 1 and 2 are coupled, and the others are not. An expanded 4x12 matrix **H**' is produced as follows:

$$\begin{bmatrix} f_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f_4 & 1 \end{bmatrix} \tag{2.26}$$

along with an expanded 12x12 matrix $\boldsymbol{\Sigma}_e$ of block diagonal form, consisting of the 4x4 matrix $\boldsymbol{\Sigma}$ duplicated three times along the diagonal. The placement of h'$_{ik}$ in H' determines if two data points are correlated within a shared block diagonal portion of $\boldsymbol{\Sigma}_e$, as is the case with data points 1 and 2 , or have separate block diagonals, and therefore uncorrelated, as with data points 3 and 4. In this example, M=12, N=4.

Defining matrix $\boldsymbol{\Lambda}'$ as the lower triangular cholesky matrix to $\boldsymbol{\Sigma}_e$ (earlier we defined $\boldsymbol{\Lambda}$ as the lower triangular cholesky: we are redefining the nomenclature for this section for convenience):

$$\boldsymbol{\Sigma}_e = \boldsymbol{\Lambda}'\boldsymbol{\Lambda} \tag{2.27}$$

It follows that we can construct

$$\mathbf{Q}' = \mathbf{H}'\boldsymbol{\Lambda}' \tag{2.28}$$

$$\mathbf{Q} = \boldsymbol{\Lambda}\mathbf{H} \tag{2.29}$$

Or, considering column vector $\mathbf{h}_i$ of the *i*th column of **H**, then

$$\mathbf{q}_i = \boldsymbol{\Lambda}\mathbf{H}_i \tag{2.30}$$

And the full intra-subject variance would be

$$\mathbf{Q}'\mathbf{Q} = \mathbf{V} = \mathbf{H}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\mathbf{H} = \mathbf{H}'\boldsymbol{\Sigma}_e\mathbf{H} \tag{2.31}$$

Furthermore, defining

$$\mathbf{X}_i = \frac{\partial(\mathbf{h}_i)}{\partial\boldsymbol{\eta}} \tag{2.32}$$

then

$$\mathbf{W}_i = \frac{\partial(\boldsymbol{\Lambda}\mathbf{h}_i)}{\partial\boldsymbol{\eta}} = \boldsymbol{\Lambda}\frac{\partial(\mathbf{h}_i)}{\partial\boldsymbol{\eta}} = \boldsymbol{\Lambda}\mathbf{X}_i \tag{2.33}$$

So,

$$Var(y_i y_j) = \mathbf{g}_i'\boldsymbol{\Omega}\mathbf{g}_j + \mathbf{h}_i'\boldsymbol{\Sigma}_e\mathbf{h}_j - (\mathbf{h}_i'\boldsymbol{\Sigma}_e\mathbf{X}_j + \mathbf{h}_j'\boldsymbol{\Sigma}_e\mathbf{X}_i)\hat{\boldsymbol{\eta}} + \mathrm{tr}\left[(\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}' + \boldsymbol{\Omega})\mathbf{X}_i'\boldsymbol{\Sigma}_e\mathbf{X}_j\right] \tag{2.34}$$

An empirical method for evaluating the population weighted residual is to perform a Monte Carlo integration over all possible $\boldsymbol{\eta}$. For a given subject, the expected population predicted values is

$$\mathbf{f}_{\bar{\eta}} = E_\eta(\mathbf{f}(\boldsymbol{\eta})) = \int_{-\infty}^{\infty} \mathbf{f}(\boldsymbol{\eta}))\, p(\boldsymbol{\eta} \mid 0, \boldsymbol{\Omega}) d\boldsymbol{\eta} \tag{2.35}$$

where

$$p(\boldsymbol{\eta} \mid 0, \boldsymbol{\Omega}) = \frac{1}{\sqrt{2\pi}|\boldsymbol{\Omega}|}\exp\left(\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Omega}\boldsymbol{\eta}\right) \tag{2.36}$$

The expected residual for an observed value is

$$\mathbf{r}_{\bar{\eta}} = E_\eta(\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})) = \int_{-\infty}^{\infty} (\mathbf{y} - \mathbf{f}(\boldsymbol{\eta}))\, p(\boldsymbol{\eta} \mid 0, \boldsymbol{\Omega}) d\boldsymbol{\eta} = \mathbf{y} - \mathbf{f}_{\bar{\eta}} \tag{2.37}$$

without using linearization methods on $\boldsymbol{\eta}$. Now, since

$$E_\varepsilon(\mathbf{y}) = \mathbf{f}(\boldsymbol{\eta}) \tag{2.38}$$

$$E_\varepsilon((\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})(\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})) = \mathbf{Q}\mathbf{Q}' \tag{2.39}$$

Then the expected population variance is (without using linearization methods on eta):

$$\mathbf{C}_{\bar{\eta}} = E_{\eta,\varepsilon}((\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})(\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})') =$$

$$E_\eta(E_\varepsilon((\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})(\mathbf{y} - \mathbf{f}(\boldsymbol{\eta})')) + E_\eta(\mathbf{f}(\boldsymbol{\eta}) - \mathbf{f}_{\bar{\boldsymbol{\eta}}})(\mathbf{f}(\boldsymbol{\eta}) - \mathbf{f}_{\bar{\boldsymbol{\eta}}})') = \mathbf{V}_{\bar{\eta}} + \mathbf{V}_{\bar{\mathbf{f}}}$$

where

$$\mathbf{V}_{\bar{\eta}} = E_{\eta}(\mathbf{Q}'(\boldsymbol{\eta})\mathbf{Q}(\boldsymbol{\eta})) = \int_{-\infty}^{\infty} \mathbf{Q}'(\boldsymbol{\eta})\mathbf{Q}(\boldsymbol{\eta})\, p(\boldsymbol{\eta} \mid 0, \boldsymbol{\Omega}) d\boldsymbol{\eta} \tag{2.40}$$

$$\mathbf{V}_{\bar{\mathbf{f}}} = E_{\eta}((\mathbf{f}(\boldsymbol{\eta}) - \mathbf{f}_{\bar{\boldsymbol{\eta}}})(\mathbf{f}(\boldsymbol{\eta}) - \mathbf{f}_{\bar{\boldsymbol{\eta}}})') = E_{\eta}(\mathbf{f}(\boldsymbol{\eta})\mathbf{f}'(\boldsymbol{\eta})) - \mathbf{f}_{\bar{\boldsymbol{\eta}}}\mathbf{f}'_{\bar{\boldsymbol{\eta}}} \tag{2.41}$$

To evaluate the expected weighted residual (EWRES),

$$\mathbf{w}_{\bar{\eta}} = \mathbf{C}_{\bar{\eta}}^{-1/2}\mathbf{r}_{\bar{\eta}}$$

where $\mathbf{C}_{\bar{\eta}}^{-1/2}$ is the inverse square root of the expected population variance matrix.

An expected conditional (without interaction) weighted residual (ECWRES) can also be evaluated if we evaluate the intra-subject residual error at the conditional mean, such that

$$\mathbf{V}_{\hat{\eta}} = \mathbf{Q}'(\hat{\boldsymbol{\eta}})\mathbf{Q}(\hat{\boldsymbol{\eta}}) \tag{2.42}$$

if $EST INTERACTION is specifed followed by $TABLE ECWRES. But all other components are Monte Carlo integrated:

$$\mathbf{C}_{\bar{\eta}} = \mathbf{V}_{\hat{\eta}} + \mathbf{V}_{\bar{\mathbf{f}}} \tag{2.43}$$

As of NONMEM 7.2, if INTERACTION in $EST was not specified, followed by $TABLE ECWRES, then

$$\mathbf{V}_0 = \mathbf{Q}'(\boldsymbol{\eta} = \mathbf{0})\mathbf{Q}(\boldsymbol{\eta} = \mathbf{0}) \tag{2.44}$$

and all other components are Monte Carlo integrated:

$$\mathbf{C}_{\bar{\eta}} = \mathbf{V}_0 + \mathbf{V}_{\bar{\mathbf{f}}} \tag{2.45}$$

In NONMEM 7.1.0 and 7.1.2, regardless of INTERACTION setting from the previous $EST command, $\mathbf{V}_{\hat{\eta}}$ is used to evaluate ECWRES.

NPDE:

The NPDE is the normalized prediction distribution error (reference [10]: takes into account within-subject correlations), also a Monte Carlo assessed diagnostic item. For the *k*th simulated vector of data $\mathbf{y}_{sk}$:

$$\mathbf{s}_{\bar{\eta}k} = \mathbf{y}_{sk} - \mathbf{f}_{\bar{\eta}} \tag{2.46}$$

its decorrelated residual vector is calculated:

$$\mathbf{w}_{\mathbf{s}\bar{\eta}k} = \mathbf{C}_{\bar{\eta}}^{-1/2}\mathbf{s}_{\bar{\eta}k} \quad (2.47)$$

and compared against the decorrelated residual vector of observed values $\mathbf{w}_{\bar{\eta}}$ such that

$$\mathbf{u} = \frac{1}{K}\sum_{k=1}^{K}\delta(\mathbf{w}_{\bar{\eta}} - \mathbf{w}_{\mathbf{s}\bar{\eta}k}) \quad (2.48)$$

For *K* random samples, where

$\delta(x) = 1$ *for* $x >= 0$

$= 0$ *for* $x < 0$

For each element in the vector. Then, an inverse normal distribution transformation is performed:

$$\mathbf{w}_{npde} = \Phi^{-1}(\mathbf{u}) \quad (2.49)$$

NPD:

The NPD is the correlated normalized prediction distribution error (reference [11]: does not take into account within-subject correlations), also a Monte Carlo assessed diagnostic item. For each vector of data **y**:

$$\mathbf{r}_{\boldsymbol{\eta}_k} = \mathbf{V}(\boldsymbol{\eta}_k)^{-1/2}\left(\mathbf{y} - \mathbf{f}(\boldsymbol{\eta}_k)\right) \quad (2.50)$$

These are then averaged over all the random samples;

$$\mathbf{u}_c = \frac{1}{K}\sum_{k=1}^{K}\Phi(\mathbf{r}_{\boldsymbol{\eta}_k}) \quad (2.51)$$

Then, an inverse normal distribution transformation is performed:

$$\mathbf{w}_{npdec} = \Phi^{-1}(\mathbf{u}_c) \quad (2.52)$$

Models Non-Linearly Modeled in Epsilon

In NONMEM, one may also model the residuals using the epsilons in a nonlinear manner, such as:

y = f*exp(eps(1))

When population analysis is performed, however, NONMEM transposes this model into its linear-epsilon residual approximate form:

y = f+f*eps(1)

and evaluates the likelihood according to this epsilon-linearization. All analysis methods (classical as well as Monte Carlo) utilize this linearization of the likelihood in epsilon. Furthermore, the assessment of NPDE, NPD, EWRES, and ECWRES as described above utilize this linearized form with respect to the epsilon model, in keeping with the way the data was analyzed.

To most properly analyze the data in a manner that is equivalent to its epsilon exponential model form, and to also properly asses the various Monte Carlo population weighted residuals, it is best to log-transform the data, and model the residual variance to the log-transformed data follows:

y = log(f)+eps(1)

The residual variance is now linear epsilon modeled, and NONMEM will analyse the data exactly according to the true distribution of the data.

Epsilon Shrinkage Evaluation

The general shrinkage evaluation of the *p*th epsilon is evaluated as

$$R(p) = 100\% \left(1 - \sqrt{\frac{\sum_{i=1}^{m} S_i(p)}{\sum_{i=1}^{m} N_i(p)}} \right) \tag{3.1}$$

summed over subjects *i* to *m*, where

$$S_i(p) = \sum_{k=1}^{M_i} \delta_{ikp} ((\mathbf{y}_{ik} - \mathbf{f}_{ik}(\hat{\eta}))' \mathbf{V}_{ik}^{-1} ((\mathbf{y}_{ik} - \mathbf{f}_{ik}(\hat{\eta})) \tag{3.2}$$

$$N_i(p) = \sum_{k=1}^{M_i} \delta_{ikp} n_k \tag{3.3}$$

Where in turn

$$n_k = \left[\sum_{j \in k} 1 \right] \tag{3.4}$$

is the number of data points of subject *i* that belong to correlated data cluster *k*,

$$\mathbf{y}_{ik} = \{ y_{ij}, j \in k \} \tag{3.5}$$

the vector of a subset of data points of subject *i*, which belong to correlated data cluster *k* of subject *i*. Also,

$$\mathbf{H}'_{ik} = \{ h'_{ijp}, j \in k, p = 1 \, to \, n \} \tag{3.6}$$

Where *n*=number of epsilons.

$$\delta_{ikp} = \delta(\sum_{j \in k} h'^2_{ijp}) \tag{3.7}$$

$$\begin{aligned} \delta(x) &= 0 \text{ for } x = 0 \\ &= 1 \text{ otherwise} \end{aligned} \tag{3.8}$$

That is, the delta function is 0 if for the *p*th epsilon no data point in *k* contributes to its evaluation. This assures that the epsilon shrinkage evaluation includes only residual terms that relate to that epsilon. For example, it assures that epsilons involved only in PK data incorporate only PK data residuals, etc. And,

$$\mathbf{V}_{ik} = \mathbf{H}'_{ik} \mathbf{\Sigma} \mathbf{H}_{ik} \tag{3.9}$$

is the residual variance matrix to subject i, data cluster *k*.

If all data points in a subject are independent, then each data point cluster k contains only its own data point k, so n_k=1 for all k=1 to M_i, the number of data points to subject i, and the above vectors and matrices are scalar quantities

Appendix A: Matrix Algebra Tools

We wish to determine the derivative of $-\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega})$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}^{-1}$. Differentiating with respect to $\boldsymbol{\mu}$ is easily done as follows:

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}\right|)}{\partial \boldsymbol{\mu}} + \frac{1}{2}\frac{\partial (\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \tag{4.1}$$

For a particular parameter,

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\phi}))}{\partial \mu_j} = \frac{1}{2}\frac{\partial (\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})}{\partial \mu_j} = -\frac{1}{2}\mathbf{i}_j'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}\mathbf{i}_j \tag{4.2}$$

where $\mathbf{i}_j$ is a vector of 0's except for the *j*th element, which is 1. But the scalar terms are equal:

$$\left((\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}\mathbf{i}_j\right)' = \mathbf{i}_j'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}) \tag{4.3}$$

so

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \mu_j} = -\mathbf{i}_j'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}) \tag{4.4}$$

and for the entire vector $\boldsymbol{\mu}$,

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\mu}} = -\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu}) \tag{4.5}$$

To differentiate $-\log(h(\boldsymbol{\theta}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega})$ with respect to $\boldsymbol{\Omega}^{-1}$ is more difficult:

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}^{-1}} = -\frac{1}{2}\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}\right|)}{\partial \boldsymbol{\Omega}^{-1}} + \frac{1}{2}\frac{\partial (\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})}{\partial \boldsymbol{\Omega}^{-1}} \tag{4.6}$$

To do so, we must develop some partial derivative relationships in linear algebra. Consider any non-singular square matrix $\mathbf{Z}$ which is related to its inverse by

$$\mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I} \tag{4.7}$$

Therefore, The partial derivative with respect to some variable *x* yields

$$\frac{\partial(\mathbf{Z}\mathbf{Z}^{-1})}{\partial x} = \frac{\partial \mathbf{Z}}{\partial x}\mathbf{Z}^{-1} + \mathbf{Z}\frac{\partial \mathbf{Z}^{-1}}{\partial x} = \frac{\partial(\mathbf{I})}{\partial x} = \mathbf{0} \tag{4.8}$$

It follows that

$$\frac{\partial \mathbf{Z}^{-1}}{\partial x} = -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial x}\mathbf{Z}^{-1} \tag{4.9}$$

Suppose

$$x = z_{jk} \tag{4.10}$$

then

$$\frac{\partial \mathbf{Z}^{-1}}{\partial z_{jk}} = -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial z_{jk}}\mathbf{Z}^{-1} = -\mathbf{Z}^{-1}\mathbf{I}_{jk}\mathbf{Z}^{-1} = \left\{-z_{mj}^{-1}z_{kp}^{-1}, \text{for all } m = 1 \text{ to } n,\ p = 1 \text{ to } n\right\} \tag{4.11}$$

where $\mathbf{I}_{jk}$ is a matrix that is 0 every where except for element *j,k* which has a value of 1. Or

$$\frac{\partial z_{mp}^{-1}}{\partial z_{jk}} = -z_{mj}^{-1}z_{kp}^{-1} \tag{4.12}$$

Similarly,

$$\frac{\partial \mathbf{Z}'^{-1}}{\partial z_{jk}} = \frac{\partial \mathbf{Z}'^{-1}}{\partial z'_{kj}} = -\mathbf{Z}'^{-1}\frac{\partial \mathbf{Z}'}{\partial z'_{kj}}\mathbf{Z}'^{-1} = -\mathbf{Z}'^{-1}\mathbf{I}_{kj}\mathbf{Z}'^{-1} = \left\{-z_{jm}^{-1}z_{pk}^{-1}, \text{for all } m = 1 \text{ to } n,\ p = 1 \text{ to } n\right\} \tag{4.13}$$

Suppose it is more convenient to differentiate a particular function with respect to the inverse of **Z**, then

$$\frac{\partial f(\mathbf{Z})}{\partial z_{jk}} = \sum_{m=1}^{n}\sum_{p=1}^{n}\frac{\partial f(\mathbf{Z})}{\partial z_{mp}^{-1}}\frac{\partial z_{mp}^{-1}}{\partial z_{jk}} = \sum_{m=1}^{n}\sum_{p=1}^{n}-z_{mj}^{-1}\frac{\partial f(\mathbf{Z})}{\partial z_{mp}^{-1}}z_{kp}^{-1} = \sum_{m=1}^{n}\sum_{p=1}^{n}-z'^{-1}_{jm}\frac{\partial f(\mathbf{Z})}{\partial z_{mp}^{-1}}z'^{-1}_{pk} \tag{4.14}$$

or in matrix notation,

$$\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}} = -\mathbf{Z}'^{-1}\frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}^{-1}}\mathbf{Z}'^{-1} \tag{4.15}$$

Furthermore, according to linear algebra, for any matrix **Z**

$$\left|\mathbf{Z}\right| = \sum_{j=1}^{n} z_{jk}Z_{jk} \qquad \text{for any } k = 1 \text{ to } n \tag{4.16}$$

where z_{jk} is an element of matrix **Z**, and Z_{jk} is its cofactor. The cofactor Z_{jk} is the determinant of the sub-matrix of **Z** that does not include row *j* and column *k*. Therefore, Z_{jk} does not contain the element z_{jk}. Again, according to linear algebra,

$$z_{jk}^{-1} = \frac{Z_{kj}}{\left|\mathbf{Z}\right|} \tag{4.17}$$

where z_{jk}^{-1} is the *j,k* element of $\mathbf{Z}^{-1}$. Therefore

$$\frac{\partial |\mathbf{Z}|}{\partial z_{lk}} = \sum_{j=1}^{n} \frac{\partial z_{jk}}{\partial z_{lk}} Z_{jk} = Z_{lk} = |\mathbf{Z}| z_{kl}^{-1} \tag{4.18}$$

or

$$\frac{\partial |\mathbf{Z}|}{\partial \mathbf{Z}} = |\mathbf{Z}| \mathbf{Z}'^{-1} \tag{4.19}$$

More generally, for **Z** raised to any power *p*,

$$\frac{\partial |\mathbf{Z}^p|}{\partial \mathbf{Z}} = \frac{\partial |\mathbf{Z}|^p}{\partial \mathbf{Z}} = p|\mathbf{Z}|^{p-1} \frac{\partial |\mathbf{Z}|}{\partial \mathbf{Z}} = p|\mathbf{Z}|^p \mathbf{Z}'^{-1} \tag{4.20}$$

It follows that for any variable *x* which influences the elements of the matrix **Z,**

$$\frac{\partial |\mathbf{Z}|^p}{\partial x} = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial |\mathbf{Z}|^p}{\partial z_{jk}} \frac{\partial z_{jk}}{\partial x} = p|\mathbf{Z}|^p \sum_{j=1}^{n}\sum_{k=1}^{n} z_{kj}^{-1} \frac{\partial z_{jk}}{\partial x} = p|\mathbf{Z}|^p \operatorname{tr}\left(\mathbf{Z}^{-1} \frac{\partial \mathbf{Z}}{\partial x} \right) \tag{4.21}$$

So also,

$$\frac{\partial \log(|\mathbf{Z}|)}{\partial z_{lk}} = \frac{1}{|\mathbf{Z}|} \frac{\partial |\mathbf{Z}|}{\partial z_{lk}} = z_{kl}^{-1} \tag{4.22}$$

so that

$$\frac{\partial \log(|\mathbf{Z}|)}{\partial \mathbf{Z}} = \mathbf{Z}'^{-1} \tag{4.23}$$

or, for any variable *x*,

$$\frac{\partial \log(|\mathbf{Z}|)}{\partial x} = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial \log(|\mathbf{Z}|)}{\partial z_{jk}} \frac{\partial z_{jk}}{\partial x} = \sum_{j=1}^{n}\sum_{k=1}^{n} z_{kj}^{-1} \frac{\partial z_{jk}}{\partial x} = \operatorname{tr}\left(\mathbf{Z}^{-1} \frac{\partial \mathbf{Z}}{\partial x} \right) \tag{4.24}$$

For any *n*x*m* matrix **Y**, any *n*x*m* matrix **X**, and any *n*x*n* matrix **Z**,

$$tr\left(\mathbf{X}'\mathbf{Z}\mathbf{Y}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} x'_{ij} z_{jk} y_{ki} = \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{m} z_{jk} y_{ki} x'_{ij} = tr\left(\mathbf{Z}\mathbf{Y}\mathbf{X}'\right) = \sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ki} x'_{ij} z_{jk} = tr\left(\mathbf{Y}\mathbf{X}'\mathbf{Z}\right) \tag{4.25}$$

If *m*=1, then

$$tr\left(\mathbf{x}'\mathbf{Z}\mathbf{y}\right) = \mathbf{x}'\mathbf{Z}\mathbf{y} = tr\left(\mathbf{Z}\mathbf{y}\mathbf{x}'\right) = tr\left(\mathbf{y}\mathbf{x}'\mathbf{Z}\right) \tag{4.26}$$

Derivatives to trace functions can be derived as follows:

$$\frac{\partial tr(\mathbf{X}'\mathbf{Z}\mathbf{Y})}{\partial x_{mp}} = tr(\mathbf{I}_{pm}) = tr(\mathbf{I}_p \mathbf{I}'_m \mathbf{Z}\mathbf{Y}) = tr(\mathbf{I}'_m \mathbf{Z}\mathbf{Y}\mathbf{I}_p) \tag{4.27}$$

Where $\mathbf{I}_j$ is a column vector of zeros for all elements except for element *j*, for which it has a value of 1. Or,

$$\frac{\partial tr(\mathbf{X}'\mathbf{Z}\mathbf{Y})}{\partial \mathbf{X}} = \mathbf{Z}\mathbf{Y} \tag{4.28}$$

Similarly (taking some shortcuts in element/matrix nomenclature),

$$\frac{\partial tr(\mathbf{X}'\mathbf{Z}\mathbf{Y})}{\partial \mathbf{Z}} = tr(\mathbf{X}'\mathbf{I}_{mp}\mathbf{Y}) = tr(\mathbf{X}'\mathbf{I}_m\mathbf{I}'_p\mathbf{Y}) = tr(\mathbf{I}'_p\mathbf{Y}\mathbf{X}'\mathbf{I}_m) = \mathbf{X}\mathbf{Y}' \tag{4.29}$$

and

$$\frac{\partial tr(\mathbf{X}'\mathbf{Z}\mathbf{Y})}{\partial \mathbf{Y}} = \mathbf{X}\mathbf{Z}' \tag{4.30}$$

Using the above relationships for the following,

$$\frac{\partial (\boldsymbol{\phi}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi}-\boldsymbol{\mu})}{\partial \boldsymbol{\Omega}^{-1}} = (\boldsymbol{\phi}-\boldsymbol{\mu})(\boldsymbol{\phi}-\boldsymbol{\mu})' \tag{4.31}$$

It follows that

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}^{-1}} = -\frac{1}{2}\boldsymbol{\Omega} + \frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\mu})(\boldsymbol{\phi}-\boldsymbol{\mu})' \tag{4.32}$$

and incidentally,

$$\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}} = -\boldsymbol{\Omega}^{-1}\left(\frac{\partial -\log(h(\boldsymbol{\phi}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}^{-1}}\right)\boldsymbol{\Omega}^{-1} = \frac{1}{2}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega} - (\boldsymbol{\phi}-\boldsymbol{\mu})(\boldsymbol{\phi}-\boldsymbol{\mu})'\right)\boldsymbol{\Omega}^{-1} \tag{4.33}$$

The above tools also allow us to evaluate the following

$$\frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}^{-1}} = -\frac{1}{2}\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}\right|)}{\partial \boldsymbol{\Omega}^{-1}} + \frac{1}{2}\frac{\partial (\hat{\boldsymbol{\phi}}_i-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\phi}}_i-\boldsymbol{\mu})}{\partial \boldsymbol{\Omega}^{-1}} + \frac{1}{2}\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|)}{\partial \boldsymbol{\Omega}^{-1}} \tag{4.34}$$

Since

$$\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|)}{\partial \omega_{jk}^{-1}} = \frac{1}{\det(\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1})}\frac{\partial \left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|}{\partial \omega_{jk}^{-1}} = \frac{1}{\left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|}\frac{\partial \left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|}{\partial (\omega_{jk}^{-1}+\hat{s}_{jki}^{-1})}\frac{\partial (\omega_{jk}^{-1}+\hat{s}_{jki}^{-1})}{\partial \omega_{jk}^{-1}}$$

$$= \left(\omega_{jk}^{-1}+\hat{s}_{jki}^{-1}\right)^{-1} \tag{4.35}$$

or

$$\frac{\partial \log(\left|\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1}\right|}{\partial \boldsymbol{\Omega}^{-1}} = (\boldsymbol{\Omega}^{-1}+\hat{\mathbf{S}}_i^{-1})^{-1} \tag{4.36}$$

It follows that

$$\frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}^{-1}} = -\frac{1}{2}\boldsymbol{\Omega} + \frac{1}{2}(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' + \frac{1}{2}\left(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1}\right)^{-1} \tag{4.37}$$

and

$$\frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}} = -\boldsymbol{\Omega}^{-1}\left(\frac{\partial L_{Ni}}{\partial \boldsymbol{\Omega}^{-1}}\right)\boldsymbol{\Omega}^{-1} \tag{4.38}$$

$$= \frac{1}{2}\boldsymbol{\Omega}^{-1}\left[\boldsymbol{\Omega} - (\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu})(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\mu})' - \left(\boldsymbol{\Omega}^{-1} + \mathbf{S}_i^{-1}\right)^{-1}\right]\boldsymbol{\Omega}^{-1} \tag{4.39}$$

Because $\boldsymbol{\Omega}$ is symmetrical, the independent parameters which must be varied to minimize the objective function consist of only half of the matrix. Let us define the lower triangular matrix **A** containing independent parameters which relate to the elements of $\boldsymbol{\Omega}$ such that

$$\begin{aligned} &\omega_{jk} = a_{jk} \\ &\omega_{kj} = a_{jk}, \text{ for } j = 1 \text{ to } n, k = 1 \text{ to } j \end{aligned} \tag{4.40}$$

or

$$\boldsymbol{\Omega} = \mathbf{A} + \mathbf{A}' - \text{diag}(\mathbf{A}) \tag{4.41}$$

It follows that

$$\begin{aligned} &\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial a_{jk}} = \frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \omega_{jk}} + \frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \omega_{kj}} \\ &-\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \omega_{jk}}\delta(j,k) = \\ &2\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \omega_{jk}} - \frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \omega_{jk}}\delta(j,k) \end{aligned} \tag{4.42}$$

or in matrix notation,

$$\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \mathbf{A}} = \text{Lower}\left[2\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}} - \text{diag}\left(\frac{\partial -\log(h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}))}{\partial \boldsymbol{\Omega}}\right)\right] \tag{4.43}$$

and equation (1.55) results. At the minimum,

$$\frac{\partial L}{\partial \mathbf{A}} = \frac{\partial L}{\partial \boldsymbol{\Omega}} = \mathbf{0} \tag{4.44}$$

Another trace relationship that appears in probability densities is:

$$tr(\boldsymbol{\Sigma}^{-1}\mathbf{R}) = tr(\boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{R}) \tag{4.45}$$

Where $\boldsymbol{\Lambda}$ is the lower triangular cholesky matrix to a symmetrical matrix $\boldsymbol{\Sigma}$, and $\mathbf{R}$ is also a symmetrical matrix. Derivatives with respect to the cholesky elements is often desired, so,

$$\frac{\partial tr(\boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{R})}{\partial \boldsymbol{\Lambda}} = -\boldsymbol{\Lambda}'^{-1}\frac{\partial tr(\boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{R})}{\partial \boldsymbol{\Lambda}^{-1}}\boldsymbol{\Lambda}'^{-1} = -\boldsymbol{\Lambda}'^{-1}\left(\boldsymbol{\Lambda}^{-1}\mathbf{R} + \boldsymbol{\Lambda}^{-1}\mathbf{R}'\right)\boldsymbol{\Lambda}'^{-1} \tag{4.46}$$
$$= -2\boldsymbol{\Lambda}'^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{R}\boldsymbol{\Lambda}'^{-1} = -2\boldsymbol{\Sigma}^{-1}\mathbf{R}\boldsymbol{\Lambda}'^{-1}$$

Appendix B: Positive Definite Properties

A matrix **A** is defined as positive definite if

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \tag{5.1}$$

for any vector $\mathbf{x} \neq \mathbf{0}$. Consider a matrix constructed as follows

$$\mathbf{A} = \mathbf{Y}\mathbf{Z}\mathbf{Y}' \tag{5.2}$$

where **Y** is any non-zero *n*x*m* matrix, **Z** is an *m*x*m* positive definite matrix, so that **A** is an *n*x*n* matrix. Then, with any non-zero *n*x1 vector **x**,

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{Y}\mathbf{Z}\mathbf{Y}'\mathbf{x} = \mathbf{v}'\mathbf{Z}\mathbf{v} \tag{5.3}$$

where

$$\mathbf{v} = \mathbf{Y}'\mathbf{x} \tag{5.4}$$

is a non-zero *m*x1 vector. It follows that

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{v}'\mathbf{Z}\mathbf{v} > 0 \tag{5.5}$$

and **A** is therefore positive definite. The sum of two positive definite matrices **A** and **B** is also positive definite. Let

$$\mathbf{C} = \mathbf{A} + \mathbf{B}$$
$$\mathbf{x}'\mathbf{A}\mathbf{x} = a > 0 \tag{5.6}$$
$$\mathbf{x}'\mathbf{B}\mathbf{x} = b > 0$$

then

$$\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{x}'\mathbf{C}\mathbf{x} = a + b > 0 \tag{5.7}$$

Thus, matrices of the form

$$\sum_{i=1}^{n} \mathbf{Y}_i \mathbf{Z}_i \mathbf{Y}_i' \tag{5.8}$$

where $\mathbf{Z}_i$ is positive definite, are also positive definite.

The second derivative of any objective function *L* evaluated at its minimum is positive definite. This can be shown by considering that the derivative of *L* evaluated at its minimum L_0 is equal to **0** (otherwise, it would not be at a minimum/maximum):

$$\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)_0 = \mathbf{0} \tag{5.9}$$

And any small perturbation $\Delta\theta$ from the minimum results in a change in *L*, called ΔL, that is positive, otherwise it would be at its maximum, not its minimum. By Taylor series expansion we have:

$$0 < \Delta L = \left(\frac{\partial L}{\partial \theta}\right)_0 \Delta\theta + \frac{1}{2}\Delta\theta' \left(\frac{\partial^2 L}{\partial \theta^2}\right)_0 \Delta\theta = \frac{1}{2}\Delta\theta' \left(\frac{\partial^2 L}{\partial \theta^2}\right)_0 \Delta\theta \quad (5.10)$$

and therefore the second derivative of the objective function evaluated at its minimum is positive definite.

Appendix C: The Fischer Score Matrix for Error Assessment in EM Problems

We wish to evaluate the expected value, over all possible data **y** and over an infinite number of subjects *m*, of the second derivative of the objective function with respect to the population parameters $\mathbf{q} = \{\boldsymbol{\theta}, \boldsymbol{\Omega}\}$. The inverse of that matrix is then the asymptotic error matrix to the parameters. We do this as follows. Noting that

$$p(\mathbf{y}_i \mid \mathbf{q}) = \int_{-\infty}^{\infty} p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}) d\boldsymbol{\phi} \tag{6.1}$$

where

$$p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}) = l(\mathbf{y}_i \mid \boldsymbol{\phi}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Omega}) \tag{6.2}$$

and

$$L = \sum_{i=1}^{m} -\log(p_i(\mathbf{y}_i \mid \mathbf{q})) = -\log(p(\mathbf{y} \mid \mathbf{q})) \tag{6.3}$$

then

$$E_{\mathbf{y}}\left(\frac{\partial^2 -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j \partial q_k} \mid \mathbf{q} \right) = \int_{\mathbf{y}} \left[\frac{\partial^2 -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j \partial q_k} \right] p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y} = \tag{6.4}$$

$$\int_{\mathbf{y}} \left[\frac{-1}{p(\mathbf{y} \mid \mathbf{q})} \frac{\partial^2 p(\mathbf{y} \mid \mathbf{q})}{\partial q_j \partial q_k} + \frac{\partial \log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j} \frac{\partial \log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_k} \right] p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y} . \tag{6.5}$$

But

$$\int_{\mathbf{y}} \frac{-1}{p(\mathbf{y} \mid \mathbf{q})} \frac{\partial^2 p(\mathbf{y} \mid \mathbf{q})}{\partial q_j \partial q_k} p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y} = -\int_{\mathbf{y}} \frac{\partial^2 p(\mathbf{y} \mid \mathbf{q})}{\partial q_j \partial q_k} d\mathbf{y} = -\frac{\partial^2 \int_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y}}{\partial q_j \partial q_k} = -\frac{\partial^2 1}{\partial q_j \partial q_k} = \mathbf{0} \tag{6.6}$$

so

$$E_{\mathbf{y}}\left(\frac{\partial^2 -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j \partial q_k} \mid \mathbf{q} \right) = \int_{\mathbf{y}} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_k} p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y} = \tag{6.7}$$

$$E_{\mathbf{y}}\left(\frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_k} \mid \mathbf{q} \right) \tag{6.8}$$

We note that

$$\int_{\mathbf{y}} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_k} p(\mathbf{y} \mid \mathbf{q}) d\mathbf{y} = \tag{6.9}$$

$$\int_{\mathbf{y}}\left(\sum_{i_1=1}^{m}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}\right)\left(\sum_{i_2=1}^{m}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}\right)\prod_{i_3=1}^{m}p_{i_3}(\mathbf{y}_{i_3}\mid\mathbf{q})d\mathbf{y}= \tag{6.10}$$

$$\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\int_{\mathbf{y}}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}\prod_{i_3=1}^{m}p_{i_3}(\mathbf{y}_{i_3}\mid\mathbf{q})d\mathbf{y} \tag{6.11}$$

Since

$$\int_{\mathbf{y}_i}p_i(\mathbf{y}_i\mid\mathbf{q})=1 \tag{6.12}$$

then

$$E_{\mathbf{y}}\left(\frac{\partial -\log(p(\mathbf{y}\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p(\mathbf{y}\mid\mathbf{q}))}{\partial q_k}\mid\mathbf{q}\right)=$$

$$\sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\int_{\mathbf{y}}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}\prod_{i_3=1}^{m}p_{i_3}(\mathbf{y}_{i_3}\mid\mathbf{q})d\mathbf{y}= \tag{6.13}$$

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\int_{\mathbf{y}_{i_1}}\int_{\mathbf{y}_{i_2}}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q})p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q})d\mathbf{y}_{i_1}d\mathbf{y}_{i_2}+ \tag{6.14}$$

$$\sum_{i=1}^{m}\int_{\mathbf{y}_i}\frac{\partial -\log(p_i(\mathbf{y}_i\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p_i(\mathbf{y}_i\mid\mathbf{q}))}{\partial q_k}p_i(\mathbf{y}_i\mid\mathbf{q})d\mathbf{y}_i \tag{6.15}$$

But

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\int_{\mathbf{y}_{i_1}}\int_{\mathbf{y}_{i_2}}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q})p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q})d\mathbf{y}_{i_1}d\mathbf{y}_{i_2}=$$

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\left[\int_{\mathbf{y}_{i_1}}\frac{\partial -\log(p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q}))}{\partial q_j}p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q})d\mathbf{y}_{i_1}\right]\left[\int_{\mathbf{y}_{i_2}}\frac{\partial -\log(p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q}))}{\partial q_k}p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q})d\mathbf{y}_{i_2}\right]= \tag{6.16}$$

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\left[\int_{\mathbf{y}_{i_1}}\frac{\partial p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q})}{\partial q_j}d\mathbf{y}_{i_1}\right]\left[\int_{\mathbf{y}_{i_2}}\frac{\partial p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q})}{\partial q_k}d\mathbf{y}_{i_2}\right]= \tag{6.17}$$

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\left[\frac{\partial\int_{\mathbf{y}_{i_1}}p_{i_1}(\mathbf{y}_{i_1}\mid\mathbf{q})d\mathbf{y}_{i_1}}{\partial q_j}\right]\left[\frac{\partial\int_{\mathbf{y}_{i_1}}p_{i_2}(\mathbf{y}_{i_2}\mid\mathbf{q})d\mathbf{y}_{i_2}}{\partial q_k}\right]= \tag{6.18}$$

$$\sum_{i_1=1}^{m}\sum_{\substack{i_2=1\\ i_2\neq i_1}}^{m}\left[\frac{\partial 1}{\partial q_j}\right]\left[\frac{\partial 1}{\partial q_k}\right]=\mathbf{0} \tag{6.19}$$

Furthermore,

$$\frac{\partial -\log(p_i(\mathbf{y}_i \mid \mathbf{q}))}{\partial q_j} = \frac{1}{p_i(\mathbf{y}_i \mid \mathbf{q})} \frac{\partial - p_i(\mathbf{y}_i \mid \mathbf{q}))}{\partial q_j} = \frac{\int_{-\infty}^{\infty}[\partial p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}) / \partial \mathbf{q}] d\boldsymbol{\phi}}{p_i(\mathbf{y}_i \mid \mathbf{q})} =$$

$$\frac{\int_{-\infty}^{\infty}[\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q})) / \partial \mathbf{q}] p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}) d\boldsymbol{\phi}}{p_i(\mathbf{y}_i \mid \mathbf{q})} = \tag{6.20}$$

$$\int_{-\infty}^{\infty}[\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q})) / \partial \mathbf{q}] z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q}) d\boldsymbol{\phi} =$$

$$E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_j} \mid \mathbf{y}_i, \mathbf{q} \right)$$

So that

$$\sum_{i=1}^{m} \int_{\mathbf{y}_i} \frac{\partial -\log(p_i(\mathbf{y}_i \mid \mathbf{q}))}{\partial q_j} \frac{\partial -\log(p_i(\mathbf{y}_i \mid \mathbf{q}))}{\partial q_k} p_i(\mathbf{y}_i \mid \mathbf{q}) d\mathbf{y}_i = \tag{6.21}$$

$$E_{\mathbf{y}}\left[\sum_{i=1}^{m} E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_j} \mid \mathbf{y}_i, \mathbf{q} \right) E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_k} \mid \mathbf{y}_i, \mathbf{q} \right) \mid \mathbf{q} \right] \approx \tag{6.22}$$

$$\sum_{i=1}^{m} E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_j} \mid \mathbf{y}_i, \mathbf{q} \right) E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_k} \mid \mathbf{y}_i, \mathbf{q} \right) \tag{6.23}$$

so

$$E_{\mathbf{y}}\left(\frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_j} \frac{\partial -\log(p(\mathbf{y} \mid \mathbf{q}))}{\partial q_k} \mid \mathbf{q} \right) =$$

$$\sum_{i=1}^{m} E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_j} \mid \mathbf{y}_i, \mathbf{q} \right) E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial q_k} \mid \mathbf{y}_i, \mathbf{q} \right) \tag{6.24}$$

We define

$$\mathbf{g}_i = E_{\boldsymbol{\phi}}\left(\frac{\partial -\log(p_i(\mathbf{y}_i, \boldsymbol{\phi} \mid \mathbf{q}))}{\partial \mathbf{q}} \mid \mathbf{y}_i, \mathbf{q} \right) \tag{6.25}$$

and is the contribution of data from individual *i* to the total gradient **g**, where

$$\mathbf{g} = \sum_{i=1}^{m} \mathbf{g}_i = \mathbf{0} \tag{6.26}$$

at the minimum.

The gradient components are evaluated by methods of differentiation of matrix algebra.

$$\mathbf{g}_{i\boldsymbol{\theta}} = -\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\phi}}_i - \boldsymbol{\mu}_i) \tag{6.27}$$

Now, for the lower triangular part of the covariance matrix,

$$\mathbf{g}_{i\,\mathrm{Lower}(\boldsymbol{\Omega})} = \mathrm{Lower}\left[\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega}-\bar{\boldsymbol{\Omega}}_i)\boldsymbol{\Omega}^{-1} - \frac{1}{2}\mathrm{diag}\left(\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega}-\bar{\boldsymbol{\Omega}}_i)\boldsymbol{\Omega}^{-1}\right)\right] \tag{6.28}$$

To summarize,

$$Var(\mathbf{qq}') = \left(\sum_{i=1}^{m}\mathbf{g}_i\mathbf{g}_i'\right)^{-1} \tag{6.29}$$

where $\sum_{i=1}^{m}\mathbf{g}_i\mathbf{g}_i'$ is known as the Fisher score matrix.

One caveat is in order. Note that the structure of the expected value second derivative can be written in the form $\mathbf{GG}'$ where $\mathbf{G}$ is a *k*x*m* matrix having *m* column vectors of $\mathbf{g}_i$ of size *k*x1, and *k* is the total number of population parameters. If the number of subjects *m* is less than *k*, then the *k*x*k* Fischer score matrix has only a rank of *m*, and the matrix is not invertible. Thus, this manner of constructing the expected value second derivative only holds true as the number of subjects *m* as well as the number of data points per subject approaches infinity. Put another way, increasing the number of data points towards infinity for each subject, while having only a limited number of subjects, especially *k*<*m*, will not lead the Fischer score matrix to approach the expected value second derivative. For such conditions, the exact second derivative should be evaluated as given in appendix D.

Appendix D: The Exact Second Derivative Matrix for Error Assessment

The second derivative matrix for the population parameters and the population variance parameters, when the population parameter density is normally distributed, is determined as in the previous appendix, but without eliminating the terms that are canceled when taking the expected value over all **y**. For the second derivative for a normal population parameter, we note the following:

$$\frac{\partial^2 O}{\partial \mathbf{q}\partial \mathbf{q}'} = \frac{\partial^2}{\partial \mathbf{q}}\left(-\frac{\partial O}{\partial \mathbf{q}}\right) = \frac{\partial^2}{\partial \mathbf{q}}\left[\sum_{i=1}^{m}\left(-\int \frac{\partial \log(p_i)}{\partial \mathbf{q}} z_i(\mathbf{q}) d\boldsymbol{\phi}\right)\right] =$$

$$\sum_{i=1}^{m}\int \frac{\partial^2 -\log(p_i)}{\partial \mathbf{q}\partial \mathbf{q}'} z_i(\mathbf{q}) d\boldsymbol{\phi} - \sum_{i=1}^{m}\int \frac{\partial -\log(p_i)}{\partial \mathbf{q}} \frac{\partial -\log(p_i)}{\partial \mathbf{q}} z_i(\mathbf{q}) d\boldsymbol{\phi} +$$
$$\sum_{i=1}^{m}\left(\int \frac{\partial -\log(p_i)}{\partial \mathbf{q}} z_i(\mathbf{q}) d\boldsymbol{\phi}\right)\left(\int \frac{\partial -\log(p_i)}{\partial \mathbf{q}} z_i(\mathbf{q}) d\boldsymbol{\phi}\right) \quad (7.1)$$

where we let **q** represent the vector of all population parameters and variances, and

$$p_i = p(\mathbf{y}_i, \boldsymbol{\phi} \mid \boldsymbol{\theta}_{\not\mu}, \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu), \boldsymbol{\Omega}) = l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\not\mu}) h(\boldsymbol{\phi} \mid \boldsymbol{\mu}_i(\boldsymbol{\theta}_\mu), \boldsymbol{\Omega}) \quad (7.2)$$

$$\mathbf{g}_i = \int \frac{\partial -\log(p_i)}{\partial \mathbf{q}} z_i(\mathbf{q}) d\boldsymbol{\phi} \quad (7.3)$$

We already know the first derivatives contributed by each individual *i*. For mu modeled thetas,

$$\frac{\partial -\log(p_i)}{\partial \theta_{\mu j_1}} = -\frac{\partial \boldsymbol{\mu}_i'}{\partial \theta_{\mu j_1}} \boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i) \quad (7.4)$$

Also,

$$\frac{\partial -\log(p_i)}{\partial \omega_{j_1 j_2}} = -c(j_1, j_2)\mathbf{I}_{j_1}' \boldsymbol{\Omega}^{-1}((\boldsymbol{\phi} - \boldsymbol{\mu}_i)(\boldsymbol{\phi} - \boldsymbol{\mu}_i)' - \boldsymbol{\Omega})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2} \quad (7.5)$$

where $\mathbf{I}_{j_1}$ is a vector of 0's for every element except for element j_1, which is valued at 1. Also,

$$c(j_1, j_2) = 1 \text{ for } j_1 \neq j_2$$
$$= 1/2 \text{ for } j_1 = j_2 \quad (7.6)$$

Finally, for non-mu modeled thetas, a finite difference formula is employed:

$$\frac{\partial -\log(p_i)}{\partial \theta_{\neq j}} \approx \frac{-[\log(p_i(\theta_{\neq j} + \Delta\theta_{\neq j})) - \log(p_i(\theta_{\neq j} - \Delta\theta_{\neq j}))]}{2\Delta\theta_{\neq j}} \tag{7.7}$$

We now derive the second derivatives

$$\frac{\partial}{\partial \theta_{\mu j_2}}\left(\frac{\partial -\log(p_i)}{\partial \theta_{\mu j_1}}\right) = \frac{\partial \boldsymbol{\mu}_i'}{\partial \theta_{\mu j_1}} \boldsymbol{\Omega}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \theta_{\mu j_2}} - \frac{\partial^2 \boldsymbol{\mu}_i'}{\partial \theta_{\mu j_2} \partial \theta_{\mu j_1}} \boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i) \tag{7.8}$$

$$\begin{aligned}\frac{\partial^2}{\partial \omega_{j_3 j_4}}\left(\frac{\partial -\log(p_i)}{\partial \omega_{j_1 j_2}}\right) &= c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Omega}^{-1}(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2} + \\ 2c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Omega}^{-1}&(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Omega}^{-1}((\boldsymbol{\phi} - \boldsymbol{\mu}_i)(\boldsymbol{\phi} - \boldsymbol{\mu}_i)' - \boldsymbol{\Omega})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2}\end{aligned} \tag{7.9}$$

$$\frac{\partial}{\partial \omega_{j_2 j_3}} \frac{\partial -\log(p_i)}{\partial \theta_{\mu j_1}} = c(j_2, j_3) \frac{\partial \boldsymbol{\mu}_i'}{\partial \theta_{\mu j_1}} \boldsymbol{\Omega}^{-1}(I_{j_2 j_3} + I_{j_3 j_2})\boldsymbol{\Omega}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}_i) \tag{7.10}$$

Also:

$$\begin{aligned}\frac{\partial^2 -\log(p_i)}{\partial \theta_{\neq j_1} \partial \theta_{\neq j_2}} &= \frac{\partial}{\partial \theta_{\neq j_1}}\left(\frac{\partial -\log(p_i)}{\partial \theta_{\neq j_2}}\right) \approx \\ &\frac{-\log(p_i(\theta_{\neq j_1} + \Delta\theta_{\neq j_1}, \theta_{\neq j_2} + \Delta\theta_{\neq j_2})) + \log(p_i(\theta_{\neq j_1} + \Delta\theta_{\neq j_1}, \theta_{\neq j_2} - \Delta\theta_{\neq j_2}))}{4\Delta\theta_{\neq j_1}\Delta\theta_{\neq j_2}} - \\ &\frac{-\log(p_i(\theta_{\neq j_1} - \Delta\theta_{\neq j_1}, \theta_{\neq j_2} + \Delta\theta_{\neq j_2})) + \log(p_i(\theta_{\neq j_1} - \Delta\theta_{\neq j_1}, \theta_{\neq j_2} - \Delta\theta_{\neq j_2}))}{4\Delta\theta_{\neq j_1}\Delta\theta_{\neq j_2}}\end{aligned} \tag{7.11}$$

And, since $\boldsymbol{\theta}_{\mu}$ show up only in the *h*() portion of the joint density, and $\boldsymbol{\theta}_{\neq}$ show up only in the *l*() portion of the joint density

$$\frac{\partial}{\partial \theta_{\mu j_1}}\left(\frac{\partial -\log(p_i)}{\partial \theta_{\neq j_2}}\right) = 0 \tag{7.12}$$

Defining the individual subject central moments:

$$\phi_{ir_1}^{(1)} = \int_{-\infty}^{\infty}(\phi_{r_1} - \mu_{ir_1})z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q})d\boldsymbol{\phi} \tag{7.13}$$

$$\phi_{ir_1r_2}^{(2)} = \int_{-\infty}^{\infty}(\phi_{r_1} - \mu_{ir_1})(\phi_{r_2} - \mu_{ir_2})z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q})d\boldsymbol{\phi} \tag{7.14}$$

$$\phi_{ir_1r_2r_3}^{(3)} = \int_{-\infty}^{\infty}(\phi_{r_1} - \mu_{ir_1})(\phi_{r_2} - \mu_{ir_2})(\phi_{r_3} - \mu_{ir_3})z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q})d\boldsymbol{\phi} \tag{7.15}$$

$$\phi_{ir_1r_2r_3r_4}^{(4)} = \int_{-\infty}^{\infty}(\phi_{r_1} - \mu_{ir_1})(\phi_{r_2} - \mu_{ir_2})(\phi_{r_3} - \mu_{ir_3})(\phi_{r_4} - \mu_{ir_4})z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q})d\boldsymbol{\phi} \tag{7.16}$$

$$\varpi_{ir_1r_2}^{(2)} = \int_{-\infty}^{\infty} s_{ir_1} s_{ir_2} z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q}) d\boldsymbol{\phi} \tag{7.17}$$

$$\sigma_{ir_1r_2}^{(2)} = \int_{-\infty}^{\infty} (\phi_{r_1} - \mu_{ir_1}) s_{ir_2} z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q}) d\boldsymbol{\phi} \tag{7.18}$$

$$\sigma_{ir_1r_2r_3}^{(3)} = \int_{-\infty}^{\infty} s_{ir_1} (\phi_{r_2} - \mu_{ir_2})(\phi_{r_3} - \mu_{ir_3}) z(\boldsymbol{\phi} \mid \mathbf{y}_i, \mathbf{q}) d\boldsymbol{\phi} \tag{7.19}$$

where

$$s_{ir_1} = \frac{\partial - \log(p_i)}{\partial \theta_{\neq r_1}} \tag{7.20}$$

and the overall central moments:

$$\phi_{r_1}^{(1)} = \frac{1}{m} \sum_{i=1}^{m} \phi_{ir_1}^{(1)} \tag{7.21}$$

$$\phi_{r_1r_2}^{(2)} = \frac{1}{m} \sum_{i=1}^{m} \phi_{ir_1r_2}^{(2)} \tag{7.22}$$

$$\phi_{r_1r_2r_3}^{(3)} = \frac{1}{m} \sum_{i=1}^{m} \phi_{ir_1r_2r_3}^{(3)} \tag{7.23}$$

$$\phi_{r_1r_2r_3r_4}^{(4)} = \frac{1}{m} \sum_{i=1}^{m} \phi_{ir_1r_2r_3r_4}^{(4)} \tag{7.24}$$

We now have the following:

$$\frac{\partial^2 O}{\partial \theta_{\mu j_2} \partial \theta_{\mu j_1}} = \sum_{i=1}^{m} \sum_{r_1}^{n} \sum_{r_2=1}^{n} \sum_{r_3=1}^{n} \sum_{r_4=1}^{n} \frac{\partial \mu_{ir_1}}{\partial \theta_{\mu j_1}} \omega_{r_1r_2}^{-1} \left(\omega_{r_2r_3} - \phi_{ir_2r_3}^{(2)} \right) \omega_{r_3r_4}^{-1} \frac{\partial \mu_{ir_4}}{\partial \theta_{\mu j_2}} - \sum_{i=1}^{m} \sum_{r_1}^{n} \sum_{r_2=1}^{n} \frac{\partial^2 \mu_{ir_1}}{\partial \theta_{\mu j_2} \partial \theta_{\mu j_1}} \omega_{r_1r_2}^{-1} \phi_{ir_2}^{(1)} + \sum_{i=1}^{m} g_{i\theta_{\mu j1}} g_{i\theta_{\mu j2}} \tag{7.25}$$

$$\frac{\partial^2 O}{\partial \omega_{j_3 j_4} \partial \omega_{j_1 j_2}} = mc(j_1, j_2)c(j_3, j_4)(\omega^{-1}_{j_1 j_3}\omega^{-1}_{j_4 j_2} + \omega^{-1}_{j_1 j_4}\omega^{-1}_{j_3 j_2})$$
$$+m2c(j_1, j_2)c(j_3, j_4)\omega^{-1}_{j_1 j_3}(\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\omega^{-1}_{j_4 r_1}(\phi^{(2)}_{r_1 r_2} - \omega_{r_1 r_2})\omega^{-1}_{r_2 j_2}$$
$$+m2c(j_1, j_2)c(j_3, j_4)\omega^{-1}_{j_1 j_4}(\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\omega^{-1}_{j_3 r_1}(\phi^{(2)}_{r_1 r_2} - \omega_{r_1 r_2})\omega^{-1}_{r_2 j_2}$$
$$-mc(j_1, j_2)c(j_3, j_4)\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\sum_{r_3=1}^{n}\sum_{r_4=1}^{n}\phi^{(4)}_{r_1 r_2 r_3 r_4}\omega^{-1}_{j_1 r_1}\omega^{-1}_{j_2 r_2}\omega^{-1}_{j_3 r_3}\omega^{-1}_{j_4 r_4} \quad (7.26)$$
$$+mc(j_1, j_2)c(j_3, j_4)\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\sum_{r_3=1}^{n}\sum_{r_4=1}^{n}\omega^{-1}_{j_1 r_1}\phi^{(2)}_{r_1 r_2}\omega^{-1}_{r_2 j_2}\omega^{-1}_{j_3 j_4}$$
$$+mc(j_1, j_2)c(j_3, j_4)\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\sum_{r_3=1}^{n}\sum_{r_4=1}^{n}\omega^{-1}_{j_1 j_2}\omega^{-1}_{j_3 r_1}\phi^{(2)}_{r_1 r_2}\omega^{-1}_{r_2 j_4}$$
$$-mc(j_1, j_2)c(j_3, j_4)\omega^{-1}_{j_1 j_2}\omega^{-1}_{j_3 j_4} + \sum_{i=1}^{m} g_{i\omega_{j_1 j_2}} g_{i\omega_{j_3 j_4}}$$

$$\frac{\partial^2 O}{\partial \omega_{j_2 j_3} \partial \theta_{\mu j_1}} = c(j_2, j_3)\sum_{r_1=1}^{n}\sum_{r_2=1}^{n}\left(\sum_{i=1}^{m}\frac{\partial \mu_{ir_1}}{\partial \theta_{\mu j_1}}\phi^{(1)}_{ir_2}\right)\left(\omega^{-1}_{r_1 j_2}\omega^{-1}_{j_3 r_2} + \omega^{-1}_{r_1 j_3}\omega^{-1}_{j_2 r_2}\right)$$
$$-c(j_2, j_3)\left(\sum_{i=1}^{m}\sum_{r_4=1}^{n}\frac{\partial \mu_{ir_4}}{\partial \theta_{\mu j_1}}\omega^{-1}_{r_4 r_1}\phi^{(3)}_{ir_1 r_2 r_3}\right)\omega^{-1}_{j_2 r_2}\omega^{-1}_{j_3 r_3} - a(j_2, j_3)g_{c_{j_1}}\omega^{-1}_{j_2 j_3} + \sum_{i=1}^{m} g_{i\theta_{\mu j_1}} g_{i\omega_{j_2 j_3}} \quad (7.27)$$

where

$$g_{\mu\theta_{j_1}} = \sum_{i=1}^{m} g_{i\theta_{\mu j_1}} \quad (7.28)$$

etcetera. And finally:

For the non-mu parameters:

$$\frac{\partial^2 O}{\partial \theta_{\neq j_1} \partial \theta_{\neq j_2}} = \sum_{i=1}^{m}\int \frac{\partial^2 - \log(p_i)}{\partial \theta_{\neq j_1} \partial \theta_{\neq j_2}} z_i(\mathbf{q}) d\boldsymbol{\phi} - \sum_{i=1}^{m}\varpi^{(2)}_{ij_1 j_2} + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{\theta_{\neq j^2}} \quad (7.29)$$

$$\frac{\partial^2 O}{\partial \theta_{\mu j_2} \partial \theta_{\neq j_1}} = \sum_{i=1}^{m}\sum_{r_1}^{n}\frac{\partial \mu_i}{\partial \theta_{\mu j_2}}\omega^{-1}_{j_2 r_1}\sigma^{(2)}_{ir_1 j_1} + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{i\theta_{\mu j_2}} \quad (7.30)$$

$$\frac{\partial^2 O}{\partial \omega_{j_2 j_3} \partial \theta_{\neq j_1}} = c(j_2, j_3)\left[\sum_{i=1}^{m}\sum_{r_1}^{n}\sum_{r_2}^{n}\omega^{-1}_{j_2 r_1}\sigma^{(3)}_{ij_1 r_1 r_2}\omega^{-1}_{r_2 j_3} - \omega^{-1}_{j_2 j_3}\sum_{i=1}^{m} g_{i\theta_{\neq j_1}}\right] + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{i\omega_{j_2 j_3}} \quad (7.31)$$

At the minimum of the objective function,

$$g_{\theta_{\mu j_1}} = 0 \quad (7.32)$$

$$g_{\theta_{\neq j_1}} = 0 \tag{7.33}$$

$$\phi^{(2)}_{r_1 r_2} = \omega_{r_1 r_2} \tag{7.34}$$

we then obtain following simplifications:

$$\frac{\partial^2 O}{\partial \theta_{\mu j_2} \partial c_{j_1}} = \sum_{i=1}^{m} g_{i\theta_{\mu j_1}} g_{i\theta_{\mu j1}} - \sum_{i=1}^{m} \sum_{r_1}^{n} \sum_{r_2=1}^{n} \sum_{r_3=1}^{n} \sum_{r_4=1}^{n} \frac{\partial \mu_{ir_1}}{\partial \theta_{\mu j_1}} \omega^{-1}_{r_1 r_2} \left(\phi^{(2)}_{ir_2 r_3} - \omega_{r_2 r_3} \right) \omega^{-1}_{r_3 r_4} \frac{\partial \mu_{ir_4}}{\partial \theta_{\mu j_2}}$$
$$- \sum_{i=1}^{m} \sum_{r_1}^{n} \sum_{r_2=1}^{n} \frac{\partial^2 \mu_{ir_1}}{\partial \theta_{\mu j_2} \partial \theta_{\mu j_1}} \omega^{-1}_{r_1 r_2} \phi^{(1)}_{ir_2} \tag{7.35}$$

$$\frac{\partial^2 O}{\partial \omega_{j_3 j_4} \partial \omega_{j_1 j_2}} = mc(j_1, j_2) c(j_3, j_4) (\omega^{-1}_{j_1 j_2} \omega^{-1}_{j_3 j_4} + \omega^{-1}_{j_1 j_3} \omega^{-1}_{j_4 j_2} + \omega^{-1}_{j_1 j_4} \omega^{-1}_{j_3 j_2})$$
$$- mc(j_1, j_2) c(j_3, j_4) \sum_{r_1=1}^{n} \sum_{r_2=1}^{n} \sum_{r_3=1}^{n} \sum_{r_4=1}^{n} \phi^{(4)}_{r_1 r_2 r_3 r_4} \omega^{-1}_{j_1 r_1} \omega^{-1}_{j_2 r_2} \omega^{-1}_{j_3 r_3} \omega^{-1}_{j_4 r_4} + \sum_{i=1}^{m} g_{i\omega_{j_1 j_2}} g_{i\omega_{j3 j4}} \tag{7.36}$$

$$\frac{\partial^2 O}{\partial \omega_{j_2 j_3} \partial \theta_{\mu j\cdot}} = c(j_2, j_3) \sum_{r_1=1}^{n} \sum_{r_2=1}^{n} \left(\sum_{i=1}^{m} \frac{\partial \mu_{ir_1}}{\partial \theta_{\mu j\cdot}} \phi^{(1)}_{ir_2} \right) \left(\omega^{-1}_{r_1 j_2} \omega^{-1}_{j_3 r_2} + \omega^{-1}_{r_1 j_3} \omega^{-1}_{j_2 r_2} \right)$$
$$- c(j_2, j_3) \left(\sum_{i=1}^{m} \sum_{r_4=1}^{n} \frac{\partial \mu_{ir_4}}{\partial \theta_{\mu j\cdot}} \omega^{-1}_{r_4 r_1} \phi^{(3)}_{ir_1 r_2 r_3} \right) \omega^{-1}_{j_2 r_2} \omega^{-1}_{j_3 r_3} + \sum_{i=1}^{m} g_{i\theta_{\mu j\cdot}} g_{i\omega_{j_2 j_3}} \tag{7.37}$$

$$\frac{\partial^2 O}{\partial \theta_{\neq j_1} \partial \theta_{\neq j_2}} = \sum_{i=1}^{m} \int \frac{\partial^2 - \log(p_i)}{\partial \theta_{\neq j_1} \partial \theta_{\neq j_2}} z_i(\mathbf{q}) d\boldsymbol{\phi} - \sum_{i=1}^{m} \varpi^{(2)}_{ij_1 j_2} + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{\theta_{\neq j2}} \tag{7.38}$$

$$\frac{\partial^2 O}{\partial \theta_{\mu j_2} \partial \theta_{\neq j_1}} = \sum_{i=1}^{m} \sum_{r_1}^{n} \frac{\partial \mu_i}{\partial \theta_{\mu j_2}} \omega^{-1}_{j_2 r_1} \sigma^{(2)}_{ir_1 j_1} + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{i\theta_{\mu j_2}} \tag{7.39}$$

$$\frac{\partial^2 O}{\partial \omega_{j_2 j_3} \partial \theta_{\neq j_1}} = c(j_2, j_3) \left[\sum_{i=1}^{m} \sum_{r_1}^{n} \sum_{r_2}^{n} \omega^{-1}_{j_2 r_1} \sigma^{(3)}_{ij_1 r_1 r_2} \omega^{-1}_{r_2 j_3} \right] + \sum_{i=1}^{m} g_{i\theta_{\neq j_1}} g_{i\omega_{j_2 j_3}} \tag{7.40}$$

Let us see that, if we take the expected value over all **y**, and as the number of subset of subjects m_i sharing a particular set of covariates (and therefore sharing the same μ_{ir}), approaches infinity, we should obtain the results in Appendix C. First, note that:

$$\int_{\mathbf{y}} \phi^{(3)}_{ir_1r_2r_3} p(\mathbf{y}\mid\mathbf{q})d\mathbf{y} = \frac{1}{m_i}\sum_{j=1}^{m_i}\int_{\mathbf{y}}\int_{-\infty}^{\infty}(\phi_{r_1}-\mu_{ir_1})(\phi_{r_2}-\mu_{ir_2})(\phi_{r_3}-\mu_{ir_3})z(\boldsymbol{\phi}\mid\mathbf{y}_j,\mathbf{q})d\boldsymbol{\phi}p(\mathbf{y}\mid\mathbf{q})d\mathbf{y} =$$

(7.41)

$$\frac{1}{m_i}\sum_{j=1}^{m_i}\int_{\mathbf{y}_j}\int_{-\infty}^{\infty}(\phi_{r_1}-\mu_{ir_1})(\phi_{r_2}-\mu_{ir_2})(\phi_{r_3}-\mu_{ir_3})l(\mathbf{y}_j\mid\boldsymbol{\phi})h(\boldsymbol{\phi}\mid\mathbf{q})d\boldsymbol{\phi}d\mathbf{y}_j = \tag{7.42}$$

$$\frac{1}{m_i}\sum_{j=1}^{m_i}\int_{-\infty}^{\infty}(\phi_{r_1}-\mu_{ir_1})(\phi_{r_2}-\mu_{ir_2})(\phi_{r_3}-\mu_{ir_3})h(\boldsymbol{\phi}\mid\mathbf{q})\int_{\mathbf{y}_j} l(\mathbf{y}_j\mid\boldsymbol{\phi})d\mathbf{y}_j = \tag{7.43}$$

$$\frac{1}{m_i}\sum_{j=1}^{m_i}\int_{-\infty}^{\infty}(\phi_{r_1}-\mu_{ir_1})(\phi_{r_2}-\mu_{ir_2})(\phi_{r_3}-\mu_{ir_3})h(\boldsymbol{\phi}\mid\mathbf{q})d\boldsymbol{\phi} = \hat{\phi}^{(3)}_{ir_1r_2r_3} \tag{7.44}$$

Since $h()$ is a multi-variate normal density, so therefore the $\hat{\phi}^{(3)}_{ir_1r_2r_3}$ is the skewness of a normally distributed variable:

$$\hat{\phi}^{(3)}_{ir_1r_2r_3} = 0 \tag{7.45}$$

Similarly,

$$\hat{\phi}^{(1)}_{ir_1} = 0 \tag{7.46}$$

and

$$\int_{\mathbf{y}} \phi^{(4)}_{ir_1r_2r_3} p(\mathbf{y}_i\mid\mathbf{q})d\mathbf{y} = \hat{\phi}^{(4)}_{ir_1r_2r_3r_4} = \omega_{r_1r_2}\omega_{r_3r_4} + \omega_{r_1r_3}\omega_{r_2r_4} + \omega_{r_1r_4}\omega_{r_2r_3} \tag{7.47}$$

is derived from the kurtosis of a normally distributed random variable. We may now make the final simplification:

$$E_{\mathbf{y}}\left(\frac{\partial^2 O}{\partial\theta_{\mu j_2}\partial\theta_{\mu j_1}}\right) = \sum_{i=1}^{m} g_{i\theta_{\mu j_1}} g_{i\theta_{\mu j_2}} \tag{7.48}$$

$$E_{\mathbf{y}}\left(\frac{\partial^2 O}{\partial\omega_{j_3 j_4}\partial\omega_{j_1 j_2}}\right) = \sum_{i=1}^{m} g_{i\omega_{j_1 j_2}} g_{i\omega_{j_3 j_4}} \tag{7.49}$$

$$E_{\mathbf{y}}\left(\frac{\partial^2 O}{\partial\omega_{j_2 j_3}\partial\theta_{\mu j_1}}\right) = \sum_{i=1}^{m} g_{i\theta_{\mu j_1}} g_{i\omega_{j_2 j_3}} \tag{7.50}$$

These results are expected based on the general proof for the expected information matrix given in Appendix C for any population parameter density.

For population mixture parameters, the second derivatives are simply:

$$\frac{\partial^2 O}{\partial\theta_a\partial\theta_a} = \sum_{i=1}^{m} g_{i\boldsymbol{\theta}_a} g'_{i\boldsymbol{\theta}_a} \tag{7.51}$$

$$\frac{\partial^2 O}{\partial\theta_a\partial\theta_{\boldsymbol{\mu}}} = \sum_{i=1}^{m} g_{i\boldsymbol{\theta}_a} g'_{i\boldsymbol{\theta}_{\boldsymbol{\mu}}} \tag{7.52}$$

$$\frac{\partial^2 O}{\partial\theta_a\partial\theta_{\boldsymbol{\mu}}} = \sum_{i=1}^{m} g_{i\boldsymbol{\theta}_a} g'_{i\boldsymbol{\theta}_{\boldsymbol{\mu}}} \tag{7.53}$$

$$\frac{\partial^2 O}{\partial\theta_a\partial\omega} = \sum_{i=1}^{m} g_{i\boldsymbol{\theta}_a} g'_{i\omega} \tag{7.54}$$

For Three hierarchical stage analysis, the second derivative matrices would have added to them:

$$\frac{\partial^2 O_P}{(\partial\boldsymbol{\theta})^2} = \boldsymbol{\Omega}_{\boldsymbol{\theta}}^{-1} \tag{7.55}$$

for all theta parameters (mu or non-mu modeled). For inter-subject variance components:

$$\frac{\partial^2}{\partial\omega_{j_3 j_4}}\left(\frac{\partial O_P}{\partial\omega_{j_1 j_2}}\right) = \rho z_W c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Omega}^{-1}(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2} + \tag{7.56}$$
$$2\rho c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Omega}^{-1}(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Omega}^{-1}(\boldsymbol{\Omega}_{\boldsymbol{\Omega}} - z_W\boldsymbol{\Omega})\boldsymbol{\Omega}^{-1}\mathbf{I}_{j_2}$$

Similarly for Sigma parameters, we add the following:

$$\frac{\partial^2}{\partial\Sigma_{j_3 j_4}}\left(\frac{\partial O_P}{\partial\Sigma_{j_1 j_2}}\right) = \rho_\Sigma z_\Sigma c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Sigma}^{-1}(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Sigma}^{-1}\mathbf{I}_{j_2} + \tag{7.57}$$
$$2\rho_\Sigma c(j_1, j_2)c(j_3, j_4)\mathbf{I}'_{j_1}\boldsymbol{\Sigma}^{-1}(\mathbf{I}_{j_3 j_4} + \mathbf{I}_{j_4 j_3})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\Sigma}} - z_\Sigma\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{I}_{j_2}$$

If we designate the exact second derivative as described in this appendix as matrix **R** , and the Fischer score matrix (appendix C) as **S**, then we can construct a variance matrix in a manner similar to NONMEM:

$$Var(\mathbf{qq}') = \mathbf{R}^{-1}\mathbf{S}\mathbf{R}^{-1} \tag{7.58}$$

The **R** matrix is not always numerically positive definite. For Monte Carlo assessed information matrices, the NONMEM program passes the matrix through a positive definiteness filter that makes small adjustments to the eigenvalues of **R**, if necessary.

Appendix E: Adjustment of Error Matrix for Constraints and Non-Positive Definiteness

The user supplied subroutine CONSTRAINT allows the user to impose constraints on the population parameters. Therefore the error matrix must be adjusted to account for these constraints, and it is done as follows.

Let the constraint matrix **W** be defined as a matrix with elements:

$$w_{ij} = \frac{\partial q_j}{\partial q_i} \tag{8.1}$$

where q_j is the jth population parameter (or Omega variance), which could have a dependence on some other population parameter q_i. If no dependence exists for parameter j, then w_{ij}=0 for $i \neq j$, and w_{ii}=1. If no dependencies are defined for any parameter, then **W**=**I**, and no correction occurs. The Error matrix is corrected as follows:

For the R matrix type covariance:

$$Var(\mathbf{qq}') = \mathbf{W}'\left(\mathbf{WRW}'\right)^{-1}\mathbf{W} \tag{8.2}$$

The logic behind this equation is as follows. Parameters that are dependent on other parameters are considered secondary parameters, in contrast to the primary parameters that are independent of other parameters. The $\left(\mathbf{WRW}'\right)^{-1}$ term creates the error matrix with rows and columns pertaining to the secondary parameters zeroed out, while the errors of the primary parameters are adjusted to account for the constraint on the model. This matrix is then multiplied on either side by **W** and **W'**, to fill the zeroed secondary parameter rows and columns of $\left(\mathbf{WRW}'\right)^{-1}$ with errors from the primary parameters, in accordance with their dependencies on the primary parameters. The resulting error matrix therefore contains errors to the primary as well as secondary parameters, and this matrix is placed in the var table by the poperr command, and the varc table by the poperr_corr command.

Similarly for the S matrix:

$$Var(\mathbf{qq}') = \mathbf{W}'\left(\mathbf{WSW}'\right)^{-1}\mathbf{W} \tag{8.3}$$

For the RSR matrix:

$$Var(\mathbf{qq}') = \left[\mathbf{W}'\left(\mathbf{WRW}'\right)^{-1}\mathbf{W}\right]\mathbf{S}\left[\mathbf{W}'\left(\mathbf{WRW}'\right)^{-1}\mathbf{W}\right] = \mathbf{W}'\left(\mathbf{WRW}'\right)^{-1}\left(\mathbf{WSW}'\right)\left(\mathbf{WRW}'\right)^{-1}\mathbf{W} \tag{8.4}$$

If a particular parameter is constrained to a fixed value, then **W** will be singular. The matrix $\mathbf{WRW}'$ is therefore inverted by the Jacobi method of extracting eignvalues and eigenvector matrices. That is, for any symmetric matrix **A**, the Jacobian process decomposes the matrix to:

$$\mathbf{A} = \mathbf{E}\Lambda\mathbf{E}' \tag{8.5}$$

where Λ is the diagonal matrix of eigenvalues and **E** is a matrix of eigenvector columns, which has the property:

$$\mathbf{E}' = \mathbf{E}^{-1} \tag{8.6}$$

The generalized inverse of **A** is then obtained as:

$$\mathbf{A} = \mathbf{E}\Lambda^{-}\mathbf{E}' \tag{8.7}$$

where Λ^{-} has diagonal elements of

$$\lambda^{-}_{i} = \frac{1}{|\lambda_i|} \; for \; |\lambda_i| > 0$$
$$= 0 \; for \; |\lambda_i| = 0 \tag{8.8}$$

In addition, the **R** matrix itself can at times be not positive definite (has negative eigenvalues), because of the imprecision of evaluating this matrix using random sampling, in the manner decribed in appendix D. It has been found in practice that using the absolute value of the eigenvalues to evaluate the inverse for Monte Carlo constructed information matrices effectively yields satisfactory error matrices. This is because negative eigenvalues are usually close to 0, and arise in the least important portions of the matrix.

Appendix F: Obtaining Analytical Derivatives of Likelihood with Respect to Cholesky of Sigma Parameters.

The following is used to provide more rapid analysis for the importance sampling, direct sampling, and SAEM methods.

We concern ourselves with the derivatives of the likelihood:

$$\log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\#})) = -\frac{1}{2}(\mathbf{y}_i - \mathbf{f}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{f}_i) + \frac{1}{2}\log\left(\left|\mathbf{V}_i\right|\right) \tag{9.1}$$

With respect to the Sigma parameters. Consider that the Sigma parameters are involved in the construction of the residual variance matrix as follows:

$$\mathbf{V} = \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \mathbf{H}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{H}' \tag{9.2}$$

Where $\boldsymbol{\Lambda}$ is the cholesky of $\boldsymbol{\Sigma}$. It is the elements of $\boldsymbol{\Lambda}$ that are actually varied to optimize the objective function, therefore we wish to determine the derivative with respect to the elements of $\boldsymbol{\Lambda}$.

First derivatives:

$$\frac{-\partial \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\#}))}{\partial \lambda_{j_1 k_1}} = \frac{1}{2}(\mathbf{y}_i - \mathbf{f}_i)'\frac{\partial \mathbf{V}_i^{-1}}{\partial \lambda_{j_1 k_1}}(\mathbf{y}_i - \mathbf{f}_i) + \frac{1}{2}\frac{\partial \log\left(\left|\mathbf{V}_i\right|\right)}{\partial \lambda_{j_1 k_1}} = \tag{9.3}$$

$$-\frac{1}{2}(\mathbf{y}_i - \mathbf{f}_i)'\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial b_{j_1 k_1}}\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{f}_i) + \frac{1}{2}tr\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial b_{j_1 k_1}}\right) \tag{9.4}$$

$$\frac{\partial \mathbf{V}}{\partial \lambda_{j_1 k_1}} = \mathbf{H}\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{j_1 k_1}}\mathbf{H}' = \mathbf{H}\frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Lambda}'}{\partial \lambda_{j_1 k_1}}\mathbf{H}' \tag{9.5}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \lambda_{j_1 k_1}} = \frac{\partial \boldsymbol{\Lambda}\boldsymbol{\Lambda}'}{\partial \lambda_{j_1 k_1}} = \frac{\partial \boldsymbol{\Lambda}}{\partial \lambda_{j_1 k_1}}\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\frac{\partial \boldsymbol{\Lambda}'}{\partial \lambda_{j_1 k_1}} = \mathbf{I}_{j_1 k_1}\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\mathbf{I}_{k_1 j_1} = \tag{9.6}$$

$$\{\delta(j, j_1)\lambda_{k k_1} + \delta(k, j_1)\lambda_{j k_1} \text{ for } j = 1 \text{ to } n, k = 1 \text{ to } n\}$$

Second derivatives:

$$\frac{-\partial^2 \log(l(\mathbf{y}_i \mid \boldsymbol{\phi}, \boldsymbol{\theta}_{\#}))}{\partial \lambda_{j_2 k_2}\partial \lambda_{j_1 k_1}} = \tag{9.7}$$

$$(\mathbf{y}_i-\mathbf{f}_i)'\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \lambda_{j_2k_2}}\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \lambda_{j_1k_1}}\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{f}_i)-\frac{1}{2}(\mathbf{y}_i-\mathbf{f}_i)'\mathbf{V}_i^{-1}\frac{\partial^2 \mathbf{V}_i}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{f}_i)+$$

$$+\frac{1}{2}tr\left(\mathbf{V}_i^{-1}\frac{\partial^2 \mathbf{V}_i}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}\right)-\frac{1}{2}tr\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \lambda_{j_2k_2}}\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \lambda_{j_1k_1}}\right) \tag{9.8}$$

$$\frac{\partial \mathbf{V}}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}=\mathbf{H}\frac{\partial^2 \mathbf{\Sigma}}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}\mathbf{H}'=\mathbf{H}\frac{\partial^2 \mathbf{\Lambda}\mathbf{\Lambda}'}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}\mathbf{H}' \tag{9.9}$$

$$\frac{\partial \mathbf{\Sigma}}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}=\frac{\partial \mathbf{\Lambda}\mathbf{\Lambda}'}{\partial \lambda_{j_2k_2}\partial \lambda_{j_1k_1}}=\frac{\partial \mathbf{\Lambda}}{\partial \lambda_{j_1k_1}}\frac{\partial \mathbf{\Lambda}'}{\partial \lambda_{j_2k_2}}+\frac{\partial \mathbf{\Lambda}}{\partial \lambda_{j_2k_2}}\frac{\partial \mathbf{\Lambda}'}{\partial \lambda_{j_1k_1}}=\mathbf{I}_{j_1k_1}\mathbf{I}_{k_2j_2}+\mathbf{I}_{j_2k_2}\mathbf{I}_{k_1j_1}= \tag{9.10}$$

$$\{\delta(j,j_1)\delta(k_1,k_2)\delta(j_2,k)+\delta(j,j_2)\delta(k_2,k_1)\delta(j_1,k) \text{ for } j=1 \text{ to } n,\ k=1 \text{ to } n\}$$

Since the Sigma parameters are only in the data likelihood portion of the conditional density, then

$$\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\boldsymbol{\theta}_{\neq\#},\boldsymbol{\mu}_i,\mathbf{\Omega}))}{\partial \boldsymbol{\theta}_{\neq\#}}=\frac{\partial -\log(p(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\lambda_{j_1k_1},\boldsymbol{\mu}_i,\mathbf{\Omega}))}{\partial \lambda_{j_1k_1}}$$
$$=\frac{\partial -\log(l(\mathbf{y}_i,\boldsymbol{\phi}\,|\,\lambda_{j_1k_1},\boldsymbol{\mu}_i,\mathbf{\Omega}))}{\partial \lambda_{j_1k_1}} \tag{9.11}$$

which is used in equation (1.48). The Sigma-like theta parameters cannot be processed in this way, because the user defines sigma-like parameters in **H**, with unpredictable functional relationships to that theta.

Appendix G: Degrees of Freedom Assessment for OMEGA Priors

The heuristic justification for Mats Karlsson's formula:

$$N = 2\left(\frac{\Omega}{E_{\Omega}}\right)^2 \tag{10.1}$$

Where N is the number of subjects of the previous analysis, Ω is an omega diagonal element, and E_{Ω} is its standard error (the error in the estimate of Ω), is as follows.

For a normally distributed random variable x, with mean 0, and variance Ω, the following holds:

$$\int xp(x \mid 0,\Omega)dx = \bar{x} = 0 \tag{10.2}$$

Define the random variable y:

$$y = (x - \bar{x})^2 \tag{10.3}$$

So

$$\int (x-\bar{x})^2 p(x \mid 0,\Omega)dx = \int yp(x \mid 0,\Omega)dx = \bar{y} = Var(x) = \Omega \tag{10.4}$$

Finally, the fourth central moment is:

$$\int (x-\bar{x})^4 p(x \mid 0,\Omega)dx = \int x^4 p(x \mid 0,\Omega)dx = 3\Omega^2 \tag{10.5}$$

Then,

$$\begin{aligned} &Var(\Omega) = STD^2(\Omega) = \int (y-\bar{y})^2 p(x \mid 0,\Omega)dx = \int y^2 p(x \mid 0,\Omega)dx - \bar{y}^2 = \\ &\int (x-\bar{x})^4 p(x \mid 0,\Omega)dx - \bar{y}^2 = 3\Omega^2 - \Omega^2 = 2\Omega^2 \end{aligned} \tag{10.6}$$

For N normal random deviates, the variance of the estimate of its average variance is

$$Var(\hat{\Omega}_N) = Var(\Omega)/N = SE^2(\hat{\Omega}_N) = E_{\Omega}^2 \tag{10.7}$$

Thus,

$$E_{\Omega} = \frac{2}{N}\Omega^2 \tag{10.8}$$

That is, the standard error of the variance is related by the above equation, as long as the N items that contribute to its assessment are normally distributed. This is the best error in the inter-subject variance that can be expected in a set of parameters from subjects with rich data for each. In a population analysis, however, some subjects with few data points will not have much information for their parameter. However, population analysis yields

empirical standard errors of Omegas $E_{\boldsymbol{\Omega}}$, that properly reflect the total information available for the Omega. Thus, given $E_{\boldsymbol{\Omega}}$, the "effective N" can be evaluated as:

$$N = 2\left(\frac{\Omega}{E_{\boldsymbol{\Omega}}}\right)^2 \tag{10.9}$$

Appendix H: Technical Note on NonParametric Analysis

Perform a standard FOCE analysis, to produce vectors of η_i, i=1 to N, at the *mode a posteriori* (MAP estimates, or empirical Bayes estimates (EBE)) for each subject i, evaluated at the final population parameters $(\boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{\Sigma})$, where N is the number of subjects, and N_s is the number of support points. These best fit etas for each subject serve as the anchors, or grid points, for the non-parametric analysis, to be evaluated by subroutine NP.

In the subroutine NP, using data of subject i, and grid point $\boldsymbol{\eta}_k$ (which may have come from EBE of subject k of the FOCE analysis for $k<=N$, or random creation of extra support points, for $N<k<=N_s$) , the data likelihood is evaluated:

$$l(\mathbf{y}_i, \boldsymbol{\eta}_k, \boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{10.10}$$

by subroutine OBJ3, and an initial prior (population density) is evaluated as

$$\pi_0(\boldsymbol{\eta}_k) = \exp\left(-\frac{1}{2}\boldsymbol{\eta}_k'\boldsymbol{\Omega}\boldsymbol{\eta}_k\right) \tag{10.11}$$

$\pi(\boldsymbol{\eta}_k)$ is stored in vnonpara(1), and file system of subroutine DAT8.

Like other optimization methods, non-parametric estimation is reiterated, until the objective function no longer changes. At any given iteration of the non-parametric optimization, the following is assessed.

For each subject i with $\mathbf{y}_i$, the set of $\boldsymbol{\eta}$ that yields the largest posterior density is evaluated,

$$l(\mathbf{y}_i, \boldsymbol{\eta}_{m_i}, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_{m_i}) \geq l(\mathbf{y}_i, \boldsymbol{\eta}_k, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_k) \text{ for all } k = 1, N_s \tag{10.12}$$

where $\boldsymbol{\eta}_{m_i}$ produces the largest value of the posterior density for subject i. $l(\mathbf{y}_i, \boldsymbol{\eta}_k, \boldsymbol{\theta}, \boldsymbol{\Sigma})$ is stored in VNONPARA(3). The m_i is stored in IC(I), and $l(\mathbf{y}_i, \boldsymbol{\eta}_{m_i}, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_{m_i})$ is stored in X79(I). The final m_i is stored in VNONPARA(5).

The objective function is evaluated for a given iteration as

$$O = -2\sum_{i=1}^{N}\log\left(\sum_{j=1}^{N_s} l(\mathbf{y}_i, \boldsymbol{\eta}_j, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_j)\right) \tag{10.13}$$

where summing the probability density over all discrete positions $\boldsymbol{\eta}_j$, j=1 to N_s, is the non-parametric or discrete density equivalent to integrating over all $\boldsymbol{\eta}$ for a continuous density function, to obtain a marginal density for each subject i. These marginal densities are in turn multiplied among all subjects i to N, to obtain the joint marginal density. Since the objective function is typically -2log(joint marginal density), it is more convenient to sum the log of the marginal densities among all subjects i, i=1 to N, as shown in the above equation. O is stored in OBJNP, and

$$\sum_{j=1}^{N_s} l(\mathbf{y}_i, \boldsymbol{\eta}_j, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_j) \tag{10.14}$$

is temporarily stored in U(I).

Normalized posterior densities are also evaluated:

$$p_i(\boldsymbol{\eta}_k) = \frac{l(\mathbf{y}_i, \boldsymbol{\eta}_k, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_k)}{\sum_{j=1}^{N_s} l(\mathbf{y}_i, \boldsymbol{\eta}_j, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_j)} \tag{10.15}$$

The $\frac{l(\mathbf{y}_i, \boldsymbol{\eta}_k, \boldsymbol{\theta}, \boldsymbol{\Sigma})}{\sum_{j=1}^{N_s} l(\mathbf{y}_i, \boldsymbol{\eta}_j, \boldsymbol{\theta}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta}_j)}$ term is stored in VNONPARA(2).

The posterior densities are normalized in the sense that

$$\sum_{k=1}^{N_s} p_i(\boldsymbol{\eta}_k) = 1 \tag{10.16}$$

as required for a proper probability density of $\boldsymbol{\eta}$. The final $p_i(\boldsymbol{\eta}_k)$ are stored in row subject i, column IPROB(K), of the .npi file. These are averaged among all subjects at a given anchor point $\boldsymbol{\eta}_k$ to obtain a posterior, or empirical, assessed "weight" at that anchor:

$$p(\boldsymbol{\eta}_k) = \frac{1}{N}\sum_{i=1}^{N} p_i(\boldsymbol{\eta}_k) \tag{10.17}$$

The final values are reported as subject 0, IPROB(K), in the .npi file.

If the following test is satisfied:

$$\delta_k = \left| \frac{p(\boldsymbol{\eta}_k) - \pi(\boldsymbol{\eta}_k)}{\pi(\boldsymbol{\eta}_k)} \right| > \varepsilon_2 \tag{10.18}$$

for some small optimization criterion ε_2, then $p(\boldsymbol{\eta}_k)$ serves as the new prior density for the next iteration:

$$\pi(\boldsymbol{\eta}_k) = p(\boldsymbol{\eta}_k) \quad (10.19)$$

If the test fails, it means that $p(\boldsymbol{\eta}_k)$ is no longer changing sufficiently with respect to its previous value $\pi(\boldsymbol{\eta}_k)$, and the updates no longer need to be performed for future iterations, for that $\boldsymbol{\eta}_k$.

When for all *k* the following is satisfied:

$$\delta_k \le \varepsilon_1 \text{ for } k = 1 \text{ to } N \quad (10.20)$$

for some small ε_1, then the non-parametric optimization is ended.

The following final information is stored:

$\pi(\boldsymbol{\eta}_k)$ is stored in VNONPARA(1), which is retrieved for each *k* via sequential calls to subroutine DAT8.

$$\bar{\boldsymbol{\eta}} = \sum_{k=1}^{N_s} \boldsymbol{\eta}_k \pi(\boldsymbol{\eta}_k) \quad (10.21)$$

is stored in EXNPETA(), EXETA().

$$\bar{\boldsymbol{\Omega}} = \sum_{k=1}^{N_s} (\boldsymbol{\eta}_k - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_k - \bar{\boldsymbol{\eta}})' \pi(\boldsymbol{\eta}_k) \quad (10.22)$$

is stored in COVNPETA(), COVETA().

The expected values EXNPETA(j), and expected covariances COVNPETA(j,k) are reported as ETA(j) and ETC(j,k), respectively, in the .npe file.

If INPETA/=0, then

$\boldsymbol{\eta}_{m_i}$ is stored in VNONPARA(2) to VNONPARA(1+*neta*), where *neta* is the eta vector length, retrieved for each subject *i* via sequential calls to subroutine DAT8, and also placed in the .npd file, labeled as ETM(j), pertaining to eta(j). That is, the grid point eta vector that best fits subject *i* is stored in record *i* of the DAT8 storage system (the entire DAT8 storage contains N_s records, where $N_s \ge N$).

If INPETA=0, then cumulative distributions are stored as follows:

$$C(\eta_{k(j)}) = \sum_{i \in (\eta_{i(j)} \le \eta_{k(j)})}^{N_s} \pi(\boldsymbol{\eta}_i) \qquad (10.23)$$

is stored in VNONPARA(1+*j*), retrieved for each support point *k* sequentially from subroutine DAT8, and also placed in the .npd file, labeled as CUM(j). Here, $\eta_{i(j)}$ is the *j*th element of the eta vector belonging to support point *i*. That is, $C(\eta_{k(j)})$ is the sum of densities $\pi(\boldsymbol{\eta}_i)$ for which the *j*th element of eta is less than or equal to the jth element of $\boldsymbol{\eta}_k$ (which is $\eta_{k(j)}$). In turn, $\eta_{k(j)}$ is stored in VETA(*j*), retrieved for each support point *k* from subroutine DAT4, and reported as eta(j) in the .npd file. The $\pi(\boldsymbol{\eta}_k)$ for support point k is reported in the .npd file in the last column, labeled as PROBABILITY.

See references [16] and [17] for information of supplementary support points and bootstrapping.

Appendix I: Note on TNPRI

The statistical basis of the frequentist method (TNPRI which stands for **T**otal **N**ormal **PRI**or)) for priors is that of sampling about some mean, with some measure of dispersion, but not requiring a rigid rule of a particular distribution, other than that it has some semblance of normal distribution where this makes sense.

Regarding THETAS:
Let THETA be the theta estimate, and SE be the standard error of the theta estimate.

When no boundaries are given in the $THETA record, then $SIML will generate random sample thetas that are normally distributed with mean THETA and variance SE^2, and with suitable correction for correlation between theta(1) and theta(2), etc., in accordance with the variance-covariance matrix of the estimates.

When boundary is set, an intermediate variable normal deviate v will be generated with mean log(THETA), and variance $(SE/THETA)^2$. Again, the random deviate v for each theta(1), theta(2), etc., is corrected to account for correlation (covariance of estimates) between theta(1) and theta(2), etc.

This log-normal deviate v is then transformed to a final theta sample w as follows:
Lower bound only:
w=exp(v)+LB
(range of v(-inf,+inf) transposes to range of w=(LB,+inf))

upper bound only:
w=ub-exp(-v)
(range of v(-inf,+inf) transposes to range of w=(-inf,UB))

Lower and upper bound:
w=LB + (UB-LB)*exp(V)/(1+exp(v))

(range of v(-inf,+inf) transposes to range of w=(LB,UB))

When SE/mean is small, then the resulting distribution of the thetas is nearly normally distributed with mean THETA and standard error SE, as reflected from the original estimates and standard errors. When SE/mean is large, this creates considerable non-normal distribution in the samples.

Regarding OMEGAS (and Sigmas):
In the case of the OMEGAS, using, the mean OMEGA and its standard error from a previous problem, the TNPRI method transposes this into its cholesky form and its equivalent standard error, so that each element in the cholesky matrix has the appropriate "mean and standard error". Transforming the mean omega elements and their standard errors into the equivalent mean and standard errors for the cholesky elements is not trivial, but it can be done using matrix algebra and the principal of propagation of errors. The principle of propagation of errors itself is accurate only as an asymptotic rule, that is, if the error is sufficiently small, then it provides reasonable results.

Also, while having omega elements be normally distributed is not reasonable, the underlying cholesky elements can be modeled as normally distributed. Further, the off-diagonal elements are allowed to be positive or negative, and these elements are sampled as a strict normal distribution. The diagonal elements of the cholesky matrix can also be normally sampled, but with the proviso that the sample be positive. Thus, as with thetas, if the standard error is small relative to the mean value, very few if any samples will be negative. If the standard error is large, then a great many samples will end up as negative. To avoid negative values altogether, the additional transformation that is done for the cholesky diagonal is to generate a random sample as x=log(cholesky estimate)+(se of cholesky estimate)/(cholesky estimate)*r, where r is created as a ~N(0,1) deviate, and then exponentiate x, and this is the cholesky diagonal sample. When se/mean is small, exp(log(mean)+SE/mean*r) is very nearly normally distributed with the appropriate mean and standard error.

Next, the cholesky matrix is multiplied by its transpose to create the random OMEGA sample, with the result of being positive definite, and having the appropriate mean and dispersion.

Appendix J: T distribution Sample Generation

The tdist6 and tdist7 examples described in intro7.pdf use the fact that two normal random deviates can be converted into a T distributed normal random deviate. The derivation is as follows.

From Numerical Recipes reference [19], the Box-Muller algorithms for creating a random normal deviate pair is as follows. Two uniform random deviate pairs u and v are generated, and modified:

$$U = 2u - 1 \tag{12.1}$$

$$V = 2v - 1 \tag{12.2}$$

$$W = U^2 + V^2 \tag{12.3}$$

Values of W>=1.0 are rejected.

Normal deviates x and y are generated as follows:

$$x = U\sqrt{-2LOG(W)/W} \tag{12.4}$$

$$y = V\sqrt{-2LOG(W)/W} \tag{13.1}$$

According to [20], a t-distribution sample of n degrees of freedom can be generated as

$$t = \frac{U}{\sqrt{W}}\sqrt{n\left(W^{-2/n} - 1\right)} \tag{13.2}$$

This suggests that rather than starting with uniform deviates u and v, one can use normal deviates x and y to generate the t deviate:

$$x^2 + y^2 = -2\log(W) \tag{13.3}$$

$$W = \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \tag{13.4}$$

$$\frac{x}{\sqrt{x^2 + y^2}} = \frac{U\sqrt{-2LOG(W)/W}}{\sqrt{-2LOG(W)}} = \frac{U}{\sqrt{W}} \tag{13.5}$$

So,

$$t = \frac{x}{\sqrt{x^2 + y^2}}\sqrt{n\left(\exp\left((x^2 + y^2)/n\right) - 1.0\right)} \tag{13.6}$$

This transformation is used in the tdist6 and tdist7 examples.

From the alternative method of the Box-Muller method ([19]) of using trigonometric functions, we note that uniform deviates u and v can be generated from the normal deviates:

$$u = \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \tag{13.7}$$

$$v = \frac{1}{2\pi}\tan^{-1}\left(\frac{y}{x}\right) = \frac{1}{2\pi}\cos^{-1}\left(\frac{x}{\sqrt{x^2 + y^2}}\right) \tag{13.8}$$

Therefore,

$$t = \cos(2\pi v))\sqrt{n\left(u^{-2/n} - 1.0\right)} \tag{13.9}$$

which is a means of generating a t sample using a uniform deviate pair, without having to reject uniform random samples, and is the method used for t sample generating function TDEV2 in ..\source\GENERAL.f90, for general degrees of freedom n.

References

[1] Beal SL, Sheiner LB. NONMEM Users Guide - Part VII. 1992.

[2] Wang Y. Derivation of various NONMEM estimation methods. J. Pharmacokinetics and Phamracodynamics, 2007; 34(5):575-93.

[3] Dennis JE, Jr., Schnabel RB. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Society for Industrial and Applied Mathematics (1996).

[4] Schumitzky A. EM algorithms and two stage methods in pharmacokinetics population analysis. In: D'Argenio DZ, ed. *Advanced Methods of Pharmacokinetic and Pharmacodynamic Systems Analysis.* Vol 2. Boston: Kluwer Academic Publishers; 1995:145-160.

[5] Bauer RJ, Guzy S. Monte Carlo parametric expectation maximization (MC-PEM) method for analyzing population pharmacokinetic/pharmacodynamic data. In: D'Argenio DZ, ed. *Advanced Methods of Pharmacokinetic and Pharmacodynamic Systems Analysis.* Vol 3. Boston: Kluwer Academic Publishers; 2004:135-163.

[6] Lavielle, M. *Monolix Users Manual* [computer program]. Version 2.1. Orsay, France: Laboratoire de Mathematiques, U. Paris-Sud; 2007.

[7] Gilks, Richardson and Spiegelhalter. Introducing Markov chain Monte Carlo. In: Markov Chain Monte Carlo in Practice. W.R. Gilks et al., Chapman & Hall (1996), chapter 1, pp 5-8.

[8] Benet, Racine-Poone, and Wakefield. MCMC for non linear hierarchical models. In: Markov Chain Monte Carlo in Practice. W.R. Gilks et al., Chapman & Hall (1996), chapter 19, pp 341-342.

[9] Hooker AC, Staatz CE, Karlsson MO. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. Pharmaceutical research 2007; 24: 2187-97.

[10] Comets E, Brendel K, Mentre F. Computing normalized prediction distribution errors to evaluate nonlinear mixed effects models: the npde add-on package for R. Computer Methods and Programs in Biometrics 2008; 90:154-166.

[11] Brendel K, Comets E, Laffont C, Laveille C, Mentre´ F. Metrics for External Model Evaluation with an Application to the Population Pharmacokinetics of Gliclazide. Pharmaceutical Research, 2006; 23(9): 2036-2049.

Further Reading:

[12] Karlsson MO and Savic RM. Diagnosing Model Diagnostics. Clinical Pharmacolocy and Therapeutics, 2007; 82(1): 17-20.

[13] Bauer RJ, Guzy S, Ng CM. A survey of population analysis methods and software for complex pharmacokinetic and pharmacodynamic models with examples. AAPS Journal 2007; 9(1):E60-83.

[14] Press WH, Teukolsky SA, Vettering WT, Flannery BP. Numerical Recipes, The Art Of Scientifc Programming. 2nd Edition, Cambridge University Press, New York, 1992, pp. 180-184.

[15] McLachlan GJ and Krishnan T. The EM Algorithm and Extensions. Wiley Series in Probability and Statistics, John Wiley and Sons, Inc. (1997).

[16] Savic RM, Karlsson MO. Evaluation of an extended grid method using nonparametric distributions. AAPS Journal. 2009; 11(3): 615-627.

[17] Baverel PG, Savic RM, Karlsson MO. Two bootstrapping routines for obtaining imprecision estimates for nonparametric parameter distributions in nonlinear mixed effects models. J. Pharmacokinetics and Pharmacodynamics 2011; 38(1):63-82.

[18] Almquist J, Leander J, Jirstrand M. Using sensitivity equations for computing gradients of the FOCE and FOCEI approximations to the population likelihood. J Pharmacokinetics and Pharmacodynamics. (2015) 42:191–209.

[19] Press WH, Teukolsky SA, Vettering WT, Flannery BP. Numerical Recipes, The Art Of Scientifc Programming. 2nd Edition, Cambridge University Press, New York, 1992, pp. 279-280.

[20] Shaw, WT, Sampling Student's T distribution – use of the inverse cumulative distribution function. Journal of Computational Finance, Volume 9/Number 4, Summer 2006. pp. 37-73.