Web-based Supplementary Materials for “Mixture joint models for event time and longitudinal data with multiple features” by Y. Huang et al.

Appendix A. HIV dynamic models and mixture of joint models related issues

Viral dynamic models can be formulated through a system of ordinary differential equations (ODE) (Guedj et al., 2010; Huang et al., 2006; Perelson et al., 1997; Wu and Ding, 1999). Under some reasonable assumptions and simplifications, two useful approximations of ODE solution, which can be used to capture viral load responses, have been proposed as follows.

$$V(t) = \log_{10}(e^{p_1 - \lambda_1 t}) \quad (\text{A.1})$$

$$V(t) = \log_{10}(e^{p_1 - \lambda_1 t} + e^{p_2 - \lambda_2 t}) \quad (\text{A.2})$$

where $V(t)$ is the $\log_{10}$ scaled plasma HIV-1 RNA levels at time t . λ_1 and λ_2 are called the first- and second-phase viral decay rates, which may represent the minimum turnover rate of productively infected cells and that of latently or long-lived infected cells, respectively (Perelson et al., 1997). The parameters p_1 and p_2 are macro-parameters; e^{p_1} and $e^{p_1} + e^{p_2}$ are the baseline viral load at time $t = 0$ in one- and two-compartment models, respectively. It is generally assumed that $\lambda_1 > \lambda_2$, which assures that the model is identifiable and is appropriate for empirical studies (Wu and Ding, 1999). Models (A.1) and (A.2) offer almost equal performance to capture the early fast-decaying segment of viral load trajectory (Wu and Ding, 1999), but model (A.2) performs better for long term of viral load trajectory (Huang and Dagne, 2011). It was noted that both models can be only applied to the early segment or longer term of the viral load response with decreasing trajectory patterns as discussed in Section 1 and shown in Figure 1(a) (two solid and two dashed lines), but the viral load trajectory may have a different shape or rebound in later stages (two dotted lines in Figure 1(a)). It was also noted that for some patients the second viral decay rate λ_2 , may vary over time because they depend on some phenomenological parameters that hide with considerable microscopic complexity and change over time. Negative values of the decay rates may correspond to viral increase and lead to viral rebound (Wu and Ding, 1999), suggesting that variation in the dynamic parameters, particularly λ_2 , may be partially associated with time-varying covariates such as repeated CD4 cell counts to capture the viral load change including viral rebound. Thus, we the following extended function are introduced

(Wu and Zhang, 2006).

$$V(t) = \log_{10}(e^{p_1 - \lambda_1 t} + e^{p_2 - \lambda_2(t)t}) \quad (\text{A.3})$$

where the second-phase decay rate $\lambda_2(t)$ is either a function of time-varying CD4 cell count (covariate) and/or an unknown smooth function. Intuitively, this model is more flexible because it assumes that the second-phase viral decay rate can vary with time as a result of drug resistance, medication adherence and other relevant clinical factors likely to affect changes in the viral load during the treatment. Therefore, all data obtained during whole study period can be used by fitting this model. The following are a couple of issues related to the proposed mixture of joint models which should be noted.

1. The models published in the literature (Huang and Dagne, 2011; Huang et al., 2011; Lu and Huang, 2014) can be treated as special cases of MNLMEJ model proposed in this article when the MNLMEJ model is tailored as follows. In comparison with the proposed MNLMEJ model which can be offered to not only estimate all parameters simultaneously, but also evaluate probabilities of class membership, Huang and Dagne (2011) considered only commonly used NLME model with mean function specified by (11) for viral load response (not a mixture) in the presence of CD4 measurement error specified by a standard linear mixed-effects model without time-to-event model involved; Huang et al. (2011) investigated longitudinal-survival data analysis based on commonly used NLME model with mean function specified by (11) for viral load response, but mixture of model structure is ignored in this sense that classification of viral load trajectories was not evaluated; although Lu and Huang (2014) studied a finite mixture of NLME models, both measurement error model for CD4 covariate and AFT model for time to decline of CD4/CD8 ratio were not considered in the sense that model in Lu and Huang (2014) is not conducted under framework of joint models.
2. Although both time-to-event and mixture of NLME response models are related to CD4 cell count, CD4 cell count was incorporated into these two models for different purposes. The mixture of NLME response model accommodated (expected) time-varying CD4 cell count as a covariate into the second-phase decay rate to predict viral trajectories, while the CD4 cell count was used to determine time to first decline in the CD4/CD8 ratio which served as an outcome in the time-to-event model.

3. For the AFT model with the Dirichlet Process (DP) prior distribution, a popular way for specifying the DP prior $DP(\eta G_0)$ is the stick-breaking prior representation. That is, the distribution of ε_i can be expressed as

$$\varepsilon_i \sim \sum_{g=1}^{\infty} w_g N(\mu_g, \kappa_g) \text{ with } \mu_g = \mu_g^* - \sum_{g=1}^{\infty} w_g \mu_g^*, (\mu_g^*, \kappa_g) \sim G_0,$$

where w_g is a random probability weight chosen to be independent of (μ_g^*, κ_g) such that $0 \leq w_g \leq 1$ and $\sum_{g=1}^{\infty} w_g = 1$. For the sake of computational convenience, we consider the following mixture formulation of the truncated approximation DP.

$$\varepsilon_i \sim \sum_{g=1}^L w_g N(\mu_g, \kappa_g), \text{ with } \mu_g = \mu_g^* - \sum_{g=1}^L w_g \mu_g^*, (\mu_g^*, \kappa_g) \sim G_0, \quad (\text{A.4})$$

where $0 \leq G < \infty$ and w_g is defined by the following stick-breaking procedure.

$$w_1 = \nu_1 \text{ and } w_g = \nu_g \prod_{l=1}^{g-1} (1 - \nu_l) \text{ for } g = 2, \dots, G, \quad (\text{A.5})$$

where $\nu_g \sim \text{Beta}(1, \eta)$ for $g = 1, 2, \dots, L-1$ and $\nu_L = 1$ so that $\sum_{g=1}^L w_g = 1$. The prior distribution for the unknown parameter η is given by $\Gamma(\eta_{10}, \eta_{20})$ with prespecified hyperparameters η_{10} and η_{20} , and the details for selecting L can refer to [Ishwaran and Zarepour \(2000\)](#).

Appendix B. Conditional posterior distributions of representative parameters

For notational convenience, let the observed data of the i th individual $\mathfrak{R}_i = \{(\mathbf{q}_i, \mathbf{d}_i, \mathbf{z}_i, z_{i0}, u_i, v_i)\}$, $\mathfrak{R} = \{(\mathbf{q}_i, \mathbf{d}_i, \mathbf{z}_i, z_{i0}, u_i, v_i), i = 1, \dots, n\}$, $\Psi = \{\beta_{ij}, i = 1, \dots, n; j = 1, \dots, n_i\}$, $\mathbf{a} = \{\mathbf{a}_i, i = 1, \dots, n\}$. The posterior distributions of latent indicators c_i , ($i = 1, \dots, n$), and mixture weights $\boldsymbol{\pi}$ are given in (17) and (18), respectively. Following the studies by [Davidian and Giltinan \(1995\)](#) and [Gelfand et al. \(1990\)](#), it can be shown from (6), (14) and (19) that the full conditional distributions of $\boldsymbol{\beta}$, σ_2^2 and $\boldsymbol{\Sigma}_b$ may be written explicitly as

$$\boldsymbol{\beta} | \boldsymbol{\alpha}, \mathbf{a}, \Psi, \boldsymbol{\Sigma}_b, \mathfrak{R} \sim N(\boldsymbol{\beta}'_0, \boldsymbol{\Lambda}'_2), \quad (\text{B.1})$$

$$\sigma_2^2 | \mathbf{c}, \Psi, \mathfrak{R} \sim IG(\omega'_3, \omega'_4), \quad (\text{B.2})$$

$$\Sigma_b | \alpha, \mathbf{a}, \Psi, \beta, \mathfrak{R} \sim IW(\Omega'_2, v'_2), \quad (\text{B.3})$$

where

$$\begin{aligned} \beta'_0 &= \Lambda'_2 \left[\sum_i n_i^{-2} \left(\sum_j \mathbf{Z}_{ij}^T \right) (\mathbf{X} \Sigma_b \mathbf{X}^T)^{-1} \left(\sum_j \beta_{ij} \right) + \Lambda_2^{-1} \tau_2 \right], \\ \Lambda'_2 &= \left[\sum_i n_i^{-2} \left(\sum_j \mathbf{Z}_{ij}^T \right) (\mathbf{X} \Sigma_b \mathbf{X}^T)^{-1} \left(\sum_j \mathbf{Z}_{ij} \right) + \Lambda_2^{-1} \right]^{-1}; \\ \omega'_3 &= \omega_3 + 2^{-1} \sum_i n_i, \quad \omega'_4 = \omega_4 + 2^{-1} \sum_i \sum_j [y_{ij} - g_{c_i}(t_{ij}, \mathbf{A}_{c_i} \beta_{ij})]^2; \\ \Omega'_2 &= \Omega_2 + \sum_i n_i^{-2} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left[\sum_j \beta_{ij} - \left(\sum_j \mathbf{Z}_{ij} \right) \beta \right] \times \left[\sum_j \beta_{ij} - \left(\sum_j \mathbf{Z}_{ij} \right) \beta \right]^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}, \\ v'_2 &= v_2 + n. \end{aligned}$$

The full conditional distribution of each $\mathbf{b}_i$, however, cannot explicitly be expressed, given the remaining parameters and data. The distribution of $\mathbf{b}_i$ has a density function that is related to $\exp\{-(2\sigma_2^2)^{-1} \sum_j [y_{ij} - g_{c_i}(t_{ij}, \mathbf{A}_{c_i}(\mathbf{Z}_{ij}\beta + \mathbf{X}\mathbf{b}_i))]^2 - 2^{-1} \mathbf{b}_i^T \Sigma_b^{-1} \mathbf{b}_i\}$ and equation (13).

The full conditional distributions of α , σ_1^2 , Σ_a and random-effects $\mathbf{a}_i$ can be derived similarly to those presented above; the full conditional distribution of γ , given the remaining parameters and data, cannot be calculated explicitly and depends on the prior, random-effects $\mathbf{b}_i$ and $\mathbf{a}_i$ as well as time-to-event model (13). Thus, their derivations are omitted to save space.

Appendix C. Prior information, Gelman-Rubin diagnostics, goodness-of-fit and model comparison

To carry out the Bayesian inference, we took weakly-informative prior distributions for the parameters. In particular, (i) fixed-effects were taken to be independent normal distribution $N(0, 100)$ for each element of the population parameter vectors α , β and γ ; (ii) we assume a noninformative inverse Gamma prior distribution $IG(0.01, 0.01)$, which has mean 1 and variance 100, for variance parameters σ_1^2 and σ_2^2 ; (iii) the priors for the variance-covariance matrices of the random-effects Σ_a and Σ_b were taken to be inverse Wishart distributions $IW(\Omega_1, \nu_1)$ and $IW(\Omega_2, \nu_2)$, where the diagonal elements for diagonal variance matrix Ω_1 and Ω_2 were 0.01, and $\nu_1 = \nu_2 = 4$; and (iv) we set hyper-parameters of Dirichlet distribution in (16), $\phi_1 = \phi_2 = \phi_3 = 1$, assuming individuals have equal probabilities of coming from any one of three classes initially. (v) The hyper-parameters for the DP prior are assigned by $\zeta_0 = 0, \eta_{10} = 0.5, \eta_{20} = 0.01$ and $\sigma_0^2 = 100$.

When the MCMC procedure was applied to the actual clinical data, convergence of the generated samples was assessed using standard tools within WinBUGS software such as trace plots and Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Figure I shows the trace plots, autocorrelation and dynamic version of GR diagnostic plots based on Model I as obtained from the WinBUGS software for the representative and important parameters α_2 , β_2 and σ_2^2 . We observe from trace plots (top panel) that the lines of three different chains mix or cross in trace, implying that convergence is ensured. For the plots GR diagnostics (bottom panel) where the three curves are given: the middle and bottom curves below the dashed horizontal line (indicated by the value one) represent the pooled posterior variance ($\hat{V}$) and average within-sample variance ($\hat{W}$), respectively, and the top curve represents their ratio ($\hat{R}$). It is seen that $\hat{R}$ is generally expected to be higher than one at the initial stage of the algorithms, but $\hat{R}$ tends to 1, and $\hat{V}$ and $\hat{W}$ stabilize as the number of iterations increase, indicating that the algorithm has approached convergence. We further monitor convergence using autocorrelation plots (middle panel) that autocorrelations are very low with a lag being 50, implying that convergence is obtained.

To assess the goodness-of-fit of the proposed models, the diagnosis plots of the fitted values versus the observed values (top panel) and normal Q-Q plots (bottom panel) from Models I, II and III are presented in Figure II. It was seen from Figure II (top panel) that Model I provides a better fit to the observed viral load data, compared with Models II and III which performs similarly. This finding can also be explained by examining the Q-Q plots of the residuals (bottom panel) that all plots show the existence of outliers, but it is clearly seen that Model I has few outliers and thus, provides a better goodness-of-fit to the data than Models II and III.

To select the best model that fits the data adequately, a Bayesian selection criterion, known as DIC suggested by Spiegelhalter et al. (2002), is used. As with other model selection criteria, we caution that DIC is not intended for identification of the “correct” model, but rather merely as a method of comparing a collection of alternative formulations. As an alternative, we also evaluate EPD formulated by $EPD = E \left\{ \sum_{i,j} (y_{rep,ij} - y_{obs,ij})^2 \right\}$ for model comparison, where the predictive value $y_{rep,ij}$ is a replicate of the observed $y_{obs,ij}$ and the expectation is taken over the posterior distribution of the model parameters θ (see Gelman et al. (2003) in detail). This criterion chooses the model where the discrepancy between predictive values and observed values

is the lowest. The estimated DICs based on Models I, II and III are 10368.9, 12572.8 and 14099.1, respectively. As mentioned previously, it is difficult to tell which model is “correct”, only which one fits the data better. Therefore, based on the DIC, the results indicate that Model I is the best fitting model and Model II is next, supporting the contention of a heterogeneous data feature. This finding is confirmed by the results of the EPD values (see Table 2). These results are also consistent with those in diagnosis of the goodness-of-fit displayed in Figure 2, indicating that Model I performs best.

Figure I. Convergence diagnostics with three Markov chains as obtained from the WinBUGS software for representative parameters based on Model I: (a) Trace plots (top panel); (b) Auto-correlation plots (middle panel); (c) Gelman-Rubin (GR) diagnostic plots (bottom panel). The middle (green) and bottom (blue) curves below the dashed horizontal line (indicated by the value one) represent the pooled posterior variance ($\hat{V}$) and average within-sample variance ($\hat{W}$), respectively, and the top (red) curve above the dashed horizontal line represents their ratio ($\hat{R}$).

Figure II. Goodness of fit: (a) Fitted values versus observed values of $\log_{10}(\text{RNA})$ (top panel). (b) Normal Q-Q plots with line (bottom panel).

Appendix D. R and WinBUGS program codes for mixture joint model

```
## R program code
library(arm)
library(R2WinBUGS)
data<-list(.....)
inits<-list() # Run three chains
inits[[1]]<-list(.....)
inits[[2]]<-list(.....)
inits[[3]]<-list(.....)
parameters<-c(.....)
mixture<-bugs(data, inits, parameters, "Winbugs.txt",n.chains=3,
               n.thin=50, n.iter=100000,n.burnin=50000,bugs.seed=654321,
               bugs.directory="C:\\Users\\WinBUGS14",DIC=TRUE,debug=TRUE)

## WinBUGS program code named Winbugs.txt
model {
for (i in 1:n)
{# Latent class membership
  c[i]~dcat(p[1:K])    # K=3
# Random effects of response model
  for(k in 1:4){b2[i,k]<-0}
  b[i,1:4]~dmnorm(b2[i,1:4],0mega2[,])
# Random effects of covariate model
  for(k in 1:3){a3[i,k]<-0}
  a[i,1:3]~dmnorm(a3[i,1:3],0mega3[,])
#(1) Event time models with DP prior distribution
  T.mean[i]<-gama[1] +gama[2]*a[i,1]+gama[3]*a[i,2]
               +gama[4]*a[i,3]+gama[5]*b[i,2]+gama[6]*b[i,4]
# Stick-breaking prior representation for DP prior distribution
  p.dp[i,1] <- r[i,1]
  for(m in 2:L)
    {p.dp[i,m] <- r[i,m]*(1-r[i,m-1])*p.dp[i,m-1]/r[i,m-1]}
  p.sum[i] <-sum(p.dp[i,])
  for (m in 1:L)
  {
    mu.b3[i,m] ~ dnorm(T.mean[i], tau4)
    r[i,m] ~ dbeta(1,alpha.r)
    pi[i,m] <-p.dp[i,m]/p.sum[i]          ## to ensure the sum is 1
  }
  group[i] ~ dcat(pi[i,])                  ## group[i]=1 to L=30
  Time[i] ~dlnorm(mu.b3[i,group[i]], tau3)I(T[i,3],T[i,5]) ## interval cesoring##
```



```

} # End of i
for(j in 1:N)
{# (2) Modelling mismeasured CD4 with normal distribution
  #Z[,j]= values of spline basis function (j=2,3,4)
  z.star[j]<-(alpha[1]+a[y[j,2],1])*Z[j,2]+(alpha[2]+a[y[j,2],2])*Z[j,3]
    +(alpha[3]+a[y[j,2],3])*Z[j,4]
  y[j,19] ~ dnorm(z.star[j],tau1) #y[j,2]=patid; y[,19]=sqrt(CD4)
# (3) Three NLME response models for mixture with N distribution
p1[j] <-beta[1] +b[y[j,2],1]
lambda1[j] <-beta[2]+beta[3]*sqrt(y[j,12]) +b[y[j,2],2]
p2[j] <-beta[4] +b[y[j,2],3]
lambda2[j] <-beta[5] +b[y[j,2],4]
lambda2TV[j]<-beta[5]+beta[6]*z.star[j] +b[y[j,2],4]
dm1[j]<-p1[j]-lambda1[j]*y[j,10] #y[,10]=day
dm11[j] <-p1[j]-step(lambda1[j]-lambda2[j])* lambda1[j] *y[j,10]
dm111[j]<-p1[j]-step(lambda1[j]-lambda2TV[j])*lambda1[j] *y[j,10]
dm2[j] <-p2[j]-step(lambda1[j]-lambda2[j])* lambda2[j] *y[j,10]
dm3[j] <-p2[j]-step(lambda1[j]-lambda2TV[j])*lambda2TV[j]*y[j,10]
dm4[j] <-exp(dm1[j])
dm44[j] <-exp(dm11[j])
dm444[j]<-exp(dm111[j])
dm5[j] <-exp(dm2[j])
dm6[j] <-exp(dm3[j])
dm7[j]<-dm44[j]+dm5[j]
dm8[j]<-dm444[j]+dm6[j]
# Limit of detection (LOD)
detect[j] <-log(25)/log(10)
detect.std[j]<-(detect[j]-m.logrna)/sd.logrna
#m.logrna=mean(logrna),sd.logrna=std(logrna)
pi[j]<-phi(detect.std[j]) #pr(y <=detect.std[j])
state[j] ~ dbern(pi[j]) #censoring indicator of LOD or not
# function 1
mu[j,1]<-detect[j]*state[j] +log(dm4[j])/log(10)*(1-state[j])
# function 2
mu[j,2]<-detect[j]*state[j] +log(dm7[j])/log(10)*(1-state[j])
# function 3
mu[j,3]<-detect[j]*state[j] +log(dm8[j])/log(10)*(1-state[j])
y[j,20]~dnorm(mu[j,c[y[j,2]]], tau2) #y[,20]=logrna
## Fitted values and residuals
fit[j]<-mu[j,c[y[j,2]]]

```

```
resid[j]<-y[j,20]-fit[j]
ssr.r[j]<-pow(resid[j],2)      #squares of residuals
}
# Prior specifications
}
```