Bayesian Parametric Modeling for the Estimation of
the Mean Window Period of HIV Infections

Marian Farah∗, Brian Tom†, Michael Sweeting‡, and Daniela De Angelis§

25 July, 2013

Abstract

In this work, we investigate a Bayesian parametric inverse-regression approach to estimate the ‘mean window period’ of HIV infections. Defined as the average time from seroconversion until a candidate biomarker crosses a specified threshold, the mean window period is a key component in cross-sectional methods for HIV incidence estimation. We work with an inverse-regression approach utilized by Sweeting et al. (2010) based on a nonlinear mixed effects model for biomarker data collected from a cohort of seroconverters. Here, we extend that approach to other parametric models using a fully Bayesian framework and derive the analytic formulas for the posterior distribution of the mean window period for each model. Our methods are illustrated on a simulated dataset of HIV seroconverters.

KEY WORDS: HIV incidence; HIV mean window period; HIV infection recency; Inverse-regression; parametric mixed effects model.

∗Marian Farah (marian.farah@mrc-bsu.cam.ac.uk) is a Career Development Fellow, MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB20SR.
†Brian Tom (brian.tom@mrc-bsu.cam.ac.uk) is a Programme Leader, MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB20SR.
‡Michael Sweeting (mjs212@medschl.cam.ac.uk) is a Senior Research Associate, Department of Public Health & Primary Care, University of Cambridge, Cambridge, CB18RN.
§Daniela De Angelis (daniela.deangelis@mrc-bsu.cam.ac.uk) is a Programme Leader, MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB20SR, and a senior statistician, Public Health England, Statistics Unit, HPA Colindale, 61 Colindale Avenue, London NW9 5EQ, UK.
1 Introduction

HIV incidence is the rate at which new HIV infections occur in a population. Accurate HIV incidence estimation is important to monitor the epidemic, evaluate the effectiveness of HIV prevention policies, and improve HIV treatment programs. A straight-forward method for incidence estimation is based on cohort studies that follow uninfected individuals over an appropriate time interval. However, due to sample size requirements and observational biases stemming from follow-up inconsistencies (Brookmeyer, 2009), researchers have been putting more effort into indirect methods for incidence estimation. In particular, the literature has concentrated on methods based on a single cross-sectional survey (e.g., Kaplan, 1997; Karon et al., 2008; Brookmeyer, 2009). Here, incidence is derived by combining the estimated prevalence of recent infections from the cross-sectional survey with the mean ‘recency period’ (or mean window period) of HIV infections, where the latter is estimated from biomarker data based on tests that identify infection recency (Brookmeyer and Quinn, 1995; Janssen et al., 1998; McDougal et al., 2006; Hargrove et al., 2008).

In general, a recency test classifies a seroconverter as recently infected if the relevant biomarker is measured below a specified threshold, \( \alpha \). The time from seroconversion until the biomarker crosses \( \alpha \) is the window period for the seroconverter. The mean window period, \( \mu(\alpha) \), is the population average time spent by the biomarker below \( \alpha \) (Sweeting et al., 2010). We focus on nonlinear regression methods that utilize a mixed effects structure such as those considered by Sweeting et al. (2010) and Sommen et al. (2011). These approaches model the evolution of the observed biomarker as noisy measurements from an underlying nonlinear curve and explicitly allow for the estimation of the seroconversion time distribution. Sweeting et al. (2010) assume the noise is purely due to measurement errors, which they model as independent Gaussian random variables, while Sommen et al. (2011) model the noise as a sum of two components: serial correlations represented by a Brownian motion, and pure measurement errors modelled as independent Gaussian random variables. Based on the posterior distributions of their model parameters, including the seroconversion times, Sweeting et al (2010) implement a fully Bayesian inverse-regression approach to obtain the posterior distribution of the mean window period. On the other hand, Sommen et al. (2011)
do not obtain an estimate of the mean window period, and instead work directly with the estimated posterior distributions of the seroconversion times, which are obtained based on the use of plug-in estimates for the parameters of the biomarker curves.

Described in Section 2, the inverse-regression approach used in Sweeting et al. (2010) is attractive because it utilizes already available posterior samples of the model parameters to obtain samples from the posterior distribution of the mean window period. To do so, a threshold is specified, and the nonlinear function describing the (true) biomarker curve for each individual is solved for the threshold crossing time. This calculation is obtained for each individual over each posterior sample of the model parameters. Then, at each posterior sample, the mean window period is taken to be the average crossing time over all individuals. In Section 2, we extend this fully Bayesian approach to the three parametric curves utilized by Sommen et al. (2011) and derive the inverse-regression expressions for the mean window period. In Section 3, we illustrate this approach using a simulated dataset of biomarker measurements from a cohort of seroconverters. In Section 4, we conclude with a discussion.

2 Methods

2.1 Notation and setup

Suppose serconverter data on \( p \) individuals consist of the following information for each individual \( i, i = 1, \ldots, p \):

- the date of the last negative HIV test result \( \ell_i^- \).
- the date of the first positive HIV test result \( f_i^+ \).
- a sequence of biomarker measurements \( y_{ij} \) observed at times \( t_{ij}, j = 1, \ldots, n_p \), since \( f_i^+ \).

Let \( u_i \) be the unknown time from seroconversion to \( f_i^+ \) for individual \( i \). Then, the growth of the biomarker since seroconversion can be described as

\[
y_{ij} = g(t_{ij} + u_i, \Phi_i) + \epsilon_{ij},
\]

(1)
where $g(t_{ij} + u_i, \phi_i)$ is the “true” underlying functional form of the biomarker curve for individual $i$, and $\phi_i$ is the associated vector of parameters governing that curve. Here, $\epsilon_{ij}$ are the error terms, typically modelled as independent Gaussian random variables with mean zero and variance $\sigma^2$ (Sweeting et al., 2010). We note that Sommen et al. (2011) model the error terms as the sum of a serial correlation component, represented by a Brownian motion, and a pure measurement error component, modelled as independent Gaussian random variables. In this work, we only consider a pure measurement error component.

Both Sweeting et al. (2010) and Sommen et al. (2011) assume $g(t_{ij} + u_i, \phi_i)$ is a monotonically increasing function of time with a plateau. Such assumption corresponds to the belief that, immediately after seroconversion, the underlying antibody response to HIV infection exhibits a period of growth until it reaches a maximum sometime later. Given a threshold $\alpha$, we are interested in estimating the mean window period, which is the average time the biomarker remains below $\alpha$ since seroconversion. The mean window period can be estimated using an inverse-regression approach as follows. For each individual $i$, set $\alpha$ equal to $g(T_i(\alpha), \phi_i)$, and solve for $T_i(\alpha)$, the time from seroconversion until the biomarker crosses $\alpha$. Then, the mean window period can be estimated by the average crossing time $T(\alpha) = \sum_{i=1}^{p} T_i(\alpha)/p$.

For the remainder of this paper, we will label the three curves used by Sommen et al. (2011) as $C_0$, $C_1$, and $C_2$, and the curve used by Sweeting et al. (2010)) as $C_3$, as specified below.

\begin{align*}
C_0 : & \quad g(t_{ij} + u_i, \phi_i) = \phi_{1i} \exp\left(-\phi_{2i} \exp\left(-\frac{(t_{ij} + u_i)}{\phi_{3i}}\right)\right), \\
C_1 : & \quad g(t_{ij} + u_i, \phi_i) = \frac{\phi_{1i}}{1 + \phi_{2i} \exp\left(-\frac{(t_{ij} + u_i)}{\phi_{3i}}\right)}, \\
C_2 : & \quad g(t_{ij} + u_i, \phi_i) = \frac{\phi_{1i} (t_{ij} + u_i)^{\phi_{3i}}}{\exp(\phi_{2i}) + (t_{ij} + u_i)^{\phi_{3i}}}, \\
C_3 : & \quad g(t_{ij} + u_i, \phi_i) = \phi_{1i} + (\phi_{2i} - \phi_{1i}) \exp\left(-(t_{ij} + u_i) \exp(\phi_{3i})\right),
\end{align*}

where $\phi_i = (\phi_{1i}, \phi_{2i}, \phi_{3i})$ for each curve.
2.2 Inverse-regression to estimate the mean window period

While inverse-regression is not a new statistical technique, Sweeting et al. (2010) were the first to use it to estimate the mean window period of HIV infections. In particular, the authors model the data assuming the underlying curve is of the form \( C_3 \); see (5). Here, \( \phi_{1i} \) is the asymptote, \( \phi_{2i} \) is the intercept, and \( \phi_{3i} \) is the log-rate for individual \( i \). To estimate the mean window period, Sweeting et al. (2010) specify a threshold \( \alpha \) and solve for the individual crossing times first, then take the average as discussed in Section 2.1. Thus, by setting \( \alpha = \phi_{1i} + (\phi_{2i} - \phi_{1i}) \exp(-T_i(\alpha) \exp(\phi_{3i})) \), we obtain

\[
T_i(\alpha) = \log \left( \frac{\phi_{2i} - \phi_{1i}}{\alpha - \phi_{1i}} \right) \exp(-\phi_{3i}). \tag{6}
\]

Note that when calculating \( T_i(\alpha) \), we must ensure that

\[
\frac{\phi_{2i} - \phi_{1i}}{\alpha - \phi_{1i}} > 0 \tag{7}
\]

in order for \( T_i(\alpha) \) to be defined. Then, the average crossing time \( T(\alpha) = \sum_{i=1}^{n} T_i(\alpha) / p \) is used as an estimate for the mean window period. In a Bayesian framework, a posterior distribution is obtained for each \( T_i(\alpha) \), and therefore, a posterior distribution is also obtained for \( T(\alpha) \).

2.3 Extending the inverse-regression to other curves

Here, we derive expressions for the mean window period associated with three parametric curves \( C_0, C_1, \) and \( C_2 \), which are used by (Sommen et al., 2011) to estimate infection times. For each of these curves, the parameters \( \phi_i = (\phi_{1i}, \phi_{2i}, \phi_{3i}) \) are specified as as random-effects to allow between individual variability, although one or more parameters might be modelled as a fixed-effect depending on the complexity of the data. For all three models, \( \phi_{1i} \) is the asymptote for individual \( i \). The individual times associated with the inflection points since seroconversion are given by \( \phi_{3i} \ln(\phi_{2i}) \) for \( C_0 \) and \( C_1 \), and by \( \left( \frac{\phi_{3i} - 1}{\phi_{3i} + 1} \right)^{1/\phi_{3i}} e^{\phi_{2i}/\phi_{3i}} \) for \( C_2 \). Setting \( \alpha = g(T_i(\alpha), \phi_i) \) for each curve, we obtain, the following expressions for the individual crossing times, respectively,
\begin{align*}
C_0 : & \quad T_i(\alpha) = -\phi_{3i} \log \left(-\frac{\log \left(\frac{\alpha}{\phi_{1i}}\right)}{\phi_{2i}}\right) \\
C_1 : & \quad T_i(\alpha) = \phi_{3i} \left(\log(\phi_{2i}) - \log \left(\frac{\phi_{1i}}{\alpha} - 1\right)\right) \\
C_2 : & \quad T_i(\alpha) = \left(\frac{\alpha \exp(\phi_{2i})}{\phi_{1i} - \alpha}\right)^{1/\phi_{3i}}
\end{align*}

(8)  

(9)  

(10)  

The posterior distribution of each \( T_i(\alpha) \) are obtained within a Bayesian framework. From those, the posterior distribution of the mean window period (average crossing time) \( T(\alpha) = \sum_{i=1}^{n} T_i(\alpha)/p \) can be constructed.

### 2.4 Prior specifications

To complete the Bayesian model, we must specify prior distributions for the model parameters \( \phi_i, u_i \), for \( i = 1, \ldots, p \), and \( \sigma^2 \). In this work, we treat the asymptote, \( \phi_{1i} \), as a fixed effect, while we treat the parameters \( \phi_{2i} \) and \( \phi_{3i} \) as random effects and specify their prior hierarchically. Let \( \tau = 1/\sigma^2 \) be the observation precision. For the results shown later, the following priors were used.

\( \tau \sim \text{Gamma}(2, 0.01) \),  

\( \phi_{1i} = \beta_1 \sim \text{N}(0, \text{var} = 1000, 000) \),  

\( \begin{pmatrix} \phi_{2i} \\ \phi_{3i} \end{pmatrix} \sim \text{iid N}_2 \left( \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}, \Sigma \right) \),  

\( \beta_1 \sim \text{N}(0, \text{var} = 1000, 000) \),  

\( \beta_2 \sim \text{N}(0, \text{var} = 1000, 000) \),  

\( \Sigma^{-1} \sim \text{Wishart}_3(R) \),

where, \( E(\Sigma^{-1}) = \nu R^{-1} = 3 R^{-1} \), and \( E(\tau) = 200 \). The specification of the covariance matrix \( \Sigma(\phi_{2i}, \phi_{3i}) \) depends on the model.
3 Results

3.1 Simulated data example

We work with a simulated dataset that includes longitudinal measurements from \( p = 50 \) individuals. Time measurements since the first positive HIV test are assumed to be equally spaced every 3 months from 0 to 2 years. For each individual, a seroconversion interval is drawn from a truncated (positive support) normal distribution with mean equal to 3 months and standard deviation equal to 1 month. Then, the seroconversion time for each individual is drawn from a uniform distribution on that individual’s seroconversion interval (see Figure 1).

Figure 1: The simulated seroconversion intervals stacked vertically for all 50 individuals, with zero being the date of the last negative test. The blue dots indicate the actual seroconversion times, each uniformly sampled within its corresponding interval.

As indicated in Section 2.1, \( u_i > 0 \) is the time from seroconversion to the first positive date for individual \( i, i = 1, \ldots, p \), and \( t_{ij} \) is the time of the \( j \)-the measurement from the first positive test date, with \( t_{i1} = 0 \). To generate the data, we assume the underlying biomarker follows the curve \( C_0 \); see (2). First, we generate \( \phi_i \) values for each of the 50 individuals as follows. To start, all individuals are given the same value for the asymptote, \( \phi_{1i} = 1 \),
\( i = 1, \ldots, p \), which means the asymptote is a \textit{fixed} effect in this simulation. Then, \( \phi_{2i} \) and \( \phi_{3i} \) are drawn from a bivariate normal distribution with mean and covariance matrices given by

\[
\mu(\phi_{2i}, \phi_{3i}) = \begin{pmatrix} 20 \\ 0.2 \end{pmatrix} \quad \text{and} \quad \Sigma(\phi_{2i}, \phi_{3i}) = \begin{pmatrix} 25 & 0.35 \\ 0.35 & 0.0053 \end{pmatrix}.
\] (17)

Thus, \( \phi_{2i} \) and \( \phi_{3i} \) are \textit{random} effects. Their values are plotted in Figure 2. Then, we generate a mean curve for each individual according to (2). Working with a threshold \( \alpha = 0.6 \), we calculate the mean window period (true average crossing time) to be approximately 0.7454 years, or 272.25 days.

Finally, the observed biomarker data are taken to be noisy measurements of the mean curve values, specifically

\[
y_{ij} = g(t_{ij} + u_i, \phi_i) + \epsilon_{ij}, \quad \text{where} \quad \epsilon_{ij} \sim N(\text{mean} = 0, \text{sd} = 0.07).
\] (18)

Adding noise results in obtaining some negative early measurements. In that case, we set the measurement equal to zero, i.e., we work with \( y_{ij} = \max(0, g(t_{ij} + u_i, \phi_i) + \epsilon_{ij}) \). The resulting biomarker dataset is plotted in Figure 3.

In the following subsections, we fit the models \( C_0-C_3 \), specified in (2-5), to this simulated dataset in a Bayesian framework to demonstrate implementation of the inverse-regression approach. Obviously, three out of the four models are misspecified for this data set. However, the three functions used by Sommen et al. (2011) are related, and we expect them to provide a reasonable fit given data generated form any one of them.
Figure 2: The simulated values of \((\phi_{2i}, \phi_{3i}), i = 1, \ldots, 50\).

Figure 3: The simulated time series biomarker data for 50 individuals.
3.2 Results under $C_0$, the model from which the data are generated

We fit the data in Figure 3 using $C_0$, the model from which the data are generated; see (2). Here, the purpose is to illustrate that accurately specifying the shape of the underlying curve provides accurate inference for the mean window period. In addition to specifying the model for the data, we must specify prior distributions for the model parameters. In particular, the inverse of the covariance matrix $\Sigma(\phi_2, \phi_3)^{-1}$ is given the prior distribution $\text{Wishart}_3(R)$, where

$$R = \begin{pmatrix} 25 & 0.35 \\ 0.35 & 0.0053 \end{pmatrix}. \quad (19)$$

The rest of the prior distribution specifications are given in Section 2.4. We fit the model in WinBUGS and obtain samples from the posterior distributions of all model parameters, including the mean window period, which is constructed as the average of (8), $i = 1, \ldots, p$. Specifically, for each model parameter, two MCMC chains were run in WinBUGS for 60,000 iteration each. For each chain, the first 10,000 iterations were discarded as burn-in. Then, thinning was applied every 25 iterations, which resulted in an posterior sample of 2,000 atoms from each chain (total of 4,000 atoms).

![Histograms of the posterior distributions of the parameters $\beta_1$, $\beta_2$, and $\beta_3$ assuming $g(t_{ij} + u_i, \phi_i) = C_0$; see (2). The true values are in red.](image)

Figure 4: Histograms of the posterior distributions of the parameters $\beta_1$, $\beta_2$, and $\beta_3$ assuming $g(t_{ij} + u_i, \phi_i) = C_0$; see (2). The true values are in red.

With regards to MCMC convergence diagnostics, the chains for $\beta_2$ and $\beta_3$ took longer to converge than the chains of the other model parameters. Additionally, the chain for $\beta_2$ was slow to mix. The rest of the MCMC chains had more favorable convergence and mixing.
properties. Figure 4 shows the posterior distributions of $\beta_1$, $\beta_2$, and $\beta_3$. Here, the posteriors are peaked around the true values indicating a good fit.

Figure 5 shows the 95% posterior region for both the biomarker curve $g(t_{ij} + u_i, \phi_i)$, given by (2), and the observed measurements, given by (1), for 6 different individuals. We chose to show inference for these 6 individuals simply to provide a graphical illustration of the model fit for different shapes of biomarker trajectories. In particular, we are able to apportion uncertainty between $g(t_{ij} + u_i, \phi_i)$ and the observation error. Invariably, uncertainty due to $g(t_{ij} + u_i, \phi_i)$ dominates in the region where the biomarker increases rapidly. Once a plateau is reached, the main source of uncertainty is the observation error.

Figure 6 includes histograms of the posterior distributions of $u_i$, the time from seroconversion until the first positive measurement, for the same six individuals. Here, prior to posterior learning is not consistently strong. This is because $u_i$, $i = 1, \ldots, p$, is in a region of the model, where observed data are not available.

Figure 7 shows posterior inference for the observation precision, $\tau$. Here, strong prior to posterior learning is obtained, with the posterior distribution estimating the true precision reasonably well. Prior sensitivity analysis for $\tau$ produced virtually the same posterior results.

Finally, posterior inference is obtained for the individual window periods according to (8), and the posterior distribution of their average (the mean window period) is obtained over all $p = 50$ individuals. Figure 8 shows the posterior distribution of the mean window period, which has a mean of 0.7602 years and a 95% probability interval (0.7310, 0.7901). The true mean window period of 0.7454 years is well within this interval, indicating a reasonable fit by the model.
Figure 5: 95% posteriors predictive intervals (dashed blue) and 95% posterior intervals of the curve \( g(t_{ij} + u_i, \phi_i) \) (dashed red) for six individuals, assuming \( g(t_{ij} + u_i, \phi_i) = C_0 \); see (2). The true biomarker curve for each individual is in Green.
Figure 6: Histograms of the posterior distributions of $u_i$ for six individuals assuming $g(t_{ij} + u_i, \phi_i) = C_0$; see (2).
Figure 7: Histogram of the posterior distribution of the observation precision, $\tau$ using Gamma(2, 0.01) prior distribution, and assuming $g(t_{ij} + u_i, \phi_i) = C_0$; see (2).

Figure 8: Histogram of the posterior distribution of the mean window period given $\alpha = 0.6$, and assuming $g(t_{ij} + u_i, \phi_i) = C_1$; see (2).
\subsection*{3.3 Results under $C_1$}

We fit the data in Figure 3 assuming the underlying curve is of the form $C_1$; see (3). The prior distributions for the model parameters are specified as discussed in Section 2.4. Here, the inverse of the covariance matrix $\Sigma(\phi_{2i}, \phi_{3i})^{-1}$ is given a Wishart$_3(R)$ prior distribution, where

$$R = \begin{pmatrix} 400 & 0 \\ 0 & 0.0025 \end{pmatrix}. \quad (20)$$

Under $C_1$, the parameter $\beta_1$ is the population asymptote, which has the same interpretation as $\beta_1$ under $C_0$, the curve from which the data were generated. Therefore, the posterior distribution for $\beta_1$ under $C_1$ may be compared to the true value of the population asymptote. However, the other population parameters of the curve, $\beta_2$ and $\beta_3$, have a different interpretation and, therefore, are not comparable to their counterparts under $C_0$.

![Figure 9: Histograms of the posterior distributions of the parameters $\beta_1$, $\beta_2$, and $\beta_3$, assuming $g(t_{ij} + u_i, \phi_i) = C_1$; see (3). The true value of $\beta_1$ is in red.](image)

With regard to MCMC convergence properties, the chains for $\beta_2$ and $\beta_3$ took longer to converge than the chains for the other model parameters, with the chain for $\beta_2$ having very slow mixing. Figure 9 shows the posterior distributions of $\beta_1$, $\beta_2$, and $\beta_3$. Although the posterior distribution of $\beta_1$ appears to underestimate the true value of the asymptote, it has an effective range of values that are very close to one, with an average of approximately 0.9930. Results reported for $\beta$ as well as other parameters are based on posterior samples of 4,000 atoms from two different MCMC chains after burn-in and thinning.
Figure 10 shows the 95% posterior region for both the biomarker curve \( g(t_{ij} + u_i, \phi_i) \), given by (3), and the observed measurements, given by (1), for 6 different individuals (the same individuals as in Section 3.2). These plots again show that uncertainty due to \( g(t_{ij} + u_i, \phi_i) \) dominates in the region where the biomarker increases rapidly. However, once a plateau is reached, the main source of uncertainty is the observation error.

Figure 6 includes histograms of the posterior distributions for \( u_i \), the time from seroconversion until the first positive measurement, for the same six individuals. Just as in Section 3.2, prior to posterior learning is not consistently strong. This is because \( u_i, i = 1, \ldots, p \), is in a region of the model, where observed data are not available.

Posterior inference for the observation precision, \( \tau \) is summarized in Figure 12. Here, strong prior to posterior learning is obtained, with the posterior distribution estimating the true precision reasonably well. Prior sensitivity analysis for \( \tau \) produced virtually the same posterior results.

Finally, posterior inference is obtained for the individual window periods according to (9), and the posterior distribution of their average (the mean window period) is obtained over all \( p = 50 \) individuals. Figure 13 shows the posterior distribution of the mean window period, which has a mean of 0.7727 years and a 95% probability interval (0.7447, 0.8009). The true mean window period of 0.7454 years is within this interval, although it is very close to its left bound indicating an overestimate of the mean window period. This is not surprising since the biomarker data are generated using a different parametric curve than the one we use to model the data.
Figure 10: 95% posteriors predictive intervals (dashed blue) and 95% posterior intervals of the curve $g(t_{ij} + u_i, \phi_i)$ (dashed red) for six individuals assuming $g(t_{ij} + u_i, \phi_i) = C_1$; see (3). The true biomarker curve for each individual is in Green.
Figure 11: Histograms of the posterior distributions of $u_i$ for six individuals assuming $g(t_{ij} + u_i, \phi_j) = C_1$; see (3).
Figure 12: Histogram of the posterior distribution of the observation precision, \( \tau \), using a Gamma(2, 0.01) prior distribution, and assuming \( g(t_{ij} + u_i, \phi_i) = C_1 \); see (3).

Figure 13: Histogram of the posterior distribution of the mean window period given \( \alpha = 0.6 \), and assuming \( g(t_{ij} + u_i, \phi_i) = C_1 \); see (3).
3.4 Results under \( C_2 \)

We fit the data in Figure 3 assuming the underlying curve is of the form \( C_2 \); see (4). The prior distributions for the model parameters are specified as discussed in Section 2.4. The inverse of the covariance matrix \( \Sigma(\phi_{2i}, \phi_{3i})^{-1} \) is given a Wishart\(_3\) \( (R) \) prior distribution, where

\[
R = \begin{pmatrix}
3.0725 & 0.1750 \\
0.1750 & 4.0625 \\
\end{pmatrix}.
\]  

(21)

Under \( C_2 \), the parameter \( \beta_1 \) is the population asymptote, which is the same the interpretation as \( \beta_1 \) under \( C_0 \), the curve from which the data were generated. Therefore, the posterior distribution for \( \beta_1 \) under \( C_2 \) may be compared to the true value of the population asymptote. However, the other population parameters of the curve, \( \beta_2 \) and \( \beta_3 \), have a different interpretation and, therefore, are not comparable to their counterparts under \( C_0 \).

![Figure 14: Histograms of the posterior distributions of the parameters \( \beta_1, \beta_2, \) and \( \beta_3 \) assuming \( g(t_{ij} + u, \phi_i) = C_2 \); see (4).](image)

With regard to MCMC convergence properties, the chains for \( \beta_2 \) and \( \beta_3 \) took longer to converge than the chains for the other model parameters. Figure 14 shows the posterior distributions of \( \beta_1, \beta_2, \) and \( \beta_3 \). Although the posterior distribution of \( \beta_1 \) appears to underestimate the true value of the asymptote, it has an effective range of values that are very close to one, with an average of approximately 1.072. Just as before, results reported for \( \beta \) as well as other parameters are based on posterior samples of 4,000 atoms from two different MCMC chains after burn-in and thinning.
Figure 15 shows the 95% posterior region for both the biomarker curve \( g(t_{ij} + u_i, \phi_i) \), given by (4), and the observed measurements, given by (1), for 6 different individuals (the same individuals as in Sections 3.2 and 3.3). Similar to results obtained in Sections 3.2 and 3.3, these plots show that uncertainty due to \( g(t_{ij} + u_i, \phi_i) \) dominates in the region where the biomarker increases rapidly. However, once a plateau is reached, the main source of uncertainty is the observation error.

Figure 16 includes the posterior distributions of \( u_i \), the time from seroconversion until the first positive measurement, for the same six individuals. Again, similar to results obtained in Sections 3.2 and 3.3, we find that prior to posterior learning is not consistently strong. This is because \( u_i, i = 1, \ldots, p \), is in a region of the model, where observed data are not available.

Posterior inference for the observation precision, \( \tau \) is summarized in Figure 17. Here, strong prior to posterior learning is obtained, with the posterior distribution estimating the true precision reasonably well. Prior sensitivity analysis for \( \tau \) produced virtually the same posterior results.

Finally, posterior inference is obtained for the individual window periods according to (10), and the posterior distribution of their average (the mean window period) is obtained over all \( p = 50 \) individuals. Figure 18 shows the posterior distribution of the mean window period, which has a mean of 0.7576 years and a 95% probability interval (0.7281, 0.7879). The true mean window period of 0.7454 years is well within this interval, indicating a reasonable fit by the model.
Figure 15: 95% posteriors predictive intervals (dashed blue) and 95% posterior intervals of the curve $g(t_{ij} + u_i, \phi_i)$ (dashed red) for six individuals assuming $g(t_{ij} + u_i, \phi_i) = C_2$; see (4).
Figure 16: Histograms of the posterior distributions of $u_i$ for six individuals assuming $g(t_{ij} + u_i, \phi_i) = C_2$; see (4).
Figure 17: Histogram of the posterior distribution of the observation precision using a Gamma(2, 0.01) prior distribution, and assuming $g(t_{ij} + u_i, \phi_i) = C_2$; see (4).

Figure 18: Histogram of the posterior distribution of the mean window period given $\alpha = 0.6$, and assuming $g(t_{ij} + u_i, \phi_i) = C_2$; see (4).
3.5 Results under $C_3$

Again, we continue to work with the same dataset in Figure 3, but we now we model the data assuming the underlying curve is of the form $C_3$. The prior distributions for the model parameters are specified as discussed in Section 2.4. The inverse of the covariance matrix $\Sigma(\phi_{2i}, \phi_{3i})^{-1}$ is given a Wishart$_3(R)$ prior distribution, where

$$R = \begin{pmatrix} 0.01 & 0.01 \\ 0.01 & 0.5625 \end{pmatrix}. \quad (22)$$

Under $C_3$, $\beta_1$ is the population asymptote, $\beta_2$ is the population intercept, and $\beta_3$ is the population log-rate. Therefore, the posterior distribution for $\beta_1$ under $C_3$ may be compared to the true value of the population asymptote. However, the other population parameters of the curve, $\beta_2$ and $\beta_3$, have a different interpretation and, therefore, may not be comparable to their counterparts under $C_0$.

![Figure 19: Histograms of the posterior distributions of the parameters $\beta_1$, $\beta_2$, and $\beta_3$ assuming $g(t_{ij} + u_i, \phi_i) = C_3$; see (5).](image)

Here, we report posterior results based on posterior samples of 4,000 atoms from two different MCMC chains after burn-in and thinning for each parameter. While $C_3$ is an extremely misspecified model for the data, we observe favorable MCMC convergence and mixing properties under this model than under the other three. Figure 19 shows the posterior distributions of $\beta_1$, $\beta_2$, and $\beta_3$. The posterior distribution of $\beta_1$ is a clear overestimate of the true value of the asymptote. In fact, even with a prior distribution more concentrated around the true value of $\beta_1$, we continue to obtain a posterior distribution that overestimates...
the true value. Only when we use an extremely informative prior concentrated around one, do we obtain an unbiased posterior distribution for this parameter.

Figure 20 shows the 95% posterior region for both the biomarker curve \( g(t_{ij} + u_i, \phi_i) \), given by (5), and the observed measurements, given by (1), for 6 different individuals (the same individuals as in Sections 3.2-3.4). Here, posterior inference is very different that that obtained in Sections 3.2–3.4. In particular, these plots show that uncertainty due to the observation error is a major contributor to the overall fit uncertainty regardless of whether a plateau is reached. In fact, the observation error uncertainty bands are much larger under \( C_3 \) than the other three models. In fact, \( C_3 \) does not change concavity, making it a less flexible curve. Thus, the model has to rely on the observation error terms to accommodate the data, which requires a larger error variance.

Figure 21 includes the posterior distributions of \( u_i \), the time from seroconversion until the first positive measurement, for the same six individuals. There are slight differences between results obtained here and results under the other parametric models, but as before, we find that prior to posterior learning is not consistently strong for these parameters. This is because \( u_i, i = 1, \ldots, p, \) is in a region of the model, where observed data are not available.

Posterior inference for the observation precision, \( \tau \) is summarized in Figure 22. Here, strong prior to posterior learning is obtained, with the posterior distribution clearly underestimating the true precision. This is a result of the model trying to accommodate the shape of the data better. Prior sensitivity analysis for \( \tau \) produced virtually the same posterior results.

Finally, posterior inference is obtained for the individual window periods according to (6), and the posterior distribution of their average (the mean window period) is obtained over all \( p = 50 \) individuals. Figure 23 shows the posterior distribution of the mean window period, which has a mean of 0.7435 years and a 95% probability interval (0.7026, 0.7846). The true mean window period of 0.7454 years is well within this interval. However, the 95% uncertainty interval is much greater than those obtained under the other models. This inflation of uncertainty in the posterior distribution of the mean window period is a direct result of the inflated uncertainty intervals associated with the posterior mean curves (Figure 20) in the region where \( \alpha = 0.6 \).
Figure 20: 95% posteriors predictive intervals (dashed blue) and 95% posterior intervals of the curve \( g(t_{ij} + u_i, \phi_i) \) (dashed red) for six individuals, assuming \( g(t_{ij} + u_i, \phi_i) = C_3 \); see (5).
Figure 21: Histograms of the posterior distributions of $u_i$ for six individuals, assuming $g(t_{ij} + u_i, \phi_i) = C_3$; see (4).
Figure 22: Histogram of the posterior distribution of the observation precision using a Gamma(2, 0.01) prior distribution, assuming $g(t_{ij} + u_i, \phi_i) = C_3$; see (4).

Figure 23: Histogram of the posterior distribution of the mean window period given $\alpha = 0.6$, and assuming $g(t_{ij} + u_i, \phi_i) = C_3$; see (5).
4 Discussion

We have implemented a Bayesian parametric inverse-regression method to estimate the ‘mean window period’ of HIV infections. This method is analogous to an approach implemented by Sweeting et al. (2010). Here, we extend that approach to three other parametric models and derive analytic expressions for the mean window period for each one. To investigate the fit under each model, we work with a simulated dataset of biomarker measurements generated from one of the parametric models. Then, we fit all four parametric models to the simulated dataset. We find that model specification has a significant impact on the posterior distribution of the mean window period. In particular, when the fitted model is not as flexible (or an under-parametrization of) the true model, the resulting posterior distribution of the mean window period has an inflated variance compared with the posterior distribution obtained using models with the same level of flexibility as the true model.

Our analysis allows for a straightforward way of apportioning uncertainty between the biomarker curve $g(t_{ij} + u_i, \phi_i)$ and the observation error. In particular, we find that the observation error is the main source of uncertainty once a plateau is reached. The results observed in this report are dependent on the threshold. Results under other thresholds are needed to generalize findings. Also, a large simulation study will further shed light on the asymptotic behavior of the posterior distribution of the mean window period.

References


