MSc Mathematics

Master's thesis

Estimation of perinatal transmission rates and calculation of the basic reproduction number for cytomegalovirus

by

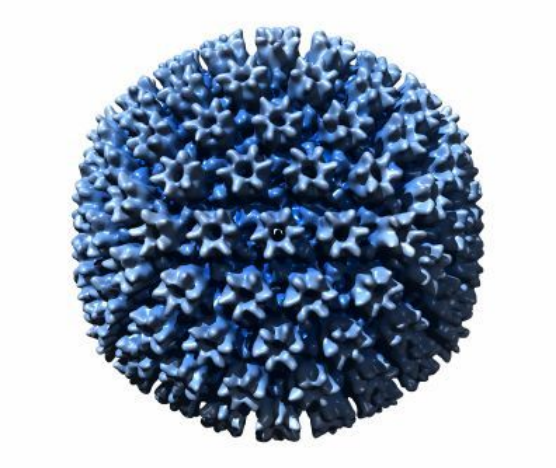
Sophia de Jong

December 20, 2017

Supervisor: Dr. M. van Boven

First examiner: Prof. dr. M.C.M. de Gunst

Second examiner: Dr. E. Belitser



Department of Mathematics
Faculty of Sciences

rivm VU VRIJE UNIVERSITEIT AMSTERDAM

Abstract

Cytomegalovirus is a herpesvirus with poorly understood transmission dynamics. Infection is often asymptomatic, but can cause serious disease in fetuses and immunocompromised individuals. A sex- and age-structured transmission model for cytomegalovirus is studied and extended with estimation of perinatal transmission and calculation of the basic reproduction number. Transmission models are fitted to serological data to find out what transmission routes are responsible for the high prevalence of the virus. Parameter samples are generated from posterior distributions using the Bayesian inference program Stan, which is based on the Hamiltonian Monte Carlo method. Results indicate that reactivation and perinatal transmission are essential for cytomegalovirus to be able to persist in the population. The study suggests that vaccination of women of fertile age with a vaccine that prevents reactivation may be most effective.

Title: Estimation of perinatal transmission rates and calculation of the basic reproduction number for cytomegalovirus
Author: Sophia de Jong, s8.de.jong@student.vu.nl, 2534437
Supervisor: Dr. M. van Boven
First examiner: Prof. dr. M.C.M. de Gunst
Second examiner: Dr. E. Belitser
Date: December 20, 2017

Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam
http://www.math.vu.nl/

Preface

From January to July, I have worked on my master's thesis at the National Institute for Public Health and the Environment (Rijksinstituut voor Volksgezondheid en Milieu, RIVM). The goal of my internship at the RIVM was to study how the dynamics of cytomegalovirus can be translated into mathematics, and to improve and extend the original transmission model, using the Bayesian inference program Stan, to get a better understanding of the dynamics of the virus. This thesis describes the development of the model, and explains the theory behind age-structured transmission models and the Hamiltonian Monte Carlo method that Stan uses to generate samples from posterior distributions. I have mainly focused on the estimation of perinatal transmission rates and the calculation of the basic reproduction number for cytomegalovirus.

I would like to thank everyone who has, in one way or another, supported me with my thesis. My thanks go out to lecturers of the Vrije Universiteit Amsterdam, my fellow students, friends and family. In particular, I would like to thank my supervisor Michiel van Boven for providing me with the opportunity to write my thesis at the RIVM, and guiding me through the project. He has been of great help and support. I have enjoyed our collaboration and I am happy to have been part of this study.

Special thanks go out to my first examiner Mathisca de Gunst for her advice and feedback. Our conversations helped me to structure my thoughts, and gave me confidence.

Additional thanks go out to Chris van Dorp, for his help with Stan and his feedback on my work, and prof. dr. Odo Diekmann, for the insightful conversation that we have had.

Summary

Human cytomegalovirus (CMV) is a common herpesvirus that infects people of all ages. The virus causes lifelong infection and has a high prevalence that increases with age. A CMV infection is often asymptomatic, but can cause serious disease in fetuses and immunocompromised individuals. Unfortunately, the transmission dynamics of the virus are poorly understood. A mathematical model can contribute to the understanding of the CMV transmission dynamics.

After primary infection, CMV is able to persist in an inactive, latent state. From this state, the virus can reactivate. It has been shown that reactivation may occur during pregnancy, which can result in transmission to the fetus. Infection of the child during the during pregnancy and the first six months of life, is called perinatal infection. Individuals can also become reinfected with a different CMV strain. One of the questions is which of these transmission routes are required to explain the high CMV seroprevalence in the Netherlands.

To gain insight in the transmission dynamics of CMV, a sex- and age-structured transmission model has been developed. The aim of this thesis is to improve this transmission model, by extending the model with estimation of perinatal transmission rates, and calculating the basic reproduction number. The basic reproduction number is of interest, since it gives insight in the ability of the virus to persist in the population.

The other main subject of this thesis is the Hamiltonian Monte Carlo method. This is a relatively new but upcoming method to generate parameter samples from a probability distribution. The Bayesian inference program Stan is based on the Hamiltonian Monte Carlo method, and is therefore able to sample more efficiently and faster than programs that are based on Markov Chain Monte Carlo methods.

To find out which transmission routes are required to explain the high CMV prevalence in the Netherlands, transmission models that include and exclude reinfection and reactivation are fitted to data using Stan. The results indicate that reactivation and perinatal transmission are essential for CMV to be able to persist in the population. It is also found that the CMV prevalence in females is higher than the prevalence in males. Females experience more reactivation than males, and for most ages, females have a higher force of infection than males. The estimated probability of perinatal transmission equals 0.37, and it is found that 17% of the children are already infected with CMV at the age of six months.

This study contributes to the understanding of the transmission dynamics of CMV. Future research can improve the model, such that the effects of vaccination on the CMV prevalence in the population can be studied.

Contents

1. Introduction

Human cytomegalovirus (CMV) is a common herpesvirus that infects people of all ages. The virus causes lifelong infection and has a high prevalence that increases with age. A CMV infection is often asymptomatic, but several studies have shown that CMV infection can cause serious disease in fetuses [8] and immunocompromised individuals [3]. CMV is also thought to play a role in exhaustion of the immune system, which may be a cause of the reduced effectiveness of influenza vaccination in elderly persons [9].

In recent years, researchers from various fields have become increasingly interested in the physiology of CMV. The intention is to develop a vaccine against the virus and to set up a vaccination program. The development of a vaccine against CMV was given high priority by the US Institute of Medicine in 2000, but no licensed vaccine has been developed yet. Development of a vaccine requires knowledge of the transmission dynamics of the virus, which are poorly understood. One of the questions is which population subgroup should be the target of vaccination, such that CMV is most efficiently combated. A mathematical model can provide insight in the transmission dynamics of CMV.

Person-to-person transmission of CMV is thought to mainly occur through transfer of saliva and urine. After primary infection, the immune system operates to bring the infection under control, but the virus is not eliminated. CMV is able to persist in an inactive, latent state. Individuals with a latent CMV infection are not infectious. However, latently infected individuals can experience reactivation of the virus, or can become reinfected with a different CMV strain. In general, it is hard to distinguish reinfection from reactivation [7]. Both can occur multiple times during life.

Various studies have shown that a latent CMV infection may reactivate during pregnancy. This can result in transmission through the placenta to the fetus [4]. Infection of the child during the perinatal period, which is defined here as the time during pregnancy and the first six months of the child's life, is called perinatal infection. Along with reactivation and reinfection, perinatal transmission is thought to play a significant role in the ability of CMV to persist in the population.

To gain insight in the transmission dynamics of CMV, researchers at the RIVM, including my supervisor Michiel van Boven, have developed a sex- and age-structured model for CMV. They have analysed CMV transmission patterns by fitting the model to cross-sectional seroprevalence data of the Dutch population. With differential equations, the transmission model specifies primary infection, reinfection and reactivation events. By fitting variants of this model that include and exclude reinfection and reactivation to data, they attempted to

find out which transmission routes are required to explain the high CMV seroprevalence in the Netherlands. This study is described in the paper *Infectious reactivation of cytomegalovirus explaining age- and sex-specific patterns of seroprevalence* (Van Boven et al., 2017).

The analyses in Van Boven et al. (2017) show that infectious reactivation of CMV is required to explain the seroprevalence data. Including infectious reinfection in the transmission model did not result in a better fit to the data. However, the model with reactivation could not explain the ability of CMV to persist in the Dutch population. Since perinatal transmission was not taken into account, it is suggested that the persistence of CMV in human populations depends on a combination of infectious reactivation and perinatal transmission.

The aim of this thesis is to study how the transmission dynamics of CMV can be translated into mathematics, and to improve the original transmission model for CMV that was constructed in Van Boven et al. (2017). This model and general age-structured transmission models are studied.

In this thesis, the original transmission model is extended with estimation of perinatal transmission rates. In the original model, the fraction of susceptibles at the age of six months is estimated separately from the other parameters, and then given as input to the transmission model. With the introduction of a perinatal transmission parameter, this fraction can be estimated within the transmission model.

In addition, the basic reproduction number for cytomegalovirus is calculated in this thesis. The basic reproduction number is of interest, since it gives insight in the ability of the virus to persist in the population, and can ultimately be used to compute the control effort that is required for elimination of the virus from the population.

The other main subject of this thesis is the Hamiltonian Monte Carlo method [2]. This is a relatively new but upcoming method to generate parameter samples from a probability distribution. The Bayesian inference program Stan is based on the Hamiltonian Monte Carlo method, and is therefore able to sample more efficiently and faster than programs that are based on Markov Chain Monte Carlo methods [11]. Markov Chain Monte Carlo, Hamiltonian Monte Carlo and Stan are extensively described in this thesis.

In Van Boven et al. (2017) parameters were initially estimated using JAGS, a Bayesian inference program that uses Markov Chain Monte Carlo methods to sample from posterior distributions [22]. Since the CMV transmission model is complex and high-dimensional, it took JAGS several hours to generate parameter samples. Therefore, the model was later programmed in Stan by Chris van Dorp, who worked on the CMV transmission model as a PhD student. Stan is able to generate 2,000 parameter samples from the model within thirty minutes.

In this thesis, the Stan code for the CMV transmission model is extended and improved. The CMV transmission model is fitted to the seroprevalence data to estimate the rates of

primary infection, reinfection, reactivation and perinatal transmission, and infectiousness of primary infected individuals and individuals that have experienced reinfection or reactivation. Fits to data of variants of the transmission model that include and exclude reinfection and reactivation are compared, in order to get a better understanding of the transmission dynamics of CMV.

The results concerning the comparison of the variants of the extended transmission model are similar to the ones that were found using the original transmission model. Reactivation is required to explain the data, in contrast to reinfection, which did not improve the fit of the transmission model to the data. Comparable to the results of the original transmission model, the CMV prevalence in females is higher than the prevalence in males. Females experience more reactivation than males, and for most ages, females have a higher force of infection than males.

In addition, the estimated probability of perinatal transmission equals 0.37, and the estimated fraction of susceptibles at the age of six months is equal to 0.83. This means that 17% of the children are already infected with CMV at the age of six months.

It is found that the basic reproduction number is smaller than 1. This was not expected. However, the calculations did not take perinatal transmission into account. Additional research is required to include perinatal transmission in the calculation of the basic reproduction number. This is one of the improvements that can be made in the future.

The sex- and age-structured transmission model that is described in this thesis contributes to the understanding of the transmission dynamics of CMV. Once the model has been further improved, and a reliable basic reproduction number has been generated, the effects of vaccination on the CMV prevalence in the population can be studied.

2. Cytomegalovirus

Human cytomegalovirus (CMV) is a common herpesvirus with poorly understood transmission dynamics. The virus causes lifelong infection and has a high prevalence that increases with age. Since the virus is often asymptomatic, individuals are not aware of the presence of the virus. Nonetheless, CMV is able to cause serious disease in fetuses and immunocompromised individuals. This chapter discusses the ability of CMV to infect humans and to persist in its hosts, leading to a high CMV prevalence in the population of the Netherlands.

2.1. CMV infection and persistence

Viral particles, named virions, consist of genetic material enclosed by a protein shell. A virus needs host cells to survive, for it does not have an own metabolism. At primary infection, a virion enters a cell and uses it to replicate its viral genome. This enables the virus to take over the host cell's mechanisms, which often leads to disease. The newly generated virions infect other cells and start to replicate. The immune system actively searches with T-cells, a type of white blood cells, for foreign particles, like virions and bacteria, that have entered the body. When a virus is detected, the immune system starts to produce antibodies that are specific for that virus. These antibodies attach to the virions present in the body fluids, thereby disabling them. In addition, the immune system actively searches and destroys infected cells. Nevertheless, viral recognition by the immune system is complicated, since viruses continuously evolve. Some viruses can even block parts of the immune system or infect its cells. However, there are also viruses that are able to harmoniously coexist with their host. The virus is then present in a latent state. Latency is defined as an infection in which the viral genome is present in host cells, but infectious virions are not produced, since not all viral genes are being expressed [7].

CMV is one of eight herpesviruses that commonly infect humans. CMV is a DNA virus containing approximately two hundred genes. Most of these genes encode for proteins [3]. The genome is protected inside virions by a protein coat. Figure 2.1 shows a reconstruction of the surface of a CMV virion.

After primary infection with CMV, the immune system of the host operates to bring the infection under control. However, CMV has evolutionarily developed various tactics to avoid elimination. Its viral genome contains genes that encode for products that counteract the host's immune system. One tactic is to disrupt the antigen presentation pathway, thereby concealing the infected cells from detection. Another tactic is to mimic gene sequences that are used by the host, thereby depressing the host's immune and inflammatory reaction [3].

As a result, the immune system fails to detect and destroy infected cells, and CMV is able to persist in the host in a latent state. In this state, the virus does not replicate and is hence inactive.

However, latent CMV can reactivate. This is actively prevented by the immune system, but when the immune system is weakened, CMV is able to reactivate. For instance, CMV reactivation has been observed during pregnancy, blood transfusion and organ transplantation. The number of latently infected cells and the number of copies of viral genomes per cell seem to influence the risk of reactivation [3].

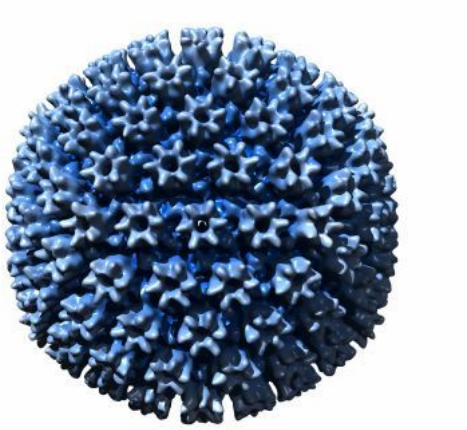


Figure 2.1.: Reconstruction of the surface of a human cytomegalovirus virion.

There exist thousands of genetic variants of CMV. This makes it possible for an already infected individual to become reinfected with a different CMV strain. In general, it is hard to distinguish reinfection from reactivation [7]. Both can occur multiple times during the host's life.

Aquisition of CMV may occur at any time during life. Person-to-person transmission, also called horizontal transmission, of CMV can happen through contact with saliva or urine, sexual contact, breastfeeding, blood transfusion or organ transplantation. The presence of active CMV during pregnancy can result in transmission through the placenta to the fetus [4]. This is called vertical transmission. After birth, the mother may infect the child through breastfeeding. Infection of the child during the perinatal period, which is defined here as the time during pregnancy and the first six months of the child's life, is called perinatal infection. An individual is potentially infectious in the period after primary infection or reinfection with CMV, and after reactivation of CMV. Then the virus is actively replicating, and hence virions are present in body fluids. This is summarized in Figure 2.2.

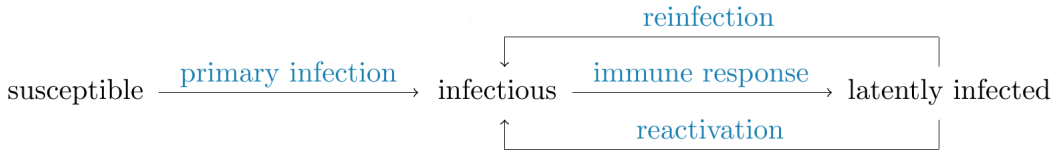


Figure 2.2.: CMV causes lifelong infection. Individuals are potentially infectious after primary infection, reinfection or reactivation. Due to activity of the immune system, the virus is controlled and remains in a latent state.

2.2. CMV prevalence

The presence of a CMV infection can be measured by the concentration of anti-CMV immunoglobulin G (IgG) antibodies in the serum of an individual. The presence of these antibodies in serum is called CMV seropositivity, and is considered to be the best method to measure CMV infection. CMV seroprevalence is the fraction of individuals in a population who are tested CMV seropositive. Worldwide studies have indicated that the CMV seroprevalence increases with age and depends on sex, ethnicity and socioeconomic status. Women, non-whites and persons of lower socioeconomic status are more likely to be CMV seropositive. This may be explained by differences in customs related to CMV transmission, like sexual activity, contact with children and crowdedness of the environment [4].

In the cross-sectional population-based study PIENTER2 [19], carried out by researchers of the RIVM, sera and questionnaire data were collected from 6,386 individuals in the Netherlands, in the period 2006 - 2007. The sera have been tested, among other things, for the presence of anti-CMV IgG antibodies. In the general population of the Netherlands, the CMV seroprevalence was found to be 45.6%. The seroprevalence in individuals of native Dutch origin was 40.1%, which is lower than the seroprevalences of 57.3% and 76.7% found respectively in Western and non-Western migrants. In Dutch and Western individuals, the seroprevalence increased from 21.8% at age six months to nine years, to 64.6% at age seventy to seventy-nine years [15]. Figure 2.3 provides a visualisation of the CMV seroprevalence data.

CMV infection usually is asymptomatic. Therefore, individuals are not aware of the presence of the virus. Nonetheless, CMV infection can cause serious disease in immunocompromised individuals, like organ transplant recipients. Research has shown that suppression of cellular immune reactions induced by immunosuppresive drugs can lead to CMV reactivation [7].

Other studies have found that CMV infection in fetuses can result in permanent hearing loss and neurological impairments [8]. In addition, it was found that the fraction of T-cells that are CMV-specific increases with age [14]. This suggests that CMV infection plays a role in exhaustion of the immune system, which may be a cause of the reduced effectiveness of influenza vaccination in elderly persons [9].

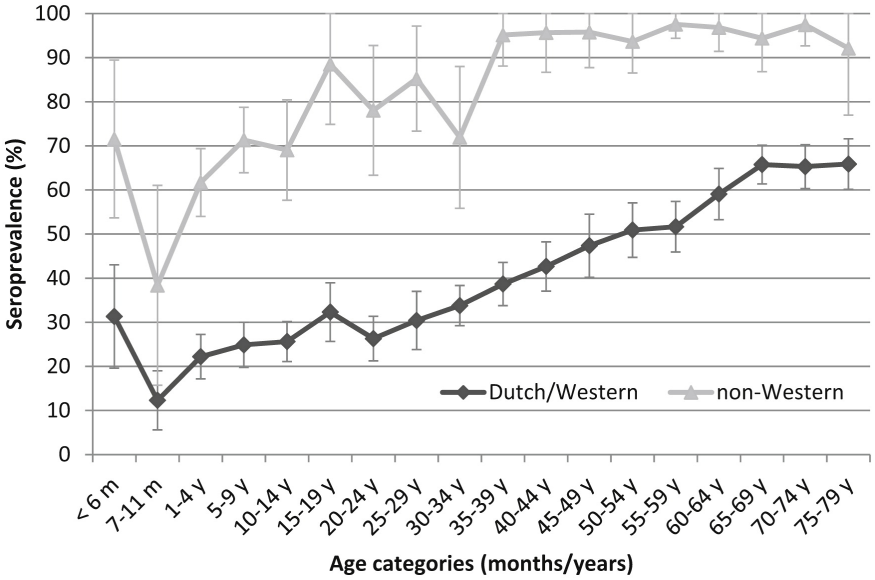


Figure 2.3.: Weighted CMV seroprevalence with 95% confidence intervals per age category for individuals of Dutch/Western origin (5,313) and non-Western origin (1,069), aged 0 - 79 years [15].

The development of a vaccine against CMV was given high priority by the US Institute of Medicine in 2000. No licensed vaccine has been developed yet, although many attempts have been made [21]. Setting up a vaccination program requires knowledge of the transmission dynamics of the virus, which are currently unclear. To combat CMV in the most efficient way, the optimal targets for vaccination must be indicated.

3. Markov Chain Monte Carlo

Monte Carlo methods are a class of algorithms that can be used when a problem is too complicated to solve analytically or with other numerical methods. Monte Carlo methods rely on repeated random sampling. For most models, like the CMV transmission model, parameter estimation cannot be done analytically. Markov Chain Monte Carlo (MCMC) methods, developed in the late 1940s, define a strategy to walk through the parameter space of a probability distribution. MCMC methods are widely used to estimate parameters of a model, given a data set, by generating parameter samples from the target probability distribution. The parameter samples can then be used to determine parameter estimates.

In the Bayesian sense, the target distribution is the posterior distribution. For complex models, posterior distributions cannot be analytically calculated, due to integration of complicated and high-dimensional expressions. However, a posterior distribution can be determined up to a constant, since it is proportional to the prior distribution times the likelihood of the data. When using MCMC methods, it is sufficient to know the target distribution up to a constant.

MCMC methods use Markov chains to sample from the target distribution. Each iteration of the MCMC algorithm results in one parameter sample. The Markov chain is constructed such that the empirical distribution of the generated samples resembles the target probability distribution. Due to the use of a Markov chain, the generated samples are dependent. This chapter describes how MCMC methods are able to produce reliable parameter estimates.

3.1. Markov chains

A Markov chain is a sequence of random variables with the property that the next state of the chain only depends on the current state of the chain, and hence not on previous states. Let (Y_n), $n \in \mathbb{N}$, be a Markov chain taking values in a measurable space $(\mathfrak{Y}, \mathscr{Y})$. If the transition probability is independent of n, the Markov chain is time-homogeneous. Then

$$Pr(Y_{n+1} \in G | Y_n = y_n, Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1) = Pr(Y_{n+1} \in G | Y_n = y_n),$$

for every measurable set $G \in \mathscr{Y}$ and states $y_1, \ldots, y_n \in \mathfrak{Y}$. This is called the Markov property.

The transition kernel of a Markov chain defines the probabilities to move from one state to any other state. Define the kernel that determines one transition by

$$Q(y, G) := Pr(Y_{n+1} \in G | Y_n = y).$$

Then the transition kernel for n transitions is given by the recursive formula

$$Q^n(y, G) = \int_{\mathfrak{Y}} Q^{n-1}(x, G)Q(y, dx),$$

since moving from state y to set G in n steps implies a transition from y to any state $x \in \mathfrak{Y}$, followed by a transition from x to set G in $n-1$ steps.

Under certain conditions, a Markov chain converges to a stationary distribution. Once the chain has reached its stationary distribution, the distribution of the states will not change under further transitions. A probability distribution Π is a stationary distribution for transition kernel Q if, for every measurable set $G \in \mathscr{Y}$,

$$\int_{\mathfrak{Y}} Q(y, G)d\Pi(y) = \Pi(G).$$

If Q and Π respectively allow densities q and π, then the detailed balance equation

$$\pi(y)q(y, z) = \pi(z)q(z, y)$$

is a sufficient condition for Π to be a stationary distribution for transition kernel Q. The detailed balance equation describes that the rate from state y to state z is equal to the rate from state z to state y, hence there is no net change in the distribution of mass over the states.

In order to generate samples from the target probability distribution, the stationary distribution of the Markov chain must be equal to the target probability distribution, and the Markov chain should converge to this distribution. In addition, it must be appropriate to use the generated samples for parameter estimation. Let $\mathfrak{Y}$ be the parameter space, and Π the probability distribution of $y \in \mathfrak{Y}$. Samples $y_1, \ldots, y_n$ are generated from the probability distribution. The required property is

$$\frac{1}{n}\sum_{i=1}^{n} f(y_i) \xrightarrow{p} \int_{\mathfrak{Y}} f(y)d\Pi(y), \tag{3.1}$$

for some measurable function $f : \mathfrak{Y} \to \mathbb{R}$. For instance, if f is the indicator function, Property (3.1) states that the sample mean converges to the true mean of the target probability distribution. By the law of large numbers, Property (3.1) holds if the samples $y_1, \ldots, y_n$ are independent. For this to hold for dependent samples, as are generated by the MCMC method, the Markov chain must satisfy certain properties.

Definition. Given a measure μ, a Markov chain with transition kernel Q is μ-irreducible if, for every measurable set $G \in \mathscr{Y}$ with $\mu(G) > 0$, there exists an $n \in \mathbb{N}$ such that $Q^n(y, G) > 0$, for every $y \in \mathfrak{Y}$.

This means that if the Markov chain is irreducible, every set with positive measure can be reached with positive probability from any state. However, to generate well-mixed samples,

the Markov chain should visit every set with positive measure often enough, which is not ensured by irreducibility. Consider

$$\eta_G := \sum_{n=1}^{\infty} \mathbb{1}_{Y_n \in G},$$

the number of times that the Markov chain visits set G. Let Y_0 denote the initial state of the Markov chain.

Definition. Given a measure μ, a Markov chain is recurrent if it is μ-irreducible and $\mathbb{E}(\eta_G | Y_0 = y) = \infty$, for every measurable set $G \in \mathscr{Y}$ with $\mu(G) > 0$, and every $y \in G$.

Hence the Markov chain is recurrent if the expected number of returns to any set in the state space is infinite, starting from any initial position. However, this still does not ensure that the Markov chain converges to a stationary distribution. Harris recurrence is a stronger version of recurrence.

Definition. Given a measure μ, a Markov chain is Harris recurrent if it is μ-irreducible and, for every set $G \in \mathscr{Y}$ with $\mu(G) > 0$ and every $y \in G$, $Pr(\eta_G = \infty | Y_0 = y) = 1$.

This means that every set with positive measure will be infinitely often visited with probability 1, starting from any initial state.

A Markov chain can be restricted in its transitions. This limits its ability to explore the whole state space. Aperiodicity is required for a well-mixing Markov chain. A state is aperiodic if the Markov chain can return to this state at any time. If the Markov chain is irreducible and has an aperiodic state, then all states are aperiodic, and hence the Markov chain is aperiodic. This is intuitively clear. From the aperiodic state, all other states can, directly or indirectly, be reached due to irreducibility. Since the chain can at any time return to the aperiodic state, all states can be visited at random times.

If the Markov chain (Y_n) is Harris recurrent, and has stationary probability distribution Π, then Property (3.1) is satisfied. If the Markov chain also is aperiodic, it converges to its stationary distribution. Markov chains with these properties are called ergodic. For a proof, see Theorem 13.0.1 in Meyn and Tweedie (2012). This implies that if the Markov chain is ergodic, and its stationary distribution equals the target distribution, then the parameter samples $Y_1, \ldots, Y_n$ generated by the MCMC method can be used to estimate the parameters.

Still, a way must be found to construct an ergodic Markov chain such that its stationary distribution equals the target probability distribution.

3.2. Metropolis-Hastings algorithm

One MCMC sampling technique is the Metropolis-Hastings algorithm. The Markov chain constructed by the Metropolis-Hastings algorithm is ergodic, and its stationary distribution equals the target probability distribution [27]. Therefore, the algorithm is able to generate samples with an empirical distribution that approximates the target probability distribution, which can hence be used to estimate the parameters of the model.

Let π be the target probability density, and q a transition density, corresponding to transition kernel Q. The algorithm consists of a proposal and a correction step. The proposed sample z is generated by the transition kernel given the current sample y, and then accepted with acceptance probability α, defined by

$$\alpha(y, z) = \min\left\{\frac{\pi(z)q(z, y)}{\pi(y)q(y, z)}, 1\right\}.$$

Note that any constant in the target distribution would have cancelled here. Therefore, it is sufficient to know the target distribution up to a constant.

The acceptance probability indicates how probable the proposed sample is with respect to the current sample. If the new sample is more probable, the acceptance probability equals 1, and the transition is made, whereas if the new sample is less probable, the transition is accepted with probability that depends on the target density. Higher acceptance is given to proposed samples that have high probability density.

Fix some initial value $Y_0 = y_0$. The Metropolis-Hastings algorithm is described below [27].

Given Y_n generate Z_{n+1} from $Q(Y_n, \cdot)$.
Generate U_{n+1} from the uniform distribution on $(0, 1)$.
If $U_{n+1} < \alpha(Y_n, Z_{n+1})$ set $Y_{n+1} = Z_{n+1}$,
else set $Y_{n+1} = Y_n$.

If the random number generated from the interval (0,1) is smaller than the acceptance probability, the proposal is accepted. If the sample is not accepted, the new sample is set equal to the previous sample.

Let P be the transition kernel of the Markov chain (Y_n). Note that transition kernel P differs from transition kernel Q, due to the if-statement in the Metropolis-Hastings algorithm. A transition from y to set G is made if the proposed sample is accepted, or if the proposed sample is rejected and $y \in G$. It follows that the transition kernel of the Markov chain is given by

$$P(y, G) = \int_G \alpha(y, z)q(y, z)d\mu(z) + \int_{\mathfrak{Y}} (1 - \alpha(y, z))q(y, z)d\mu(z)\mathbb{1}_G(y).$$

There is no trivial choice for the transition kernel Q. The performance of the algorithm depends on the transition kernel. Samples should be proposed throughout the support of the target probability distribution, and should not be rejected too often. If the overall acceptance probability is high, the Markov chain may be taking too small steps and may therefore not be able to explore the whole target distribution in the available time. On the other hand, if the overall acceptance probability is low, the algorithm frequently performs the else step, and the Markov chain gets stuck. Both extremes result in poorly mixed parameter samples. Therefore, the acceptance rate is a measure for the performance of the Metropolis-Hastings algorithm. Research has indicated that under general conditions, the optimal acceptance probability is 0.234 [24]. However, when the conditions are not met, the optimal acceptance probability can be quite different [1]. With this in mind, the transition kernel can be adapted, such that the optimal acceptance probability is reached.

A frequently used transition kernel density is $q(y, z) = f(z - y)$, for some density f. If the transition kernel takes this form, the algorithm is called the random walk Metropolis-Hastings algorithm, since $Z_{n+1} = Y_n + \varepsilon_n$, with $\varepsilon_n \sim f$ [27]. The density f is often chosen as a normal density. In that case, one could, for instance, change the variance of the normal distribution to achieve an acceptance probability around 0.234. By increasing the variance, samples can be proposed that are more distant to the current sample. This reduces the acceptance probability.

3.3. MCMC algorithm performance

In theory, a properly designed Markov chain will eventually explore any probability distribution. If the Markov chain behaves in an optimal way, it converges from the initial position to the target probability distribution. This is illustrated by Figure 3.1. The number of iterations that this takes, depends on the initial position of the chain, and the choice of the transition kernel Q. The samples that are generated before the Markov chain has converged, should not be included in the output of the algorithm. Therefore, the user can specify a warm-up phase. During this phase, variables that determine the behaviour of the transition kernel, such as the acceptance probability, are being updated. After the warm-up, the transition variables are optimized, and the stationary distribution should be reached. The Markov chain explores the stationary distribution, and the empirical distribution of the generated parameter samples resembles the target probability distribution.

This ideal behaviour requires that the transition kernel Q can smoothly follow the structure of the probability distribution. For instance, problems occur when the probability distribution has regions that are highly curved. The Markov chain may have problems sampling from these regions, resulting in biased estimates. The Markov chain can also become stuck on highly curved regions, generating samples at the same position, and hence overestimating these regions. These problems especially occur when the probability distribution is high-dimensional [2].

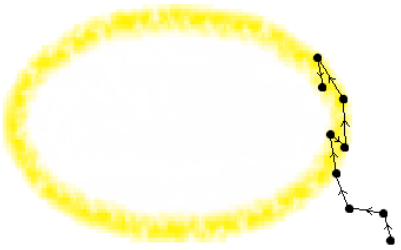


Figure 3.1.: A simplified representation of the convergence of the Markov chain (black) to the target probability distribution (yellow) [2].

A possible solution is to decrease the range of the transition kernel Q, such that the Markov chain takes smaller steps and better recognizes the curves of the probability distribution. As a consequence, it will take longer to explore the whole distribution. If the process is stopped before samples have been generated from the whole distribution, the parameter estimates will not be reliable.

In order to generate reliable samples from a complex and high-dimensional distribution, the geometry of the distribution must be taken into account. Transitions are required that can follow the contours of the distribution, and can take large jumps into unexplored regions. The transitions must be directed in some way.

4. Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a relatively new but upcoming method to generate parameter samples from a probability distribution. The method was developed in the late 1980s, but only recently has it been implemented in software for statistical computations. Like MCMC methods, the HMC method constructs ergodic Markov chains. Using Hamiltonian dynamics, Markov transitions with directed movement are created, thereby avoiding inefficient random walk behaviour. It is sufficient to know the target probability distribution up to a constant.

To clarify the idea, imagine a ball on the frictionless surface of a probability distribution. The position of the ball corresponds to the parameter sample. Each position variable is associated with a momentum variable, a measure for the direction and velocity of the ball. Its potential energy is the height of the surface at the position of the ball. Its kinetic energy is related to its momentum. The total energy, named the Hamiltonian, equals the sum of the potential and kinetic energy. The Hamiltonian is invariant, since during the motion of the ball, potential energy is continuously transferred into kinetic energy, and vice versa. For if the ball loses height, it gains velocity, and if the ball gains height, it loses velocity. Given some initial momentum, the ball will evaluate some predefined part of the probability distribution.

The Hamiltonian is defined in terms of the target probability distribution. In addition to the position variables, momentum variables are introduced. This chapter describes how the HMC method makes use of momentum and energy to efficiently explore the target probability distribution.

4.1. Hamiltonian dynamics

Consider a D-dimensional parameter space, containing position vectors x. This space can be extended to a $2D$-dimensional phase space, containing pairs of position and momentum vectors (x,m). Let the target probability distribution be denoted by $\pi(x)$. The joint distribution function of the pair (x,m), called the canonical distribution, can then be defined by

$$\pi(x, m) = \pi(m|x)\pi(x),$$

according to Bayes's rule. The canonical density can be written in terms of the Hamiltonian H,

$$\pi(x, m) = e^{-H(x,m)}.$$

The value of the Hamiltonian in a point is called the energy of that point. It follows that the Hamiltonian can be written as

$$H(x,m) = -\log \pi(m|x) - \log \pi(x) \coloneqq K(x,m) + V(x),$$

where $K(x,m)$ denotes the kinetic energy and $V(x)$ the potential energy [18]. The potential energy is known, since the target probability distribution $\pi(x)$ is known, at least up to a constant. The kinetic energy, related to the conditional distribution $\pi(m|x)$, must be specified. The kinetic energy can be defined in many ways.

The partial derivatives of the Hamiltonian determine how position x and momentum m change over time, according to Hamilton's equations

$$\frac{dx}{dt} = \frac{\partial H}{\partial m}, \quad \frac{dm}{dt} = -\frac{\partial H}{\partial x}.$$

Hamilton's equations define paths that move through the phase space. By projecting these paths onto the parameter space, the target probability distribution can be explored [2]. The Hamiltonian paths are obtained by integrating Hamilton's equations.

4.2. Choice of kinetic energy

The momentum variables are defined by the kinetic energy. There are multiple ways to define the kinetic energy, but not all forms will work well. Every choice of kinetic energy and integration time yields a transition that will interact differently with a given parameter distribution.

From Hamilton's equations it follows that the Hamiltonian is conserved,

$$\frac{dH}{dt} = \frac{dx}{dt}\frac{\partial H}{\partial x} + \frac{dm}{dt}\frac{\partial H}{\partial m} = \frac{\partial H}{\partial m}\frac{\partial H}{\partial x} - \frac{\partial H}{\partial x}\frac{\partial H}{\partial m} = 0.$$

Consequently, every Hamiltonian path is restricted to a set with a constant value of the Hamiltonian. This value is called the energy E of the path, and the set is called the energy level set, defined by

$$S \coloneqq \{x, m | H(x,m) = E\}.$$

Each pair (x,m) can be specified by the energy E of its level set and its position x_E within the level set. Therefore, the canonical distribution can be decomposed into a microcanonical distribution and a marginal energy distribution,

$$\pi(x,m) = \pi(x_E|E)\pi(E).$$

A Markov transition defines jumps between the level sets, part of which are then explored using Hamiltonian dynamics. The efficiency of the exploration of a level set is determined by how completely the set is explored, and how long this takes. With the optimal integration time, just enough of each level set is explored. To increase the integration speed, the

level sets must be simple. The ideal kinetic energy $K(x, m)$ will interact with the parameter distribution $\pi(x)$ in a way that the level sets are as uniform as possible. Nonetheless, there still remain infinitely many possible choices for the kinetic energy [2].

The kinetic energy is typically chosen to be the negative log density of a Gaussian random vector with zero mean. Note that in this simplified case, the kinetic energy is independent of the current position x. In this context, the covariance matrix of the Gaussian distribution is known as the mass matrix. The effect of the mass matrix is to transform the target distribution in such a way that it has a simpler geometry for sampling. If the mass matrix approximates the covariance of the target distribution, the variances in the mass matrix give all parameters of the target distribution the same scale, and the covariances make the parameters approximately independent. This results in energy level sets that are more uniform and hence easier to explore. It is possible to use other forms of kinetic energy, but Gaussian kinetic energies tend to perform well [18].

4.3. HMC algorithm

The HMC algorithm consists of two steps. First, momentum variables m^* are generated from the conditional distribution $\pi(m|x)$, defined by the kinetic energy, using the initial values of the position variables x. Using Hamiltonian dynamics with these momentum variables, the level set is explored (explained below) and new position variables x^* are generated. This is followed by an acceptance step. The new state (x^*, m^*) is accepted, given the current state (x, m), with probability

$$\alpha((x^*, m^*)|(x, m)) = \min\left\{\frac{\pi(m^*|x^*)\pi(x^*)}{\pi(m|x)\pi(x)}, 1\right\}.$$

The acceptance probability indicates how probable the new state is with respect to the current state. If the new state is more probable, the acceptance probability equals 1, and the transition is made, whereas if the new state is less probable, the transition is accepted with probability that depends on the target density. Higher acceptance is given to proposed states that have high probability density. The Hamiltonian Monte Carlo algorithm is described below.

Given x_n generate m_{n+1} from $\pi(\cdot|x_n)$.
With Hamiltonian dynamics, propose x'_{n+1}.
Generate U_{n+1} from the uniform distribution on $(0, 1)$.
If $U_{n+1} < \alpha((x'_{n+1}, m_{n+1})|(x_n, m_n))$ set $x_{n+1} = x'_{n+1}$,
else set $x_{n+1} = x_n$.

Figure 4.1 visualises the exploration of the level sets. When a momentum variable is generated, a transition is made to another level set. The level set is explored for some time, and then a position variable is generated, and possibly accepted.

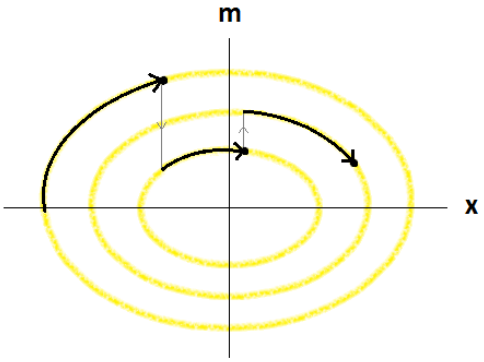


Figure 4.1.: Each Markov transition (grey) lifts the current state onto a random energy level set (yellow), which is then explored by a Hamiltonian path (black). The projection of the dots to the x-axis gives the generated position samples [2].

In practice, the Hamiltonian paths cannot be analytically computed. Hamilton's equations must be approximated by discretizing time, using some small step size. The HMC algorithm generates trajectories using symplectic integrators. A symplectic integrator numerically solves Hamilton's equations, while the symplectic structure of the Hamiltonian dynamics is preserved. This means that numerical trajectories that are generated by symplectic integrators preserve phase space volume. As a consequence, the error between the numerical trajectory and the exact path is limited. Betancourt (2017) provides a detailed description of symplectic integrators.

One symplectic integrator is the leapfrog integrator. Given a small step size ϵ, the leapfrog integrator simulates the Hamiltonian path by the algorithm below [18].

$$
\begin{aligned}
m\left(t+\frac{\epsilon}{2}\right) &= m(t) - \frac{\epsilon}{2}\frac{\partial V}{\partial x}(x(t)) \\
x(t+\epsilon) &= x(t) + \epsilon\frac{\partial K}{\partial m}\left(m\left(t+\frac{\epsilon}{2}\right)\right) \\
m(t+\epsilon) &= m\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial V}{\partial x}(x(t+\epsilon))
\end{aligned}
$$

First, a half step is taken to compute the momentum variables. Then, using these momentum variables, a full step is taken to determine the new position variables. At last, again a half step is taken to determine new momentum variables, using the current values of the position and momentum variables. This is repeated for a defined number of steps with size ϵ. More steps of smaller size lead to a better approximation of the continuous path, but this increases the run time. When too few steps are taken, the algorithm will show random walk behaviour. On the other hand, when too many steps are taken, the trajectory explores position variables that it had already explored, which is a waste of time. Figure 4.2 shows the approximation of one path by a trajectory. The red part indicates that the trajectory explores position values that it had already reached. In this case, too many steps were taken.

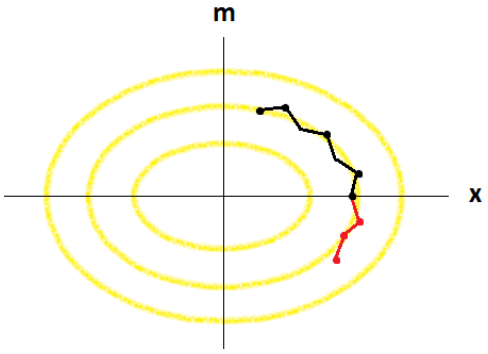


Figure 4.2.: A level set is explored by a trajectory that has been calculated by the leapfrog integrator. The red part indicates that the trajectory has already reached these positions x. This should be prevented. The new position is sampled from the positions that the leapfrog algorithm has generated (indicated by dots).

Approximation errors cause the trajectory to deviate from the continuous path. As a consequence, the Hamiltonian is not constant over time. Therefore, it can occur that the numerical trajectory drifts away from the energy level set. Then divergent samples are generated.

After some defined integration time, the new position is sampled from the positions on the generated trajectory, and is possibly accepted. This is repeated for each iteration of the HMC algorithm. Figure 4.3 illustrates how the resulting numerical Hamiltonian trajectories explore the level sets.

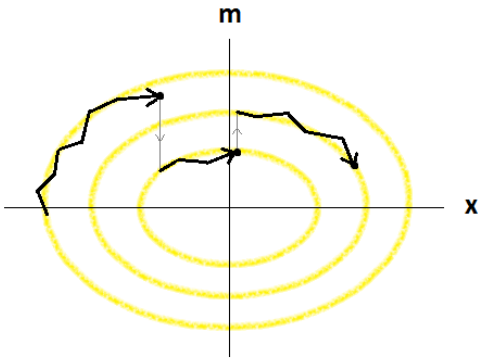


Figure 4.3.: Each Markov transition (grey) lifts the current state onto a random energy level set (yellow), which is then explored by a Hamiltonian trajectory (black), calculated by the leapfrog integrator.

The HMC algorithm makes it possible to jump to regions with different probability density. In an iteration, any value can be sampled for the momentum variables, which then affect the position variables in a controlled way. As a consequence, the algorithm shows no random walk behaviour, like the MCMC methods, but systematically moves through the target distribution. Furthermore, the HMC algorithm is ergodic. The Markov chain asymptotically converges to its stationary distribution, which is equal to the target probability distribution, and Property (3.1) holds. Therefore, the HMC algorithm generates position vectors that can be used to estimate parameters of a model. For a proof of ergodicity of the HMC method, see Livingstone et al. (2016).

5. Stan

Stan is a program that performs Bayesian inference and optimization for a broad class of models. Stan was motivated by the desire to solve problems that could not be solved in reasonable time using existing programs. The Stan user defines a log likelihood function over the parameters, conditioned on data, and defines priors for the parameters. Stan then computes the log posterior density up to a constant. Stan uses Hamiltonian Monte Carlo to sample from the log posterior density, and uses the No-U-Turn Sampler to tune the HMC algorithm. The program is named after Stanislaw Ulam, a mathematician who was one of the developers of the Monte Carlo method in the 1940s. The first version of Stan was released in 2012. Stan currently is in version 2.14, and is still improving. Stan is faster for complex and high-dimensional models than other programs, and scales better for large data sets [11].

A Stan program is translated to a C++ program by the Stan compiler for efficient computation. Stan programs can be called from many platforms, including R, Python, Matlab and Julia. Stan programs are very structured. A program contains seven blocks, one for functions, data, parameters, transformed data, transformed parameters, the model, and generated quantities. The functions block is optional and contains user-defined functions. In the data block, data is declared that is input for the Stan program. The parameters block states the parameters that must be estimated. The transformed data and parameters blocks are optional and contain functions of data and parameters, to make the model better readable. In the model block, priors for the parameters are defined, together with the log probability density. The generated quantities block is optional, and contains values that the user wants to have returned. Parameters and transformed parameters can be restricted. For instance, one can define a parameter to be positive. See Appendix A for the Stan program code of the CMV transmission model.

5.1. Reverse-mode automatic differentiation

Sampling using HMC requires calculation of gradients. For each leapfrog step, gradients of the potential energy and the kinetic energy need to be evaluated at the position corresponding to the current parameter values of the log posterior with respect to the parameters in the leapfrog algorithm. Stan computes gradients using reverse-mode automatic differentiation. This is where most of the time in the Stan algorithm is spent.

Reverse-mode automatic differentiation works as follows. Let f be a differentiable function, and let x_i be its variables, for $i = 1, \ldots, n$. The goal is to evaluate the partial derivatives $\frac{\partial f}{\partial x_i}$. The function f can be split into parts g_j, $j = 1, \ldots, k, k > n$, that have relatively simple

derivatives. These parts build up f in such a way that $g_1 = x_1$, $g_2 = x_2$, ..., $g_n = x_n$, and g_{n+1}, ..., g_k are functions of previously defined parts g_j, such that at last $g_k = f$.

The derivatives of the parts are calculated in reverse, starting with $\frac{\partial f}{\partial g_k}$ (which equals 1), and ending with $\frac{\partial f}{\partial g_1}$. Let h_p denote differentiable functions. The partial derivatives can be calculated using the chain rule,

$$\frac{\partial f}{\partial g_i} = \sum_{p:g_p=h_p(g_i)} \frac{\partial f}{\partial g_p} \frac{\partial g_p}{\partial g_i}.$$

Note that $\frac{\partial z}{\partial g_i} = \frac{\partial f}{\partial x_i}$, for $i = 1, \ldots, n$. Hence, the desired gradient is obtained.

In order to calculate gradients, the kinetic energy must be defined. By default, Stan uses the $-\log$ of the Gaussian distribution with mean 0 and a diagonal mass matrix. A mass matrix of this form accounts for the difference in scale between the parameters, but not for correlation. The mass matrix is estimated by adjusting it to the sample covariance of the latter half of the warm-up samples. Mass matrices with non-zero covariance terms are available in Stan, but are not commonly used. The sampling is most efficient if the mass matrix approximates the covariance of the posterior, since this simplifies the energy level sets, but often there is not enough information to determine this. Moreover, using non-diagonal mass matrices will slow the leapfrog integration even further [25].

5.2. No-U-Turn Sampler

HMC requires the right tuning to be efficient. Two variables that have to be tuned are the number of steps and the step size that are used in the leapfrog integrator to approximate the Hamiltonian paths. Together, these two variables determine the trajectory length. If the trajectory length is too short, it is impossible to propose distant samples. This will result in an inefficient random walk. On the other hand, if the trajectory length is too long, the trajectory will retrace itself, which is a waste of time. Fewer steps result in a faster algorithm, but this must be compensated by a larger step size to get the desired trajectory length. This leads to more variation in the Hamiltonian. As a consequence, the approximation error may accumulate, and divergent samples are generated. Until recently, the number of steps and the step size had to be tuned by the user. The No-U-Turn Sampler (NUTS) has taken over this task [13].

During the warm-up phase, NUTS determines the step size by adapting it to a target acceptance rate. This step size is then used for all iterations. The target acceptance rate is the only tuning variable that is left to the user. By default the acceptance rate equals 0.8. By increasing this rate, the step size decreases and the algorithm becomes more accurate. This may help when many divergences occur, but it slows down the algorithm.

NUTS determines the optimal number of steps using a tree building algorithm. In the first iteration, a single leapfrog step is taken from the current state so the trajectory has a total of two steps. Then two more steps are added, then four more, then eight more, and so on. Hence, the length of the trajectory doubles with each iteration. This continues until the trajectory returns to position values that it has already reached (a U-turn occurs, as in Figure 4.2), or diverges. The number of doublings is known as the tree depth. This tree building algorithm automatically creates trajectories that are neither too short nor too long. This means that trajectory lengths vary among iterations.

5.3. Stan performance

The main limitation of Stan is that it does not work for models with discrete parameters. This is due to the use of the Hamiltonian Monte Carlo method. HMC can only be used to sample from continuous distributions for which the density function can be evaluated. The partial derivatives of the log of this density function must exist. Discrete parameters simply do not have gradients. In addition, there are models that are slow to run or do not converge in Stan, but overall, Stan performs better and faster on most models than statistical inference programs that are based on MCMC methods, such as JAGS [11].

Using Hamiltonian dynamics, Stan is able to generate well-mixed samples from complex and high-dimensional posterior density functions. However, high curvature regions of the posterior can still cause problems. A trajectory in such a region may diverge, resulting in a sample that is off the posterior density. Stan indicates which samples have diverged. Often, the number of divergent samples can be reduced by reducing the step size in the leapfrog integrator. This can be done by changing the value of the target acceptance rate when calling the Stan program.

A Stan program can run multiple chains in parallel. This is advised, since the generated samples may converge to different distributions, or may not converge at all. If this is the case, there is probably something wrong with the program code, and the generated samples should not be used for analysis. Moreover, the run time decreases if the total number of iterations is divided over multiple chains that run at the same time. However, the number of chains that can run in parallel depends on the number of cores of the computer.

To check whether the Markov chains have converged to the stationary distribution, the samples that have been generated by the chains can be compared. Stan returns the potential scale reduction statistic $\hat{R}$, which measures the ratio of the variance of the samples generated by each chain to the variance of all generated samples together. If all chains have converged to the stationary distribution, the variances will be the same and $\hat{R}$ will be close to 1. Stan splits each chain in half before calculating $\hat{R}$, in order to detect whether the first half of each chain mixes well with the second half.

ShinyStan is a software tool that provides visual and numerical diagnostics of the generated Stan fit [26]. It visualises the chains, such that the user can easily see whether the chains have converged and are well-mixed. ShinyStan also shows which transitions have diverged, and presents summaries of the parameter estimates. Therefore, ShinyStan is a useful tool to get a global indication of the model's behaviour.

The Stan team is currently working on improving Stan. The team is involved with the Stan users. Questions in the Stan Google Group are quickly answered, and suggestions from users to improve Stan are taken into consideration. The team works on stabilising Stan and making it work on models with very complex and high-dimensional posteriors, while keeping the program run fast. In the future, the team also wants to extend Stan, and to allow models with discrete parameters [25].

6. Age-structured transmission models

Age plays an important role in the transmission dynamics of many infectious diseases. Due to different behaviour and state of the immune system, individuals of different age may not have the same risk of infection. An age-structured transmission model takes this into account.

Throughout this chapter, it is assumed that the population is in endemic equilibrium. This means that the infection has established itself, such that the fractions of susceptible and infected individuals of all ages do not change in time. Therefore, all variables are time-independent. This chapter discusses the design of an age-structured transmission model, as has been described in Diekmann et al. (2013). Reproduction numbers, which are important measures for the potential of an infection to become an epidemic, are also discussed.

6.1. The next-generation operator

Age is a continuous variable. One could define a maximum age, but for now take age $a \in \mathbb{R}_{\geq 0}$. Define $\mathcal{F}(a)$ by the probability that an arbitrary individual of age 0, will at least reach age a. It is assumed that infected and uninfected individuals have the same life expectancy. This is a reasonable assumption, unless the infection significantly increases the risk of death.

A contact between two individuals can be defined in many ways, depending on how the infection is transmitted through the population. For instance, a contact can be two persons having a conversation, but can also be sexual. The expected number of contacts per unit of time that an individual of age a has with individuals of age α depends on the definition of the contact, and the size and composition of the population. Let $\tilde{c}(a, \alpha)$ be the contact rate, defining the contact that one individual of age a has with one individual of age α, per unit of time. Let $N(a)$ denote the number of individuals in the population with age a. It follows that an individual of age α is contacted by $\int_0^\infty \tilde{c}(a, \alpha)N(a)da$ individuals of all ages, per unit of time.

When a susceptible individual has contact with an infected individual, the infection is not necessarily transmitted. Define $h(\tau, \alpha)$ by the infectiousness of the infection. This function describes the probability of transmission of the infection, given a contact at time $\alpha + \tau$ between a susceptible of any age and an individual that was infected at age α. Parameter τ denotes the time after this individual was infected. This probability depends on α and τ, since the infectiousness of an individual may depend on the individual's age, and may change with the passage of time.

The next-generation operator describes how the distribution of the infected population changes from one generation to the next generation. Here, the population is distributed over the ages. Let $\phi(\alpha)$ be a function that describes the generations, such that $\int_A \phi(\alpha)d\alpha$ equals the expected number of infected individuals with ages in $A \subset \mathbb{R}_{\geq 0}$. The next-generation operator is defined by its kernel $k(a, \alpha)$, which describes the expected number of newly infected individuals with age a, caused by one individual that was infected at age α, during the time that this individual is infectious. The next-generation operator K is defined by

$$(K\phi)(a) = \int_0^\infty k(a, \alpha)\phi(\alpha)d\alpha. \tag{6.1}$$

The kernel $k(a, \alpha)$ can be expressed in terms of $\mathcal{F}$, $\tilde{c}$, N and h. An individual that became infected at age α will τ units of time later have age $\alpha + \tau$, provided that the individual survives from age α to age $\alpha + \tau$. The probability that the individual reaches age $\alpha + \tau$, given that the individual has reached age α, is given by $\mathcal{F}(\alpha+\tau)/\mathcal{F}(\alpha)$. The probability that the infection is transmitted from this individual to a susceptible individual is $h(\tau, \alpha)$. At the time of transmission, the infected individual has age $\alpha + \tau$, and therefore has contact with $\tilde{c}(a, \alpha+\tau)N(a)$ individuals of age a, per unit of time. Hence, the kernel of the next-generation operator is given by

$$k(a, \alpha) = \int_0^\infty h(\tau, \alpha)\tilde{c}(a, \alpha + \tau)N(a)\frac{\mathcal{F}(\alpha + \tau)}{\mathcal{F}(\alpha)}d\tau. \tag{6.2}$$

6.1.1. The short-disease approximation

If a disease has an infectious period that is at least three orders of magnitude shorter than the average life expectancy, it is reasonable to make the short-disease approximation [6]. The short-disease approximation implies that the probability that an individual dies during the infectious period is small, and that the contact rate does not change during this short period of time. Hence

$$\frac{\mathcal{F}(\alpha + \tau)}{\mathcal{F}(\alpha)} \approx 1 \text{ and } \tilde{c}(a, \alpha + \tau) \approx \tilde{c}(a, \alpha).$$

In addition, a transmission probability that depends only on the age of infection can be defined by

$$H(\alpha) \coloneqq \int h(\tau, \alpha)d\tau.$$

As a consequence, the kernel of the next-generation operator, as stated by Equation (6.2), takes an easier form,

$$k(a, \alpha) = H(\alpha)\tilde{c}(a, \alpha)N(a).$$

For notational convenience, define the contact intensity by

$$c(a, \alpha) \coloneqq \tilde{c}(a, \alpha)N(a), \tag{6.3}$$

such that

$$k(a, \alpha) = H(\alpha)c(a, \alpha).$$

The contact intensity $c(a, \alpha)$ is the total number of contacts that one individual of age α has with individuals of age a, per unit of time.

6.1.2. The force of infection

The force of infection is defined by the probability per unit of time for a susceptible to become infected. Let $S(a)$ denote the fraction of susceptibles of age a. It is assumed that the infection does not affect population numbers. The force of infection for an individual of age a, denoted by $\Lambda(a)$, is described by the contact intensity with individuals of age $\alpha + \tau$, times the fraction of individuals that were infected at age α, given by $\Lambda(\alpha)S(\alpha)$, and that are still alive τ units of time later, times the probability that this results in transmission of the infection. Hence the force of infection is given by

$$\Lambda(a) = \int_0^\infty \int_0^\infty h(\tau, \alpha)c(a, \alpha + \tau)\Lambda(\alpha)S(\alpha)\frac{\mathcal{F}(\alpha + \tau)}{\mathcal{F}(\alpha)}d\tau d\alpha.$$

When the short-disease approximation is applied, this becomes

$$\Lambda(a) = \int_0^\infty H(\alpha)c(a, \alpha)\Lambda(\alpha)S(\alpha)d\alpha.$$

6.2. The basic reproduction number

Reproduction numbers are measures for the ability of an infection to spread through a population. The basic reproduction number R_0 is defined by the expected number of individuals that one infected individual will infect during its entire period of infectiousness, in a susceptible population [6]. When $R_0 > 1$, it is expected that the infected individual infects at least one other individual, and the infection is able to invade the population. When $R_0 < 1$, it is expected that the infected individual infects less than one other individual, and the infection cannot establish itself. The infection can be controlled by vaccination of the population. This reduces the number of infectious individuals, and therefore reduces R_0. The fraction of the population that should be vaccinated such that $R_0 < 1$ is called the critical control effort.

6.2.1. Calculation of R_0

Calculating R_0 is not straightforward. The next-generation operator defined by Equation (6.1) works on an infinite-dimensional and uncountable vector space. In order to introduce the basic reproduction number, consider the following simplified situation.

Suppose that the vector space is countable. One could, for instance, consider age classes of a certain range. The next-generation operator K can then be represented by a next-generation matrix M. By definition, the next-generation matrix is non-negative. For simplicity, let M

be an $n \times n$ non-negative matrix with real, distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $x_1, \ldots, x_n$. Since the eigenvectors are independent, any vector $\phi \in \mathbb{R}$ can be written as a linear combination of these eigenvectors, with constants $c_1, \ldots, c_n \in \mathbb{R}$. Let vector ϕ describe the distribution of the current generation of infected individuals over the age classes. Then the mth generation is given by $M^m\phi$. It follows that

$$\begin{aligned} M^m\phi &= M^m(c_1x_1 + c_2x_2 + \cdots + c_nx_n) \\ &= c_1\lambda_1^m x_1 + c_2\lambda_2^m x_2 + \cdots + c_n\lambda_n^m x_n \\ &= c_1\lambda_1^m \left(x_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^m x_2 + \cdots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^m x_n \right). \end{aligned}$$

Now suppose that $\lambda_1 > |\lambda_i|$, for $i = 2, \ldots, n$. Then λ_1 is called the dominant eigenvalue of M. It follows that $\lim_{m\to\infty} M^m\phi = \lim_{m\to\infty} c_1\lambda_1^m x_1$. So the dominant eigenvalue of the next-generation matrix determines whether the infected population grows or declines. If $\lambda_1 > 1$, the infection is able to invade the population, whereas if $\lambda_1 < 1$, the infection cannot establish itself. In this case, the basic reproduction number R_0 is equal to the dominant eigenvalue of the next-generation matrix.

More generally, R_0 is defined as the spectral radius of the next-generation operator K [12]. If the next-generation operator can be described by a matrix, the spectrum $\sigma(K)$ is equal to the set of eigenvalues of this matrix. The eigenvalues of a matrix M can be found by solving the characteristic equation $M - \lambda I = 0$, where I denotes the identity matrix. Similarly, the spectrum of a linear operator K is the set of all $\lambda \in \mathbb{R}$ for which $K - \lambda I = 0$, where I denotes the identity operator. The spectral radius of K is defined by

$$\rho(K) \coloneqq \sup\{|\lambda| : \lambda \in \sigma(K)\},$$

the radius of the smallest circle in the complex plane that is centred at the origin and contains the whole spectrum of K.

In 1907, the German mathematician Perron proved that the spectral radius of a positive, square matrix is the dominant eigenvalue of this matrix. This eigenvalue is algebraically and geometrically simple, and has a positive eigenvector [20]. Hence, if the next-generation matrix suffices these properties, R_0 is equal to the dominant eigenvector of the next-generation matrix. However, the next-generation matrix is non-negative, and not necessarily positive. In 1912, the German mathematician Frobenius extended Perron's theorem for non-negative matrices [10]. As an additional requirement, the matrix must be irreducible.

Definition. A non-negative matrix A is irreducible if there exist integers m for every index pair (i, j), such that $(A^m)_{ij} > 0$. If there exists one integer m such that $(A^m)_{ij} > 0$ for all pairs (i, j), then A is primitive.

The Perron-Frobenius theorem states that the spectral radius of a non-negative, square and irreducible matrix A is the strictly dominant eigenvalue of A. This eigenvalue is algebraically and geometrically simple, and the corresponding eigenvector is strictly positive. Moreover, if A is primitive, the dominant eigenvalue is the only eigenvalue with a positive eigenvector [10].

Hence, if the next-generation operator can be described by an irreducible matrix, R_0 is equal to the dominant eigenvalue of this matrix. However, there is no general way to find the spectral radius of a linear operator on an uncountable vector space, where the next-generation operation cannot be described by a matrix. In practice, assumptions are made to define the next-generation operator on a countable vector space such that R_0 can be computed.

6.2.2. Critical control effort

If $R_0 > 1$, the infection is expected to invade the population. To prevent this, individuals can be vaccinated. This reduces the expected number of infections per infected individual.

Consider a homogeneous population and suppose that a perfect vaccine is available. When individuals are vaccinated with this vaccine, they cannot be infected and are not infectious. Let a fraction v of the population be vaccinated. Then a fraction v of all contacts does no longer lead to infection. Hence, the expected number of infections per infected individual is

$$R_v \coloneqq (1-v)R_0.$$

When $R_v < 1$, the infection is under control and will not invade the population. It follows that a fraction $v > 1 - 1/R_0$ of the whole population should be vaccinated to reach this effect. The fraction $v_c = 1 - 1/R_0$ is called the critical control effort.

Consider a heterogeneous population, with next-generation matrix M that has dominant eigenvalue R_0. Let x be the dominant eigenvector of M, hence $Mx = R_0x$. Suppose that the same fraction v of each population subgroup is vaccinated. This yields a new next-generation matrix $M_v \coloneqq (1-v)M$. It follows that $M_vx = (1-v)R_0x = R_vx$. So R_v is an eigenvalue of M_v, but is this the dominant eigenvalue?

Suppose that there exists an eigenvalue $\lambda > R_v$ of M_v, with corresponding eigenvector y. Then $M_vy = (1-v)My = \lambda y$, so $My = \lambda/(1-v)y$. It follows that y is an eigenvector of M with corresponding eigenvalue $\lambda/(1-v)$. This is a contradiction, since $\lambda/(1-v) > R_0$, while R_0 is the dominant eigenvalue of M. Hence R_v is the dominant eigenvalue of M_v.

If not all population subgroups are vaccinated, or are vaccinated with different fractions, the critical control effort v_c is not equal to $1 - 1/R_0$. A new next-generation matrix must be constructed, wherein the vaccination fractions, in terms of v, are incorporated per entry. Then the dominant eigenvalue can be determined, and the critical control effort v_c can be calculated such that the dominant eigenvalue is smaller than 1.

7. A transmission model for CMV

This chapter describes the sex- and age-structured model for the transmission of CMV, that has been constructed in Van Boven et al. (2017). The model follows the structure described in Chapter 6. In Van Boven et al. (2017) statistical models are fitted to data in order to define classes of susceptible and infected individuals. In that paper, transmission of CMV is modelled by differential equations, which are discretized in order to estimate the parameters with the available data.

This thesis adds the estimation of perinatal transmission of CMV and the calculation of the basic reproduction number to the model described in Van Boven et al. (2017). In this chapter, the equations for perinatal transmission are stated and, in order to calculate the basic reproduction number, the continuous next-generation operator is transformed into a next-generation matrix by discretization.

The analyses make use of data from the cross-sectional population-based study PIENTER2 [19], carried out in the Netherlands in 2006 and 2007. In a cross-sectional study, data are collected from many subjects, without regard to differences in time. Forty districts distributed over five geographic regions of the Netherlands were randomly selected with probabilities proportional to their population size, and individuals were sampled proportional to the age distribution. Blood samples and questionnaires were obtained from 6,382 participants. Only data of Dutch persons and migrants of Western ethnicity are analysed, since CMV seroprevalence of individuals with Western and non-Western ethnicity significantly differ (Figure 2.3). Furthermore, analysis is restricted to sera from individuals older than six months, since antibodies of the mother are still present in the blood of the child before that age. This yields a total of 5,179 observations that are fit for analysis.

The extended CMV transmission model is implemented in Stan, in order to use the HMC method to sample from the posterior distribution. The original Stan code for the transmission model was provided by Chris van Dorp. In this thesis, the code is extended and improved.

The CMV transmission model is fitted to data to estimate the rates of primary infection, reinfection, reactivation and perinatal transmission, and infectiousness of primary infected individuals and individuals that have experienced reinfection or reactivation. The data contain CMV titer values, and information about sex, age and ethnicity. It is assumed that CMV titer values provide information on CMV infection status, and also on whether or not reinfection or reactivation has taken place. Persons with low, intermediate and high titer values are considered susceptible, latently infected and infected with reinfection or reactivation episodes, respectively.

7.1. Mixture distribution for classification

In Van Boven et al. (2017), different statistical models are fitted to the serological data set, in order to classify the CMV titer values. The CMV titer values were transformed such that they can be described by normal mixture distributions. This was done using a Box-Cox transformation with parameter 0.3, given by

$$\tilde{T}_j = \frac{T_j^{0.3} - 1}{0.3},$$

for each titer value T_j. The unit of a transformed titer value is denoted by U/ml. The transformed CMV titer values are presented in Figure 7.1.

It is found in Van Boven et al. (2017) that the data are well described by a mixture of three normal distributions, corresponding to low, intermediate and high titer values. Therefore, three classes, S, L and B are defined, describing the susceptible individuals, latently infected individuals and infected individuals with boosted titer values, respectively.

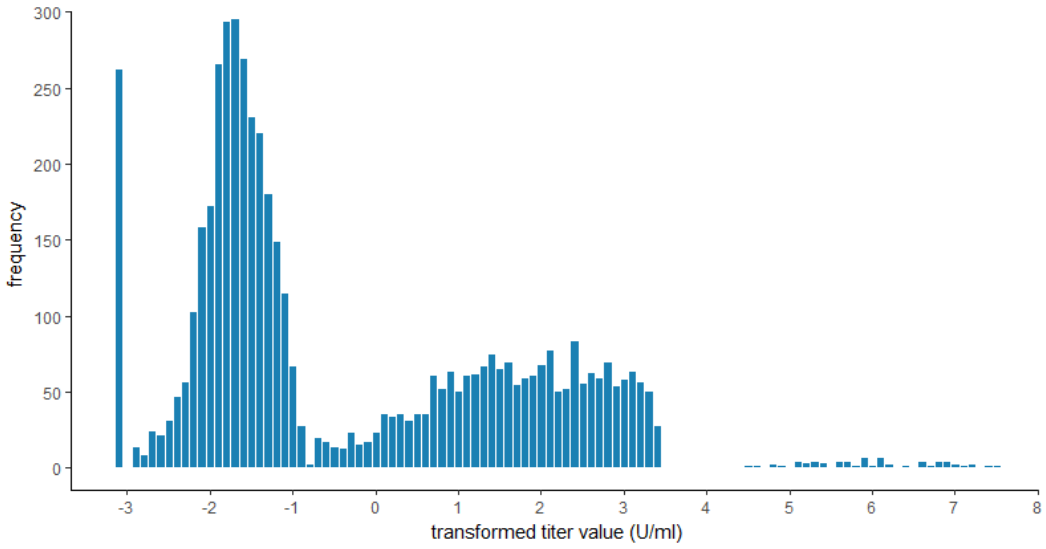


Figure 7.1.: Histogram of the 5,179 transformed CMV titer values [29].

The normal mixture distribution is defined by six parameters, namely the means μ_S, μ_L and μ_B, and the standard deviations σ_S, σ_L and σ_B of the three normal distributions. The parameters were initially estimated in Van Boven et al. (2017) using R [23] and JAGS [22]. Normal prior distributions $\mathcal{N}(0, 0.001)$ were set on the means, and flat Gamma prior distributions were set on the standard deviations. The MCMC method was used to sample from the posterior distribution.

Figure 7.2 shows the normal distributions that provide the best fit to the data. It can be seen that susceptible individuals and infected individuals are better distinguished than latently infected individuals and individuals that have experienced reinfection or reactivation. The mixture distribution is given as input to the transmission model.

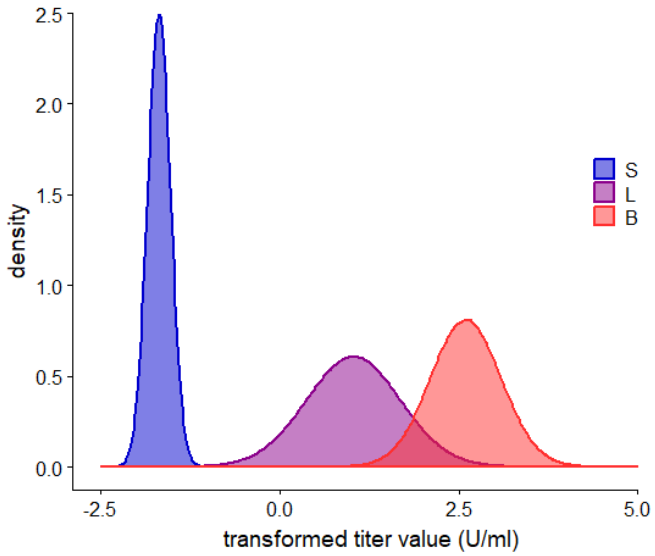


Figure 7.2.: The normal mixture distribution for the classification of the titer values to the classes S (susceptible), L (latently infected) and B (infected with boosted titer values). The distributions for these classes are $\mathcal{N}(-1.68, 1/6.27)$, $\mathcal{N}(1.03, 1/1.52)$ and $\mathcal{N}(2.60, 1/2.03)$ respectively [29].

A significant number of transformed titer values are concentrated near -2.91 U/ml. This spike can be seen in Figure 7.1. Therefore, class S is modelled by a mixture of a normal distribution and this spike, to ensure that titers smaller than or equal to this value belong to class S with probability 1. This spike is not included in Figure 7.2.

Figure 7.1 also shows a clear shift in frequency between 3 and 4 U/ml. This is due to the limited capability of the measuring instrument. Right-censoring is applied to the transformed titer values above 3.41 U/ml. These titers are out of the range of the measuring instrument, therefore they could be 3.41 U/ml or higher. This is taken into account when the log likelihood is computed.

7.2. The model's equations

The transmission model assumes that the population is in endemic equilibrium. This means that the fractions of susceptible and infected individuals do not change in time, so that all variables are time-independent. According to the mixture distribution, the population can be divided into three classes. Individuals can be either susceptible (S), latently infected (L), or infected with boosted titer values (B). Latently infected individuals can become infectious again by reinfection or reactivation, as is illustrated in Figure 2.2. The assumption is made that the titer values of these individuals increase after reinfection or reactivation, and hence these individuals are placed in class B in this model.

Person-to-person transmission of CMV to an uninfected person can occur from a primary infected person, or from a person who has experienced a reactivation episode or has been reinfected. As the infectious period is short relative to the lifespan of the host (weeks versus decades), the short-disease approximation is made, as is described in Section 6.1.1. The infectious periods are not explicitly modelled, and it is assumed that the transitions are instantaneous. Therefore, the model does not include an infectious state, as in Figure 2.2.

Transition from class S to class L occurs with force of infection $\Lambda^i(a)$ per year, for individuals of sex i and age a. Transition from class L to class B occurs with rate $\rho^i(a)+z\Lambda^i(a)$ per year, where $\rho^i(a)$ denotes the rate of reactivation, and $z \in (0,1)$ is the susceptibility to reinfection of latently infected individuals relative to the susceptibility to primary infection of uninfected persons. It is assumed that individuals can experience reinfection or reactivation only once. Figure 7.3 summarizes the model's transitions.

$$S^i(a) \xrightarrow[\text{primary infection}]{\Lambda^i(a)} L^i(a) \xrightarrow[\text{reactivation / reinfection}]{\rho^i(a)+z\Lambda^i(a)} B^i(a)$$

Figure 7.3.: Schematic of the model [29].

The classes and transition rates are sex- and age-specific. For sexes $i \in \{♀, ♂\}$, the age-dependent differential equations are given by

$$\begin{aligned}
\frac{dS^i(a)}{da} &= -\Lambda^i(a)S^i(a) \\
\frac{dL^i(a)}{da} &= \Lambda^i(a)S^i(a) - (\rho^i(a)+z\Lambda^i(a))L^i(a) \\
\frac{dB^i(a)}{da} &= (\rho^i(a)+z\Lambda^i(a))L^i(a).
\end{aligned} \tag{7.1}$$

$S^i(a)$, $L^i(a)$ and $B^i(a)$ denote the fractions of individuals of sex i and age a in class S, L and B respectively. Note that $S^i(a)+L^i(a)+B^i(a)=1$, for all i and a.

Equations (7.1) form a linear system of differential equations with variable coefficients. The solutions of the system can be obtained using integration by parts. If the coefficients would have been constant, the system could have been analytically solved. Since the coefficients are variable, numerical methods are required in order to solve the system.

By considering the data, it is assumed that at age 0, no individuals have experienced reinfection or reactivation, and a non-zero fraction is latently infected. There is no distinction between females and males at this age. This yields the initial conditions $S^i(0) = S_0$, $L^i(0) = 1 - S_0$ and $B^i(0) = 0$, for $i \in \{\female, \male\}$.

The solutions of the differential equations are

$$\begin{aligned} S^i(a) &= S_0 e^{-\int_0^a \Lambda^i(t)dt} \\ L^i(a) &= (1-S_0)e^{-\int_0^a \rho^i(t)+z\Lambda^i(t)dt} + S_0 \int_0^a \Lambda^i(\tau) e^{-\left(\int_0^\tau \Lambda^i(t)dt + \int_\tau^a \rho^i(t)+z\Lambda^i(t)dt\right)} d\tau \\ B^i(a) &= 1 - S^i(a) - L^i(a). \end{aligned} \tag{7.2}$$

The solutions have a clear interpretation. The fraction of susceptibles of sex i and age a, $S^i(a)$, is given by the fraction of susceptibles of age 0, times the probability that no infection occurs until at least age a. The fraction of latently infected individuals of sex i and age a, $L^i(a)$, is described by the fraction of latently infected individuals of age 0 times the probability that no reinfection or reactivation takes place until at least age a, plus the fraction of susceptibles of age 0 times the probability that primary infection has taken place somewhere in the age interval $(0, a)$, and that in that interval, no reinfection or reactivation has occurred.

The force of infection $\Lambda^i(a)$ is the probability per year for a susceptible of sex i and age a to become infected. In this case, individuals in class L are also susceptible, since they can be reinfected, or the latent virus can reactivate. For simplicity, it is assumed that at most one reinfection or reactivation per individual can take place. Let M be the maximum age of the individuals. Note that $\Lambda^\ell(\alpha)S^\ell(\alpha)$ is the probability of primary infection, and $(\rho^\ell(\alpha) + z\Lambda^\ell(\alpha))L^\ell(\alpha)$ is the probability of reinfection and reactivation, for individuals of sex $\ell \in \{\female, \male\}$ and age $\alpha \in [0, M]$. Per day, individuals of sex i and age a have $c^{i\ell}(a, \alpha)$ contacts with an individual of sex ℓ and age α. Define infectiousness of CMV by parameters β_1 and β_2. These parameters denote the probability of transmission of primary infection, and reinfection or reactivation respectively, per contact with an infectious individual. Then the term $\beta_1\Lambda^\ell(\alpha)S^\ell(\alpha)$ denotes the infectious output that is generated by primary infection, and the term $\beta_2(\rho^\ell(\alpha) + z\Lambda^\ell(\alpha))L^\ell(\alpha)$ denotes the infectious output that is generated by reinfection and reactivation. Hence, the sex- and age-specific force of infection is given by

$$\Lambda^i(a) = \sum_{\ell \in \{\female, \male\}} \int_0^M c^{i\ell}(a, \alpha) \left[\beta_1 \Lambda^\ell(\alpha) S^\ell(\alpha) + \beta_2(\rho^\ell(\alpha) + z\Lambda^\ell(\alpha))L^\ell(\alpha)\right] d\alpha. \tag{7.3}$$

In order to calculate the basic reproduction number, the kernel of the next-generation operator must be defined. The kernel describes the expected number of newly infected individuals

with sex i and age a, caused by one individual of sex ℓ that became infected at age α, during its period of infectiousness, in a susceptible population. Since all individuals in the population are defined to be susceptible, except for one infected individual, reinfection of the infected individual will not occur. Hence, the susceptible individuals become infected with probability β_1 through contact with the primary infected individual, and with probability β_2 through contact with the individual once it has experienced reactivation. Hence, the kernel of the next-generation operator is given by

$$k^{i\ell}(a,\alpha) = \beta_1 c^{i\ell}(a,\alpha) + \beta_2 \int_\alpha^M c^{i\ell}(a,\tau)\rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}d\tau. \tag{7.4}$$

Note that this kernel does not take perinatal transmission into account.

7.2.1. Discretization of the model's equations

The solutions (7.2) can be inserted in Equation (7.3). This yields equations for the force of infection that cannot be solved analytically. In order to numerically estimate the parameters of the model with the available data, the equations are discretized. It is assumed that the transition and contact rates are constant in n predefined age-intervals of size Δa. The age classes are defined by the vector $\mathsf{a} = (0, \Delta a, 2\Delta a, \ldots, n\Delta a)$, where $n = M/\Delta a$, such that the jth age class contains all individuals with age in the interval $[\mathsf{a}_{[j]}, \mathsf{a}_{[j+1]})$. The continuous rates $\Lambda^i(a)$ and $\rho^i(a)$ with $a \in [\mathsf{a}_{[j]}, \mathsf{a}_{[j+1]})$ are replaced by Λ^i_j and ρ^i_j, where $i \in \{\mathrm{♀}, \mathrm{♂}\}$ still denotes sex and $j = 0, \ldots, n-1$ denotes the age class.

Consider the solutions (7.2). These equations were discretized by Chris van Dorp. Since Λ^i_j is piecewise constant on the age interval $[\mathsf{a}_{[j]}, \mathsf{a}_{[j+1]})$, for $j = 0, \ldots, n-1$, integration over the interval $(0, a)$ can be replaced by the sum of the integrations over the intervals $[\mathsf{a}_{[j]}, \mathsf{a}_{[j+1]})$, which have length Δa. Define the index of the age interval of an individual with age a by $j_a \coloneqq \lfloor a/\Delta a \rfloor$. Note that age a is $a - j_a\Delta a$ years removed from the left boundary of the j_ath age interval. Therefore, the discretization of $S^i(a)$ is given by

$$\begin{aligned} \bar{S}^i(a) &= S_0 e^{-\sum_{j=0}^{j_a-1}\int_{j\Delta a}^{(j+1)\Delta a}\Lambda^i_j dt - \int_{j_a\Delta a}^{a}\Lambda^i_{j_a}dt} \\ &= S_0 e^{-\Delta a\sum_{j=0}^{j_a-1}\Lambda^i_j - (a-j_a\Delta a)\Lambda^i_{j_a}} \\ &= S^i_{j_a} e^{-(a-j_a\Delta a)\Lambda^i_{j_a}}, \end{aligned} \tag{7.5}$$

where $S^i_j \coloneqq S_0 e^{-\Delta a\sum_{l=0}^{j-1}\Lambda^i_l}$.

For the discretization of $L^i(a)$, define

$$X^i(a) \coloneqq (1 - S_0)e^{-\int_0^a \pi^i(t)dt},$$

where $\pi^i(a) \coloneqq \rho^i(a) + z\Lambda^i(a)$. Also define

$$Y^i(a) \coloneqq \int_0^a \Lambda^i(\tau)\frac{S^i(\tau)}{X^i(\tau)}d\tau,$$

such that

$$L^i(a) = X^i(a) + X^i(a)Y^i(a). \tag{7.6}$$

The discretization of $X^i(a)$ is similar to the discretization of $S^i(a)$,

$$\bar{X}^i(a) = X^i_{j_a} e^{-(a-j_a\Delta a)\pi^i_{j_a}}, \tag{7.7}$$

where $X^i_j := (1-S_0)e^{-\Delta a \sum_{l=0}^{j-1}\pi^i_l}$.

The discretization of $Y^i(a)$ is given by

$$\bar{Y}^i(a) = \sum_{j=0}^{j_a-1} \Lambda^i_j \frac{S^i_j}{X^i_j} \int_{j\Delta a}^{(j+1)\Delta a} e^{-(\tau-j\Delta a)(\Lambda^i_j-\pi^i_j)} d\tau + \Lambda^i_{j_a} \frac{S^i_{j_a}}{X^i_{j_a}} \int_{j_a\Delta a}^{a} e^{-(\tau-j_a\Delta a)(\Lambda^i_{j_a}-\pi^i_{j_a})} d\tau \tag{7.8}$$

$$= \sum_{j=0}^{j_a-1} \frac{\Lambda^i_j}{\Lambda^i_j-\pi^i_j}\frac{S^i_j}{X^i_j}\left(1-e^{-\Delta a(\Lambda^i_j-\pi^i_j)}\right) + \frac{\Lambda^i_{j_a}}{\Lambda^i_{j_a}-\pi^i_{j_a}}\frac{S^i_{j_a}}{X^i_{j_a}}\left(1-e^{-(a-j_a\Delta a)(\Lambda^i_{j_a}-\pi^i_{j_a})}\right).$$

Equations (7.6), (7.7) and (7.8) lead to the discretization of $L^i(a)$. The discretization of $B^i(a)$ can be obtained by

$$\bar{B}^i(a) = 1 - \bar{S}^i(a) - \bar{L}^i(a).$$

For the discretization of the force of infection $\Lambda^i(a)$, the contact intensities $c^{i\ell}(a,\alpha)$ are replaced with contact intensities $c^{i\ell}[j_a, j_\alpha]$, where $j_a = \lfloor a/\Delta a \rfloor$ is the index of the age interval of an individual with age a. These discrete contact intensities define contacts between the n age classes of size Δa. The construction of the contact intensities is explained in Section 7.2.3. The discretization of the force of infection is given by

$$\Lambda^i_j = \sum_{\ell\in\{\female,\male\}} \sum_{k=0}^{n-1} c^{i\ell}[j,k]\left(\beta_1\Lambda^\ell_k S^\ell(k\Delta a) + \beta_2\pi^\ell_k L^\ell(k\Delta a)\right), \tag{7.9}$$

for $j = 0,\ldots,n-1$. Note that values of S and L at the left boundaries of the age intervals are used in the computation of the forces of infection.

7.2.2. Discretization of the next-generation operator

In order to compute the basic reproduction number, the continuous next-generation operator is transformed into a next-generation matrix. Recall that the kernel of the next-generation operator for the CMV model is given by Equation (7.4). The next-generation operator does not take perinatal transmission into account.

It follows by Equations (6.1) and (7.4) that the next-generation operator is given by

$$(K\phi)^i(a) = \sum_{\ell\in\{\female,\male\}} \int_0^\infty \left(\beta_1 c^{i\ell}(a,\alpha) + \beta_2 \int_\alpha^M c^{i\ell}(a,\tau)\rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv} d\tau\right)\phi^\ell(\alpha)d\alpha. \tag{7.10}$$

For notational convenience, define $I_j \coloneqq [\mathsf{a}_{[j]}, \mathsf{a}_{[j+1]})$. In order to discretize the next-generation operator and construct a matrix, the contact intensity from the interval I_k to the interval I_l is written as $c^{i\ell}(a,\alpha) = c^{i\ell}[k,l]$, for $a \in I_k$ and $\alpha \in I_l$. This can be notated by

$$c^{i\ell}(a,\alpha) = \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} c^{i\ell}[k,l]\mathbb{1}_{I_k}(a)\mathbb{1}_{I_l}(\alpha),$$

where $\mathbb{1}_I$ denotes the indicator function, which is defined by

$$\mathbb{1}_I(a) = \begin{cases} 1, & \text{if } a \in I \\ 0, & \text{otherwise.} \end{cases}$$

Hence, one can write

$$c^{i\ell}(a,\alpha) = \sum_{k=0}^{n-1} f_k(a)g_k(\alpha), \tag{7.11}$$

where $f_k(a) = \mathbb{1}_{I_k}(a)$ and $g_k(\alpha) = \sum_{l=0}^{n-1} c^{i\ell}[k,l]\mathbb{1}_{I_l}(\alpha)$. For notational convenience, the sex indices are ignored in the next part.

Substitution of Equation (7.11) in Equation (7.10) gives

$$(K\phi)(a) = \sum_{k=0}^{n-1}\left[\int_0^\infty \left(\beta_1 g_k(\alpha) + \beta_2 \int_\alpha^M g_k(\tau)\rho(\tau)e^{-\int_\alpha^\tau \rho(v)dv}d\tau\right)\phi(\alpha)d\alpha\right] f_k(a).$$

Note that integration over α results in a constant vector, so the range of the next-generation operator K is spanned by $\{f_k\}_{k=0}^{n-1}$. Since the next-generation distribution of the infected population is formed by applying the next-generation operator to the current distribution,

$$\phi^{n+1} = K\phi^n,$$

ϕ is in the span of $\{f_k\}_{k=0}^{n-1}$. Therefore, one can write $\phi = \sum_{j=0}^{n-1} b_j f_j$, for some constant vector b. Hence

$$K\phi = \sum_{k=0}^{n-1} d_k f_k,$$

where the vector d is given by

$$d_k = \sum_{j=0}^{n-1} b_j \int_0^\infty \left(\beta_1 g_k(\alpha) + \beta_2 \int_\alpha^M g_k(\tau)\rho(\tau)e^{-\int_\alpha^\tau \rho(v)dv}d\tau\right) f_j(\alpha)d\alpha.$$

It follows that the vector d is formed by a matrix multiplication with vector b. Let M be the next-generation matrix corresponding to the next-generation operator K. The next-generation matrix M is given by

$$M[k,j] = \int_0^\infty \left(\beta_1 g_k(\alpha) + \beta_2 \int_\alpha^M g_k(\tau)\rho(\tau)e^{-\int_\alpha^\tau \rho(v)dv}d\tau\right) f_j(\alpha)d\alpha. \tag{7.12}$$

Let $j_a = \lfloor a/\Delta a \rfloor$ denote the index of the age interval of an individual with age a. Substitution of $f_{j_\alpha}(\alpha) = \mathbb{1}_{I_{j_\alpha}}(\alpha)$ and $g_{j_a}(\alpha) = \sum_{j=0}^{n-1} c^{i\ell}[j_a, j]\mathbb{1}_{I_j}(\alpha)$ in Equation (7.12) results in

$$\begin{aligned}
M^{i\ell}[j_a, j_\alpha] &= \int_0^\infty \beta_1 \sum_{j=0}^{n-1} c^{i\ell}[j_a, j]\mathbb{1}_{I_j}(\alpha)\mathbb{1}_{I_{j_\alpha}}(\alpha)d\alpha \\
&\quad + \int_0^\infty \left(\beta_2 \int_\alpha^M \sum_{j=0}^{n-1} c^{i\ell}[j_a, j]\mathbb{1}_{I_j}(\tau)\rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}d\tau \right) \mathbb{1}_{I_{j_\alpha}}(\alpha)d\alpha \\
&= \int_0^\infty \beta_1 \; c^{i\ell}[j_a, j_\alpha] \; \mathbb{1}_{I_{j_\alpha}}(\alpha)d\alpha \\
&\quad + \int_0^\infty \left(\beta_2 \sum_{j=0}^{n-1} c^{i\ell}[j_a, j] \int_\alpha^M \rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}\mathbb{1}_{I_j}(\tau)d\tau \right) \mathbb{1}_{I_{j_\alpha}}(\alpha)d\alpha \\
&= \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_1 \; c^{i\ell}[j_a, j_\alpha] \; d\alpha \\
&\quad + \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \left(\beta_2 \sum_{j=j_\alpha}^{n-1} c^{i\ell}[j_a, j] \int_{j\Delta a}^{(j+1)\Delta a} \rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}d\tau \right) d\alpha \\
&\quad - \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \left(\beta_2 \; c^{i\ell}[j_a, j_\alpha] \int_{j_\alpha \Delta a}^{\alpha} \rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}d\tau \right) d\alpha.
\end{aligned}$$

Define

$$Z^\ell(\alpha) \coloneqq e^{-\int_0^\alpha \rho^\ell(v)dv}.$$

The discretization of $Z^\ell(\alpha)$ is similar to the discretization of $S^i(a)$, as is described by Equations (7.5), and is given by

$$\bar{Z}^\ell(\alpha) = e^{-\Delta a \sum_{l=0}^{j_\alpha - 1} \rho_l^\ell - (\alpha - j_\alpha \Delta a)\rho_{j_\alpha}^\ell}. \tag{7.13}$$

Also define

$$W_j^\ell(\alpha) \coloneqq \int_{j\Delta a}^{(j+1)\Delta a} \rho^\ell(\tau)e^{-\int_\alpha^\tau \rho^\ell(v)dv}d\tau.$$

On the interval I_j, $\rho^\ell(\tau)$ is constant. It follows that the discretization of $W_j^\ell(\alpha)$ is given by

$$\bar{W}_j^\ell(\alpha) = \frac{1}{\bar{Z}^\ell(\alpha)} \int_{j\Delta a}^{(j+1)\Delta a} \rho_j^\ell e^{-\Delta a \sum_{l=0}^{j-1} \rho_l^\ell - (\tau - j\Delta a)\rho_j^\ell}d\tau.$$

Substitution of this equation in $M^{i\ell}[j_a, j_\alpha]$ gives

$$\begin{aligned}
M^{i\ell}[j_a, j_\alpha] = &\int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_1 \; c^{i\ell}[j_a, j_\alpha] \; d\alpha \\
&+ \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \left(\beta_2 \sum_{j=j_\alpha}^{n-1} \frac{c^{i\ell}[j_a, j]}{\bar{Z}^\ell(\alpha)} \int_{j\Delta a}^{(j+1)\Delta a} \rho_j^\ell e^{-\Delta a \sum_{l=0}^{j-1} \rho_l^\ell - (\tau - j\Delta a)\rho_j^\ell} d\tau \right) d\alpha \\
&- \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \left(\beta_2 \; \frac{c^{i\ell}[j_a, j_\alpha]}{\bar{Z}^\ell(\alpha)} \int_{j_\alpha \Delta a}^{\alpha} \rho_{j_\alpha}^\ell e^{-\Delta a \sum_{l=0}^{j_\alpha - 1} \rho_l^\ell - (\tau - j_\alpha \Delta a)\rho_{j_\alpha}^\ell} d\tau \right) d\alpha.
\end{aligned}$$

Integration over τ results in

$$\begin{aligned}
M^{i\ell}[j_a, j_\alpha] = &\int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_1 \; c^{i\ell}[j_a, j_\alpha] \; d\alpha \\
&+ \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_2 \sum_{j=j_\alpha}^{n-1} \frac{c^{i\ell}[j_a, j]}{\bar{Z}^\ell(\alpha)} e^{-\Delta a \sum_{l=0}^{j-1} \rho_l^\ell} \left(1 - e^{-\Delta a \rho_j^\ell}\right) d\alpha \\
&- \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_2 \; \frac{c^{i\ell}[j_a, j_\alpha]}{\bar{Z}^\ell(\alpha)} e^{-\Delta a \sum_{l=0}^{j_\alpha - 1} \rho_l^\ell} \left(1 - e^{-(\alpha - j_\alpha \Delta a)\rho_{j_\alpha}^\ell}\right) d\alpha.
\end{aligned}$$

When $\bar{Z}^\ell(\alpha)$, given by Equation (7.13), is substituted, this becomes

$$\begin{aligned}
M^{i\ell}[j_a, j_\alpha] = &\int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_1 \; c^{i\ell}[j_a, j_\alpha] \; d\alpha \\
&+ \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_2 \sum_{j=j_\alpha}^{n-1} c^{i\ell}[j_a, j] e^{-\Delta a \sum_{l=j_\alpha}^{j-1} \rho_l^\ell} e^{(\alpha - j_\alpha \Delta a)\rho_{j_\alpha}^\ell} \left(1 - e^{-\Delta a \rho_j^\ell}\right) d\alpha \\
&- \int_{j_\alpha \Delta a}^{(j_\alpha+1)\Delta a} \beta_2 \; c^{i\ell}[j_a, j_\alpha] \left(e^{(\alpha - j_\alpha \Delta a)\rho_{j_\alpha}^\ell} - 1\right) d\alpha.
\end{aligned}$$

Finally, integration over α results in an equation for the entries of the next-generation matrix,

$$\begin{aligned}
M^{i\ell}[j_a, j_\alpha] = &\beta_1 c^{i\ell}[j_a, j_\alpha] \Delta a \\
&+ \beta_2 \sum_{j=j_\alpha}^{n-1} c^{i\ell}[j_a, j] e^{-\Delta a \sum_{l=j_\alpha}^{j-1} \rho_l^\ell} \left(1 - e^{-\Delta a \rho_j^\ell}\right) \frac{1}{\rho_{j_\alpha}^\ell} \left(e^{\Delta a \rho_{j_\alpha}^\ell} - 1\right) \\
&- \beta_2 c^{i\ell}[j_a, j_\alpha] \left(\frac{1}{\rho_{j_\alpha}^\ell} \left(e^{\Delta a \rho_{j_\alpha}^\ell} - 1\right) - \Delta a \right).
\end{aligned}$$

Rewriting this equation as follows

$$
\begin{aligned}
M^{i\ell}[j_a, j_\alpha] &= \beta_1 c^{i\ell}[j_a, j_\alpha]\Delta a \\
&\quad + \beta_2 c^{i\ell}[j_a, j_\alpha]\left(\Delta a - \frac{1}{\rho^\ell_{j_\alpha}}\left(1 - e^{-\Delta a \rho^\ell_{j_\alpha}}\right)\right) \\
&\quad + \beta_2 \sum_{j=j_\alpha+1}^{n-1} c^{i\ell}[j_a, j] e^{-\Delta a \sum_{l=j_\alpha}^{j-1} \rho^\ell_l}\left(1 - e^{-\Delta a \rho^\ell_j}\right)\frac{1}{\rho^\ell_{j_\alpha}}\left(e^{\Delta a \rho^\ell_{j_\alpha}} - 1\right),
\end{aligned}
\tag{7.14}
$$

gives a clearer interpretation of the next-generation matrix. The first term describes the expected number of individuals in the susceptible population that the infected individual infects while this individual is infectious after primary infection. The second and third term describe the expected number of individuals that the infected individual infects while being infectious after reactivation. In the second term, the individual experiences reactivation in age interval j_α, which is the same age interval in which the individual became primary infected. The third term describes the expected number of individuals that are infected by the individual if the individual experiences reactivation in one of the subsequent age intervals.

7.2.3. Contact rates

Contact rates are assumed to be symmetric. This means that the total number of contacts from individuals in group A to individuals in group B is equal to the total number of contacts from individuals in group B to individuals in group A. Contact intensities are not symmetric, for these depend on the population size of the group that makes the contact. Let $\tilde{c}^{i\ell}[j,k]$ denote the contact rate and $c^{i\ell}[j,k]$ the contact intensity per day, for contacts from individuals of sex i and age class j to individuals of sex ℓ and age class k. Let w^i_j be the total number of individuals in the population with sex i and age j. Then

$$c^{i\ell}[j,k] = \tilde{c}^{i\ell}[j,k]\ w^i_j.$$

In other words, the sex- and age-specific contact intensities $c^{i\ell}[j,k]$ are defined by the product of the contact rate and the demographic composition of the population, as in Equation (6.3).

The contact intensities were estimated in Van de Kassteele et al. (2016) using Bayesian models. They have analysed contact data from the Netherlands that was gathered in 2006 and 2007. A total of 825 Dutch participants recorded 11,225 unique contacts during one day. Demographic statistics from the same period were retrieved from Statistics Netherlands (CBS).

In Van de Kassteele et al. (2016), a contact is defined by a conversation of at least one sentence between two individuals that are physically close, or by two individuals touching each other's skin. Contact data are available from individuals of age 0 to 80 years. Consider female participants in age class k in the study. They have reported $Y^{\female\male}_{kj}$ contacts with males

in age class j. Let $t_k^{\female}$ denote the total number of female participants in age class k. The contact intensity is estimated by the mean number of contacts with males in age class j that are reported by one female participant in age class k. This is given by

$$c^{\male\female}[j,k] = \frac{\mathbb{E}\left(Y_{kj}^{\female\male}\right)}{t_k^{\female}}.$$

Note that this is the male-to-female contact intensity, since it gives the number of contacts that one female in age class k receives from males in age class j, per day.

Figure 7.4 presents a colour plot of the estimated contact intensity matrix. It can be seen that individuals have most contact with individuals of their own sex and age class, and that women make more contact than men.

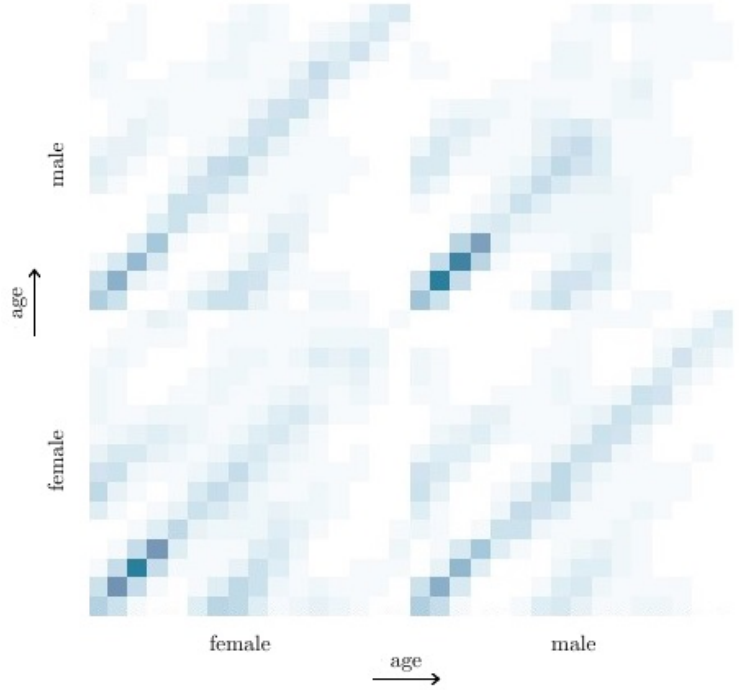


Figure 7.4.: Colour plot of the contact intensity matrix $c^{i\ell}[j,k]$ that was estimated in Van de Kassteele et al. (2016). The entries denote contact intensities from individuals of sex i and age class j (placed on the rows) to individuals of sex ℓ and age class k (placed on the columns). Contact intensities are shown between females and males from age class [0,5) up to age class [75,80). A dark entry means that the contact intensity between individuals of that age interval is high.

7.2.4. Log likelihood

The log likelihood of the data is given by the normal mixture distribution with parameters μ_S, μ_L, μ_B, σ_S, σ_L and σ_B, that is shown in Figure 7.2. If an individual of sex i and age a has a titer value between -2.91 U/ml and 3.41 U/ml, its contribution to the log likelihood is equal to the log of the sum of the probabilities for the individual with that sex and age to be in class S, L or B, times the probability to find its titer value T in that class. Let $\phi_{\mu,\sigma}$ denote the probability density function of the normal distribution with mean μ and standard deviation σ. The contribution to the log likelihood is given by

$$\log\left[\bar{S}^i(a)\ \phi_{\mu_S,\sigma_S}(T) + \bar{L}^i(a)\ \phi_{\mu_L,\sigma_L}(T) + \bar{B}^i(a)\ \phi_{\mu_B,\sigma_B}(T)\right]. \tag{7.15}$$

If an individual has a titer value below -2.91 U/ml, then this individual belongs to class S with probability 1. Hence, its contribution to the log likelihood is equal to the log of the probability to find an individual with this sex and age in class S. Its contribution is given by

$$\log\left[\bar{S}^i(a)\right]. \tag{7.16}$$

For individuals with titer values above 3.41 U/ml, right-censoring is applied. These titers can be 3.41 U/ml or higher, so the complementary cumulative distribution function at 3.41 is used, instead of the probability density function at the individual's titer value. Let $\Phi_{\mu,\sigma}$ denote the cumulative distribution function of the normal distribution with mean μ and standard deviation σ. The contribution to the log likelihood for these individuals is given by

$$\begin{aligned}\log\big[&\bar{S}^i(a)\ (1-\Phi_{\mu_S,\sigma_S}(3.41)) + \bar{L}^i(a)\ (1-\Phi_{\mu_L,\sigma_L}(3.41)) \\ &+ \bar{B}^i(a)\ (1-\Phi_{\mu_B,\sigma_B}(3.41))\big].\end{aligned} \tag{7.17}$$

The total log likelihood is equal to the sum of the contributions of all individuals in the data set.

7.3. Perinatal transmission

Perinatal transmission is thought to play a significant role in the transmission of CMV. During pregnancy, a latent CMV infection may reactivate and infect the unborn child, which can lead to serious disease. After birth, the mother may infect the child through breastfeeding. In order to estimate the probability of perinatal transmission, the transmission model constructed in Van Boven et al. (2017) is extended with a parameter for the probability of perinatal transmission.

In Van Boven et al. (2017) S_0, the fraction of susceptibles at age six months, was estimated separately from the other parameters and given as input to the transmission model. Using birth statistics, S_0 can be estimated within the transmission model. In this way, a parameter can be defined to estimate the probability of perinatal transmission of CMV using Stan.

The transmission model uses data from the PIENTER2 study in 2006 and 2007, therefore birth data from this same period are used. The numbers of children born, distributed over

the ages of the mothers, are collected. By dividing these numbers over the total number of births in 2006 and 2007, the relative contributions to the birth rate shown in Table 7.1 are obtained. Birth statistics were retrieved from Statistics Netherlands (CBS) [5].

age mother	<20	20-25	25-30	30-35	35-40	40-45	45>
contribution to birth rate	0.0101	0.0893	0.2680	0.3788	0.2173	0.0351	0.0014

Table 7.1.: Contributions of women of fertile age to the birth rate.

Sixteen age classes of size five years are defined. It is assumed that women give birth at ages 15 to 50. In the model, the contributions to the birth rate for ages <20 and 45> are assigned to the age classes 15-20 and 45-50 respectively. Let F be a vector containing the relative contributions to the birth rate for each age class. The entries F_3 to F_9 are given by the table, the remaining entries are equal to 0. This vector is provided as data to the Stan model. The fraction susceptibles at age six months, denoted by S_0, is defined to be a parameter in Stan. A parameter ν is introduced to indicate the probability of perinatal transmission.

As before, S_0 is used to compute the fraction of susceptibles of sex i and age a,

$$\bar{S}^i(a) = S_0 e^{-\Delta a \sum_{l=0}^{j_a-1} \Lambda_l^i - (a - j_a \Delta a)\Lambda_{j_a}^i},$$

where $j_a = \lfloor a/\Delta a \rfloor$ denotes the index of the age interval of an individual with age a.

Let $\hat{S}_0$ be a transformed parameter in Stan. This parameter will be used as the mean of the prior for S_0. The contribution of the female age classes to the birth rate is used to compute the fraction of children that will not be infected by their mothers during pregnancy and the first six months of their lives. It is assumed that the children of susceptible women are susceptible at the age of six months, and that the children of infected women are latently infected at the age of six months with perinatal transmission probability ν. It follows that S_0 can be estimated by

$$\hat{S}_0 = \sum_{k=0}^{n-1} F_k \left[S^{\text{♀}}(k\Delta a) + (1-\nu)\left(1 - S^{\text{♀}}(k\Delta a)\right)\right]. \tag{7.18}$$

In this case $n = 16$. Note that values of S at the left boundaries of the age intervals are used in the computation of $\hat{S}_0$.

At last, S_0 is given the prior $\mathcal{N}(\hat{S}_0, 10^{-4})$. As a result, S_0 will resemble $\hat{S}_0$. The perinatal transmission parameter ν is given a non-informative $\mathcal{U}[0,1]$ prior.

7.4. Parameter estimation using Stan

The transmission model is implemented in Stan in order to obtain Bayesian estimates for the parameters using the HMC method. The model takes 5,179 data points containing

transformed titer values, sexes and ages as input. Age intervals of size $\Delta a = 5$ years are used. The maximum age in the data set is $M = 79$ years, so this yields sixteen age classes $[0, 5), [5, 10), \ldots, [75, 80)$. The contact intensity matrix is constructed using the intensities estimated in Van de Kassteele et al. (2016), such that it defines contacts between these age classes. The means and standard deviations that define the mixture distribution are also input to the model. The data and all other constant values that the model takes as input, are defined in the data block of the Stan model.

In total, the model contains forty-three parameters, namely β_1, β_2, z, S_0, ν, thirty-two forces of infection Λ_j^i (defined on each age interval), and six reactivation rates ρ_j^i. Based on preliminary analyses, the reactivation rates ρ_j^i are modelled by constant rates on the age intervals $[0, 20)$, $[20, 50)$ and $[50, 80)$. This yields six parameters $p^i_{[0,20)}$, $p^i_{[20,50)}$ and $p^i_{[50,80)}$, with $i \in \{\text{♀}, \text{♂}\}$. The parameters are defined in the parameters block of the Stan model. The parameters block also contains two hyperparameters that are used to estimate the reactivation rates ρ_j^i. The parameters are all restricted to be non-negative.

The transformed parameters block contains quantities that are not data or parameters. For instance, the equations that describe the fractions S, L and B are defined in this compartment. The transformed data block contains vectors that are used to make the program more efficient. Equations are given in vectorised form, since this decreases the run time of Stan.

In the model block, the priors on the parameters are defined. The parameters β_1 and β_2 are given a $\mathcal{N}(0, 0.1)$ prior. For z and ν, $\mathcal{U}(0, 1)$ priors are chosen. The reactivation parameters ρ_j^i all have the same hyperprior $\mathcal{N}(\mu_\rho, \sigma_\rho)$, where hyperparameters μ_ρ and σ_ρ have $\mathcal{N}(0, 0.1)$ priors. In addition to the prior distributions, the model block contains the log likelihood, as defined by Equations (7.15), (7.16) and (7.17).

Since the forces of infection Λ_j^i and the fraction of susceptibles at age six months S_0 are defined as parameters, they cannot be defined by an equation in Stan. Therefore, transformed parameters $\hat{\Lambda}_j^i$ and $\hat{S}_0$ are introduced in the transformed parameters block. The forces of infection Λ_j^i are given $\mathcal{N}(\hat{\Lambda}_j^i, 1/p)$ priors, and S_0 is given a $\mathcal{N}(\hat{S}_0, 1/p)$ prior, where p denotes a penalty. In this way, the forces of infection are estimated according to Equation (7.9), and S_0 is estimated according to Equation (7.18). By increasing p, the run time increases, but the estimation becomes more accurate. The penalty p is input to the Stan model.

In the generated quantities block, the log likelihood is again defined, since the quantities that are defined in the model block are not saved. The contribution of each individual to the log likelihood is saved, such that these values can be used to calculate information criteria, in order to compare models.

8. Results

The results were obtained using the Bayesian inference program Stan. The Stan code for the CMV transmission model with reinfection, reactivation and estimation of perinatal transmission can be found in Appendix A.

The Stan model was run for 4,000 iterations, distributed over 4 chains. By default, half of the samples were generated during the warm-up. The target acceptance rate was set equal to 0.99, to prevent generation of divergent samples. This results in an output of 2,000 parameter samples. It took 30 minutes to generate the samples.

8.1. Model comparison

In order to find out which transmission routes are required to explain the data, the fits of the variants of the CMV transmission model with and without reinfection and reactivation are compared. For each model, the Watanabe-Akaike information criterion (WAIC) is computed. The WAIC is based on the series expansion of leave-one-out cross-validation (LOO), and closely approximates Bayesian cross-validation [30]. Models with lower WAIC fit the data better. The R loo package is used to compute the WAIC for the transmission models [31].

The transmission model that contains reinfection and reactivation, and estimates perinatal transmission, is referred to as the full model. In the model without reinfection, the parameter z is set equal to 0. In the model without reactivation, the reactivation rates $\rho^i_{[0,20)}$, $\rho^i_{[20,50)}$ and $\rho^i_{[50,80)}$ for $i \in \{♀, ♂\}$, are set equal to 0.

model	n	WAIC	ΔWAIC
full	43	22,156.3	0.3
no reinfection	42	22,156.0	0
no reactivation	37	22,341.2	185.2

Table 8.1.: The WAIC for three variants of the transmission model. The number of parameters is denoted by n, and the difference in WAIC with the best fitting model is given by ΔWAIC.

Table 8.1 presents the WAIC for each model. It can be seen that the model that does not contain reactivation has the highest WAIC. This suggests that reactivation is required to explain the data. The model that does not contain reinfection has the lowest WAIC, suggesting that reinfection is not required to explain the data. However, the difference in WAIC between this model and the full model is small. These results are similar to the ones found in Van Boven et al. (2017).

Since the difference in WAIC between the full model and the model without reinfection is small, results of both models are considered.

8.2. Parameter estimates

Parameters are estimated by the medians of the obtained samples. The posterior medians and 95% credible intervals of the parameters of the full model and the model without reinfection are presented in Table 8.2. The estimates of the thirty-two forces of infection of the full model are shown in Figure 8.1. The forces of infection of the model without reinfection are similar, and are therefore not shown.

As can be seen in Table 8.2, the female reactivation rates $\rho_j^{\female}$ are higher than the male reactivation rates $\rho_j^{\male}$. The estimated reactivation rates are higher in the model without reinfection, but the other parameter estimates are similar.

Figure 8.1 shows that the force of infection is higher in females than in males in the first and last age classes. The force of infection decreases at the age of 15, but increases again at age 25. This can be explained by the increasing contact intensity with small children, as was shown by Figure 7.4. The low force of infection in males of older age can also be explained by their low contact intensity. Similar results were found in Van Boven et al. (2017).

For both models, the estimated probability of perinatal transmission ν equals 0.37. The estimated fraction of susceptibles at age six months S_0 is equal to 0.83. This means that 17% of all children are already infected at the age of six months.

Most credible intervals are acceptable, but the reinfection parameter z has an exceptionally large credible interval, and the posterior distribution of β_1 seems to be skewed. This is further analysed in Section 8.4.

	full model			model without reinfection		
	posterior median	95% CrI		posterior median	95% CrI	
β_1	0.0025	(0.00015,	0.011)	0.0025	(0.000032,	0.012)
β_2	0.067	(0.054,	0.074)	0.067	(0.053,	0.074)
z	0.29	(0.013,	0.81)	-	-	
$\rho^{\female}_{[0,20)}$	0.014	(0.0048,	0.021)	0.019	(0.013,	0.024)
$\rho^{\female}_{[20,50)}$	0.022	(0.015,	0.029)	0.025	(0.021,	0.031)
$\rho^{\female}_{[50,80)}$	0.025	(0.016,	0.035)	0.028	(0.020,	0.038)
$\rho^{\male}_{[0,20)}$	0.0057	(0.00024,	0.013)	0.0097	(0.0047,	0.015)
$\rho^{\male}_{[20,50)}$	0.011	(0.0041,	0.018)	0.015	(0.010,	0.019)
$\rho^{\male}_{[50,80)}$	0.011	(0.0037,	0.019)	0.014	(0.0077,	0.020)
ν	0.37	(0.31,	0.42)	0.37	(0.31,	0.43)
S_0	0.83	(0.81,	0.86)	0.83	(0.80,	0.86)

Table 8.2.: Posterior medians and 95% credible intervals of the parameters of the full model and the model without reinfection.

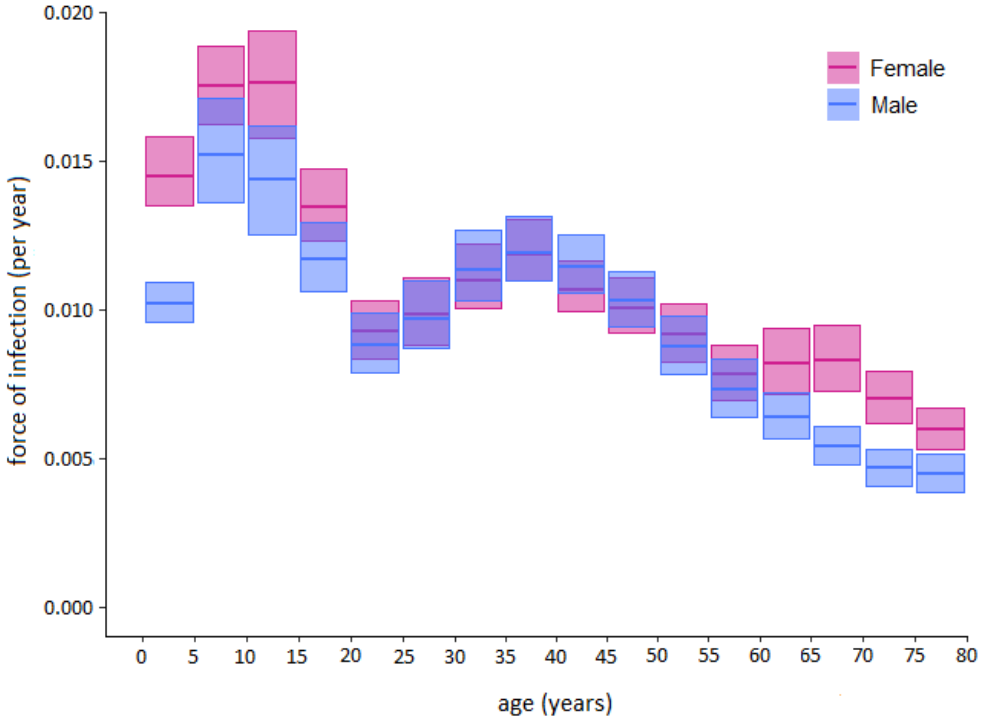


Figure 8.1.: Estimates of the female and male forces of infection in the full model. The thick lines denote posterior medians, and are surrounded by 95% credible intervals.

8.3. Estimation of CMV prevalence

Stan also returns fractions of susceptible individuals (S), latently infected individuals (L) and individuals that have experienced reinfection or reactivation (B). These fractions are computed for 80 ages. Figure 8.2 shows the estimated development of the CMV prevalence in females and males. The shown results are obtained from the full model. The estimated prevalences of the model without reinfection are similar and are therefore not shown.

It is clear that the CMV prevalence increases with age. At age 80 years, only 35% of the females and 39% of the males are still susceptible. It can also be seen that the fraction of females in class B increases faster than the fraction of males in class B. At age 80 years, 48% of the females have experienced reinfection or reactivation, against 32% of the males.

Note that the credible intervals for the fraction of individuals in class S are smaller than the intervals for the classes L and B. This is caused by the classification of the samples by the mixture distribution that is shown in Figure 7.2. Susceptible individuals and infected individuals are better distinguished than latently infected individuals and individuals that have experienced reinfection or reactivation. The credible intervals for class B are small for low ages and increase with age, for the fraction of individuals in class B at age six months is initially set equal to 0.

8.4. Convergence analysis

Table 8.2 indicates that there may be some problems concerning the convergence of the Hamiltonian Monte Carlo algorithm for parameter β_1 (the probability of transmission of primary infection) and parameter z (the susceptibility to reinfection of latently infected individuals relative to the susceptibility to primary infection of uninfected persons). For comparison, the posterior distributions of the parameter samples for β_1, β_2 (the probability of transmission of reinfection or reactivation), z and ν (the probability of perinatal transmission) are presented in Figure 8.3.

Most parameters have Gaussian-shaped posteriors, like the perinatal transmission probability ν, but the parameters β_1, β_2 and z deviate from this.

Parameter z was given a $\mathcal{U}(0,1)$ prior. It can be seen that the resulting posterior completely covers the interval $(0,1)$. This might mean that the data are not informative enough to estimate z.

The parameters β_1 and β_2, that denote the probability of transmission of primary infection, and reinfection or reactivation respectively, were given $\mathcal{N}(0,0.1)$ priors and were restricted to be larger than 0. Both posteriors are skewed. The skewness of the posterior of β_1 is probably due to the boundary at 0.

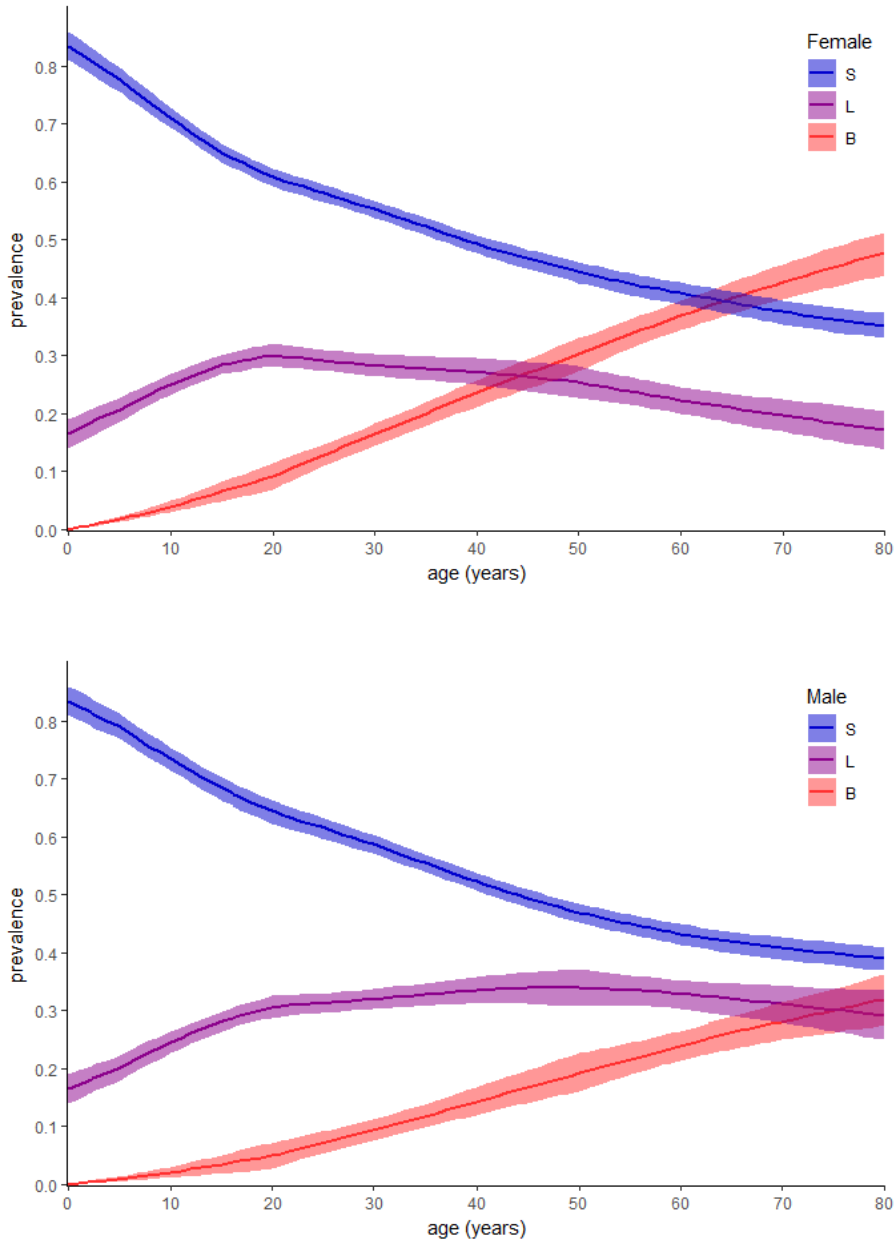


Figure 8.2.: Age-specific CMV prevalence in females and males, estimated by the full model. Posterior medians (lines) and 95% credible intervals of the fractions of susceptible individuals (S), latently infected individuals (L) and individuals that have experienced reinfection or reactivation (B) are shown.

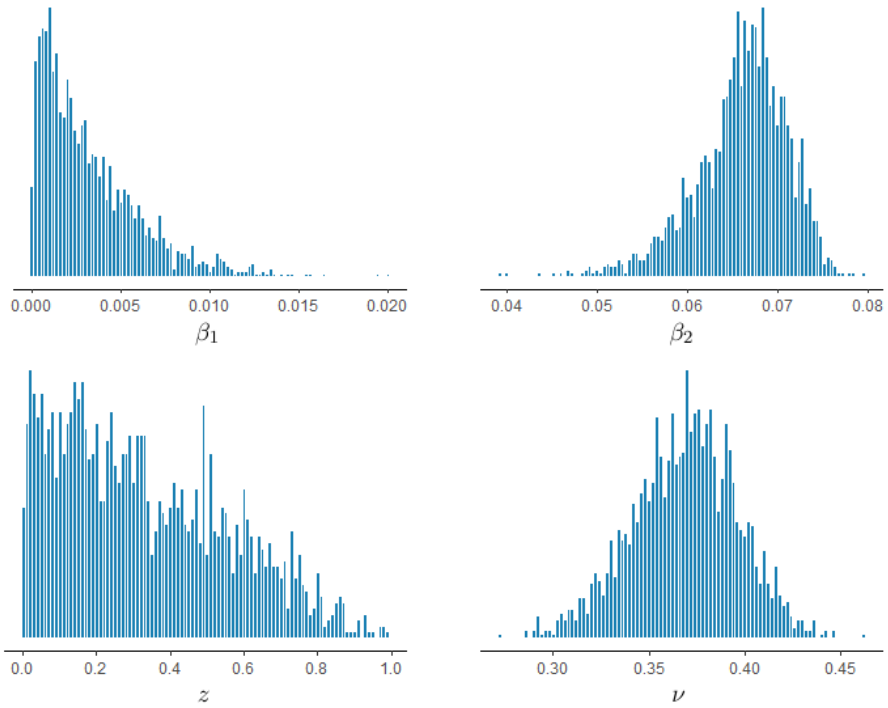


Figure 8.3.: Posteriors of the parameters β_1, β_2, z and ν of the full model, given as output by Stan.

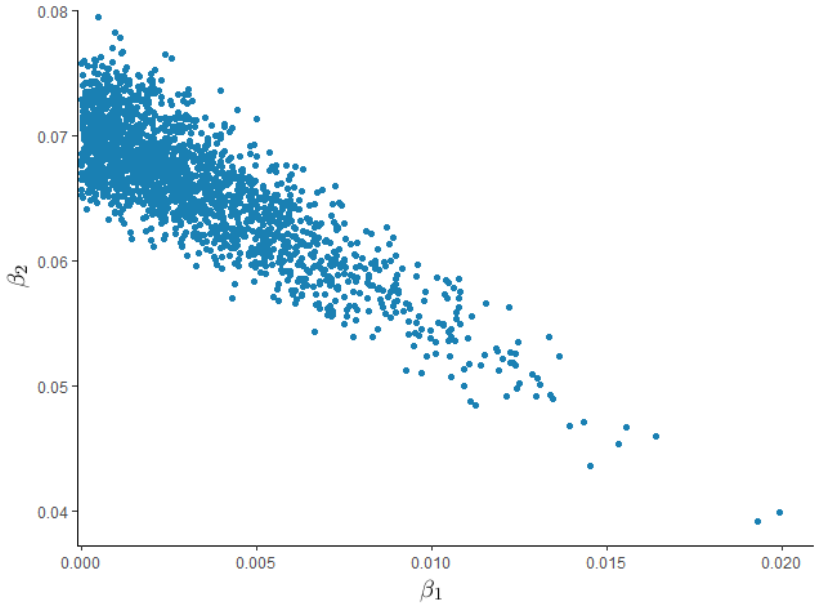


Figure 8.4.: Scatterplot of β_1 against β_2.

The skewness of β_2 could be a result of the skewness of β_1. Figure 8.4 illustrates the strong correlation between β_1 and β_2. Low estimates of β_1 are paired with high estimates of β_2, and vice versa.

To check whether the chains have converged to a stationary distribution, the potential scale reduction statistic $\hat{R}$ is considered. The $\hat{R}$ statistic measures the ratio of the variance of samples generated by each chain to the variance of all generated samples together. If all chains have converged to the stationary distribution, this ratio will be equal to 1. Stan returns $\hat{R}$ for each parameter [25].

For all parameters, $\hat{R}$ is close to 1, ranging from 1.01 to 1.06. The $\hat{R}$ statistic of β_1, β_2 and z equals 1.02, 1.03 and 1.03 respectively, indicating that the deviating shapes of the posteriors are not caused by a convergence problem.

8.5. The basic reproduction number

The basic reproduction number R_0 is calculated for the full model and the model without reinfection using the next-generation matrix that is given by Equation (7.14). For all parameter samples, the next-generation matrix is computed and the dominant eigenvalue is obtained. The R code can be found in Appendix B.

The estimated basic reproduction numbers are presented in Table 8.3. The basic reproduction numbers for both models are smaller than 1. This was not expected, since the CMV prevalence in the Dutch population is high. However, the perinatal transmission route has not been included in the construction of the next-generation matrix.

The basic reproduction number R_0 is higher for the model without reinfection. This is due to the higher estimates for the reactivation rates ρ_j^i (see Table 8.2).

R_0	posterior median	95% CrI
full model	0.91	(0.66, 1.07)
model without reinfection	0.99	(0.89,1.15)

Table 8.3.: The posterior medians and the 95% credible intervals of the basic reproduction number R_0 for the full model and the model without reinfection.

Figure 8.5 shows a colour plot of the next-generation matrix of the full model. The individuals that are infected in a certain age class are distributed over the columns. The entries give the expected number of individuals, distributed over the rows, that they infect. Dark entries mean that many infections occur. The upper halves of each of the four matrix blocks are darker than the lower halves. This is a result of the high contact intensity between the same age classes, and the reactivation events that occur. For if an individual is infected in a certain

age class, it will mainly infect other individuals in that age class. When the individual is in a higher age class, the CMV infection may reactivate, and the individual will mainly infect other individuals in this higher age class.

It can also be seen that men in the highest age classes cause the least infections, and that women of fertile age infect young children. This can be explained by the structure of the contact intensity matrix, presented by Figure 7.4.

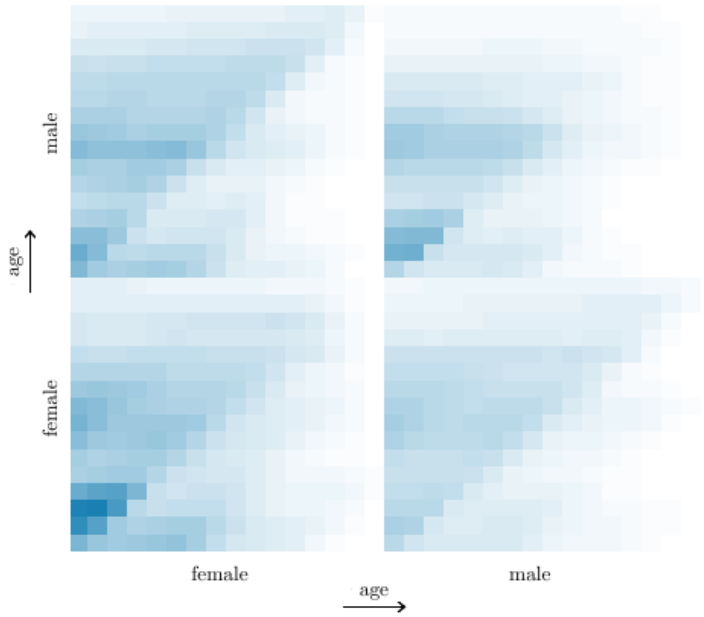


Figure 8.5.: Colour plot of the next-generation matrix of the full model. Females and males that are infected in one of the sixteen age classes are distributed over the columns. The colour of the entries indicates the expected number of individuals, distributed over the rows, that they infect. The darker the entry, the more individuals are expected to be infected.

9. Discussion

The constructed sex- and age-structured transmission model contributes to our understanding of the transmission dynamics of CMV. The model indicates that reactivation is essential to explain the serological data set. This implicates that most CMV infections may be caused by individuals who go through a reactivation episode.

It is also found that the CMV prevalence in females is higher than in males. For most ages, females have a higher force of infection, but the difference in prevalence between females and males is mainly due to the high reactivation rate in females. These results are similar to the results that were found in Van Boven et al. (2017). A possible explanation for the observed differences may be found in differences between female and male immune systems. It is known that the immune system weakens during pregnancy. However, reactivation does not seem to primarily occur in women of fertile age.

This study also provides an estimate for the perinatal transmission rate. The probability that a mother infects her child during pregnancy or in the six months after birth equals 0.37. As a result, 17% of all children are already infected at the age of six months.

The results were generated by the Bayesian inference program Stan, that is based on the Hamiltonian Monte Carlo method. The theory behind Hamiltonian Monte Carlo is complicated, but Stan itself is relatively easy to use. Stan generates samples from the complex and high-dimensional CMV transmission model incredibly fast. This enabled us to extend the model and to try things that the program JAGS, based on Markov Chain Monte Carlo, could not handle. I believe that within a few years, Stan will be the main program to use for generating parameter samples from posterior distributions.

Overall, Stan behaves well, and biologically plausible parameter samples are generated. However, some points deserve closer examination. One of the parameters that is not estimated well, is the susceptibility to reinfection relative to the susceptibility to primary infection, denoted by z. As can be seen in Figure 8.3, the estimated posterior of z is not Gaussian-shaped and completely covers the interval $(0, 1)$. This might indicate that the data are not informative enough to estimate z. This is not that surprising, since the model contains no less than forty-three parameters.

Moreover, the difference in WAIC between the full model and the model without reinfection is very small. It can not be concluded which model performs better. The estimation of parameter z should first be improved, before conclusions about the performances of the

different variants of the transmission model can be made. The model or method, or both, should be changed in order to obtain a valid estimate for z.

The posterior distributions of the parameters β_1 and β_2, the probabilities of transmission of primary infection, and reinfection or reactivation respectively, are skewed. The parameter β_1 was given a $\mathcal{N}(0, 0.1)$ prior, and was restricted to be non-negative. The skewness of the posterior of β_1 is probably caused by this restriction. Actually, it is not appropriate to use $\mathcal{N}(0, 0.1)$ as a prior for β_1 and add the non-negative restriction to it. It would be better to use a prior that has this boundary embedded, like a uniform or beta prior. Another option is to transform the parameter such that a normal prior can be used.

The skewness of β_2 may be a result of the skewness of β_1, since these parameters are strongly correlated. Low estimates of β_1 are paired with high estimates of β_2, and vice versa. There is no biological explanation for this. The correlation is probably due to a parameter identification problem. The question is whether one parameter for the probability for transmission of primary infection, and transmission of reinfection and reactivation would suffice. Future studies should investigate the relation between the parameters β_1, β_2 and z, and address the identification problem for β_1 and β_2, and the estimation problem for z.

Additional research is also required to compute a basic reproduction number that takes perinatal transmission into account. Currently, the estimated basic reproduction number R_0 is smaller than 1, while we expect it to be larger than 1, since CMV is clearly able to persist in the population.

An intuitive way to add perinatal transmission to the calculation of the basic reproduction number, is to compute the number of children that a woman in a certain age class is expected to give birth to, using data from the Statistics Netherlands database, and to multiply this number by the estimated perinatal transmission probability. This gives the expected number of children of age 0 to 6 months that a woman in a certain age class is expected to infect. I have added the resulting numbers to the first age class of the next-generation matrix. The resulting next-generation matrix is shown in Figure 9.1.

It is clear that including perinatal transmission in the next-generation matrix makes a difference. The entries that denote transmission from mother to child are darker than all other entries. Moreover, the basic reproduction numbers have increased, and the basic reproduction number of the model without reinfection is even larger than 1. The basic reproduction number for the full model and the model without reinfection equal 0.98 and 1.08 respectively. This substantiates the hypothesis that perinatal transmission is essential for CMV to be able to persist in the population.

However, this intuitive method is not precise, and should be improved. The expected number of children of age up to six months that a woman in a certain age class is expected to infect, is added to the first age class, which actually accounts for individuals in the age range of six months to five years. An additional row and column should be constructed for the age class up to six months.

Apart from these findings, the transmission model itself can be improved. One of the main assumptions is that individuals can experience reinfection and reactivation only once. In practice, these events can occur multiple times during the individual's life. My supervisor, Michiel van Boven, has constructed a transmission model that allows multiple reinfection and reactivation events to happen. Using JAGS, it was not possible to run this more complicated model, but Stan can manage it. It would be interesting to calculate the perinatal transmission rate and the basic reproduction number for this model.

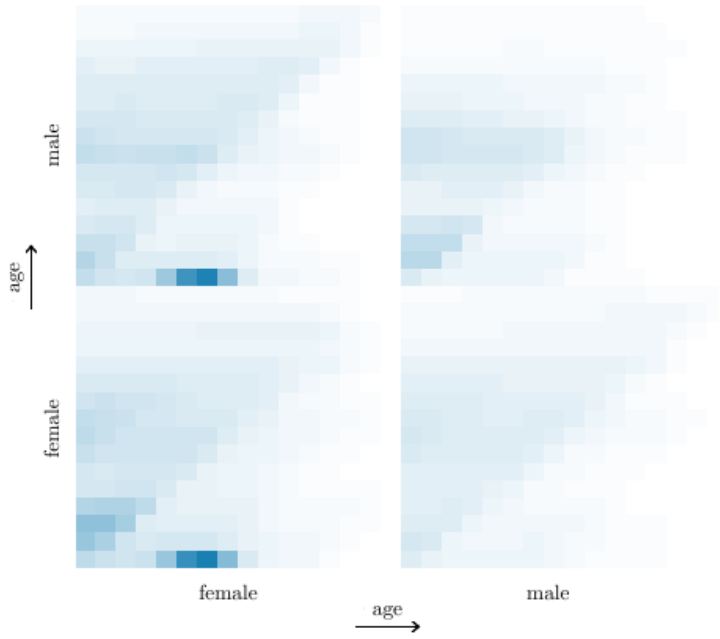


Figure 9.1.: Colour plot of the next-generation matrix of the full model, including perinatal transmission. Females and males that are infected in one of the sixteen age classes are distributed over the columns. The colour of the entries indicates the expected number of individuals, distributed over the rows, that they infect. The darker the entry, the more individuals are expected to be infected.

Another assumption that is made, is that transmission of CMV is proportional to social human contact patterns. The contact intensities that are used in the analyses are defined by conversations between individuals, or by individuals touching each other's skin. Although this assumption is often made, transmission of CMV does not completely correspond to these contacts, since transmission requires transfer of saliva or urine, and more complex contact patterns may be necessary. Unfortunately, such contact patterns are not available. The contact intensities could be altered by constants that give, for instance, more weight to contacts with young children, since the probability that transfer of saliva or urine occurs is higher for these contacts. However, it is questionable whether this would significantly change the results.

Unfortunately, Stan cannot handle all improvements and extensions that are made. For instance, it would be more accurate to estimate the means and standard deviations of the normal mixture distribution (see Figure 7.2) together with the other parameters in the transmission model, instead of estimating these parameters separately. I have modelled this in Stan, but Stan was not able to produce reliable estimates for the parameters. The estimated parameter values were highly dependent on the initial conditions and priors. Probably the data set is not informative enough to estimate an additional six parameters.

I also tried to distinguish between perinatal transmission from latently infected mothers, and mothers that thave experienced reinfection or reactivation, by introducing an additional parameter. However, this also resulted in estimates that were highly dependent on initial conditions and priors. Hence, this distinction is left out of the model.

This study contributes to our understanding of the transmission dynamics of CMV. The results suggest that reactivation and perinatal transmission are essential for CMV to persist in the population. Future research should first solve the estimation problems of parameters β_1, β_2 and z, and then focus on estimation of the perinatal transmission rate within the extended model in which multiple reinfection and reactivation events can occur, and calculation of the basic reproduction number that takes perinatal transmission into account. When a reliable basic reproduction number is generated, the effects of vaccination can be measured. Since most CMV infections seem to be caused by individuals who go through a reactivation episode, a vaccine that prevents reactivation may be most effective, instead of a vaccine that prevents primary infection. This also means that individuals of older age are the potential targets for vaccination, instead of young children. Prevention of reactivation in pregnant women is mainly preferable, since perinatal transmission seems to be common.

References

[1] Bédard M. 2006. Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234. *Stochastic Processes and their Applications.* **118(12)**:2198-2222.

[2] Betancourt M. 2017. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat.ME].*

[3] Boeckh M, Geballe AP. 2011. Cytomegalovirus: Pathogen, Paradigm, and Puzzle. *The Journal Clinical Investigation.* **121(5)**:1673-1680.

[4] Cannon MJ, Schmid DS, Hyde TB. 2010. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in Medical Virology.* **20**:202-213.

[5] Statistics Netherlands (CBS). 2016. CBS StatLine - Birth; key figures. Retrieved from http://statline.cbs.nl/Statweb/publication/?VW=T&DM=SLEN&PA=37422eng&D1=7-14&D2=56-57&HD=170328-1110&LA=EN&HDR=T&STB=G1

[6] Diekmann O, Heesterbeek H, Britton T. *Mathematical Tools for Understanding Infectious Disease Dynamics.* Princeton University Press; 2013.

[7] Diosi P, Kazanjian P. 2003. Transmission or Recurrence? A Historical Dilemma of Iatrogenic Infections due to Cytomegalovirus. *Journal of the History of Medicine and Allied Sciences.* **58(1)**:56-78.

[8] Dollard SC, Grosse SD, Ross DS. 2007. New estimates of the prevalence of neurological and sensory sequelae and mortality associated with congenital cytomegalovirus infection. *Reviews in Medical Virology.* **17(5)**:355-363.

[9] Frasca D, Blomberg BB. 2016. Aging, cytomegalovirus (CMV) and influenza vaccine responses. *Human Vaccines & Immunotherapeutics.* **12(3)**:682-690.

[10] Frobenius G. 1912. Über Matrizen aus nicht negativen Elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften.* 456-477.

[11] Gelman A, Lee D, Guo J. 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Education and Behavioral Statistics.* **40(5)**:530-543.

[12] Heesterbeek JAP, Dietz K. 1996. The concept of R_0 in epidemic theory. *Statistica Neerlandica.* **50(1)**:89-110.

[13] Hoffman MD, Gelman A. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research.* **15**:1351-1381.

[14] Klenerman P, Oxenius A. 2016. T cell responses to cytomegalovirus. *Nature Reviews Immunology.* **16(6)**:367-377.

[15] Korndewal MJ, Mollema L, Tcherniaeva I, Van der Klis F, Kroes ACM, Oudesluys-Murphy AM, Vossen ACTM, De Melker HE. 2015. Cytomegalovirus infection in the Netherlands: seroprevalence, risk factors, and implications. *Journal of Clinical Virology.* **63**:53-58.

[16] Livingstone S, Betancourt M, Byrne S, Girolami M. 2016. On the Geometric Ergodicity of Hamiltonian Monte Carlo. *arXiv:1601.08057 [stat.CO].*

[17] Meyn SP, Tweedie RL. *Markov chains and stochastic stability.* Springer Science & Business Media; 2012.

[18] Neal RM. Handbook of Markov Chain Monte Carlo. In: Neal RM. *MCMC Using Hamiltonian Dynamics.* Chapman & Hall/CRC Press; 2011.

[19] Mollema L, De Melker HE, Hahne SJM, Van Weert JWM, Berbers GAM, Van der Klis FRM. 2009. PIENTER 2-project: second research project on the protection against infectious diseases offered by the national immunization programme in the Netherlands. *RIVM Rapport 230421001.*

[20] Perron O. 1907. Zur Theorie der Matrices. *Mathematische Annalen.* **64(2)**:248-263.

[21] Plotkin S. 2015. The history of vaccination against cytomegalovirus. *Medical Microbiology and Immunology.* **204**:247-254.

[22] Plummer M. 2003. Jags: A program for analysis of Bayesian graphical models using Gibbs sampling.

[23] R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[24] Roberts GO, Gelman A, Gilks WR. 1997. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability.* **7**:110-120.

[25] Stan Development Team. 2016. Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0. Retrieved from http://mc-stan.org.

[26] Stan Development Team. 2017. ShinyStan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models. R package version 2.3.0. Retrieved from http://mc-stan.org.

[27] Szabó BT, Van der Vaart AW. 2017. *Bayesian Statistics* [Lecture notes]. Mastermath.

[28] Van de Kassteele J, Van Eijkeren J, Wallinga J. 2017. Efficient estimation of age-specific social contact rates between men and women. *The Annals of Applied Statistics*. **11(1)**:320-339.

[29] Van Boven M, Van de Kassteele J, Korndewal MJ, Van Dorp CH, Kretzschmar MEE, Van der Klis F, De Melker HE, Vossen AC, Van Baarle D. 2017. Infectious reactivation of cytomegalovirus explaining age- and sex-specific patterns of seroprevalence. *PLoS Comput Biol*. **13(9)**: e1005719.

[30] Vehtari A, Gelman A. 2014. WAIC and cross-validation in Stan. Retrieved from http://www.stat.columbia.edu.

[31] Vehtari A, Gelman A, Gabry J. 2016. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 1.0.0. Retrieved from https://github.com/stan-dev/loo.

A. Stan code for the full CMV model

```
data {
  int<lower=0> N;                              // number of subjects
  int<lower=1> DeltaA;                         // length of age intervals
  int<lower=1> A;                              // number of age intervals
  int<lower=1> A_rho;                          // number of age intervals for reactivation
  int<lower=1, upper=A_rho> RhoClasses[A];     // indices for transforming shortRho into rho
  matrix<lower=0>[A, A] Contact_MM;            // gender- and age-specific contact matrices
  matrix<lower=0>[A, A] Contact_FM;
  matrix<lower=0>[A, A] Contact_MF;
  matrix<lower=0>[A, A] Contact_FF;
  int<lower=0, upper=DeltaA*A> Ages[N];        // ages of the subjects
  real Titers[N];                              // antibody titers
  int Censor[N];                               // 0 = normal, 1 = censored, 2 = spike
  real RightCensor;                            // titers above this value are censored
  real MuS;                                    // means of classification mixture
  real MuL;
  real MuB;
  real<lower=0> SigmaS;                        // standard deviations of the
  real<lower=0> SigmaL;                        // classification mixture
  real<lower=0> SigmaB;
  real<lower=0> Penalty;                       // accuracy for estimating forces of infection
  int<lower=0, upper=1> Gender[N];             // 0 = female, 1 = male
  vector[7] BirthContribution;                 // contribution to birth rate
}


transformed data {
  // vectors used in vector manipulations
  vector[1] Zero;
  vector[DeltaA*A] LongOnes;
  vector[DeltaA*A+1] LlongOnes;

  // indices for transforming lambda into longLambda
  int<lower=1, upper=A> ExtendIdxs[DeltaA*A];
  int<lower=1, upper=DeltaA*A+1> ReduceIdxs[A];
  int<lower=1, upper=DeltaA*A+1> ReduceIdxsRightShift[A];

  Zero = rep_vector(0.0, 1);                   // for cumulative sums starting at 0
  LongOnes = rep_vector(1.0, DeltaA*A);        // same length as longLambda and longPi
  LlongOnes = rep_vector(1.0, DeltaA*A+1);     // same length as S, L, and B
```

```
  for ( aa in 1:DeltaA*A ) {                  // aa-1 = 0,1,2,3,4, 5,6,7,8,9, 10,...
    int j;
    j = (aa-1)/DeltaA + 1;                    // j = 1,1,1,1,1, 2,2,2,2,2, 3,...
    ExtendIdxs[aa] = j;
  }
  for ( j in 1:A ) {                          // j = 1,2,3,...
    ReduceIdxs[j] = 1 + (j-1)*DeltaA;         // x-1 = 0,5,10,...,DeltaA*(A-1)
  }
  for ( j in 1:A ) {                          // j = 1,2,3,...
    ReduceIdxsRightShift[j] = 1 + j*DeltaA;   // x-1 = 5,10,15,...,DeltaA*A
  }
}


parameters {
  vector<lower=0>[A] lambda_f;                // force of infection (female)
  vector<lower=0>[A] lambda_m;                // force of infection (male)
  real<lower=0> beta1;                        // infection rate after primary infection
  real<lower=0> beta2;                        // infection rate after reinfection/reactivation
  real<lower=0, upper=1> z;                   // reduction in susceptibility for reinfection
  vector<lower=0>[A_rho] shortRho_f;          // reactivation rate (female)
  vector<lower=0>[A_rho] shortRho_m;          // reactivation rate (male)
  real<lower=0> muRho;                        // hyperprior for rho
  real<lower=0> sigmaRho;
  real<lower=0,upper=1> S0;                   // fraction susceptibles at age 0
  real<lower=0,upper=1> nu;                   // perinatal transmission rate
}


transformed parameters {
  vector<lower=0>[A] rho_f;                   // reactivation rate (female)
  vector<lower=0>[A] rho_m;                   // reactivation rate (male)
  vector<lower=0>[A] pi_f;                    // shorthand: pi = rho + z * lambda
  vector<lower=0>[A] pi_m;
  vector<lower=0, upper=1>[DeltaA*A+1] S_f;   // susceptible (female)
  vector<lower=0, upper=1>[DeltaA*A+1] S_m;   // susceptible (male)
  vector<lower=0, upper=1>[DeltaA*A+1] L_f;   // latently infected (female)
  vector<lower=0, upper=1>[DeltaA*A+1] L_m;   // latently infected (male)
  vector<lower=0, upper=1>[DeltaA*A+1] B_f;   // infected with reinfection/reactivation (female)
  vector<lower=0, upper=1>[DeltaA*A+1] B_m;   // infected with reinfection/reactivation (male)
  vector<lower=0>[7] S_birth;

  // vectors for calculating S, L, and B
  vector<lower=0, upper=1>[DeltaA*A+1] X_f;   // latently infected at six months (female)
  vector<lower=0, upper=1>[DeltaA*A+1] X_m;   // latently infected at six months (male)
  vector<lower=0>[DeltaA*A+1] Y_f;            // (L_f-X_f)/X_f
  vector<lower=0>[DeltaA*A+1] Y_m;            // (L_m-X_m)/X_m
```

```
// lambda should be similar to lambda_hat
vector<lower=0>[A] lambda_hat_f;
vector<lower=0>[A] lambda_hat_m;

// long versions of lambda and pi
vector<lower=0>[DeltaA*A] longLambda_f;
vector<lower=0>[DeltaA*A] longLambda_m;
vector<lower=0>[DeltaA*A] longPi_f;
vector<lower=0>[DeltaA*A] longPi_m;

// for notational convenience
rho_f = shortRho_f[RhoClasses];
rho_m = shortRho_m[RhoClasses];
pi_f = rho_f + z * lambda_f;
pi_m = rho_m + z * lambda_m;
longLambda_f = lambda_f[ExtendIdxs];
longLambda_m = lambda_m[ExtendIdxs];
longPi_f = pi_f[ExtendIdxs];
longPi_m = pi_m[ExtendIdxs];

// define S, L, B, X, and Y
S_f = S0 * exp(-cumulative_sum(append_row(Zero, longLambda_f)));
S_m = S0 * exp(-cumulative_sum(append_row(Zero, longLambda_m)));

X_f = (1.0 - S0) * exp(-cumulative_sum(append_row(Zero, longPi_f)));
X_m = (1.0 - S0) * exp(-cumulative_sum(append_row(Zero, longPi_m)));

Y_f = cumulative_sum(append_row(Zero, longLambda_f .* (S_f[:DeltaA*A] ./ X_f[:DeltaA*A]) .*
      (LongOnes - exp(-(longLambda_f - longPi_f))) ./ (longLambda_f - longPi_f)));
Y_m = cumulative_sum(append_row(Zero, longLambda_m .* (S_m[:DeltaA*A] ./ X_m[:DeltaA*A]) .*
      (LongOnes - exp(-(longLambda_m - longPi_m))) ./ (longLambda_m - longPi_m)));

L_f = X_f .* (Y_f + LlongOnes);
L_m = X_m .* (Y_m + LlongOnes);

B_f = LlongOnes - S_f - L_f;
B_m = LlongOnes - S_m - L_m;

// define lambda_hat
lambda_hat_f = Contact_FF * (beta1 * (S_f[ReduceIdxs] - S_f[ReduceIdxsRightShift]) +
                             beta2 * (B_f[ReduceIdxsRightShift] - B_f[ReduceIdxs])) +
               Contact_MF * (beta1 * (S_m[ReduceIdxs] - S_m[ReduceIdxsRightShift]) +
                             beta2 * (B_m[ReduceIdxsRightShift] - B_m[ReduceIdxs]));
lambda_hat_m = Contact_MM * (beta1 * (S_m[ReduceIdxs] - S_m[ReduceIdxsRightShift]) +
                             beta2 * (B_m[ReduceIdxsRightShift] - B_m[ReduceIdxs])) +
               Contact_FM * (beta1 * (S_f[ReduceIdxs] - S_f[ReduceIdxsRightShift]) +
                             beta2 * (B_f[ReduceIdxsRightShift] - B_f[ReduceIdxs]));
```

```
  // estimate S0
  for ( a in 1:7 ) {
    S_birth[a] = BirthContribution[a]*(S_f[A+1+DeltaA*(a-1)] +
                 (1-nu)*(1-S_f[A+1+DeltaA*(a-1)]));
  }
  S_f[1] = sum(S_birth);
}


model {
  // priors on the (hyper)parameters
  beta1 ~ normal(0, 0.1);
  beta2 ~ normal(0, 0.1);
  z ~ uniform(0, 1);

  muRho ~ normal(0, 0.1);
  sigmaRho ~ normal(0, 0.1);

  S0 ~ normal(S_f[1],1/Penalty);

  shortRho_f[1] ~ normal(muRho, sigmaRho);
  shortRho_m[1] ~ normal(muRho, sigmaRho);

  for ( a in 2:A_rho ) {
    shortRho_f[a] ~ normal(shortRho_f[a-1], sigmaRho);
    shortRho_m[a] ~ normal(shortRho_m[a-1], sigmaRho);
  }

  // penalise the difference between lambda and lambda_hat
  lambda_f ~ normal(lambda_hat_f, 1/Penalty);
  lambda_m ~ normal(lambda_hat_m, 1/Penalty);

  // likelihood of the data
  for ( i in 1:N ) {
    int aa;
    real pS; real pL; real pB;
    aa = Ages[i] + 1; // the index for S, L and B

    // compute the probability to be in class S, L or B given age individual
    if ( Gender[i] == 0 ) {
      pS = S_f[aa];
      pL = L_f[aa];
      pB = B_f[aa];
    }
    else {
      pS = S_m[aa];
      pL = L_m[aa];
      pB = B_m[aa];
    }
```

```
    // define the likelihood
    if ( Censor[i] == 0 ) { // normal data
      target += log( pS * exp(normal_lpdf(Titers[i] | MuS, SigmaS)) +
                     pL * exp(normal_lpdf(Titers[i] | MuL, SigmaL)) +
                     pB * exp(normal_lpdf(Titers[i] | MuB, SigmaB)) );
    }
    else if ( Censor[i] == 1 ) { // right censored
      target += log( pS * exp(normal_lccdf(RightCensor | MuS, SigmaS)) +
                     pL * exp(normal_lccdf(RightCensor | MuL, SigmaL)) +
                     pB * exp(normal_lccdf(RightCensor | MuB, SigmaB)) );
    }
    else if ( Censor[i] == 2 ) { // spiked
      target += log(pS);
    }
  }
}


generated quantities {
  // save the likelihood
  vector[N] log_lik;

  for ( i in 1:N ) {
    int aa;
    real pS; real pL; real pB;
    aa = Ages[i] + 1; // the index for S, L and B

    // compute the probability to be in class S, L or B given age individual
    if ( Gender[i] == 0 ) {
      pS = S_f[aa];
      pL = L_f[aa];
      pB = B_f[aa];
    }
    else {
      pS = S_m[aa];
      pL = L_m[aa];
      pB = B_m[aa];
    }

    // define the likelihood
    if ( Censor[i] == 0 ) { // normal data
      log_lik[i] = log( pS * exp(normal_lpdf(Titers[i] | MuS, SigmaS)) +
                        pL * exp(normal_lpdf(Titers[i] | MuL, SigmaL)) +
                        pB * exp(normal_lpdf(Titers[i] | MuB, SigmaB)) );
    }
    else if ( Censor[i] == 1 ) { // right censored
      log_lik[i] = log( pS * exp(normal_lccdf(RightCensor | MuS, SigmaS)) +
                        pL * exp(normal_lccdf(RightCensor | MuL, SigmaL)) +
                        pB * exp(normal_lccdf(RightCensor | MuB, SigmaB)) );
```

```
    }
    else if ( Censor[i] == 2 ) { // spiked
      log_lik[i] = log(pS);
    }
  }
}
```

B. R code for calculation of the basic reproduction number

```
library(rstan)
library(loo)
library(readr)

# constants
N = 16
DeltaA = 5

# contact intensity matrix
ContactData = read_csv("contact_intensities_aggregated.csv")
FF = ContactData$mFF
FM = ContactData$mFM
MF = ContactData$mMF
MM = ContactData$mMM

# obtain samples from Stan output
S_beta1 = extract_log_lik(FIT_full, parameter_name = 'beta1')
S_beta2 = extract_log_lik(FIT_full, parameter_name = 'beta2')
S_rho_f = extract_log_lik(FIT_full, parameter_name = 'rho_f')
S_rho_m = extract_log_lik(FIT_full, parameter_name = 'rho_m')
S_lambda_f = extract_log_lik(FIT_full, parameter_name = 'lambda_f')
S_lambda_m = extract_log_lik(FIT_full, parameter_name = 'lambda_m')
S_nu = extract_log_lik(FIT_full, parameter_name = 'nu')

# function for notational convenience
Z <- function(k,h,rho) {
    if (k > h-1){
        Z = 1
    }
    else {
        Z = exp(-DeltaA*sum(rho[k:(h-1)]))
    }
    return(Z)
}
```

```
# calculate next-generation matrix entries
m <- function(i,l,j,k,beta1,beta2,rho_f,rho_m,lambda_f,lambda_m) {
    if (i == 'F' & l == 'F'){
        Contact = matrix(FF,nrow=16,ncol=16,byrow=TRUE)
        rho = rho_f
    }
    else if (i == 'F' & l == 'M'){
        Contact = matrix(MF,nrow=16,ncol=16,byrow=TRUE)
        rho = rho_m
    }
    else if (i == 'M' & l == 'F'){
        Contact = matrix(FM,nrow=16,ncol=16,byrow=TRUE)
        rho = rho_f
    }
    else if (i == 'M' & l == 'M'){
        Contact = matrix(MM,nrow=16,ncol=16,byrow=TRUE)
        rho = rho_m
    }
    else {
        print('invalid input')
    }

    term = rep(0,N)
    if (k < N){
        for (h in (k+1):N){
            term[h] = Contact[j,h] * Z(k,h,rho) * (1-exp(-DeltaA*rho[h]))
        }
    }

    m = DeltaA * beta1 * Contact[j,k] +
        beta2 * sum(term) * (1/rho[k]) * (exp(DeltaA*rho[k])-1) +
        beta2*Contact[j,k]*(DeltaA - (1/rho[k]) * (1 - exp(-DeltaA*rho[k])))
    return(m)
}

# construct next-generation matrix and calculate R0
M = matrix(rep(0,(2*N)^2),nrow=2*N,ncol=2*N)
R <- function(beta1,beta2,rho_f,rho_m,lambda_f,lambda_m,nu){
    for (k in 1:N) {
        for (j in 1:N){
            M[j,k] = m('F','F',j,k,beta1,beta2,rho_f,rho_m,lambda_f,lambda_m)
            M[j+N,k] = m('M','F',j,k,beta1,beta2,rho_f,rho_m,lambda_f,lambda_m)
            M[j,k+N] = m('F','M',j,k,beta1,beta2,rho_f,rho_m,lambda_f,lambda_m)
            M[j+N,k+N] = m('M','M',j,k,beta1,beta2,rho_f,rho_m,lambda_f,lambda_m)
        }
    }
    R = max(Re(eigen(M)$values))
    return(R)
}
```

```
# calculate R0 for all samples in Stan output
R0 = rep(0,length(S_beta1))
for (s in 1:length(S_beta1)) {
    beta1 = S_beta1[s]
    beta2 = S_beta2[s]
    rho_f = S_rho_f[s,]
    rho_m = S_rho_m[s,]
    lambda_f = S_lambda_f[s,]
    lambda_m = S_lambda_m[s,]
    nu = S_nu[s]
    R0[s] = R(beta1,beta2,rho_f,rho_m,lambda_f,lambda_m,nu)
}
quantile(R0, c(0.025,0.5,0.975))
```