The effect of household factor on HIV transmission

Marco Maninetti

February 4, 2023

Abstract

The understanding of the factors that drive the spread of HIV is crucial for the development of effective prevention and treatment strategies. In recent years, technological advancements in viral deep sequencing have enabled the use of phylogenetic source attribution techniques, which can help identify the potential transmission pairs of the virus. In this study, we aim to expand upon previous research by utilizing a Bayesian semi-parametric Poisson flow model to gain a more comprehensive understanding of the distinctions between HIV infections that occur within and outside of the household. This approach is applied to data collected from the Rakai community in Uganda, a region with a high prevalence of HIV. By using a Bayesian method, we aim to obtain solid and robust results that can contribute to the ongoing efforts to combat the spread of HIV. This paper is intended as a follow-up to the previous study conducted in [Xi et al., 2022] and builds on their research, but it shifts the focus on the household factor.

1 Introduction

1.1 Scope of the paper

The Rakai District in Uganda has long been a focal point for the study of HIV transmission dynamics. Imperial College, among other research institutions, has contributed extensively to the understanding of the epidemiology of HIV in sub-Saharan Africa. In particular, past research has revealed that women under the age of 25 are primarily infected by older men between the ages of 25 and 40, and subsequently play a role in the spread of the virus among men of their own age ([de Oliveira et al., 2017]). Furthermore, there is little evidence of male-to-male transmissions ([Ratmann et al., 2019]), and female-to-female transmissions are extremely rare. With this in mind, our research focuses on differences in HIV transmissions within and outside of households within the Rakai Community, specifically considering male-female and female-male infections. In 1.2, we present our data, in 2, we conduct an exploratory data analysis to identify notable patterns, in 3, our Bayesian model is described, and in 4, we discuss our findings and conclusions.

1.2 Data

Data are collected by the Rakai Community Cohort Study (RCCS). The study is organised in different rounds, in the Rakai Community, in Uganda. Each round lasts for about one year and an half. The first round included in our study is round 10 (insert dates here), while the last one is round 18 (insert dates here). Blood samples are collected, on which a phylogenetic analysis is then conducted, to identify potential transmission pairs. This analysis outputs an hypotetical transmission pair, the direction of infection, and a transmission score, between 0 and 1. The score represents the probability that the transmission did actually happen. Thresholding is applied (i.e., only pairs with more than 50% of probability of transmission are considered). Each transmission is identified to either have happened within the same household, or out of household. We compare household identifiers in the rounds where participants were interviewed, to check whether they have ever been reported to have lived together (i.e, in the same household in the same round). To infer the source's and recipient's age at the time of infection, we use TSI (time since infection) estimates (see e.g, [Golubchik et al., 2022]). They enable us to estimate a date of transmission, for each pair. Rakai community comprehends 4 communities (on the Lake Victoria) and 36 inland communities. These two community types have well documented epidemiological differences, as investigated for example in [Ratmann et al., 2020]).

2 Exploratory data analysis

The exploratory data analysis is an important step to identify notable patterns. We are interested in investigating how these patterns vary between transmissions within household and out of household. We will stratify the infections by recipient community, round, age, and sex.

2.1 Infections by recipient community

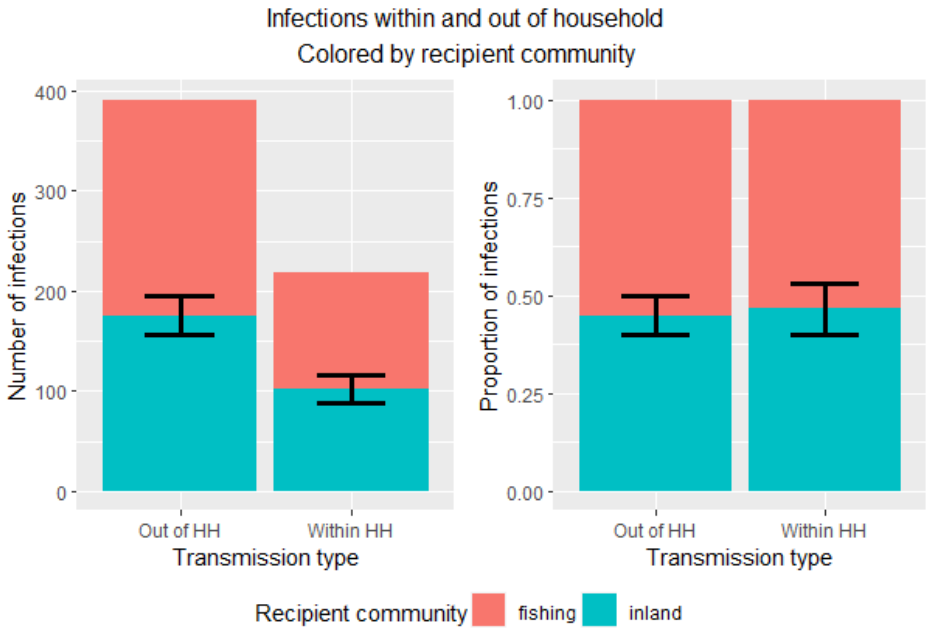


Figure 1: Infections within and out of household, colored by recipient community. The 95% binomial confidence intervals are displayed in black.

Figure 2: Infections by recipient community, colored by transmission type. The 95% binomial confidence intervals are displayed in black.

We are interested in understanding the effect of recipient community and transmission type, and their potential correlation. In figure 1 and figure 2, the number and proportion of infections is displayed for transmissions within and out of household, stratified by recipient community (inland and fishing).

	Fishing	Inland
Out of household	216(192-240)	175(154-198)
Within household	117(99-137)	102(85-121)

Table 1: Infections by recipient community and transmission type.

In table 1, the raw numbers with the 95% binomial confidence intervals are reported. As we can see, there are more observed infections in fishing communities (54.5%) than in inland (45.5%). Also, there are more infections out of household (64%) than within household (36%). (ADD INFO ABOUT PERCENTAGES OF PEOPLE IN INLAND/FISHING AND MARRIED/NOT MARRIED, TO MAKE THESE DATA MEANINGFUL). The interaction of these two factors doesn't seem to play an important role here. The proportion of transmissions within household is approximately the same in inland and fishing communities. So, we could conclude that from this point of view, the behaviour is really similar across different communities.

2.2 Infections by round

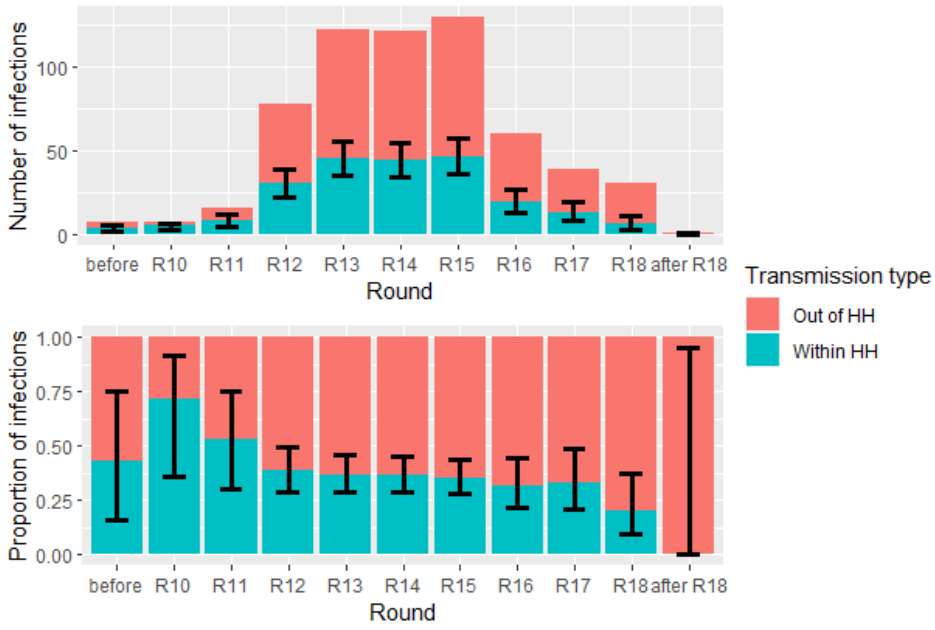


Figure 3: Number and proportion of detected infections, grouped by round. The 95% binomial confidence intervals are displayed in black.

In 3, we present the number and proportion of detected infections by round, grouping by transmission type. The total number of detected infections changes substantially across rounds. The proportion of infections within household is higher in earlier rounds. There may be a difference in migration rate between married and unmarried people, which leads to a higher rate of emigration from the Rakai Community ([**?**]). People who stay in the community likely have a family. Our blood sample analysis covers rounds 15 to 18 and thus, infections from round 10 involve individuals who participated in the study in round 10 and after round 15. It's not surprising that these older infections mainly occurred within households. It is also interesting to note (and it is worth further investigation) that most of the detected infections happened between round 12 and 15. It is very likely that some other rounds are underrepresented in the sample. Comprehensive sequencing started in round 15, so we don't have as much data for earlier rounds.

2.3 Infections by age

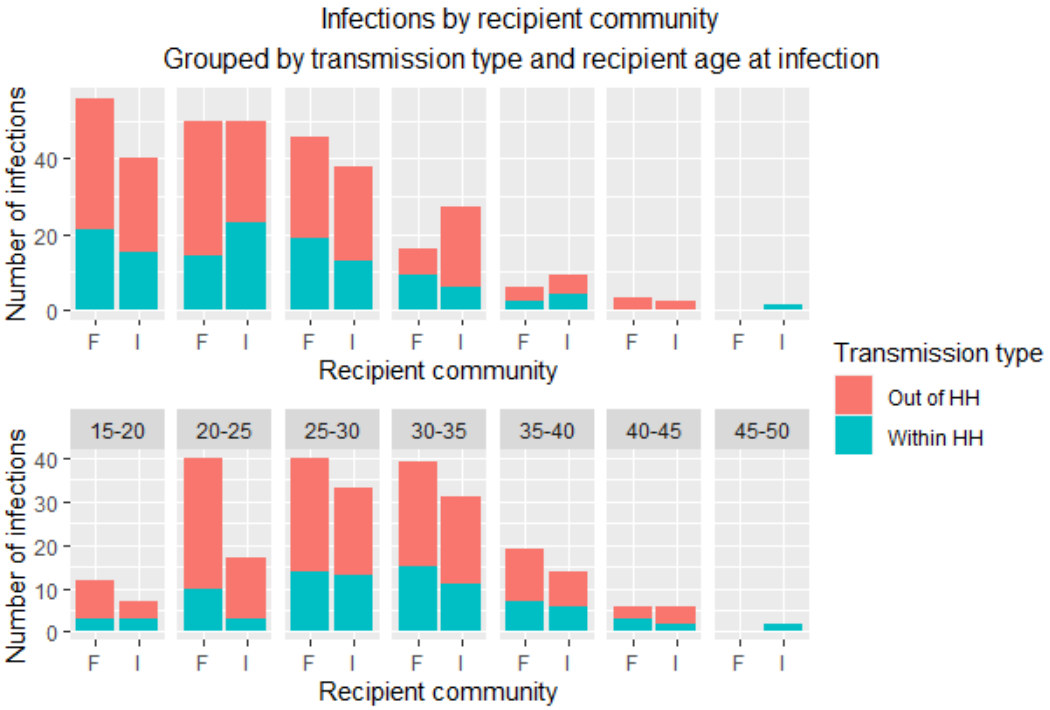
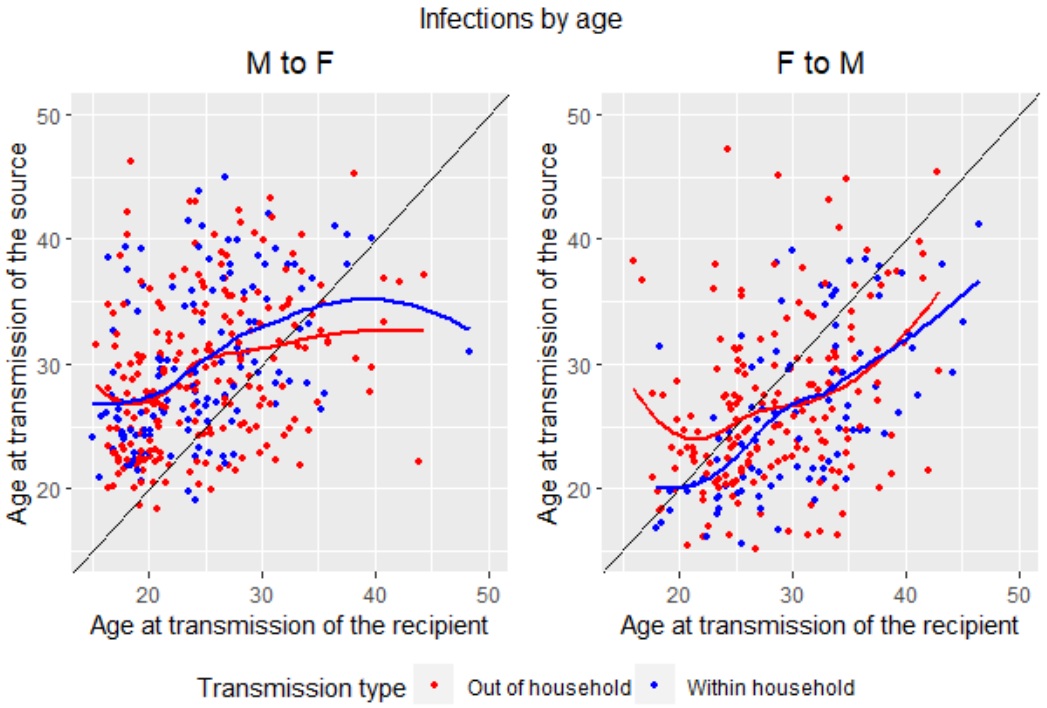


Figure 4: Upper panel: infections from male to female. Lower panel: infections from female to male. F indicates fishing communities, and I indicates inland communities.

Here, we present the number of detected infections, grouped by recipient's sex, community, and age at infection. Transmissions from male to female tend to have younger recipients. This has already been extensively discussed in previous research (see [de Oliveira et al., 2017]). It is interesting to note that the proportion of transmission within and out of household changes significantly across different age bands. There are noticeable differences also in the proportion of transmissions within and out of household in fishing and inland communities. In some age bands, this proportion greatly differs for fishing and inland communities. In 2.1, we concluded that there is not a relevant correlation between recipient community and transmission type. Nonetheless, both these covariates seem to be very correlated with the recipient age. In particular, infections from female to male appear to happen within the same household more often for recipients aged 25-35 than younger. For infections male to female, the proportion of transmissions within household appears to be more constant across recipient age bands.



The 2D plots of figure 2.3 represent the age at infection of the source and the recipient, stratified by sex and transmission type. The blue and red line are a smooth estimate of the relation between recipient's and source's age, computed using local polynomial regression (loess in R).

The most noticeable difference between transmissions within and out of household appears for young male recipients. In these cases, the female sources tend to be much older for transmissions happening out of household. The younger men are, the less likely it is for them to be married. They can have sexual contacts with similar-aged or older women. The latter have an high prevalence in the data, and probably less chances to be married. So, it is not surprising that out of household sources for young male recipients are older than within household sources. However, the initial decreasing pattern in sources' age is interesting and unexpected.

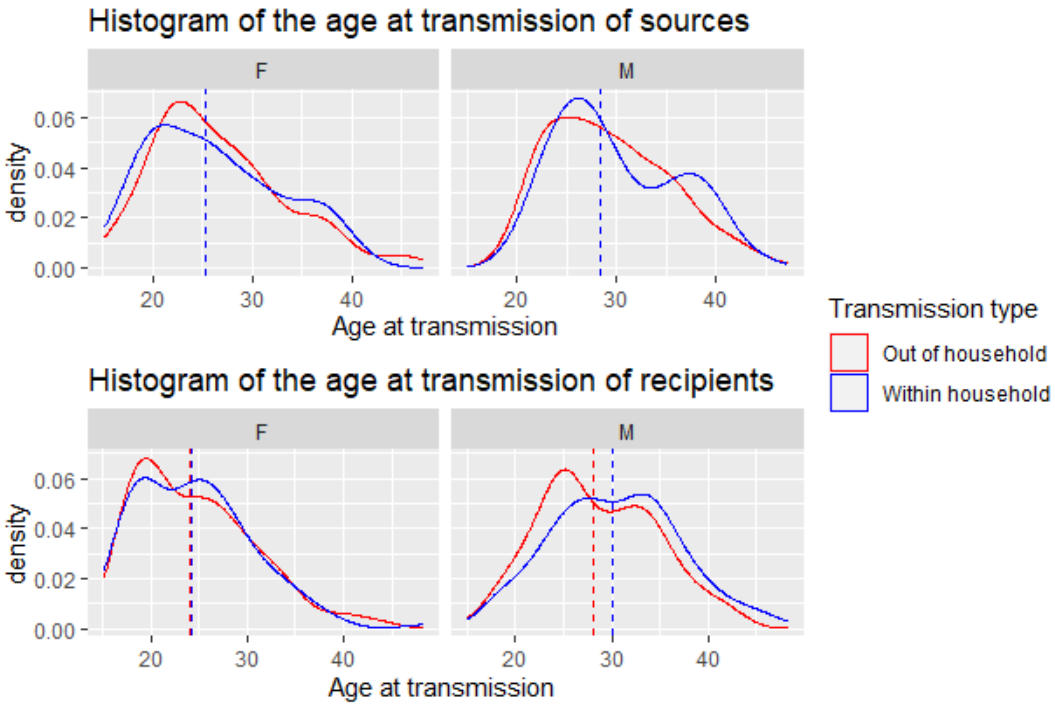


Figure 5: Histograms of sources' and recipients' age at transmission, stratified by sex and transmission type. The dashed line represents the median.

Figure 5 represents histograms of sources' and recipients' age at transmission, stratified by sex and transmission type.

As expected, males tend to be older, both as sources and recipients. The median age of female sources is basically the same for both transmissions within and out of household. The same applies to male sources, and to female recipients. Instead, male recipients tend to be older for transmissions in the same household. This could be due to the fact that people who get married at a young age are less exposed to risky sexual behaviour. Interestingly, there are some hints to bimodality, for all the 8 distributions. We do not know whether this is just due to noise or lack of data, or it is an epidemiological feature.

2.4 Infections by sex

HH	M to F	F to M
Out of HH	217(194-241)	174(153-197)
Within HH	127(108-148)	92(76-111)

Table 2: Infections by recipient community and HH, with the binomial 95% confidence intervals.

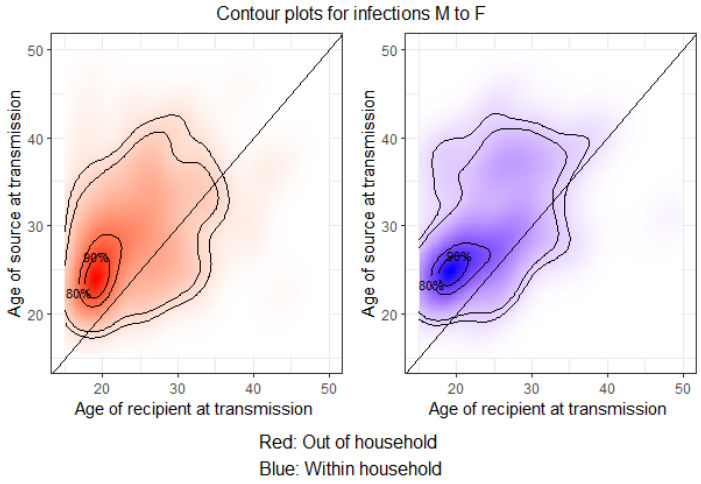


Figure 6: Contour plots for sources' and recipients' ages for infections male to female

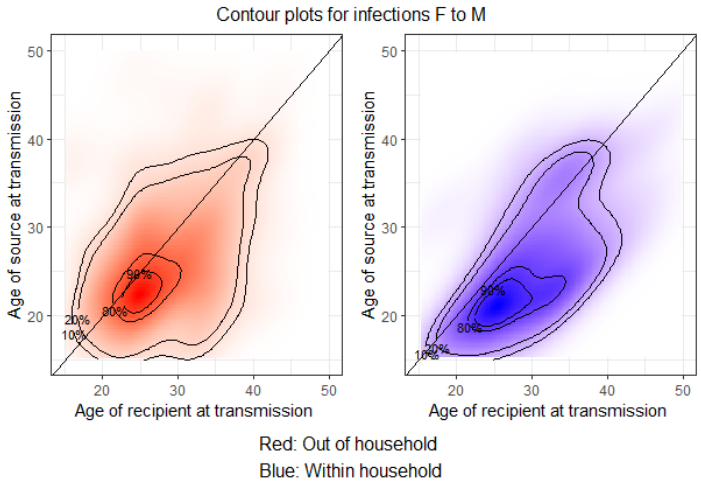


Figure 7: Contour plots for sources' and recipients' ages for infections female to male

Figure 2.4 and figure 2.4 are contour plots of the ages at infection for the source and the recipient, stratifies by sex and transmission type. These are obtained by using a 2-dimensional kernel estimator (function kde in the package ks in R). We can see that there are similar patterns for transmission within and out of household. For transmissions female to male, the ones out of household seem to be more centered, while the ones within household tend to have an older recipient.

Table 2 represents raw numbers. It is interesting to notice that the ratio of M to F transmissions is higher for infections within household (0.6 versus 0.57).

3 Methodology

3.1 Introduction

Our main interest is in investigating the effect of the household factor. We are seeking an answer to the question "How do infections within household differ from the ones out of household?". Given our exploratory data analysis, the type of community (inland or fishing) seems not to play a crucial role. It does not have any noticeable interaction with the transmission type, and with other factors. So, we decided to exclude it from the analysis. There are two other factors which are fundamental: direction of transmission (infections from male to female have different characteristics than infections from female to male) and age (both of the source and of the recipient). These are the features we will include in our analysis, together with the household factor.

3.2 Statistical model

Our model builds on the one presented in [Xi et al., 2022]. They present a semiparametric Bayesian Poisson flow model. Due to the possible interaction between direction and type of transmission (within and out of household), we decided to consider these 2 variables together, and not separately. So, 4 categories for infections arise. We aim to estimate transmission flows among age groups (for the sake of simplicity, we will use the subscript 'w' to indicate 'within household' and 'o' to indicate 'out of household'). We divide sources and recipients in 2-years age bands, from 15 to 50 years old. We define K as the number of age bands (18, in our case). We define $\pi_{a,b}^{s,h}$ as the proportion of transmissions from age band a to age band b, with direction s and type h. Similarly, $Z_{a,b}^{s,h}$ indicates the count of transmissions from age band a to age band b, with direction s and type h. We define the flow matrix as:

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}^{mf,w} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\pi}^{mf,o} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\pi}^{fm,w} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\pi}^{fm,o} \end{pmatrix}, \quad \boldsymbol{\pi}^{mf,w} = \begin{pmatrix} \pi_{11}^{mf,w} & \cdots & \pi_{1K}^{mf,w} \\ \vdots & \ddots & \vdots \\ \pi_{K1}^{mf,w} & \cdots & \pi_{KK}^{mf,w} \end{pmatrix} \quad (2)$$

It is important to notice that we have some structural zeros. We exclude same sex transmissions. Obviously, it does not make sense to talk about a transmission from an individual in the same household to an individual in a different household. Equation (2) has $4K^2$ non-zero entries to estimate, which amounts to 1296 parameters.

Based on [Xi et al., 2022], we propose a Poisson model for the likelihood:

$$Z_{a,b}^{s,h} \sim Poisson(\lambda_{a,b}^{s,h})$$

So,

$$p(\boldsymbol{Z}|\boldsymbol{\lambda}) \propto \prod_{a,b,s,h} (\lambda_{ab}^{s,h})^{Z_{ab}^{s,h}} \exp\left(-\lambda_{ab}^{s,h}\right) = \left[\prod_{a,b,s,h} (\pi_{ab}^{s,h})^{Z_{ab}^{s,h}}\right] [\eta^{Z^+} \exp\left(-\eta\right)],$$

where Z^+ is the total count of transmissions, λ_{ab} represents the flow intensity from group a to group b, $\eta = \sum_{a,b} \lambda_{ab}$, and π_{ab} are recovered through $\pi_{ab} = \lambda_{ab}/\eta$. So, λ is a vector, with $4K^2$ entries. The first K^2 entries of $\boldsymbol{\lambda}$ correspond to flows in the same household, in the direction M to F. The second

K^2 entries correspond to flows within different households, in the direction M to F. The third K^2 entries of $\boldsymbol{\lambda}$ correspond to flows in the same household, in the direction F to M. The fourth K^2 entries of correspond to flows within different households, in the direction F to M.

We propose the following priors:

$$\log \boldsymbol{\lambda} = \mu \mathbf{1} + \nu_1 \mathbb{1}_{mf,w} + \nu_2 \mathbb{1}_{mf,o} + \nu_3 \mathbb{1}_{fm,w} + \boldsymbol{f}$$

$$\boldsymbol{f} = \left(\boldsymbol{f}_{mf,w}^T,\ \boldsymbol{f}_{mf,o}^T,\ \boldsymbol{f}_{fm,w}^T,\ \boldsymbol{f}_{fm,o}^T\right)^T$$

$$\boldsymbol{f}_{mf,w} \sim \mathcal{GP}\left(0, k_{mf,w}\right) \quad \boldsymbol{f}_{mf,o} \sim \mathcal{GP}\left(0, k_{mf,o}\right)$$

$$\boldsymbol{f}_{fm,w} \sim \mathcal{GP}\left(0, k_{fm,w}\right) \quad \boldsymbol{f}_{fm,o} \sim \mathcal{GP}\left(0, k_{fm,o}\right)$$

$$k_{mf,s}\left((a_1,b_1),(a_2,b_2)\right) = \sigma_{mf,w}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{mf,w,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{mf,w,b}^2}\right]\right)$$

$$k_{mf,o}\left((a_1,b_1),(a_2,b_2)\right) = \sigma_{mf,o}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{mf,o,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{mf,o,b}^2}\right]\right)$$

$$k_{fm,w}\left((a_1,b_1),(a_2,b_2)\right) = \sigma_{fm,w}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{fm,w,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{fm,w,b}^2}\right]\right)$$

$$k_{fm,o}\left((a_1,b_1),(a_2,b_2)\right) = \sigma_{fm,o}^2 \exp\left(-\left[\frac{(a_2-a_1)^2}{2\ell_{fm,o,a}^2} + \frac{(b_2-b_1)^2}{2\ell_{fm,o,b}^2}\right]\right)$$

$$\sigma_{mf,w}^2\ ,\ \sigma_{mf,o}^2\ ,\ \sigma_{fm,w}^2\ ,\ \sigma_{fm,o}^2 \sim \text{Half-Normal(0,10)}$$

$$\ell_{g,i} \sim \text{Inv-Gamma}\left(\alpha_{g,i}, \beta_{g,i}\right)$$

$$\mu\ ,\ \nu_1\ ,\ \nu_2\ ,\ \nu_3 \sim \text{Normal(0,100)}$$

μ is the baseline log-transmission intensity, ν_1, ν_2, and ν_3 are scalars. $k_{mf,s}$, $k_{mf,d}$, $k_{fm,s}$, and $k_{fm,d}$ are gender and household specific squared exponential kernels, with variance parameters $\sigma_{mf,s}^2, \sigma_{mf,d}^2, \sigma_{fm,s}^2, \sigma_{fm,d}^2$ and lenght scales $\ell_{mf,s,a}, \ell_{mf,s,b}, \ell_{mf,d,a}, \ell_{mf,d,b}, \ell_{fm,s,a}, \ell_{fm,s,b}, \ell_{fm,d,a}, \ell_{fm,d,b}$.

TODO: how to define α and β here?

4 Conclusion

References

[de Oliveira et al., 2017] de Oliveira, T., B M Kharsany, A., Gräf, T., Cawood, C., Khanyile, D., Grobler, A., Puren, A., Madurai, S., Baxter, C., Abdool Karim, Q., and Abdool Karim, S. S. (2017). Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study .

[Golubchik et al., 2022] Golubchik, T., Abeler-Dörner, L., Hall, M., and Wymant, C. (2022). HIV-phyloTSI: Subtype-independent estimation of time since HIV-1 infection for cross-sectional measures of population incidence using deep sequence data.

[Monod et al., 2021] Monod, M., Blenkinsop, A., Brizzi, A., Chen, Y., Cardoso Correia Perello, C., Jogarah, V., Wang, Y., Flaxman, S., Bhatt, S., and Ratmann, O. (2021). Regularised B-splines projected Gaussian Process priors to estimate time-trends of age-specific COVID-19 deaths related to vaccine roll-out .

[Ratmann et al., 2019] Ratmann, O., Grabowski, M. K., Hall, M., Golubchik, T., Wymant, C., Abeler-Dörner, L., Bonsall, D., Hoppe, A., Leigh Brown, A., de Oliveira, T., Gall, A., Kellam, P., Pillay, D., Kagaayi, J., Kigozi, G., C. Quinn, T., J. Wawer, M., Laeyendecker, O., Serwadda, D., H. Gray, R., and Fraser, C. (2019). Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis .

[Ratmann et al., 2020] Ratmann, O., Kagaayi, J., Hall, M., Golubchik, T., Kigozi, G., and Xi, X. (2020). Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda .

[Xi et al., 2022] Xi, X., EF Spencer, S., Hall, M., Grabowski, M. K., Kagaayi, J., and Ratmann, O. (2022). Inferring the sources of HIV infection in Africa from deep sequence data with semi-parametric Bayesian Poisson flow models.