Cohort-based smoothing methods for age-specific contact rates

Yannick Vandendijck[a] ,Oswaldo Gressani [a,*] ,Christel Faes[a] , Carlo G. Camarda[b], Niel Hens[a,c]

[a] Interuniversity Institute for Biostatistics and statistical Bioinformatics (IBioStat), Hasselt University, Belgium

[b] French Institute for Demographic Studies (INED), Aubervilliers, France

[c] Centre for Health Economics Research and Modelling Infectious Diseases, Vaxinfectio, University of Antwerp, Belgium

Abstract

The use of social contact rates is widespread in infectious disease modeling since it has been shown that they are key driving forces of important epidemiological parameters. Quantification of contact patterns is crucial to parametrize dynamic transmission models and to provide insights on the (basic) reproduction number. Information on social interactions, can (for instance) be obtained from population-based contact surveys, such as the European Commission project POLYMOD. Estimation of age-specific contact rates from these studies is often done using a piecewise constant approach or bivariate smoothing techniques. For the latter, typically, smoothing is done in the dimensions of the respondent's and contact's

*Corresponding author. E-mail address: *oswaldo.gressani@uhasselt.be*

age. We propose a new flexible strategy based on a smoothing constrained approach - taking into account the reciprocal nature of contacts - where the contact rates are assumed smooth from a cohort perspective as well as from the age distribution of contacts. This is achieved by smoothing over the diagonal components (including all subdiagonals) of the social contact matrix. This approach is supported by the fact that people age with time and thus motivates smoothly varying contact rates from a cohort angle. Two approaches that allow for smoothing of social contact data over cohorts are proposed namely, (1) reordering of the diagonal components of the social contact matrix; and (2) reordering of the penalty matrix associated with the diagonal components. Parameter estimation is done in the likelihood framework by using constrained penalized iterative reweighted least squares (C-PIRLS), under Poisson and negative Binomial distributional assumptions for the observed contacts. A simulation study underlines the benefits of cohort-based smoothing based on two scalar measures of performance. Finally, the proposed methods are illustrated on the Belgian POLYMOD data of 2006. Code to reproduce the results of the article can be downloaded on this Github repository https://github.com/oswaldogressani/Cohort_smoothing.

Keywords: Penalized iterative reweighted least squares, Penalized likelihood, Constrained smoothing, Social contact matrix.

1 Introduction

Understanding the spread of infectious diseases in an epidemic context is a challenging task for mathematical modelers. It is especially made difficult by the complexities and intricacies of demography dynamics and rich social contact networks. Social contact mixing patterns play a key role in assessing disease transmission and are known to be crucial determinants of important epidemiological parameters such as the basic reproduction number and the force of infection (see e.g., Vynnycky and White, 2010; Hens et al., 2009). One approach to account for mixing patterns is by the use of the so-called "Who Acquires Infection From Whom" (WAIFW) matrix and the use of serological data to estimate the

WAIFW parameters (Anderson and May, 1991; Greenhalgh and Dietz, 1994; Farrington et al., 2001; Van Effelterre et al., 2009). Another approach proposed by Farrington and Whitaker (2005) is to model contact rates as a continuous surface and estimate parameters from serologic survey data. The main limitations of both approaches is that they rely on structural assumptions on the WAIFW matrix and on an arbitrary choice of the parametric model used for the continuous contact surface. Alternatively, over the last two decades or so, several studies have reported on ways of collecting data on social mixing behaviour relevant to the spread of close contact infections directly from individuals through self-reported number of contacts (Wallinga et al., 2006; Beutels et al., 2006; Edmunds et al., 1997, 2006; Mikolajczyk et al., 2007). The POLYMOD initiative can arguably be counted among the most important studies in infectious disease epidemiology in Europe, providing a large and representative popoulation based survey on social contacts (Mossong et al., 2008). The estimation of smooth age-specific contact rates from the POLYMOD project data is typically performed by applying a negative Binomial model on the aggregated number of contacts. To ensure enough flexibility, a bivariate frequentist smoothing method is implemented by using a tensor product spline as a function of the respondent's and contact's ages as a smooth interaction term (Mossong et al., 2008; Hens et al., 2009; Goeyvaerts et al., 2010). From a Bayesian perspective, van de Kassteele et al., 2017 estimate social contact rates by means of a Gaussian Markov Random Field (GMRF) and use Integrated Nested Laplace Approximations (INLA) Rue et al. (2009) as the main tool for inference.

We propose a new smoothing constrained approach, where contact rates are assumed to be smooth both from a cohort perspective and from the age distribution of contacts. This means that smoothing in the direction of the age of contacts will remain. However, smoothing over the dimension of the age of respondents will be replaced by smoothing contact rates from a cohort perspective by focusing on the diagonal compontents (including all subdiagonals) of the social contact matrix. Under the likelihood framework and assuming Poisson or negative Binomial models for the aggregated number of contacts, diagonal smoothing of contact matrices is achieved through two alternative approaches: (1) reordering of the

diagonal components yielding a rectangular grid; and (2) reordering of the penalty matrix to translate a penalization scheme over the diagonal components. The first approach builds further upon work published by two of the co-authors in a proceedings paper (*reference not provided in reviewing process*).

The article is organized as follows. Section 2 aims at presenting three competing approaches to smoothly estimate social contact rates. Section 3 investigates the statistical performance of the proposed approaches through a numerical study and Section 4 illustrates the methodology on the Belgian POLYMOD data. Finally, Section 5 concludes with a discussion and prospects for future research.

2 Smoothing social contact data

In this section, we present three competing smoothing constrained approaches (SCAs) to infer social contact rates. First, we describe the classic approach where smoothing is performed in the dimensions of the respondent's and contact's ages, thus ignoring the cohort effect. The latter baseline model will be referred to $\mathcal{M}_0$. Second, we present the new competing models, namely the SCA where contact rates are assumed smooth from a cohort perspective. Two approaches are investigated both in terms of performance and computational speed, namely model $\mathcal{M}_1$, where a reordering of the diagonal components is considered to reproduce a rectangular contact matrix; and model $\mathcal{M}_2$, where a reordering of the components of the penalty matrix yields a penalization scheme targeting the diagonal components of the social contact matrix.

2.1 Absence of smoothing over cohorts

Let $\mathbf{Y} = (y_{ij})$ be a square $(m \times m)$ matrix where the ijth entry is the total number of contacts made by the respondents of age $i-1$ with individuals of age $j-1$, with indices $i = 1, \ldots, m$ and $j = 1, \ldots, m$. This information can be extracted from the self-reported contact diaries of the participants and in our specific case $m = 77$ for the Belgian POLYMOD data. Let $\mathbf{y}$ be the $m^2 \times 1$ vector obtained by arranging the matrix $\mathbf{Y}$ by row

order into a vector. Furthermore, let the $m \times 1$ vector $\mathbf{r} = (r_i)$ contain the total number of respondents of age $i-1$. Define the $m \times m$ matrix $\mathbf{E} = \mathbf{r}\mathbf{1}_m$, where $\mathbf{1}_m$ is a $1 \times m$ vector of ones, and define $\mathbf{e}$ as the $m^2 \times 1$ vector obtained by arranging the matrix $\mathbf{E}$ by row order into a vector. Let the $m \times 1$ vector $\mathbf{p} = (p_i)$ denote the population size of individuals of age $i-1$ and define the $m \times m$ matrix $\mathbf{P} = \mathbf{p}\mathbf{1}_m$. The supplementary materials provide examples of how to construct these vectors and matrices for the specific case $m = 4$. The expected number of contacts made by participants of age $i-1$ with contacts of age $j-1$ is denoted by $\mathrm{E}(y_{ij}) = \mu_{ij} = r_i\gamma_{ij}$, where γ_{ij} is the actual contact rate of individuals of age $i-1$ with contacts of age $j-1$. In other words, γ_{ij} is the average number of contacts an individual of age $i-1$ makes with an individual of age $j-1$. Define the so-called social contact matrix $\mathbf{\Gamma}$ as the $m \times m$ matrix with elements γ_{ij} (see Figure 1 left panel) and let $\boldsymbol{\gamma}$ be the $m^2 \times 1$ vector obtained by arranging the matrix $\mathbf{\Gamma}$ by row order into a vector.

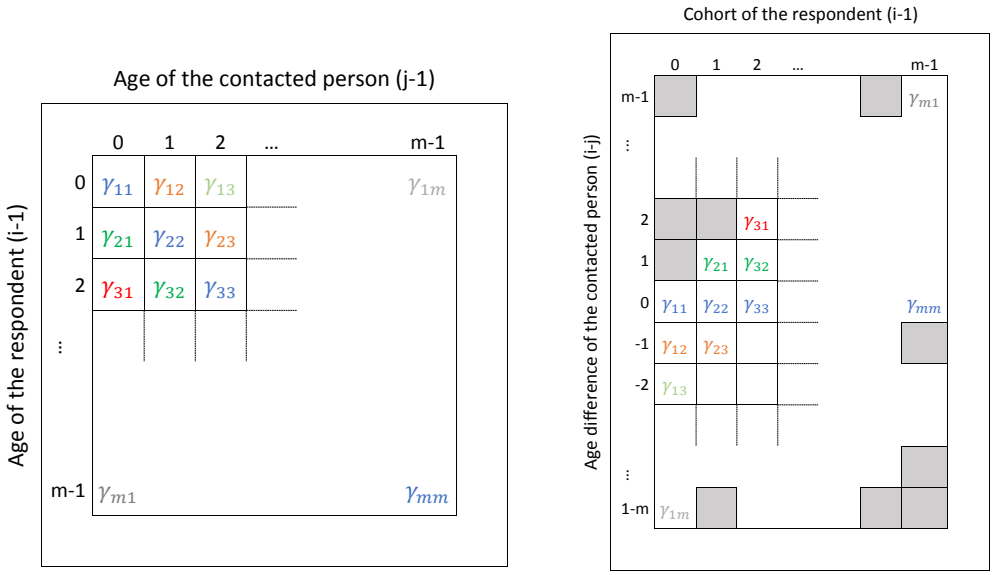

Figure 1: Schematic representation of the original data structure of $\mathbf{\Gamma}$ over ages of respondents and ages of contacts (left panel) and the restructured matrix $\breve{\mathbf{\Gamma}}$ over cohorts of the respondents and age differences of the contacted persons (right panel). Cells with nuisance parameters in $\breve{\mathbf{\Gamma}}$ are depicted with gray squares.

The expected number of contacts can also be written as $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{e} \odot \boldsymbol{\gamma}$, where $\odot$ denotes component-wise multiplication (also known as the Hadamard product). The interest lies in estimating the unknown contact parameters γ_{ij} from data $\mathbf{y}$ in a smooth way, such that the

important signal in the mixing patterns is captured. For this purpose, we assume that the observed contacts (y_{ij}) are realizations from a Poisson distribution, i.e. $\mathbf{y} \sim \text{Poiss}(\boldsymbol{\mu})$. For modeling purposes, a log-link function is specified, namely $\log(\boldsymbol{\gamma}) = \boldsymbol{\eta}$, so that $\log(\boldsymbol{\mu}) = \log(\mathbf{e}) + \log(\boldsymbol{\gamma}) = \log(\mathbf{e}) + \boldsymbol{\eta}$.

Let $\mathbf{H}$ be the $m \times m$ matrix with ijth element η_{ij}. Interest is in the estimation of the m^2 unknown parameters $\boldsymbol{\eta}$. It can be readily seen that the maximum likelihood estimates are given by $\hat{\boldsymbol{\eta}} = \log(\mathbf{y}/\mathbf{e})$, and thus $\hat{\boldsymbol{\gamma}} = \mathbf{y}/\mathbf{e}$, in case the parameters can be estimated freely. However, these estimates do not yield a smooth contact rate surface and hence are only of interest for exploratory purposes. We prefer to work with a modeling approach that yields social contact rates that are smooth and reciprocal. Reciprocity of contacts in this context means that the total number of contacts on the population level from age i to age j must equal the total number of contacts from age j to age i. The latter reciprocal nature can be expressed mathematically as $\gamma_{ij}p_i = \gamma_{ji}p_j$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$ and can be written as the difference $\log(\gamma_{ij}) - \log(\gamma_{ji}) = \log(p_j) - \log(p_i)$ and thus:

$$\eta_{ij} - \eta_{ji} = \log(p_j) - \log(p_i). \tag{1}$$

In matrix form:

$$\mathbf{L}\boldsymbol{\eta} = \boldsymbol{\nu}, \tag{2}$$

where $\mathbf{L}$ is a $\frac{m(m-1)}{2} \times m^2$ allocation matrix with entries $+1$ and -1 to suit the left-hand side of (1) and vector $\boldsymbol{\nu}$ is given by:

$$\begin{aligned}\boldsymbol{\nu}^T = \big(&\log(p_2) - \log(p_1), \log(p_3) - \log(p_1), \ldots, \log(p_n) - \log(p_1),\\ &\log(p_3) - \log(p_2), \log(p_4) - \log(p_2), \ldots, \log(p_n) - \log(p_2),\\ &, \ldots,\\ &\log(p_n) - \log(p_{m-1})\big).\end{aligned}$$

Estimation of the smoothed parameters $\boldsymbol{\eta}$ that satisfy the reciprocal constraints is per-

formed through constrained penalized iterative reweighted least squares (C-PIRLS) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Eilers and Marx, 1996; Wood, 2006). Given current estimates $\hat{\boldsymbol{\eta}}^{[k]}$ and $\hat{\boldsymbol{\mu}}^{[k]}$ at iteration k, parameter estimates $\hat{\boldsymbol{\eta}}^{[k+1]}$ at iteration $k+1$ are obtained by solving the set of linear equations:

$$\begin{pmatrix} \mathbf{W}^{[k]} + \mathbf{P} & \mathbf{L}^T \\ \mathbf{L} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\eta}}^{[k+1]} \\ \boldsymbol{\zeta}^{[k+1]} \end{pmatrix} = \begin{pmatrix} \mathbf{W}^{[k]}\mathbf{z}^{[k]} \\ \boldsymbol{\nu} \end{pmatrix}. \tag{3}$$

The parameter estimates $\hat{\boldsymbol{\gamma}}^{[k+1]}$ are obtained by exponentiation $\hat{\boldsymbol{\gamma}}^{[k+1]} = \exp\left(\hat{\boldsymbol{\eta}}^{[k+1]}\right)$. In (3), $\boldsymbol{\zeta}^{[k+1]}$ is a $\frac{m(m-1)}{2} \times 1$ vector of Lagrange multipliers, $\mathbf{W}^{[k]}$ is a $m^2 \times m^2$ diagonal matrix with entries $W_{ll}^{[k]} = \mu_l^{[k]} = \mathrm{e}_l \exp\left(\eta_l^{[k]}\right)$ and $\mathbf{z}^{[k]}$ is a $m^2 \times 1$ vector of the so-called *pseudodata* given by:

$$z_l^{[k]} = \eta_l^{[k]} + \left(\frac{\mathrm{y}_l}{\mu_l^{[k]}} - 1\right). \tag{4}$$

To enforce smoothness over two dimensions, the penalty term $\mathbf{P}$ in (3) is a $m^2 \times m^2$ matrix given by (see Marx and Eilers, 2005):

$$\mathbf{P} = \lambda_1 \mathbf{I}_m \otimes (\mathbf{D}_h^T \mathbf{D}_h) + \lambda_2 (\mathbf{D}_v^T \mathbf{D}_v) \otimes \mathbf{I}_m, \tag{5}$$

where $\otimes$ denotes the Kronecker product and λ_1 and λ_2 are smoothing parameters for, respectively, the horizontal and vertical dimension in Figure 1 (left panel). The matrices $\mathbf{D}_h$ and $\mathbf{D}_v$ are second order difference matrices and $\mathbf{I}$ is the identity matrix. The above iterative process is repeated until convergence, namely until $\max \mid \hat{\boldsymbol{\eta}}^{[k+1]} - \hat{\boldsymbol{\eta}}^{[k]} \mid < 10^{-4}$. The optimal smoothing parameters are chosen based on minimization of the Akaike Information Criterion (Akaike, 1973) via a grid search:

$$\mathrm{AIC} = -2\log(\hat{L}) + 2\widehat{ED}, \tag{6}$$

where $\hat{L}$ is the maximized value of the likelihood function and the effective degrees of freedom, $\widehat{ED}$, is the trace of the hat matrix given by (see Wood, 2006):

$$\mathbf{A} = \mathbf{W}^{1/2} \left(\mathbf{W} + \mathbf{P}\right)^{-1} \mathbf{W}^{1/2}. \tag{7}$$

2.2 Cohort smoothing by reordering the contact matrix ($\mathcal{M}_1$ model)

In the previous section, contact rate parameters are smoothed in the dimensions of the respondent's and contact's ages. We now describe a new strategy where contact rates are smoothed over the diagonal components and thus over cohorts. In addition, we also smooth over the dimension of the contact's age since the distribution of the age of (grand)parents can in general be assumed smooth (e.g., children will meet their parents and grandparents who are, for example, $\pm$ 30 and $\pm$60 years older). We describe how this can be achieved by restructuring the data and contact matrix over the cohorts and the contacts' ages.

The contact matrix $\mathbf{\Gamma}$ is restructured in such a way that each diagonal (the main diagonal and all sub-diagonals) is present as a row in the restructured matrix. The restructured matrix is denoted $\breve{\mathbf{\Gamma}}$. Figure 1 (right panel) gives a graphical representation of this restructured matrix. The matrix $\breve{\mathbf{\Gamma}}$ has dimension $(2m-1) \times m$ and is constructed by entering row i of $\mathbf{\Gamma}$ in column i of $\breve{\mathbf{\Gamma}}$ at positions $m-i+1$ to $2m-i$. In that manner, all subsequent diagonal elements are present in the same row. By construction, matrix $\breve{\mathbf{\Gamma}}$ contains nuisance contact rate parameters that are not directly of interest. Restructured matrices $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$, constructed from $\mathbf{Y}$ and $\mathbf{E}$, are created similarly as $\breve{\mathbf{\Gamma}}$. Missing cell entries are present for $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$ at the same cells where the nuisance parameters are present for $\breve{\mathbf{\Gamma}}$. To handle these missing entries, we impute arbitrary values (say, 9999) in $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$ and construct a $(2m-1) \times m$ weight matrix $\breve{\mathbf{W}}$, where the ijth entry of $\breve{\mathbf{W}}$ equals zero if the ijth entry in $\breve{\mathbf{\Gamma}}$ is a nuisance parameter and equals one otherwise. This weight matrix avoids that the imputed values for the missing entries influence parameter estimation.

Let $\breve{\mathbf{y}}$, $\breve{\mathbf{e}}$, $\breve{\mathbf{w}}$ and $\breve{\boldsymbol{\gamma}}$ be the $(2m^2-m) \times 1$ vectors obtained by arranging the matrices $\breve{\mathbf{Y}}$, $\breve{\mathbf{E}}$, $\breve{\mathbf{W}}$ and $\breve{\mathbf{\Gamma}}$ by column order into a vector. Again, we assume that $\mathrm{E}(\breve{\mathbf{y}}) = \breve{\boldsymbol{\mu}} = \breve{\mathbf{e}} \odot \breve{\boldsymbol{\gamma}} \odot \breve{\mathbf{w}}$ and that the observations come from a Poisson distribution, namely $\breve{\mathbf{y}} \sim \mathrm{Pois}(\breve{\boldsymbol{\mu}})$ and $\log(\breve{\boldsymbol{\gamma}}) = \breve{\boldsymbol{\eta}}$. The objective is now to estimate the $2m^2 - m$ unknown parameters $\breve{\boldsymbol{\eta}}$. How-

ever, only the m^2 parameters of $\breve{\boldsymbol{\eta}}$ corresponding to the non-nuisance parameter entries in $\breve{\boldsymbol{\Gamma}}$ are of interest. The reciprocity assumption of the contacts, namely $\breve{\gamma}_{ij}p_i = \breve{\gamma}_{ji}p_j$, can again be written in matrix form as:

$$\mathbf{L}\breve{\boldsymbol{\eta}} = \boldsymbol{\nu}, \tag{8}$$

where $\mathbf{L}$ is an $(\frac{m(m-1)}{2}) \times (2m^2 - m)$ allocation matrix to accommodate the reciprocity constraints. Estimation of the smoothed parameters $\breve{\boldsymbol{\eta}}$ is again performed through C-PIRLS. Updated parameter estimates are now obtained by solving the set of linear equations:

$$\begin{pmatrix} \mathbf{W}^{[k]} + \mathbf{P} & \mathbf{L}^T \\ \mathbf{L} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\breve{\boldsymbol{\eta}}}^{[k+1]} \\ \boldsymbol{\zeta}^{[k+1]} \end{pmatrix} = \begin{pmatrix} \mathbf{W}^{[k]}\mathbf{z}^{[k]} \\ \boldsymbol{\nu} \end{pmatrix}. \tag{9}$$

In (9), $\boldsymbol{\zeta}^{[k+1]}$ are again Lagrange multipliers, $\mathbf{W}^{[k]}$ is an $(2m^2 - m) \times (2m^2 - m)$ diagonal matrix with entries $W_{ll}^{[k]} = \breve{\mu}_l^{[k]} = \breve{\mathrm{e}}_l \exp(\breve{\eta}_l^{[k]})\breve{w}_l$ and $\mathbf{z}^{[k]}$ is an $(2m^2 - m) \times 1$ vector of pseudovalues given by:

$$z_l^{[k]} = \breve{\eta}_l^{[k]} + \left(\frac{\breve{\mathrm{y}}_l}{\breve{\mu}_l^{[k]}} - 1 \right). \tag{10}$$

Here, the penalty term $\mathbf{P}$ in (9) is a $(2m^2 - m) \times (2m^2 - m)$ matrix given by:

$$\mathbf{P} = \lambda_1 \mathbf{I}_m \otimes (\mathbf{D}_v^T \mathbf{D}_v) + \lambda_2 (\mathbf{D}_h^T \mathbf{D}_h) \otimes \mathbf{I}_{2m-1}, \tag{11}$$

where λ_1 and λ_2 are smoothing parameters for, respectively, the vertical and horizontal dimension in Figure 1 right panel, i.e. age and cohort of the original data structure. Optimal smoothing parameters are again computed via grid search using the AIC.

2.3 Cohort smoothing by reordering the penalty matrix ($\mathcal{M}_2$ model)

An altenative approach to smooth over cohorts is to work from the perspective of the penalty matrix without rearranging the original social contact matrix. The methodology

is very similar as the one described in Section 2.1. Matrices $\mathbf{Y}$, $\mathbf{E}$, $\mathbf{P}$, $\mathbf{\Gamma}$ and vectors $\mathbf{y}$, $\mathbf{e}$, $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ are defined similarly as in Section 2.1. Again, a Poisson distribution is assumed for the observed contact rates and the reciprocity constraint is written in matrix form as $\mathbf{L}\boldsymbol{\eta} = \boldsymbol{\nu}$. C-PIRLS is used once more to solve the set of linear equations in (3) for parameter estimation. The penalty matrix $\mathbf{P}$, constructed differently as the penalty term in (5), is now a $m^2 \times m^2$ matrix given by:

$$\mathbf{P} = \lambda_1 \mathbf{I}_m \otimes (\mathbf{D}_h^T \mathbf{D}_h) + \lambda_2 \mathbf{P}_d, \tag{12}$$

where λ_1 and λ_2 are smoothing parameters for, respectively, the horizontal and the diagonal dimension in Figure 1 (left panel), with optimal values chosen by the AIC. The $m^2 \times m^2$ matrix $\mathbf{P}_d$ is responsible for the penalization of the parameters of the cohorts (all diagonals and subdiagonals). For example, in the specific case where $\mathbf{\Gamma}$ is a 4×4 matrix (i.e. $\boldsymbol{\gamma} = \{\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{21}, \ldots, \gamma_{44}\}$), the penalty matrix $\mathbf{P}_d$ is a 16×16 matrix (see appendix A).

A major advantage of using the penalty matrix $\mathbf{P}_d$ to achieve cohort smoothing is the absence of nuisance parameters in the matrix $\mathbf{\Gamma}$ (cf. the approach in the previous section using $\breve{\mathbf{\Gamma}}$). This entails a non-negligible computational gain, since only m^2 parameters in $\mathbf{\Gamma}$ need to be estimated, whereas the $\mathcal{M}_1$ approach requires estimation of $2m^2 - m$ parameters in $\breve{\mathbf{\Gamma}}$ (thus including $m^2 - m$ nuisance parameters). However, the modified penalty matrix $\mathbf{P}_d$ is less trivial to construct. Whereas the penalty in (11) is easily obtained using standard matrix multiplication, the construction of $\mathbf{P}_d$ requires an algorithm (see supplementary materials). This is not a major disadvantage as the construction of $\mathbf{P}_d$ is performed only once and outside the C-PIRLS algorithm without requiring a too high computational budget.

2.4 Kink on the main diagonal of the social contact matrix

The use of smoothing approaches for estimating social contact rates can lead to estimates that are oversmoothed for individuals of the same age, meaning that the estimated contact rate is smaller than the true one in the population. For example, students make an above

average number of contacts with individuals of their own age (e.g., in school, sport clubs, etc.). Smoothing approaches thus, potentially, lead to an underestimation of the social contact rates on the main diagonal of the contact matrix, especially for children and young adults. To take this into account, we introduce the use of a so-called *kink* on the main diagonal of the social contact matrix for $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_2$, that can force a sudden increase (or decrease) of the estimated social contact rates for children and young adults of the same age.

The kink is introduced through a small adjustment in the penalty matrices in (11) and (12). More specifically, in the dimension of the contact's age, the social contact rates that belong to the main diagonal, i.e. η_{ii} and γ_{ii}, are not penalized. In (11) this is achieved by changing the $(2m-3) \times (2m-1)$ matrix $\mathbf{D}_v$ as follows:

$$\mathbf{D}_v^* = \begin{array}{c} \\ \vdots \\ m-3 \\ m-2 \\ m-1 \\ m \\ m+1 \\ m+2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccccc} \ldots & m-3 & m-2 & m-1 & m & m+1 & m+2 & m+3 & \ldots \end{array} \\ \begin{bmatrix} & & & & & & & & \\ & 1 & -2 & 1 & & & & & \\ & & 1 & -1 & 0 & & & & \\ & & & 1 & 0 & -1 & & & \\ & & & & 0 & -1 & 1 & & \\ & & & & & 1 & -2 & 1 & \\ & & & & & & 1 & -2 & \\ & & & & & & & & \end{bmatrix} \end{array}. \tag{13}$$

From the above matrix $\mathbf{D}_v^*$, it is clear that the social contact rates that belong to the main diagonal, namely, η_{ii} and γ_{ii}, are not penalized since column m only has zero values. The penalty matrix in (11) is now reformulated as follows:

$$\mathbf{P} = \lambda_1 \left(\mathbf{I}_m^{(1)} \otimes (\mathbf{D}_v^{*T} \mathbf{D}_v^*) + \mathbf{I}_m^{(2)} \otimes (\mathbf{D}_v^T \mathbf{D}_v) \right) + \lambda_2 (\mathbf{D}_h^T \mathbf{D}_h) \otimes \mathbf{I}_{2m-1}, \tag{14}$$

where $\mathbf{I}_m^{(1)}$ and $\mathbf{I}_m^{(2)}$ are diagonal indicator matrices given by:

$$\mathbf{I}_m^{(1)} = \{ \underbrace{1, \ldots, 1}_{\times \text{ max.kink.age}} , \underbrace{0, \ldots, 0}_{\times \text{ m-max.kink.age}} \} \text{ and}$$

$$\mathbf{I}_m^{(2)} = \{ \underbrace{0, \ldots, 0}_{\times \text{ max.kink.age}} , \underbrace{1, \ldots, 1}_{\times \text{ m-max.kink.age}} \},$$

where *max.kink.age* indicates the maximum age at which a kink on the main diagonal is possible. In this paper, we calibrate *max.kink.age* = 31 (i.e., $\{0, \ldots, 30\}$ years), although a sensitivity analysis with higher values for *max.kink.age* yielded quantitatively similar results. In penalty matrix (12), a similar adjustment is applied to the matrix $\mathbf{D}_h$. It is worth noting that social contact rates on the main diagonal that are adjusted by the kink are still penalized in the dimension of the cohort to assure that smooth contact rates are obtained on the diagonals of the contact matrix. The introduction of this kink thus leads to a smoothed contact surface that is non-differentiable on the main diagonal in the dimension of the contact's age.

2.5 Negative Binomial Likelihood

Using the Poisson distribution for the observed contacts y_{ij} implies that the mean and the variance are equal, i.e. $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij})$, while in practice, contact data often display overdispersion. Not accounting for possible overdispersion can lead to biased results. We therefore also impose a negative Binomial distribution for the observed contacts, namely $y_{ij} \sim \mathrm{NegBin}(\mu_{ij}, \alpha_{ij})$. The use of a negative Binomial distribution implies that $\mathrm{E}(Y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}^2 \alpha_{ij}^{-1}$. Here, we consider the parameterization with $\alpha_{ij} = \mu_{ij}\phi^{-1}$, where $\phi > 0$ denotes the disperion parameter and the variance is given by $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1+\phi)$. In the limiting case where ϕ tends to zero, the mean and variance will be equal. Note that the variance term resembles the error term of an overdispersed Poisson distribution (Nelder and Lee, 1992). The alternative parameterization with $\alpha_{ij} = \phi^{-1}$ (leading to $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1 + \phi\mu_{ij})$) was also explored but not further described since it performed worse in terms of AIC for the application on the Belgian contact data.

In case ϕ is fixed at a certain value, parameter estimates $\hat{\boldsymbol{\eta}}$ are again obtained through C-PIRLS. The only adaptation is that the entries of $\mathbf{W}^{[k]}$ are given by $W_{ll}^{[k]} = \mu_l^{[k]}/(1+\phi)$. Rather than fixing ϕ at a certain value, we are also interested in obtaining a data-driven estimate of ϕ. In that endeavor, a two-stage iteration scheme is undertaken, namely by iterating and cycling between holding ϕ fixed and holding $\boldsymbol{\eta}$ fixed at its current estimate. More specifically, by holding ϕ fixed at the current estimate $\hat{\phi}^{[k]}$, estimates $\hat{\boldsymbol{\eta}}^{[k+1]}$ are obtained through C-PIRLS. Next, $\boldsymbol{\eta}$ is fixed at $\hat{\boldsymbol{\eta}}^{[k+1]}$ and an updated estimate $\hat{\phi}^{[k+1]}$ is obtained using the moment estimator (Breslow, 1984). This process is iterated until convergence. Moment estimation of ϕ is based on the Pearson's chi-squared statistic (Breslow, 1984), namely:

$$\sum_{i,j=1}^{m} \frac{\left(y_{ij} - \mu_{ij}^{[k]}\right)^2}{(1+\phi)\mu_{ij}^{[k]}} = m^2 - \widehat{ED}, \tag{15}$$

where $\widehat{ED}$ is the trace of the matrix given in (7). This leads to a straightforward estimate of $\hat{\phi}^{[k]}$:

$$\hat{\phi}^{[k]} = \frac{1}{m^2 - \widehat{ED}} \sum_{i,j=1}^{m} \frac{\left(y_{ij} - \mu_{ij}^{[k]}\right)^2}{\mu_{ij}^{[k]}}. \tag{16}$$

Optimal smoothing parameters λ_1 and λ_2 are again chosen via a grid search using the criterion AIC $= -2\log(\hat{L}) + 2(\widehat{ED} + 1)$. Adding 1 to $\widehat{ED}$ accounts for the estimation of the additional ϕ parameter in the negative Binomial setting.

2.6 Quantifying the uncertainty of estimates

In order to quantify the uncertainty of the estimate $\hat{\boldsymbol{\eta}}$, we need to compute its associated variance-covariance matrix. For this purpose, we follow Wood (2006) and use a Bayesian approach to determine the posterior variance-covariance matrix by:

$$\mathbf{V}_{\boldsymbol{\eta}} = (\mathbf{W} + \mathbf{P})^{-1}. \tag{17}$$

Moreover, as justified by large sample results, the corresponding posterior distribution is

taken to be multivariate normal:

$$\boldsymbol{\eta} \sim \mathcal{N}\left(\hat{\boldsymbol{\eta}}, \mathbf{V}_{\boldsymbol{\eta}}\right). \tag{18}$$

The above (approximate) posterior distribution can be used to calculate confidence intervals for parameters η_{ij} or for non-linear functions of these parameters (such as γ_{ij}). An estimate of $\mathbf{V}_{\boldsymbol{\eta}}$ can be obtained by plugging in $\mathbf{W}$ at convergence together with the estimated optimal smoothing parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in $\mathbf{P}$. The result in (18) can also be used to generate *new* social contact matrices by sampling from the obtained multivariate Gaussian distribution. This can be extremely useful to acknowledge the variability originating from social contact data in the estimation of epidemiological parameters and/or health economic evaluations (Bilcke et al., 2011). Further computational and algorithmic considerations related to C-PIRLS are given in appendix B.

3 Simulation Study

A comparison of the methods introduced in Sections 2.1-2.3 is implemented via a simulation study, where the observed contact rates are generated from a Poisson and negative Binomial distribution respectively. We investigate a scenario in which no kink is needed on the main diagonal, and another scenario in which the kink is specified.

Our data generating process is based on a so-called true social contact matrix, denoted by $\boldsymbol{\Gamma}^*$, from which data is simulated. To obtain such a matrix, a non-parametric regression is applied to the Belgian social contact data. More specifically, the observed contacts rates (see Figure 2 left panel), y_{ij}/r_i are smoothed using local linear regression. Using a local linear regression approach, there is no guarantee that $\mathbf{K}^* \equiv \boldsymbol{\Gamma}^* \odot \mathbf{P}$ is symmetric. Therefore, we derive a simple symmetric matrix from $\mathbf{K}^*$, denoted by $\breve{\mathbf{K}}^*$, computed as:

$$\left(\breve{\mathbf{K}}^*\right)_{ij} = \left(\breve{\mathbf{K}}^*\right)_{ji} = \frac{(\mathbf{K}^*)_{ij} + (\mathbf{K}^*)_{ji}}{2}.$$

The true contact surface, $\breve{\mathbf{\Gamma}}^*$ that is used for data simulation is obtained by $\breve{\Gamma}^*_{ij} = \breve{K}^*_{ij}/P_{ij}$. Finally, we denote the log-transformed matrix by $H^*_{ij} = \log\left(\breve{\Gamma}^*_{ij}\right)$. In Figure 3, the true social contact matrices used to generate the data for the simulation study $\breve{\mathbf{\Gamma}}^*$ and $\mathbf{H}^*$ are shown. To account for a kink in the simulation study, we proceed as follows. Let $\breve{\mathbf{\Gamma}}^\dagger$ denote the true social contact matrix with a kink on the main diagonal. Matrix $\breve{\mathbf{\Gamma}}^\dagger$ is similar as matrix $\breve{\mathbf{\Gamma}}^*$, with the exception that the values of $\breve{\Gamma}^\dagger_{ii}$, for $i = 1, \ldots, 24$, are artificially increased in the following manner:

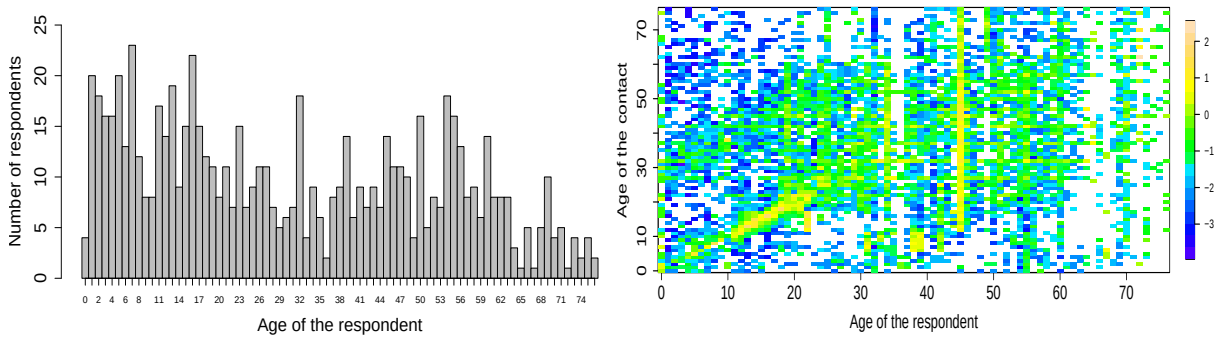

Figure 2: The number of respondent per age (left panel) and the observed log-contact rates ($\log(y_{ij}/r_i)$) (right panel) of the Belgian social contact data. A white cell indicates that there were no contacts observed for those particular ages of the respondents and contacts.

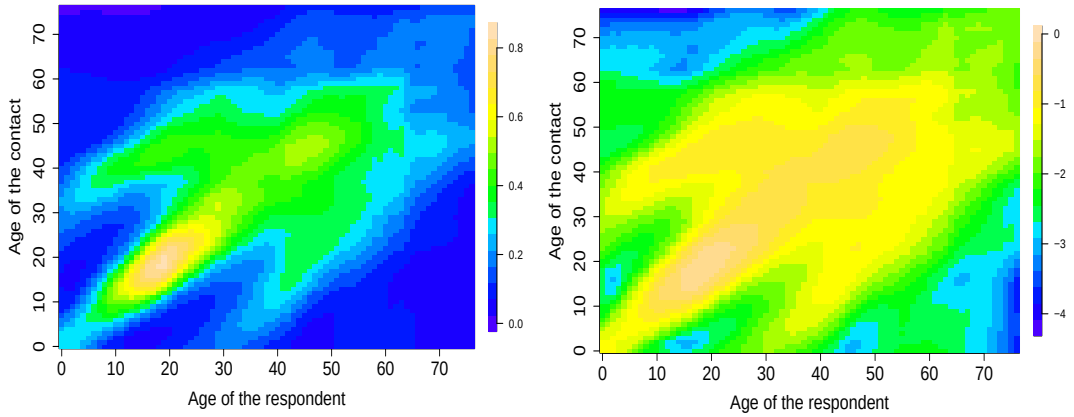

Figure 3: True social contact matrices $\breve{\mathbf{\Gamma}}^*$ (left) and $\mathbf{H}^*$ (right) used for the data generating process in the simulation study. The true social contact surfaces are obtained from a non-parametric regression using a local linear fit to the Belgian social contact data.

$$\breve{\Gamma}^{\dagger}_{ii} = \begin{cases} \breve{\Gamma}^{*}_{ii}\left(1 + \frac{1}{11}(i-1)\right) & i \in \{1,\ldots,12\}, \\ \breve{\Gamma}^{*}_{ii}\left(2.0 - \frac{1}{11}(i-13)\right) & i \in \{13,\ldots,24\}, \\ \breve{\Gamma}^{*}_{ii} & i > 24. \end{cases}$$

Thus for ages between 0 and 23 a higher number of contacts is obtained on the main diagonal. Data is simulated using the same participant distribution as in the Belgian social contact data with sample size $n = 745$ (see Figure 2 left). For the Poisson distribution, data is simulated as follows:

$$y^{*}_{ij} \sim \text{Pois}\left(r_i \breve{\Gamma}^{*}_{ij}\right). \tag{19}$$

For the negative Binomial distribution (with $\phi = 2$), the observed number of contacts are obtained as:

$$y^{*}_{ij} \sim \text{NegBin}\left(\mu_{ij} = r_i \breve{\Gamma}^{*}_{ij}, \alpha_{ij} = \mu_{ij} 2^{-1}\right). \tag{20}$$

We simulate $S = 100$ datasets for each distributional setting (i.e. Poisson and negative Binomial) and fit models $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_3$ to each dataset with and without consideration of a kink. Optimal smoothing parameters are obtained via grid search using the AIC. This yields estimated social contact matrices $\hat{\mathbf{\Gamma}}^{(s)}$ and $\hat{\mathbf{H}}^{(s)}$, for $s = 1,\ldots,S$. The estimation performance of the different methods are compared using the squared bias and mean square error (MSE). These scalar measures of performance are given by:

$$\text{Bias}^2_{\Gamma} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\breve{\Gamma}^{*}_{ij} - \hat{\Gamma}^{(s)}_{ij}\right)\right)^2, \tag{21}$$

$$\text{Bias}^2_{H} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(H^{*}_{ij} - \hat{H}^{(s)}_{ij}\right)\right)^2, \tag{22}$$

$$\text{MSE}_{\Gamma} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\breve{\Gamma}^{*}_{ij} - \hat{\Gamma}^{(s)}_{ij}\right)^2\right), \tag{23}$$

$$\text{MSE}_{H} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(H^{*}_{ij} - \hat{H}^{(s)}_{ij}\right)^2\right). \tag{24}$$

Besides the performance of pointwise estimators given in Tables 1 and 2, we also assess the accuracy with which uncertainty is quantified by looking at the coverage performance of 95% pointwise confidence intervals (CIs) of η_{ij} in Table 3. Using the approximate posterior distribution in (18), 95% pointwise CIs are easily calculated (i.e., $\pm 1.96\times$ the square root of the Bayesian posterior variance). The reported nominal coverages of the CIs are calculated by averaging over the entries of the entire social contact matrix.

For all simulation settings, we observe that models that smooth over cohorts ($\mathcal{M}_1$ and $\mathcal{M}_2$) are performing better in terms of MSE than $\mathcal{M}_0$, and this holds for both $\mathbf{H}^*$ and $\breve{\mathbf{\Gamma}}^*$. In terms of bias, the results are less clear, but overall model $\mathcal{M}_2$ is performing best. When comparing models $\mathcal{M}_1$ and $\mathcal{M}_2$, we observe that the latter model has better performance. In the simulation settings in which no kink is introduced on the main diagonal, we observe that models with a kink on the main diagonal perform slightly worse than those without a kink. However, in the simulation settings with a kink, a more pronounced difference is observed in favour of the models with a kink on the main diagonal, especially for $\breve{\mathbf{\Gamma}}^*$. The better performance of models including a kink is mainly due to the better estimation of the main diagonal components of the social contact matrix. No meaningful differences are observed outside the main diagonal region.

Table 1: Squared bias of the social contact matrices $\mathbf{H}^*$ and $\breve{\mathbf{\Gamma}}^*$ over $S = 100$ simulations using $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_3$ with and without a kink on the main diagonal.

Squared bias		Models without kink on main diagonal					
		bias^2 of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			bias^2 of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)		
Simulation setting		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Poisson	w/o kink	69.62	58.62	49.41	1.53	1.39	1.32
NegBin	w/o kink	93.16	91.53	77.76	2.50	2.63	2.52
Poisson	w kink	57.67	60.86	51.79	3.32	3.37	3.31
NegBin	w kink	96.52	82.14	70.44	4.77	4.38	4.31
Squared bias		Models with kink on main diagonal					
		bias^2 of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			bias^2 of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)		
Simulation setting		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Poisson	w/o kink	/	58.92	49.66	/	1.49	1.43
NegBin	w/o kink	/	93.64	79.42	/	3.05	2.92
Poisson	w kink	/	58.57	49.33	/	1.53	1.47
NegBin	w kink	/	80.70	68.63	/	2.62	2.53

Table 2: Mean square error of the social contact matrices $\mathbf{H}^*$ and $\breve{\mathbf{\Gamma}}^*$ over $S = 100$ simulations using $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_3$ with and without a kink on the main diagonal.

MSE		Models without kink on main diagonal					
		MSE of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			MSE of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)		
Simulation setting		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Poisson	w/o kink	90.81	74.45	68.05	2.41	1.98	1.94
NegBin	w/o kink	154.73	130.41	123.72	4.79	3.99	3.96
Poisson	w kink	82.36	77.78	71.85	4.33	3.98	3.95
NegBin	w kink	156.94	123.59	120.50	7.11	5.86	5.87
MSE		Models with kink on main diagonal					
		MSE of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			MSE of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)		
Simulation setting		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
Poisson	w/o kink	/	75.13	68.67	/	2.18	2.14
NegBin	w/o kink	/	133.15	126.00	/	4.57	4.51
Poisson	w kink	/	75.72	69.63	/	2.28	2.24
NegBin	w kink	/	122.71	119.25	/	4.41	4.40

In the negative Binomial simulation setting, the overdisperion parameter ϕ is estimated well. In the simulation setting without a kink, model $\mathcal{M}_2$ without a kink has an average estimate for ϕ of 1.92 with 95% of the estimated overdispersion parameters between 1.74 and 2.22. For the simulation setting with a kink, we find 1.93 (1.71 - 2.20) for model $\mathcal{M}_2$.

Table 3: Nominal coverage of 95% pointwise confidence intervals of the social contact matrices $\mathbf{H}^*$ ($\mathbf{H}^\dagger$) over $S = 100$ simulations using $\mathcal{M}_0, \mathcal{M}_1$ and $\mathcal{M}_3$ with and without a kink on the main diagonal. The nominal coverage is calculated by averaging over the entire social contact matrix.

		Models without kink on main diagonal			Models with kink on main diagonal	
Simulation setting		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$
Poisson	w/o kink	92.06	93.86	95.47	93.57	95.16
NegBin	w/o kink	95.10	94.51	95.92	93.90	95.36
Poisson	w kink	94.79	93.17	94.76	93.48	95.06
NegBin	w kink	95.01	96.26	97.26	96.22	97.26

Table 3 highlights the nominal coverage results of the different simulation settings. We observe that all methods produce pointwise CIs with close to 95% nominal coverage. In the last simulation setting (the negative Binomial distribution with a kink on the main diagonal), a slight overcoverage is observed for models $\mathcal{M}_1$ and $\mathcal{M}_2$. In the latter scenarios, the average lengths of the 95% CIs are 0.65, 0.61 and 0.60, for $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ with a kink, respectively. This implies that the overcoverage is not directly associated with wider

CIs. Finally, the results in Table 3 indicate that the large sample result in (18) can be used to construct CIs with appropriate nominal coverage.

4 Application: Belgian Social Contact Data

The proposed smoothing methods are illustrated on the POLYMOD social contact data of Belgium, obtained through a population-based contact survey carried out over the period of March to May 2006. Participants kept a paper diary with information on their contacts over one day. A contact was defined as a two-way conversation of at least three words in each other's proximity and the gathered information included the age of the contact, gender, location, duration, frequency, and whether or not touching was involved. Sampling weights – the inverse of the probability that an observation is included because of the sampling design – are available for each participant, based on official age and household size data of the year 2000 census published by Eurostat (Mossong et al., 2008). To estimate population-related social contacts, these sampling weights are included in the analysis. We consider the contact data of all participants aged between 0 and 76 years (both included). In total, we have information on 745 participants from which 399 (53.6%) are females and 345 (46.3%) are males (the information on gender was omitted for one participant). The mean age of the respondents is 31 years. We also restrict to contacts made with individuals between 0 and 76 years (both included), resulting in a total of 13 493 contacts. This gives a crude mean of 18.1 contacts per participant. Furthermore, the age structure of the general population in which the contact survey is conducted in 2006 is obtained from Eurostat (Eurostat, 2017), where the population size in the 0-76 years interval is N=9 777 488.

In Figure 2 (right panel), the observed log-contact rates $\log(y_{ij}/r_i)$ of the POLYMOD Belgian social contact data are shown. To estimate the social contact rates from these data, we use models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ under a Poisson and negative Binomial assumption on the number of contacts with and without a kink on the main diagonal. Let w_k^* denote the normalized sampling weight of respondent k, with $k = 1, \ldots, 745$. The ijth input of $\mathbf{Y}$ is constructed as $(\mathbf{Y})_{ij} = y_{ij} = \sum_{k \in \text{ age } (i-1)} w_k^* y_{k,(i-1,j-1)}$ and corresponds to a weighted

sum of the number of contacts made by respondents of age $i-1$ with contacts of age $j-1$. It follows that the inputs of the vector $\mathbf{r}$ are given by $r_i = \sum_{k \in \text{ age } (i-1)} w_k^*$.

In Table 4, summary results of the fitted models are given. It can be seen that the negative Binomial distribution performs better in terms of the AIC (smaller AIC is "better"), so that the assumption of a variance that is linearly dependent on the mean is preferred. The effective degrees of freedom ($\widehat{ED}$) for the Poisson case are higher, indicating that the Poisson distribution tries to explain the observed variability through the mean. From here, we focus on the results of the negative Binomial model. It can be observed that approaches $\mathcal{M}_1$ and $\mathcal{M}_2$ including a kink are performing better in terms of the AIC as compared to those without a kink. Regarding the estimated smoothing parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$, an interesting difference is observed between $\mathcal{M}_0$ and models $\mathcal{M}_1$ and $\mathcal{M}_2$. In $\mathcal{M}_0$, the optimal values for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are similar, while for the models accounting for cohort smoothing, the optimal value for $\hat{\lambda}_2$ is larger than $\hat{\lambda}_1$, indicating that more penalization is needed in the direction of the cohorts. In terms of computational speed, fitting model $\mathcal{M}_2$ is much faster ($\approx$ 4 times faster) as compared to model $\mathcal{M}_1$, as for the latter model $2m^2 - m = 11\ 781$ parameters (including $m^2 - m$ nuisance parameters) need to be estimated, as compared to $m^2 = 5\ 929$ parameters for $\mathcal{M}_2$.

Table 4: Summary results of the fitted models to the Belgian social contact data. Estimated smoothing parameters, effective degrees of freedom, -2 times log-likelihood, AIC and ϕ are provided.

Model	$\hat{\lambda_1}$	$\hat{\lambda_2}$	$\widehat{ED}$	$-2\log(\hat{L})$	AIC	$\hat{\phi}$
$\mathcal{M}_0$ Poisson	0.46	0.46	1 979.5	20 151.7	24 110.7	-
$\mathcal{M}_0$ NegBin	15.17	16.64	181.5	20 732.7	21 097.7	3.08
$\mathcal{M}_1$ Poisson	0.50	0.32	2 085.6	20 318.4	24 489.6	-
$\mathcal{M}_1$ NegBin	22.76	1714.91	55.8	20 988.5	21 102.1	3.76
$\mathcal{M}_2$ Poisson	0.50	0.32	2 125.0	20 287.6	24 537.7	-
$\mathcal{M}_2$ NegBin	27.36	1564.02	59.5	20 994.2	21 115.2	3.76
With kink						
$\mathcal{M}_1$ NegBin	30.00	1584.89	54.3	20 967.6	21 078.2	3.70
$\mathcal{M}_2$ NegBin	40.00	1584.89	55.4	20 973.2	21 086.0	3.70

In Figures 4 and 5, the estimated log contact rate surfaces, $\hat{\mathbf{H}}$, and the mixing at the population level, $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ under the negative Binomial distribution without a kink are shown. Generally, the surfaces are able to capture important features of human contact behaviour. There is a clear difference in the estimated surfaces for model $\mathcal{M}_0$ and models $\mathcal{M}_1$ and $\mathcal{M}_2$ in the sense that diagonal components are more pronounced for models accounting for cohort smoothing. The shifted diagonal between children and parents is also more clearly visible.

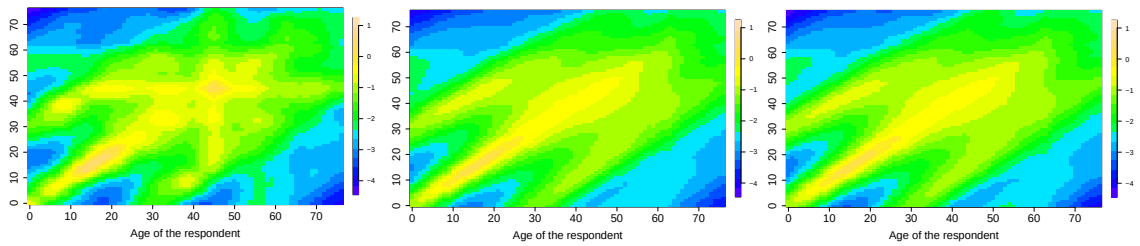

Figure 4: The estimated log contact rates surface, $\hat{\mathbf{H}}$, for models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ without kink (left to right) with the negative Binomial distribution.

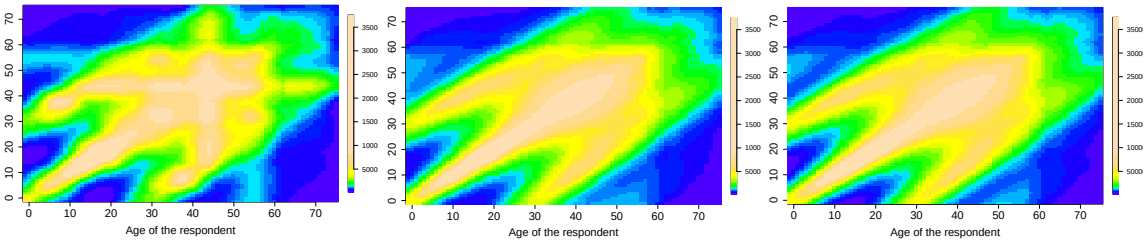

Figure 5: The estimated mixing at the population level, $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$ without kink (left to right) with the negative Binomial distribution.

Based on the AIC values in Table 4, we see that the models including a kink are preferred. In addition, based on the results of the simulation study in Section 3, the estimated contact rates are very similar for models $\mathcal{M}_1$ and $\mathcal{M}_2$, so that we prefer the use of model $\mathcal{M}_2$ for the POLYMOD Belgian social contact data as it is less computationally intensive.

In Figure 6, estimated contact surfaces are shown for model $\mathcal{M}_2$ with the negative Binomial distribution and a kink on the main diagonal. From the figure on the bottom, it is observed that the main diagonal has higher values for younger ages for the model including the kink and this yields higher values on the main diagonal of $\hat{\mathbf{H}}$ and $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$. For the model where the kink is absent, the values in the estimated matrix $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$ range from

1 496.2 to 162 986.5, whereas for the model with kink, the values range from 1 608.4 to 375 371.5. The kink thus allows for a huge increase in the estimated number of contacts for children and young adults with individuals of the same age. These results enforce the fact that a kink is needed to capture the non-smooth effect of mixing with people of the same age, especially for the children and young adults.

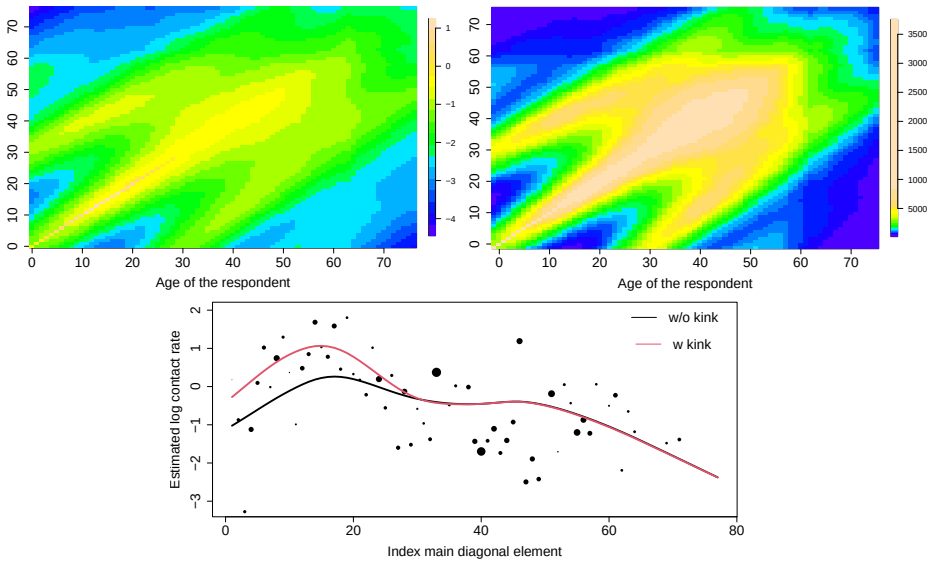

Figure 6: The estimated log contact rates surface (left), $\hat{\mathbf{H}}$, and the mixing at the population level (right), $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for model $\mathcal{M}_2$ with the negative Binomial distribution including an additional kink on the main diagonal. The diagonal elements of $\hat{\mathbf{H}}$ for the model with and without a kink (bottom), together with the observed log-contact rates.

5 Discussion

Quantifying contact behaviour contributes to a better understanding of how infectious diseases spread (Anderson and May, 1991; Edmunds et al., 1997). Social contact rates play a major role in mathematical models used to model infectious disease transmission. In this paper, we describe a smoothing constrained approach to estimate social contact rates from self-reported social contact data. The proposed approach assumes that the contact rates are smooth from a cohort perspective as well as from the age distribution of contacts. Thus, besides smoothing in the direction of the age of contacts, we propose to smooth contact rates from a cohort perspective by following two alternative strategies.

The simulation study and the data application show that approach $\mathcal{M}_2$, in which the penalty matrix is reordered (and penalization is performed over the diagonal components), is performing better. It was observed that this method yielded the smallest MSE over all simulation settings. Additionally, confidence intervals with nominal coverage close to 95% were obtained. In the Belgian data application, the computation time of method $\mathcal{M}_2$ is three to four times faster than method $\mathcal{M}_1$, and so we recommend the use of the former approach for the estimation of social contact rates. The true social contact surface used in the data generating process of the simuation study was obtained through local linear regression of the raw social contact rates of the Belgian POLYMOD study. This approach is preferred for two reasons. First, by using the same data in the simulation study as in the application presented in Section 4, a better view of the performance of the different approaches can be obtained. Second, we are not aware of any easy applicable mathematical formula or fully parametric model of a two dimensional surface that would be suitable to represent a contact rate surface. A search is needed to calibrate the smoothing parameters λ_1 and λ_2. This is a disadvantage compared to the approach by van de Kassteele et al. (2017) in which the amount of smoothing is directly estimated together with model parameters from the information in the data. However, with the availability of fast parallel computing and multi-core machines, the grid search can be performed relatively fast.

In this paper, the contact rates are assumed indifferent for men and women. Recently, van de Kassteele et al. (2017) presented a Bayesian model for estimating social contact rates for men and women, with results suggesting that different contact patterns exist and thus that there is a gender effect. Future work could investigate how the methodology proposed in this paper can be extended to estimate social contact rates between both sexes without increasing the computational burden. A comparison with other methods used to smooth social contact data was not done in this paper. Future extensions could focus on the impact of social contact matrices obtained from different methods on key epidemiological parameters. In general, age-specific contact rates are also used as an input in the comparison and evaluation of vaccination schedules via future projections (Beutels

et al., 2013). Most evaluations assume a fixed social contact rate matrix and thus that no uncertainty is related to this input. The result derived in equation (18) offers a tool to account for the variability associated with the estimation of social contact rates. By simulation of new contact matrices from (18), the associated variability can be taken into account in the evaluation of vaccination strategies and related health economic evaluations.

Finally, our proposed methodology does not employ any regression basis such as B-splines because an exact link between the constraints and linear predictors is needed. We are exploring whether the proposed methodology can be extended to make use of basis functions that will likely lead to a reduction of the computational cost. Alternative ways of incorporating the reciprocal nature of the phenomenon will thus be necessary.

Acknowledgments

For the simulation study with the negative Binomial model, we used the infrastructure of the VSC –Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government– department EWI. Support from the University of Antwerp scientific chair in Evidence-Based Vaccinology, financed in 2009-2017 by a gift from Pfizer, is acknowledged [to NH]. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 682540 - TransMID). Oswaldo Gressani, Niel Hens and Christel Faes would also like to thank the European Union's Research and Innovation Action EpiPose (grant number 101003688) for funding this work.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability statement

Code to reproduce the results of this paper is available at `https://github.com/oswaldogressani/Cohort_smoothing`.

Appendix

Appendix A. Penalty matrix $\mathbf{P}_d$

$$\mathbf{P}_d = \begin{array}{c} \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{14} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \\ \gamma_{24} \\ \gamma_{31} \\ \gamma_{32} \\ \gamma_{33} \\ \gamma_{34} \\ \gamma_{41} \\ \gamma_{42} \\ \gamma_{43} \\ \gamma_{44} \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} & \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{array} \\ \begin{bmatrix} 1 & & & & & -2 & & & & & 1 & & & & & \\ & 1 & & & & & -2 & & & & & 1 & & & & \\ & & 1 & & & & & -1 & & & & & & & & \\ & & & 0 & & & & & & & & & & & & \\ & & & & 1 & & & & & -2 & & & & & 1 & \\ -2 & & & & & 5 & & & & & -4 & & & & & 1 \\ & -2 & & & & & 4 & & & & & -2 & & & & \\ & & -1 & & & & & 1 & & & & & & & & \\ & & & & & & & & 1 & & & & & -1 & & \\ & & & & -2 & & & & & 4 & & & & & -2 & \\ 1 & & & & & -4 & & & & & 5 & & & & & 2 \\ & 1 & & & & & -2 & & & & & 1 & & & & \\ & & & & & & & & & & & & 0 & & & \\ & & & & & & & & -1 & & & & & 1 & & \\ & & & & 1 & & & & & -2 & & & & & 1 & \\ & & & & & 1 & & & & & -2 & & & & & 1 \end{bmatrix} \end{array}.$$

Appendix B. Computational considerations.

`R` version 4.1.2 is used to fit the proposed models. To enhance convergence of the proposed C-PIRLS fitting scheme, we first perform parameter estimation using penalized iterative reweighted least squares without using the symmetry constraint and use the obtained estimated parameters as starting values in the C-PIRLS algorithm. To initiate the estimation of PIRLS without the symmetry constraint, starting values $\hat{\boldsymbol{\eta}}^{[0]}$ are needed. These can, for example, be set at $\hat{\boldsymbol{\eta}}^{[0]} = \log\left((\mathbf{y}+1)/(\mathbf{e}+1)\right)$.

The same number of parameters are estimated as there are entries in the matrices $\boldsymbol{\Gamma}$ or $\breve{\boldsymbol{\Gamma}}$. For instance, in our application ($m = 77$) in Section 4, we need to estimate $m^2 = 5\,929$ and $2m^2 - m = 11\,781$ parameters, respectively. This is practically challenging on a regular personal computer. Therefore, we make use of sparse matrix implementations by using the `R`-package `Matrix` (Bates and Maechler, 2017).

To choose the optimal smoothing parameters λ_1 and λ_2, a grid search is performed with

both $\lambda_1, \lambda_2 \in \{0.5, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000\}$. This initial grid search gives an indication (based on minimization of the AIC) of the values of the optimal smoothing parameters. In a second step, a greedy grid search is performed an a denser grid using the `cleversearch` function in the `R`-package `svcm` (Heim, 2007).

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60:255–265.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Bates, D. and Maechler, M. (2017). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-8.

Beutels, P., Shkedy, Z., Aerts, M., and Van Damme, P. (2006). Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiology and Infection*, 134(6):1158–1166.

Beutels, P., Vandendijck, Y., Willem, L., Goeyvaerts, N., Blommaert, A., Van Kerckhove, K., Bilcke, J., Hanquet, G., Neels, P., Thiry, N., Liesenborgs, J., and Hens, N. (2013). Seasonal influenza vaccination: prioritizing children or other target groups? Part II: Cost-effectiveness analysis. *KCE Report 204, Health Technology Assessment*.

Bilcke, J., Beutels, P., Brisson, M., and Jit, M. (2011). Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: A practical guide. *Medical Decision Making*, 31(4):675–692.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(1):38–44.

Edmunds, W. J., Kafatos, G., Wallinga, J., and Mossong, J. R. (2006). Mixing patterns and the spread of close-contact infectious diseases. *Emerging Themes in Epidemiology*, 3(1):10.

Edmunds, W. J., O'callaghan, C. J., and Nokes, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1384):949–957.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eurostat (2017). Population table for belgium 2006. *Eurostat, Luxembourg. (Available from http://epp.eurostat.ec.europa.eu/)*.

Farrington, C. P., Kanaan, M. N., and Gay, N. J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society. Series C - Applied Statistics*, 50(3):251–283.

Farrington, C. P. and Whitaker, H. J. (2005). Contact surface models for infectious diseases. *Journal of the American Statistical Association*, 100(470):370–379.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Van Damme, P., and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):255–277.

Greenhalgh, D. and Dietz, K. (1994). Some bounds on estimates for reproductive ratios derived from the age-specific force of infection. *Mathematical Biosciences*, 124(1):9 – 57.

Heim, S. (2007). *svcm: 2d and 3d space-varying coefficient models in R*. R package version 0.1.2.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P., and Beutels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in belgium. *BMC infectious diseases*, 9(1):1–18.

Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

McCullagh, p. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd edition.

Mikolajczyk, R., Akmatov, M., Rastin, S., and Kretzschmar, M. (2007). Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology and Infection*, 136(6):813–822.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Medicine*, 5(3):1–1.

Nelder, J. A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):273–284.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

van de Kassteele, J., van Eijkeren, J., and Wallinga, J. (2017). Efficient estimation of age-specific social contact rates between men and women. *Ann. Appl. Stat.*, 11(1):320–339.

Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P., and Beutels, P. (2009). Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidemiology & Infection*, 137(1):48–57.

Vynnycky, E. and White, R. (2010). *An Introduction to Infectious Disease Modelling*. New York: Oxford University Press.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, 164:936–944.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.