bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

last update: March 28, 2018

Cohort-based Smoothing Methods for Age-specific Contact Rates

Yannick Vandendijck[1], Carlo G. Camarda[2], Niel Hens[1,3]

[1]Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium

[2]Institut National d'Études Démographiques (INED), Paris, France

[3]Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

Corresponding author: yannick.vandendijck@uhasselt.be

Abstract

The use of social contact rates is widespread in infectious disease modelling, since it is known that they provide proxies of crucial determinants of epidemiological and disease transmission parameters. Information on social contact rates can, for example, be obtained from a population-based contact diary survey, such as the POLYMOD study. Estimation of age-specific contact rates from these studies is often done using bivariate smoothing techniques. Typically, smoothing is done in the dimensions of the respondent's and contact's age. In this paper, we introduce 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 contact rates should vary smoothly by cohorts. Two approaches that allow for smoothing of social contact data over cohorts are proposed: (1) reordening of the diagonal components of the social contact rate matrix; and (2) reordening of the penalty matrix associated with the diagonal components. Parameter estimation is done using constrained penalized iterative reweighted least squares. A simulation study is presented to compare methods. The proposed methods are illustrated on the Belgian POLYMOD data of 2006.

Key words: Contact rates; Penalized iterative reweighted least squares; Penalized likelihood; Smoothing; Social contact data

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

The use of social contact mixing patterns is omnipresent in statistical and mathematical models of infectious disease transmission. Mixing patterns 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). 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 contact surface and estimate parameters from serological data. Both approaches are some what *ad hoc* in the sense that they require assumptions about the structure of the WAIFW matrix and the parametric model used for the continuous contact surface.

Alternatively, over the last two decades, 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; Van Hoang et al., 2018). The European commission project POLYMOD is arguably one of the most important studies to date (Mossong et al., 2008). The study is a large and representative population based survey on social contacts recorded on a randomly assigned day in 8 European countries (Belgium, England and Wales, Finland, Germany, Italy, Luxembourg, Poland and The Netherlands). Social contact rates estimated from these self-reported data have been used in ample studies to investigate the spread of infectious disease transmission (see e.g., Goeyvaerts et al., 2010; Abrams and Hens, 2015).

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. For smoothing purposes, but in addition to ensure enough flexibility, a bivariate smoothing approach (within the likelihood framework) is undertaken using a tensor product spline as a function of the respondent's and contact's age as a smooth interaction term (Mossong et al., 2008; Hens et al., 2009; Goeyvaerts et al., 2010). Recently, also hierarchical Bayesian models have been used for social contact rates estimation (van de Kassteele et al., 2017). When estimating the social contact rates, the reciprocal nature of contacts needs to be taken into account (which 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) and this will be achieved by the proposed method. We propose a smoothing constrained approach where the contact rates are assumed smooth from a cohort perspective as well as from the age distribution of contacts. Thus, smoothing in the direction of the age of contacts will remain, however instead of smoothing over the dimension of the age of respondents, we will smooth contact rates from a cohort

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

perspective. This is achieved by smoothing of the social contact rates over the diagonal components (including all subdiagonals). In this manner the social contact rates are modelled from a cohort perspective, namely people age through time and we assume that contact rates for consecutive time points are similar. The maximum likelihood framework is used for parameter estimation.

In the current paper, we describe two approaches that allow for smoothing of social contact data over the diagonal components: (1) reordening of the diagonal components to reproduce a rectangular grid; and (2) reordening of the penalty matrix such that penalization is performed over the diagonal components. The first approach builds further upon work published by two of the co-authors in a proceedings paper (Camarda et al., 2013). Poisson or negative binomial distributions are assumed for the aggregated number of contacts.

The use of smoothing approaches for estimating social contact rates could lead to contact rate estimates that are oversmoothed for individuals of the same age, meaning that the estimated contact rate is smaller than 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 social 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 that can force a sudden increase (or decrease) of the estimated social contact rates for children and young adults of the same age.

To illustrate the proposed methods, we apply these to the POLYMOD social contact data of Belgium. The methods can also be applied to the POLYMOD data of other countries or studies. We refer to `http://www.socialcontactdata.org/` for a website on sharing social contact data from different countries and studies. The Belgian POLYMOD data were obtained via a population-based contact survey that has been carried out over the period March-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. The contact information included the age of the contact, gender, location, duration, frequency, and whether or not touching was involved. In this paper, 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 31.0 years. There is at least one participant at each age between the range of 0 and 76. We also restrict to the contacts made with individuals between 0 and 76 years (both included). In total, there is information on 13 493 contacts. This thus gives a crude mean of 18.1 contacts per participant. The age structure of the general population in which the contact survey is conducted in the year 2006 is obtained from (Eurostat, 2017). The size of the population aged 0-76 years in Belgium 2006

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was N=9 777 488.

The proposed methodology is described in details in Section 2. A simulation study to investigate the performance of the proposed methodology is presented in Section 3. The application of the methods to the Belgian POLYMOD data is presented in Section 4. We end with a discussion in Section 5.

2 Methodology: a Smoothing Constrained Approach

In this section, we present the smoothing constrained approach (SCA) used in this paper to estimate smooth social contact rates. First, we describe the approach when smoothing is performed in the dimensions of the respondent's and contact's ages. Smoothing in these dimensions is typically done when estimating smooth social contact rates. Next, the most important contribution of this paper is described, namely the SCA where the contact rates are assumed smooth from a cohort perspective which is achieved by smoothing of the social contact rates over the diagonal components (including all subdiagonals). Two approaches are investigated both in terms of performance and speed: (1) reordening of the diagonal components to reproduce a rectangular grid; and (2) reordening of the penalty matrix such that penalization is performed over the diagonal components.

2.0 No Smoothing over Cohorts

Let $\mathbf{Y} = (y_{ij})$ be a $m \times m$ matrix where the ijth element is the total number of contacts made by the respondents of age $i-1$ with individuals of age $j-1$, with $i = 1, \ldots, m$ and $j = 1, \ldots, m$. This information can be extracted from the self-reported contact diaries of the participants. In our specific case $m = 77$. 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 $i-1$ (see Supplementary Materials) and define the $m \times m$ matrix $\mathbf{P} = \mathbf{p}\mathbf{1}_m$. In the Supplementary Materials we present all these vectors and matrices for an example with $m = 4$.

Define the expected number of contacts made by participants of age $i-1$ with contacts of age $j-1$ as $\mathrm{E}(y_{ij}) = \mu_{ij} = r_i \gamma_{ij}$, where γ_{ij} is the actual contact rate of individuals of age $i-1$ with contacts of age $j-1$. The interpretation of 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. 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.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

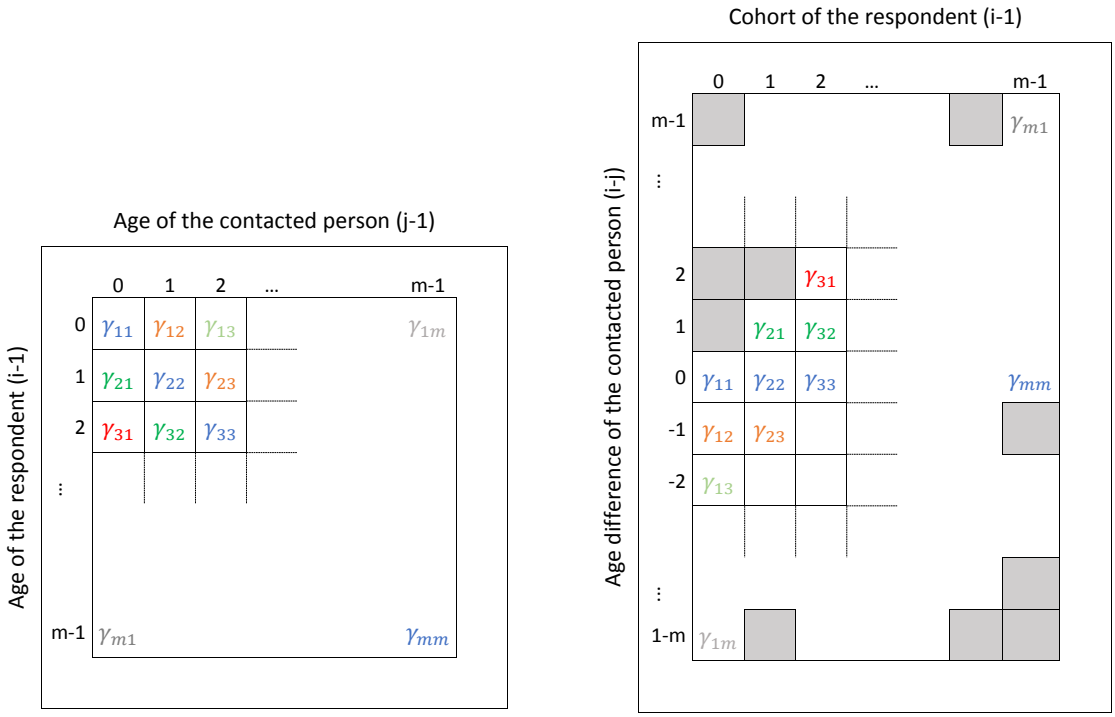

Figure 1: Schematic representation of the original data structure of $\mathbf{\Gamma}$ over ages of respondents and ages of contacts (left panel) and the restructured matrix $\breve{\mathbf{\Gamma}}$ over cohorts of the respondents and age differences of the contacted persons (right panel). Cells with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

The interest is in the estimation of the unknown contact rates 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, namely $\mathbf{y} \sim \text{Pois}(\boldsymbol{\mu})$. For modelling purposes a log-link function is used, namely $\log(\boldsymbol{\gamma}) = \boldsymbol{\eta}$ which yields $\log(\boldsymbol{\mu}) = \log(\mathbf{e}) + \log(\boldsymbol{\gamma}) = \log(\mathbf{e}) + \boldsymbol{\eta}$. 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 is, therefore, only of interest for exploratory analysis. We prefer to work with a modelling approach that yields social contact rates that are smooth and reciprocal. The reciprocal nature of contacts can be expressed as $\gamma_{ij}p_i = \gamma_{ji}p_j$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$. This can be written as $\log(\gamma_{ij}) - \log(\gamma_{ji}) = \log(p_j) - \log(p_i)$ and thus

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

In matrix form:

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 the vector $\boldsymbol{\nu}$ is given by

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

Estimation of the smoothed parameters $\boldsymbol{\eta}$ that satisfy the reciprocal constraints is performed 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 found 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 easily obtained by $\hat{\boldsymbol{\gamma}}^{[k+1]} = \exp(\hat{\boldsymbol{\eta}}^{[k+1]})$. 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 $W_{ll}^{[k]} = \mu_l^{[k]} = \mathrm{e}_l \exp(\eta_l^{[k]})$ 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 which is given by (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 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. We iterate this process until convergence, namely until $\max \mid \hat{\boldsymbol{\eta}}^{[k+1]} - \hat{\boldsymbol{\eta}}^{[k]} \mid < 10^{-4}$.

A grid search is done to find the optimal smoothing parameters. The optimal smoothing parameters are chosen based on minimization of the Akaike Information Criterion (AIC) (Akaike, 1973):

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

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 which is given by (Wood, 2006)

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.1 Smoothing over Cohorts: Reordening of the Contact Matrix

In the section above, we smooth the contact rates parameters in the dimensions of the respondent's and contact's ages. Next, we describe how the contact rates can be smoothed over the diagonal component 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 age of the contacts.

We explain the restructuring for the contact matrix $\boldsymbol{\Gamma}$ in detail. The contact matrix $\boldsymbol{\Gamma}$ is restructured such that each diagonal (the main diagonal and all sub-diagonals) is present as a row in the restructured matrix. The restructured matrix is denoted $\breve{\boldsymbol{\Gamma}}$. Figure 1 (right panel) presents a graphical representation of this restructured matrix. The matrix $\breve{\boldsymbol{\Gamma}}$ is of dimension $(2m-1) \times m$ and is constructed by entering row i of $\boldsymbol{\Gamma}$ in column i of $\breve{\boldsymbol{\Gamma}}$ at positions $m-i+1$ to $2m-i$. From Figure 1 (right panel), it can be observed that in this manner all subsequent diagonal elements are present in the same row. By construction, the matrix $\breve{\boldsymbol{\Gamma}}$ contains nuisance contact rates parameters which are not of interest.

Restructured matrices $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$, constructed from $\mathbf{Y}$ and $\mathbf{E}$, are created similarly as $\breve{\boldsymbol{\Gamma}}$. Missing cell entries are present for $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$ at the same cells where the nuisance parameters are present for $\breve{\boldsymbol{\Gamma}}$. To handle these missing entries, we impute arbitrary values (*e.g.*, 9999) in $\breve{\mathbf{Y}}$ and $\breve{\mathbf{E}}$ and we 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 $\breve{\boldsymbol{\Gamma}}$ is a nuisance parameter and one otherwise. This matrix weight is used to avoid 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 $(2m^2-m) \times 1$ vector obtained by arranging the matrices $\breve{\mathbf{Y}}$, $\breve{\mathbf{E}}$, $\breve{\mathbf{W}}$ and $\breve{\boldsymbol{\Gamma}}$ 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 result a Poisson distribution, namely $\breve{\mathbf{y}} \sim \mathrm{Pois}(\breve{\boldsymbol{\mu}})$. For modelling purposes, we set $\log(\breve{\boldsymbol{\gamma}}) = \breve{\boldsymbol{\eta}}$. Interest is the estimation of the $2m^2-m$ unknown parameters $\breve{\boldsymbol{\eta}}$. However, 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 a $(\frac{m(m-1)}{2}) \times (2m^2-m)$ allocation matrix to suit the reciprocity constraints.

Estimation of the smoothed parameters $\breve{\boldsymbol{\eta}}$ is again performed through C-PIRLS. Updated parameter estimates are now updated by solving the set of linear equations

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In (9), $\boldsymbol{\zeta}^{[k+1]}$ are again Lagrange multipliers, $\mathbf{W}^{[k]}$ is a $(2m^2 - m) \times (2m^2 - m)$ 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 a $(2m^2 - m) \times 1$ vector of pseudodata given by

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

Here, the penalty term $\mathbf{P}$ in (9) is a $(2m^2 - m) \times (2m^2 - m)$ matrix which is 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). Optimal smoothing parameters are again chosen via grid search using AIC.

2.2 Smoothing over Cohorts: Reordening of the Penalty Matrix

In this section, we describe how we can smooth over the cohorts and over the dimension of the contact's age by reordening of the penalty matrix. The methodology is very similar as the one described in Section 2.0. The matrices $\mathbf{Y}$, $\mathbf{E}$, $\mathbf{P}$, $\boldsymbol{\Gamma}$ and the vectors $\mathbf{y}$, $\mathbf{e}$, $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ are defined similar as in Section 2.0. Again, we assume a Poisson distribution for the observed contact rates and the reciprocity constraint is written in matrix form as $\mathbf{L}\boldsymbol{\eta} = \boldsymbol{\nu}$, while C-PIRLS by solving the set of linear equations in (3) is used for parameter estimation. The penalty matrix $\mathbf{P}$, constructed differently as the penalty term in (5), is 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). 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 $\boldsymbol{\Gamma}$ is a 4×4

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matrix (thus $\boldsymbol{\gamma} = \{\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{21}, \ldots, \gamma_{44}\}$) the penalty matrix $\mathbf{P}_d$ is a 16×16 matrix given by

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

Optimal smoothing parameters are again chosen via grid search using AIC.

The advantage of using the penalty matrix $\mathbf{P}_d$ to achieve cohort smoothing is the fact that no nuisance parameters need to be constructed in the matrix $\boldsymbol{\Gamma}$ (cfr. the approach in the previous section using $\breve{\boldsymbol{\Gamma}}$). This speeds up computation time since only m^2 parameters in $\boldsymbol{\Gamma}$ need to be estimated, whereas the approach in the previous section needs estimating of $2m^2 - m$ parameters in $\breve{\boldsymbol{\Gamma}}$ (thus including $m^2 - m$ nuisance parameters). The disadvantage of using the penalty matrix $\mathbf{P}_d$ is the fact that its construction is non-trivial. Whereas the penalty in (11) is easily obtained using standard matrix multiplication, the construction of $\mathbf{P}_d$ requires a more computer-intensive algorithm (see Supplementary Materials).

2.3 Kink on the Main Diagonal of the Social Contact Matrix

Using the SCA methodology described in Sections 2.0-2.2, it can be argued that the main diagonal of the matrix $\mathbf{H}$, or equivalently $\boldsymbol{\Gamma}$, is oversmoothed. In other words, the contact rates between respondents and contacts of the same age are oversmoothed, which leads to an underestimation of the social contact rates on the main diagonal. This effect could especially be present for children and young adults who make an above average number of contacts with persons of the same age (*i.e.*, in

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

school, sports club, ...). Therefore, we introduce a so-called *kink* that can force a sudden increase or decrease of the estimated social contact rates for children and young adults on the main diagonal.

We introduce this kink for the methods described in Sections 2.1-2.2. Allowing for the kink is done by 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

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

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

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

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

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

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

where *max.kink.age* indicates the maximum age at which a kink on the main diagonal is possible. In this paper, we take *max.kink.age* $= 31$ (i.e., $\{0,\dots,30\}$ years). 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 $\mathbf{D}_h$.

We note that the 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 rates. 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. More details on the implementation and example code of the kink are given in the Supplementary Materials.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.4 Negative Binomial Likelihood

In case the Poisson distribution is used for the observed contacts (y_{ij}), it is assumed that the mean and the variance are equal: $\mathrm{E}(Y_{ij}) = \mathrm{Var}(Y_{ij})$. However, in practice, contact data often display overdispersion meaning that the variance of the responses exceed the mean. Not accounting for this possible overdispersion can lead to erroneous results. To accommodate overdispersion a negative binomial distribution can be assumed 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 $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \mu_{ij}^2 \alpha_{ij}^{-1}$. Different parametrizations for α_{ij} lead to different negative binomial distributions (Lawless, 1987). Here, we will consider the parametrization with $\alpha_{ij} = \mu_{ij}\phi^{-1}$, where $\phi > 0$ denotes the disperion parameter. In this parametrization, the variance of the negative binomial distribution is given by $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1+\phi)$. If ϕ tends to zero, the mean and variance will be equal. If $\phi > 0$, the variance will exceed the mean and thus accounting for overdisperion in the data. We note that the variance term, $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1+\phi)$, resembles the error term of an overdispersed Poisson distribution (Nelder and Lee, 1992). The parametrization with $\alpha_{ij} = \phi^{-1}$ (leading to $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1+\phi\mu_{ij})$) was also explored by the authors but not further described since this parametrization performed worse in terms of AIC for the application in Section 4.

In case ϕ would be fixed at a certain value, parameter estimates $\hat{\boldsymbol{\eta}}$ are again obtained through C-PIRLS. The methodology for the C-PIRLS estimation is similar as described in Sections 2.0-2.2 with only one adaptation, namely the entries of $\mathbf{W}^{[k]}$ are given by $W_{ll}^{[k]} = \mu_l^{[k]}/(1+\phi)$. However, rather than fixing ϕ at a certain value, we are also interested in obtaining an estimate for ϕ using the available data. For this, a two-stage iteration scheme is undertaken, namely by iterating and cycling between holding ϕ fixed and holding $\boldsymbol{\eta}$ fixed at its current estimates, the estimates $(\hat{\phi}, \hat{\boldsymbol{\eta}})$ will be obtained. More specifically, by holding ϕ fixed at the current estimate $\hat{\phi}^{[k]}$, estimates $\hat{\boldsymbol{\eta}}^{[k+1]}$ are obtained through C-PIRLS estimation. Next, $\boldsymbol{\eta}$ is fixed at the estimates $\hat{\boldsymbol{\eta}}^{[k+1]}$ and an updated estimate $\hat{\phi}^{[k+1]}$ is obtained using the moment estimation (Breslow, 1984). This process is iterated until convergence. Moment estimation of ϕ is based on the Pearson chi-squares 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 $\hat{\phi}^{[k]}$, namely

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

Optimal smoothing parameters λ_1 and λ_2 are again chosen via grid search using AIC, which is given by $\mathrm{AIC} = -2\log(\hat{L}) + 2(\widehat{ED}+1)$ for the negative binomial distribution. The probability density

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

function of the negative binomial distribution is used to calculate the maximized value of the likelihood function. The plus one term in the calculation of the AIC is included to account for the estimation of the ϕ parameter.

2.5 Uncertainty of the Estimates

In the previous sections we have discussed how the model parameters $\boldsymbol{\eta}$ can be estimated using C-PIRLS. It is also of interest to quantify the uncertainty of the obtained estimates. In particular, we are interested in the variance-covariance matrix associated with the $\hat{\boldsymbol{\eta}}$ estimates. For this purpose a Bayesian posterior covariance matrix which is given by

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

can be used (Wood, 2006). Furthermore, the corresponding posterior distribution is the multivariate normal

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

This last results is only approximate and is justified by large sample results (Wood, 2006). The result in (18) can be used to calculate confidence intervals for parameters η_{ij} or for non-linear functions of these parameters η_{ij} (e.g., 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 normal distribution. This can be extremely useful when one wants to acknowledge for the variability originating from social contact data in the estimation of epidemiological parameters and/or health economic evaluations (Bilcke et al., 2011).

2.6 Computational Note

`R` version 3.4 is used to fit the proposed models. To enhance convergence of the proposed C-PIRLS fitting scheme, we first perform parameters 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 fitting. 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 $\hat{\boldsymbol{\eta}}^{[0]} = \log\left((\mathbf{y}+1)/(\mathbf{e}+1)\right)$.

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. This, however, implies that 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 $2m^2 - m = 11\ 781$

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

parameters, respectively. This is practically challenging on a regular personal computer. Therefore, we make use of sparse matrix implementations 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 a good indication (based on minimization of the AIC) of the values of the optimal smoothing parameters. In a second step, a grid search on a more narrow grid is performed.

More details on the implementation and example code are given in the Supplementary Materials. Code of the algorithm used to construct the penalty matrix $\mathbf{P}_d$ is also provided there. In Section 4, computing times of the data application are reported. All software code is freely available from https://github.com/yannickvandendijck/.

3 Simulation Study

In order to compare the proposed methods in Section 2.0-2.2, a simulation study is used. We perform the simulation study once assuming that the observed contact rates are realizations from the Poisson distribution and once that they arise from the negative binomial distribution. Furthermore, we shall investigate scenarios in which no kink is needed on the main diagonal, and a scenario in which the kink is needed. This thus yields the investigation of four simulation scenarios.

To establish a so-called *true* social contact matrix, denoted by $\mathbf{\Gamma}^*$, from which data will be simulated, a non-parametric regression is applied to the Belgian social contact data. More specifically, the observed contacts rates (see Figure 2), y_{ij}/r_i, of the Belgian social contact data are smoothed using local linear regression. Using the local linear regression there is no guarantee that $\mathbf{K}^* \equiv \mathbf{\Gamma}^* \odot \mathbf{P}$ is symmetric. Therefore, using a simple solution, a symmetric matrix from $\mathbf{K}^*$, denoted by $\breve{\mathbf{K}}^*$, is calculated by $\left(\breve{\mathbf{K}}^*\right)_{ij} = \left(\breve{\mathbf{K}}^*\right)_{ji} = \frac{(\mathbf{K}^*)_{ij}+(\mathbf{K}^*)_{ji}}{2}$. The true contact surface, $\breve{\mathbf{\Gamma}}^*$, which is used for data simulation is obtained by $\breve{\Gamma}^*_{ij} = \breve{K}^*_{ij}/P_{ij}$. Finally, let $H^*_{ij} = \log\left(\breve{\Gamma}^*_{ij}\right)$. In Figure 3 the true social contact matrices $\breve{\mathbf{\Gamma}}^*$ and $\mathbf{H}^*$ are shown.

In two of the four simulation settings, a kink is introduced on the main diagonal of the social contact matrix. Let $\breve{\mathbf{\Gamma}}^\dagger$ denote the true social contact matrix with a kink on the main diagonal. The matrix $\breve{\mathbf{\Gamma}}^\dagger$ is exactly 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

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

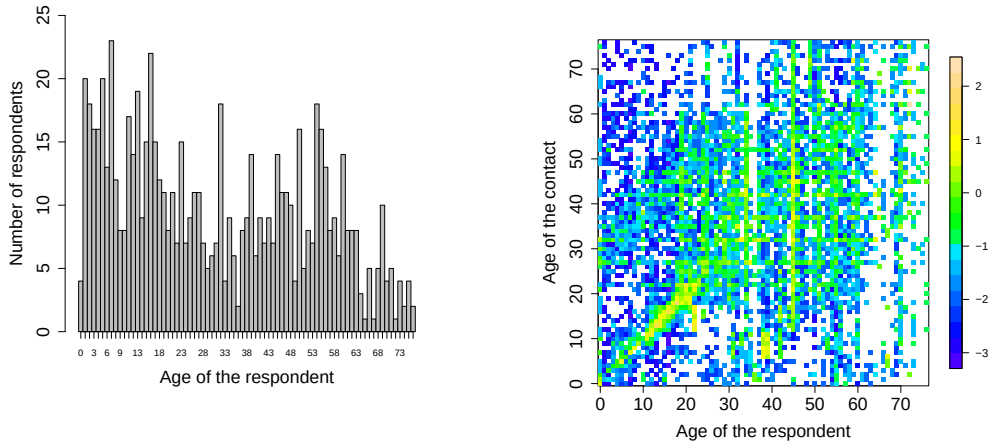

Figure 2: The number of respondent per age (left) and the observed log-contact rates ($\log(y_{ij}/r_i)$) (right) of the Belgian social contact data. A white cell indicates that there were no contacts observed for those particular ages of the respondents and contacts.

Thus for ages between 0 and 23 a higher number of contacts is obtained on the main diagonal. The main diagonal elements of $\breve{\mathbf{\Gamma}}^*$ and $\breve{\mathbf{\Gamma}}^\dagger$ are presented in Figure 3.

Data is simulated using the same participant distribution as in the Belgian social contact data ($n = 745$) (see Figure 2). For the Poisson distribution, data is simulated as

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

For the negative binomial distribution (using $\phi = 2.0$) data is simulated as

$$y^*_{ij} \sim \text{NegBin}\left(\mu_{ij} = r_i \breve{\Gamma}^*_{ij}, \alpha_{ij} = \mu_{ij}(2.0)^{-1}\right). \tag{20}$$

For each setting $S = 100$ simulated datasets are obtained. Next, each simulated dataset is analysed using the methods described in Sections 2.0-2.2. In all simulation settings, both models with and without a kink are used to allow for comparison. Data simulated from the Poisson distribution are analyzed using a Poisson likelihood, and similar for the negative binomial distribution. Optimal smoothing parameters are obtained via grid search using AIC. This yields estimated social contact matrices $\hat{\mathbf{\Gamma}}^{(s)}$ and $\hat{\mathbf{H}}^{(s)}$, for $s = 1, \ldots, S$. The estimation performances of the different methods are compared using squared bias and mean squared error. These measures of performance are calculated as

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

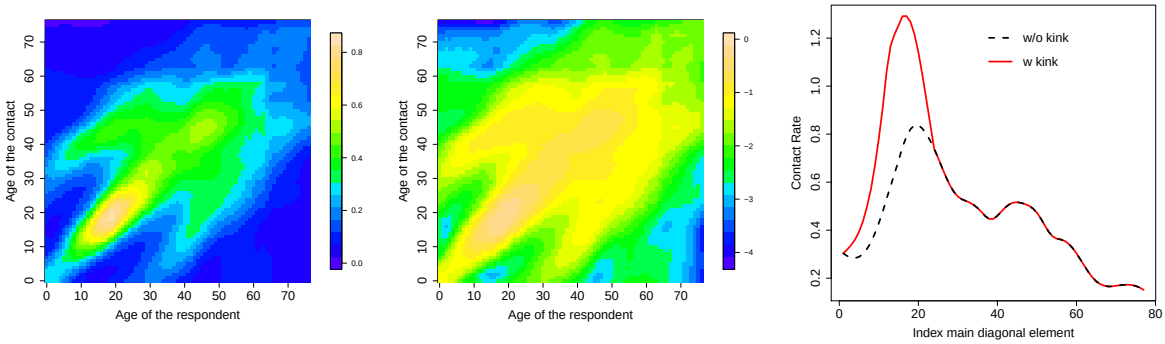

Figure 3: The true social contact matrices $\breve{\mathbf{\Gamma}}^*$ (left) and $\mathbf{H}^*$ (middle) used in the simulation study. The true social contact surfaces are obtained from a non-parametric regression using a local linear fit of the Belgian social contact data. The main diagonal elements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

$$\text{MSE} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\breve{\Gamma}^*_{ij} - \hat{\Gamma}^{(s)}_{ij}\right)^2\right) \quad \text{and} \quad \text{MSE} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{S}\sum_{s=1}^{S}\left(H^*_{ij} - \hat{H}^{(s)}_{ij}\right)^2\right). \tag{22}$$

A similar approach is followed for the social contact matrix with a kink on the main diagonal.

In addition to the estimation performance, the estimation of the uncertainty (see Section 2.5) is evaluated by calculating the nominal coverage of the 95% pointwise confidence intervals (CIs) of η_{ij}. Using result (18), 95% pointwise CIs can be 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 entire social contact matrix (thus for $i = 1, \ldots, m$ and $j = 1, \ldots, m$).

For all simulation settings, we observe (see Table 1) that models that smooth over the cohorts (models M1 and M2) are performing better in terms of MSE than the model that does not smooth over the cohorts (model M0). This holds for both $\mathbf{H}^*$ and $\breve{\mathbf{\Gamma}}^*$. In terms of bias, the results are less clear, but overall model M2 is performing best. When comparing models M1 and M2, we observe that M2 (based on the methodology described in Section 2.2) is performing better. In the simulation settings in which no kink is introduced on the main diagonal, it is observed that the models with a kink on the main diagonal perform slightly worse than the models 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 the models with 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 (see Supplementary Materials). Other graphical results are also presented in the Supplementary Materials.

In the negative binomial simulation setting, the overdisperion parameter ϕ is estimated well. In the

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Squared bias and MSE of the social contact matrices $\mathbf{H}^*$ and $\breve{\mathbf{\Gamma}}^*$ over 100 simulation using the methods described in Sections 2.0-2.2 with and without a kink on the main diagonal. Data is simulated using four simulation settings.

bias^2 results

		Models without kink on main diagonal						Models with kink on main diagonal			
		bias^2 of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			bias^2 of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)			bias^2 of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)		bias^2 of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)	
Simulation setting		M0	M1	M2	M0	M1	M2	M1	M2	M1	M2
Poisson	w/o kink	69.62	58.62	49.41	1.53	1.39	1.32	58.92	49.66	1.49	1.43
NegBin	w/o kink	93.16	91.53	77.76	2.50	2.63	2.52	93.64	79.42	3.05	2.92
Poisson	w kink	57.67	60.86	51.79	3.32	3.37	3.30	58.57	49.33	1.53	1.47
NegBin	w kink	96.52	82.14	70.44	4.77	4.38	4.31	80.70	68.63	2.62	2.53

MSE results

		Models without kink on main diagonal						Models with kink on main diagonal			
		MSE of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)			MSE of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)			MSE of $\mathbf{H}^*$ ($\mathbf{H}^\dagger$)		MSE of $\breve{\mathbf{\Gamma}}^*$ ($\breve{\mathbf{\Gamma}}^\dagger$)	
Simulation setting		M0	M1	M2	M0	M1	M2	M1	M2	M1	M2
Poisson	w/o kink	90.81	74.45	68.05	2.41	1.98	1.94	75.13	68.67	2.18	2.14
NegBin	w/o kink	154.73	130.41	123.72	4.79	3.99	3.96	133.15	126.00	4.57	4.51
Poisson	w kink	82.36	77.78	71.85	4.33	3.98	3.95	75.72	69.63	2.28	2.24
NegBin	w kink	156.94	123.59	120.50	7.11	5.86	5.87	122.71	119.25	4.41	4.40

M0: Based on the methodology described in Section 2.0; M1: Based on the methodology described in Section 2.1; M2: Based on the methodology described in Section 2.2

simulation setting without a kink, model M2 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 M2 with a kink.

Table 2 presents 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) an overcoverage is observed for methods M1 and M2. In this simulation setting, the average lengths of the 95% CIs are 0.65, 0.61 and 0.60, for M0, and M1 and M2 with a kink, respectively. This implies that the overcoverage is not directly associated with wider CIs. The results in Table 2 indicate that the large sample result in (18) can be used to construct CIs of appropriate nominal coverage.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Nominal coverage (in %) of the 95% point-wise confidence intervals of the social contact matrices $\mathbf{H}^*$ ($\mathbf{H}^\dagger$) over 100 simulation using the methods described in Sections 2.0-2.2 with and without a kink on the main diagonal. The nominal coverage is calculated by averaging over the entire social contact matrix. Data is simulated using four simulation settings.

		Models without kink on main diagonal			Models with kink on main diagonal	
Simulation setting		M0	M1	M2	M1	M2
Poisson	w/o kink	92.06	93.86	94.47	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	95.07
NegBin	w kink	95.01	96.26	97.26	96.22	97.26

M0: Based on the methodology described in Section 2.0

M1: Based on the methodology described in Section 2.1

M2: Based on the methodology described in Section 2.2

4 Application: Belgian Social Contact Data

In Figure 2, 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 the three different modelling approaches described in Section 2.0-2.2 are applied. A Poisson and negative binomial distribution is assumed. In addition, models with a kink on the main diagonal are also investigated.

In Table 3, summary results of the fitted models are given. When comparing the distributions, it can be observed that the negative binomial distribution performs better in terms of AIC. Thus implies that the assumption of a variance that is linearly dependent on the mean is preferred. The effective degrees of freedom for the Poisson case are also higher indicating that the Poisson distribution tries to explain the observed variability through the mean. We now discuss in more details the results of the negative binomial distribution. It can be observed that approaches M1 and M2 are performing somewhat better in terms of AIC as compared to M0. The models with the kink on the main diagonal are performing slightly better than the models without kink. For the estimated smoothing parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ an interesting difference is observed between model M0 and models M1 and M2. In M0 the optimal values for $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are similar. For M1 and M2, however, the optimal value for $\hat{\lambda}_2$ is larger than $\hat{\lambda}_1$, which indicates that more penalization is needed in the direction of the cohorts for M1 and M2.

In Table 3, computation times to fit the different models are provided. It is clear that parameter estimation in model M2 is much faster (4 times faster) when compared to model M1. This difference in computation time can be explained by the difference in parameters that need to be estimated.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Summary results of the nine fitted models to the Belgian social contact data. Estimated smoothing parameters, effective degrees of freedom, -2 times log-likelihood, AIC and ϕ are provided. Computation time (T_{comp}) in seconds is also given.

Model	Distribution	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\widehat{ED}$	$-2\log(\hat{L})$	AIC	$\hat{\phi}$	T_{comp}
M0	Poisson	1	1	1 378.8	19 579.1	22 336.7	-	12.9
M0	NegBin	20	20	182.4	19 777.9	20 144.8	1.95	94.3
M1	Poisson	1	1	1 377.1	19 909.4	22 663.6	-	39.3
M1	NegBin	10	1000	86.3	19 948.7	20 123.2	2.20	260.5
M2	Poisson	1	1	1 411.9	19 890.1	22 713.9	-	11.4
M2	NegBin	10	1500	87.5	19 957.3	20 134.2	2.20	57.1
Models with kink on main diagonal (see Section 2.3)								
M1	NegBin	20	800	77.6	19 948.0	20 105.2	2.17	237.4
M2	NegBin	20	1000	80.3	19 950.3	20 112.9	2.18	55.3

M0: Based on the methodology described in Section 2.0

M1: Based on the methodology described in Section 2.1

M2: Based on the methodology described in Section 2.2

For model M1, $2m^2 - m = 11\ 781$ parameters (including $m^2 - m$ nuisance parameters) need to be estimated, whereas for model M2, $m^2 = 5\ 929$ parameters are estimated.

In Figure 4, the estimated log contact rates surfaces, $\hat{\mathbf{H}}$, and the mixing at the population level, $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for models M0, M1 and M2 with the negative binomial distribution and without kink are shown. In general the surfaces capture the important features of human contact behaviour. There is a clear difference in the estimated surfaces of model M0 and models M1 and M2. It can observed that diagonal components are more pronounced for models M1 and M2. The shifted diagonal between children and parents is also more clearly observed. The estimated surfaces of models M1 and M2 are very similar.

Based on the results of the simulation study in Section 3, the computation times and the fact that the estimated contact rates are very similar for models M1 and M2, we prefer the use of model M2 (thus based on the methodology described in Section 2.2) for the POLYMOD Belgian social contact data.

In Figure 5, estimated contact surfaces are shown for model M2 with the negative binomial distribution with a kink on the main diagonal. From the figure on the right hand-side, it is observed that the main diagonal has higher values for younger ages for the model with the kink. This yields higher values on the main diagonal of $\hat{\mathbf{H}}$ and $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$ for the model with the kink. For the model without the kink, the values in the estimated matrix $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$ range from 2 287.4 to 253 511.6, whereas for the model with kink the values range from 2 350.8 to 397 637.7. The kink thus allows for a huge increase

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

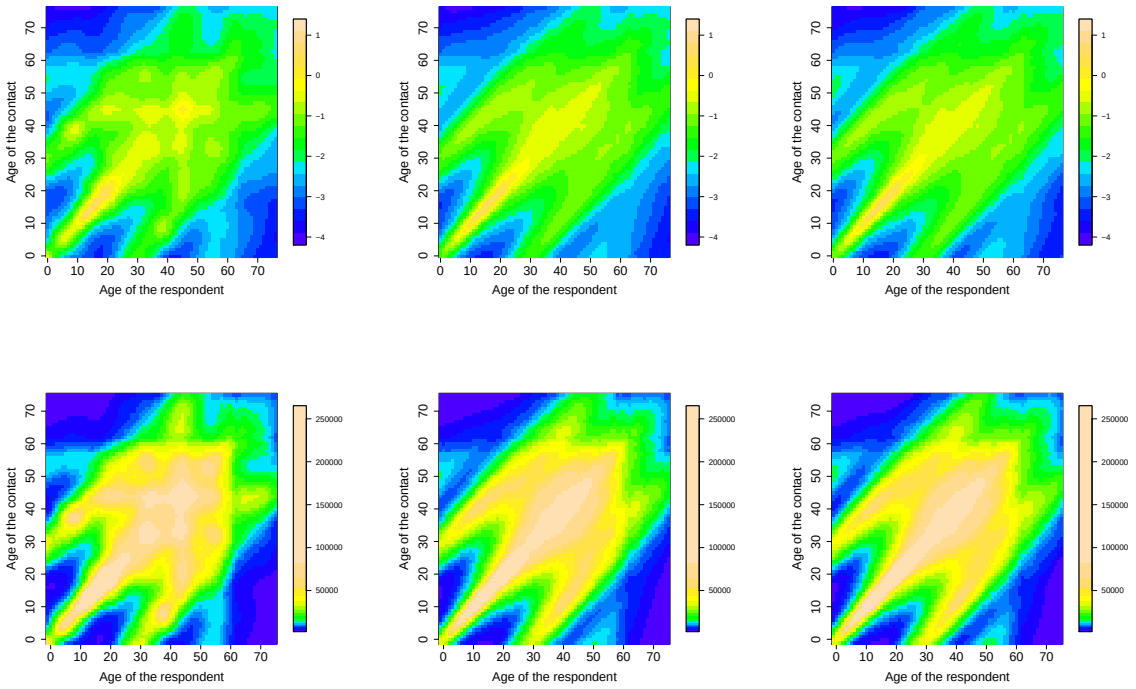

Figure 4: The estimated log contact rates surface (top), $\hat{\mathbf{H}}$, and the mixing at the population level (bottom), $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for models M0, M1 and M2 without kink (left to right) with the negative binomial distribution.

in the estimated number of contacts for children and young adults with individuals of the same age. Based on the AIC-values in Table 3, the models with a kink on the main diagonal are preferred. 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.

Additional results for this data application are provided in the Supplementary Materials.

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

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

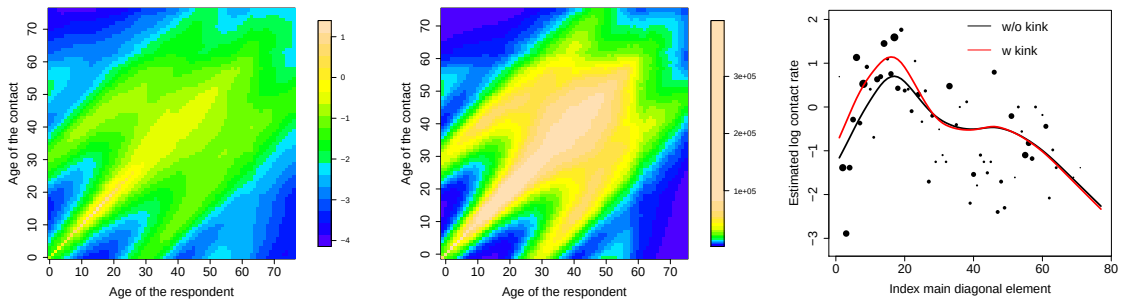

Figure 5: The estimated log contact rates surface (left), $\hat{\mathbf{H}}$, and the mixing at the population level (middle), $\hat{\mathbf{\Gamma}} \odot \mathbf{P}$, for model M2 with the negative binomial distribution with an additional kink on the main diagonal. The diagonal elements of $\hat{\mathbf{H}}$ for the model without and with a kink (right), together with the observed log-contact rates.

propose to smooth contact rates from a cohort perspective. This is achieved by smoothing of the social contact rates over the diagonal components. Social contact rates are thus modelled from a cohort perspective, namely people age through time and thus contact rates for consecutive time points are similar. Two possible models to achieve this cohort smoothing were described: (1) reordening of the diagonal components to reproduce a rectangular grid; and (2) reordening of the penalty matrix such that penalization is performed over the diagonal components.

The maximum likelihood framework was used and parameter estimation was done through constrained penalized iterative reweighted least squares. The proposed approach has the advantage that smooth contact rates are obtained over the cohorts. Typical smoothing approaches for social contact data smooth in the direction of the respondent's and contact's age which can lead to less smooth results over the cohorts (Goeyvaerts et al., 2010). Second, the reciprocal nature of contacts can be explicitly taken care of in the parameter estimation through Lagrange multipliers. Third, the described methods are easily adjusted such that a kink can be introduced for the main diagonal contact rates. This adjustment is desirable because underestimation of the social contact rates on the main diagonal can be present since, especially, children and young adults make an above average number of contacts with persons of the same age. The epidemiological interpretation and the impact on key epidemiological parameters of such a kink is an interesting topic for future investigation.

The results of the simulation study and the data application show that approach (2), in which the penalty matrix is reordered such that penalization is performed over the diagonal components, is performing better. In the simulation study it was observed that this method yielded the smallest MSE over all simulation settings. Additionally, confidence intervals with nominal coverage of close to 95%

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were obtained. In the application study, the computation time of method (2) is three to four times faster than method (1). Therefore, we recommend the use of approach (2) for the estimation of social contact rates.

The true social contact surface used in the simulation study was obtained through local linear regression of the raw social contact rates of the Belgian POLYMOD study. This approach was 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 was obtained. Second, the authors are not aware of any easy applicable mathematical formula or fully parametric model of a two dimensional surface that would represent a contact rates surface.

A grid search is needed to estimate 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 a grid search can be performed fast. To determine the optimal smoothing parameters the Akaike information criterion (Akaike, 1973) was used in this paper. We also investigated the use of the Bayesian information criterion (BIC) (Schwarz, 1978) for smoothing parameter selection, however, we noticed that the use of BIC leads to overly-smoothed and thus non-satisfactory social contact rates.

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. Their results reveal that different contact patterns exist between men and women. Future work could investigate how the proposed methodology in this paper can be extended to estimate social contact rates between both sexes without increasing the computational burden.

In the application, it was observed that the negative binomial distribution is better to describe the POLYMOD data at hand. In this paper, the dispersion parameter for the negative binomial distribution is assumed constant across ages and is treated as a nuisance parameter. It would be interesting in future work to allow for the dispersion parameter to depend on age as well. Although this would be computationally challenging since the moment estimation of ϕ applied in this paper could not be used. The paper of Perperoglou and Eilers (2010) on modelling individual deviance effects could be a good starting point to investigate this further. The association between the dispersion parameter and age could be of direct interest to infectious disease modellers.

A comparison with other methods used to smooth social contact data was not done in this paper. Future work of the authors will 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 assume that no

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

uncertainty is related to this input. Result (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.

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 adjusted such that basis functions can be deployed which likely will lead to a reduction of the computational cost.

SUPPLEMENTARY MATERIAL: (xxxxxxxx.pdf) This file contains additional information with respect to the notations used in the paper. Software code is also presented. Additional results of the simulation study in Section 3 and the applications in Section 4 are provided.

ACKNOWLEDGEMENTS: For the simulation study 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 2009-2014 by a gift from Pfizer, is acknowledged [to NH]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. 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).

References

Abrams, S. and Hens, N. (2015). Modeling individual heterogeneity in the acquisition of recurrent infections: an application to parvovirus B19. *Biostatistics*, 16(1):129–142.

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.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Camarda, C. G., Hens, N., and Eilers, P. H. C. (2013). Modelling social contact data: a smoothing constrained approach. In Muggeo, V. M. R., Capursi, V., Boscaino, G., and Lovison, G., editors, *Proceedings of the 28th International Workshop on Statistical Modelling. Palermo, Italy, 8-12 July 2013*.

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.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

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):5.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

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.

Perperoglou, A. and Eilers, P. (2010). Penalized regression with individual deviance effects. *Computational Statistics*, 25(2):341–361.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

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.

bioRxiv preprint doi: https://doi.org/10.1101/290551; this version posted March 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 and Infection*, 137(1):4857.

Van Hoang, T., Coletti, P., Melegaro, A., Wallinga, J., Grijalva, C., Edmunds, J., Beutels, P., and Hens, N. (2018). A systematic review of social contact surveys to inform transmission models of close contact infections. *Submitted to PLOS Medicine*.

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.