On the Identifiability of the Fuzzy-Clustering-Based Multi-Population Mortality Model

The mortality model we investigate is given by the equation

$$\log m_{x,t}^i = \alpha_x^i + \left(\sum_{l=1}^{k} \omega^{i,l} \beta_x^l \right) \kappa_t^i. \tag{1}$$

Here, each cluster $l \in \{1, \dots, k\}$ has a distinct age effect β_x^l, and the weight parameter $\omega^{i,l}$ indicates for every population i how similar its age effect is to that of cluster l. This is a special case of a k-factor CAE model

$$\log m_{x,t}^i = \alpha_x^i + \sum_{l=1}^{k} \beta_x^l \kappa_t^{i,l} \tag{2}$$

with $\kappa_t^{i,l} = \omega^{i,l} \kappa_t^i$. For the model to be easily interpretable, it is desirable that $\omega^{i,l} \in [0,1]$ and $\sum_{l=1}^{k} \omega^{i,l} = 1$ for all $i \in \{1, \dots, P\}$. In this case, the age effect of each population is a convex combination of the age effects of the clusters.

Let Θ be the parameter space and $\theta := \left(\left(\alpha_x^i\right)_x^i, \left(\beta_x^l\right)_x^l, \left(\kappa_t^i\right)_t^i, \left(\omega^{i,l}\right)^{i,l} \right) \in \Theta$ a vector of model parameters. We also use the notations $\alpha := \left(\alpha_x^i\right)_x^i \in \mathbb{R}^{A \times P}$, $\beta := \left(\beta_x^l\right)_x^l \in \mathbb{R}^{A \times k}$, $\kappa := \left(\kappa_t^i\right)_t^i \in \mathbb{R}^{Y \times P}$ and $\omega := \left(\omega^{i,l}\right)^{i,l} \in \mathbb{R}^{P \times k}$. The log-likelihood function is given by

$$L(\theta) = \sum_{i=1}^{P} \sum_{x=x_1}^{x_A} \sum_{t=t_1}^{t_Y} \left(D_{x,t}^i \cdot \log\left(m_{x,t}^i\right) - E_{x,t}^i \cdot m_{x,t}^i \right) + K, \tag{3}$$

with some constant $K \in \mathbb{R}$ which only depends on the data.

It is in principle possible to numerically maximize L using a gradient-based optimization algorithm such as L-BFGS-B and thereby obtain a maximum likelihood estimate for θ. However, if we do not impose any constraints on the optimization, θ is obviously not unique. Thus, the model is not identifiable, which is problematic both from a statistical and a practical point of view. We will discuss this issue extensively in the following and start by giving a formal definition of identifiability in this context.

Definition 1. *Let $f_p^{(eq)}, f_q^{(ie)} : \Theta \to \mathbb{R}$ for $p = 1, \dots, n_{eq}, q = 1, \dots, n_{ie}$ be real-valued functions. We say that the fuzzy maximum likelihood clustering model (1) is identifiable under the constraints*

$$\begin{aligned} f_p^{(eq)}(\theta) &= 0, p = 1, \dots, n_{eq}, \\ f_q^{(ie)}(\theta) &\leq 0, q = 1, \dots, n_{ie}, \end{aligned} \tag{4}$$

if for two parameter vectors $\theta, \tilde{\theta} \in \Theta$, which fulfill

$$\alpha_x^i + \left(\sum_{l=1}^{k} \omega^{i,l} \beta_x^l \right) \kappa_t^i = \tilde{\alpha}_x^i + \left(\sum_{l=1}^{k} \tilde{\omega}^{i,l} \tilde{\beta}_x^l \right) \tilde{\kappa}_t^i \tag{5}$$

for all $x \in \{x_1, \ldots, x_A\}, t \in \{t_1, \ldots, t_Y\}, i \in \{1, \ldots, P\}$ *and satisfy the constraints (4), it follows that*

$$\alpha_x^i = \tilde{\alpha}_x^i, \kappa_t^i = \tilde{\kappa}_t^i \tag{6}$$

for all $x \in \{x_1, \ldots, x_A\}, t \in \{t_1, \ldots, t_Y\}, i \in \{1, \ldots, P\}$ *and*

$$\omega = \tilde{\omega} S, \beta = \tilde{\beta} S, \tag{7}$$

where $S \in \mathbb{R}^{k \times k}$ *is a permutation matrix.*

Remark. *(i) In words, if certain constraints make the fuzzy maximum likelihood clustering model identifiable, this means that two sets of parameters* θ *and* $\tilde{\theta}$ *which both maximize the log-likelihood and fulfill the constraints must be identical (up to permutation of columns in the case of* β *and* ω*). Put more simply, the corresponding constrained maximization problem has then exactly one solution, implying that we can uniquely identify the model parameters.*

(ii) We do not require ω *and* β *to be identifiable exactly but only up to permutation of columns. We deem this to be sufficient as the order of the columns of* β *and* ω *corresponds to the numbering of the clusters and is, thus, not relevant for the interpretation of the model.*

We will consider two sets of constraints, which differ in the requirements on the weight matrix ω. With the first set of constraints we require that the first k rows of ω, which we denote by $\omega^{1:k,1:k}$, equal the identity matrix I_k. This means that the first k populations each get their own cluster, i.e., $\omega^{i,j} = 1$ if $i = j$ and 0 otherwise for $i, j \in \{1, \ldots, k\}$, and the remaining populations are subsequently assigned cluster weights "relative" to this initialization when the model is fit. Of course, via a renumbering of the populations, any k populations can be the ones which initially get their own cluster, which means that the choice is up to the modeler. This choice should ensure that the chosen populations have sufficiently different age effects. Therefore, it could be based on some a priori knowledge or analysis on which populations might exhibit distinct, prototypic age effects. For numerical studies, we have implemented the following simple heuristic: We start with the 2 populations whose individual Lee-Carter (ILC) age effects have the largest Euclidean distance. Then, we successively choose populations whose ILC age effects maximize the sum of Euclidean distances to the ILC age effects of all the populations we have already chosen, until we reach k populations. We call our first set of constraints the *identity matrix initialization* (IMI) constraints. Note that we do not demand $\omega^{i,l} \geq 0$ for $i > k$ in this case.

With the second, alternative set of constraints we require that all entries of ω are non-negative – which, by the additional constraint $\sum_{l=1}^{k} \omega^{i,l} = 1$, implies that they are at most 1 – and that among all matrices fulfilling the remaining constraints ω maximizes the sum of within-cluster variances. Therefore, we call these the *non-negativity variance-maximizing* (NNVM) constraints.

Expressed in formulas, to fit the model, we solve

$$\sup_{\theta\in\Theta} L(\theta) \tag{8}$$

subject to

$$\begin{aligned}\sum_{x=x_1}^{x_A} \beta_x^l &= 1 \text{ for all } l \in \{1,\dots,k\},\\ \sum_{t=t_1}^{t_Y} \kappa_t^i &= 0 \text{ for all } i \in \{1,\dots,P\},\end{aligned} \tag{9}$$

and, furthermore, either to

$$\begin{aligned}\sum_{l=1}^{k} \omega^{i,l} &= 1 \text{ for all } i \in \{k+1,\dots,P\},\\ \omega^{1:k,1:k} &= I_k,\end{aligned} \tag{IMI}$$

or, alternatively, to

$$\begin{aligned}\sum_{l=1}^{k} \omega^{i,l} &= 1 \text{ for all } i \in \{1,\dots,P\},\\ \omega^{i,l} &\geq 0 \text{ for all } i \in \{1,\dots,P\}, l \in \{1,\dots,k\},\\ f_\omega(R) &\leq f_\omega(I_k) \text{ for all } R \in \mathcal{D}_\omega,\end{aligned} \tag{NNVM}$$

where

$$\mathcal{D}_\omega := \{R \in \mathrm{GL}(k) : \omega R \succcurlyeq \mathbf{0}_{P\times k} \text{ and } R\mathbb{1}_k = \mathbb{1}_k\} \tag{10}$$

and $f_\omega : \mathcal{D}_\omega \to \mathbb{R}$ is the sum of within-cluster variances,

$$f_\omega(R) := \sum_{l=1}^{k} \frac{1}{P-1} \sum_{i=1}^{P} \left((\omega R)^{i,l} - \left(\overline{\omega R}\right)^{\bullet,l} \right)^2. \tag{11}$$

Here, we have used the notation

$$\left(\overline{\omega R}\right)^{\bullet,l} := \frac{1}{P} \sum_{j=1}^{P} (\omega R)^{j,l} \tag{12}$$

for the column means. It is obvious that the constraints (NNVM) do not determine the order of the columns of β and ω but this is in accordance with Definition 1.

We present some other ways to express the function f_ω and give an interpretation of what it means to maximize this function.

Remark (Other expressions for and interpretation of f_ω). *(i) By the definition of matrix multiplication, we have*

$$f_\omega(R) = \sum_{l=1}^{k} \frac{1}{P-1} \sum_{i=1}^{P} \left(\sum_{q=1}^{k} r^{q,l} \left(\omega^{i,q} - \overline{\omega}^{\bullet,q} \right) \right)^2 . \tag{13}$$

(ii) Assuming that $\sum_{q=1}^{k} \omega^{i,q} = 1$ for all $i = 1, \ldots, P$, we calculate

$$f_\omega(R) = \sum_{l=1}^{k} \frac{1}{P-1} \sum_{i=1}^{P} \left(\sum_{q=1}^{k-1} \left(r^{q,l} - r^{k,l} \right) \cdot \left(\omega^{i,q} - \overline{\omega}^{\bullet,q} \right) \right)^2 . \tag{14}$$

(iii) Expanding the square in the definition of f_ω and using the definition of the Frobenius norm $\| \cdot \|$, we get

$$f_\omega(R) = \frac{1}{P-1} \left(\|\omega R\|^2 - \frac{1}{P} \| \mathbb{1}_P^\top \omega R \|^2 \right). \tag{15}$$

(iv) As the rows of ωR sum up to 1 for any $R \in \mathcal{D}_\omega$, we have

$$\left(\overline{\omega R}\right)^{\bullet,\bullet} := \frac{1}{Pk} \sum_{i=1}^{P} \sum_{l=1}^{k} (\omega R)^{i,l} = \frac{1}{k} \tag{16}$$

for its overall average. By partitioning the sum of squares as known from analysis of variance, we get

$$f_\omega(R) = \frac{1}{P-1} \sum_{l=1}^{k} \left(\sum_{i=1}^{P} \left((\omega R)^{i,l} - \frac{1}{k} \right)^2 - P \left(\left(\overline{\omega R}\right)^{\bullet,l} - \frac{1}{k} \right)^2 \right). \tag{17}$$

From (17), we immediately see that maximizing f_ω amounts to choosing R such that the entries of ωR differ as much as possible from $\frac{1}{k}$, which means that populations have a clearer tendency to which cluster they belong, while the column means of ωR are as close to $\frac{1}{k}$ as possible, which means that clusters tend to have similar sizes.

In order to show that the NNVM constraints imply identifiability, we need existence and uniqueness (up to permutations of columns) of the solution of the optimization problem

$$\sup_{R \in \mathcal{D}_\omega} f_\omega(R), \tag{18}$$

where ω has full rank and fulfills $\omega \mathbb{1}_k = \mathbb{1}_P$. The following proposition explicitly gives the solution for the case $k = 2$.

Proposition 1. *If $k = 2$, $\operatorname{rank}(\omega) = 2$ and $\omega \mathbb{1}_2 = \mathbb{1}_P$, the optimization problem (18) is solved by*

$$R^* := \frac{1}{\omega^{max} - \omega^{min}} \cdot \begin{pmatrix} \omega^{max} - 1 & 1 - \omega^{min} \\ \omega^{max} & -\omega^{min} \end{pmatrix}, \tag{19}$$

where $\omega^{max} := \max_{i=1,\dots,P} \omega^{i,1}$ *and* $\omega^{min} := \min_{i=1,\dots,P} \omega^{i,1}$. *The solution is unique up to permutation of columns.*

Proof. Note that the fact that ω has full rank and fulfills $\omega \mathbb{1}_k = \mathbb{1}_P$ implies that none of its columns can be constant. So we have $\omega^{\max} > \omega^{\min}$ and the definition of R^* makes sense. Also, R^* lies in $\mathcal{D}_\omega$:

- By direct calculation, we see that $R^* \mathbb{1}_k = \mathbb{1}_k$.
- The determinant of R^* is easily checked to equal -1, which implies $R^* \in \mathrm{GL}(k)$.
- It will become clear below that $\omega R^* \succcurlyeq \mathbf{0}_{P\times k}$.

In the following, we consider only matrices $R \in \mathrm{GL}(k)$ which satisfy $R\mathbb{1}_k = \mathbb{1}_k$. We write $r^1 := r^{1,1}$ and $r^2 := r^{2,1}$ for the entries of their first column; the corresponding entries of the second column are then $1 - r^1$ and $1 - r^2$. As the first step, we restrict ourselves to matrices R which additionally fulfill $r^1 < r^2$.

For $k = 2$, Equation (14) describing the objective function f_ω simplifies to

$$f_\omega(R) = \left(r^1 - r^2\right)^2 \frac{2}{P-1} \sum_{i=1}^{P} \left(\omega^{i,1} - \overline{\omega}^{\cdot,1}\right)^2, \tag{20}$$

which shows that solving the constrained optimization problem amounts to maximizing the distance between r^1 and r^2 in such a way that the corresponding matrix R still lies in the feasible set.

For $k = 2$, the constraint $\omega R \succcurlyeq \mathbf{0}_{P\times k}$ can equivalently be written as

$$\begin{aligned} \omega^{i,1}(r^1 - r^2) + r^2 &\geq 0, \\ \omega^{i,1}(r^2 - r^1) + 1 - r^2 &\geq 0, \end{aligned} \tag{21}$$

for $i = 1, \dots, P$, where we have used that $\omega \mathbb{1}_k = \mathbb{1}_P$. Due to $r^1 < r^2$, this system of $2P$ inequalities is equivalent to the following system of 2 inequalities:

$$\omega^{\max}(r^1 - r^2) + r^2 \geq 0, \tag{22}$$

$$\omega^{\min}(r^2 - r^1) + 1 - r^2 \geq 0. \tag{23}$$

We distinguish the following six cases:

(i) $\omega^{\min} > 1$: Applying first (22) and then (23), we get

$$r^1 \geq \frac{\omega^{\max} - 1}{\omega^{\max}} \cdot r^2 \geq \frac{\omega^{\max} - 1}{\omega^{\max}} \cdot \frac{\omega^{\min} r^1 - 1}{\omega^{\min} - 1}, \tag{24}$$

which is equivalent to

$$r^1 \leq \frac{\omega^{\max} - 1}{\omega^{\max}(\omega^{\min} - 1)} \cdot \frac{1}{\frac{\omega^{\min}}{\omega^{\min} - 1} \cdot \frac{\omega^{\max} - 1}{\omega^{\max}} - 1} = \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}}. \tag{25}$$

From this, it follows with (22) that

$$\begin{aligned} r^2 - r^1 &\leq \left(\frac{\omega^{\max}}{\omega^{\max} - 1} - 1 \right) r^1 \leq \left(\frac{\omega^{\max}}{\omega^{\max} - 1} - 1 \right) \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}} \\ &= \frac{1}{\omega^{\max} - \omega^{\min}}. \end{aligned} \tag{26}$$

(ii) $\omega^{\min} = 1$: From (23) we get

$$r^1 \leq 1 = \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}}, \tag{27}$$

and, additionally using (22),

$$r^2 - r^1 \leq \left(\frac{\omega^{\max}}{\omega^{\max} - 1} - 1 \right) r^1 \leq \frac{\omega^{\max}}{\omega^{\max} - 1} - 1 = \frac{1}{\omega^{\max} - \omega^{\min}}. \tag{28}$$

(iii) $1 > \omega^{\min} \geq 0$: Applying first (23) and then (22), we get

$$r^2 \leq \frac{\omega^{\min} r^1 - 1}{\omega^{\min} - 1} \leq \frac{\omega^{\min}}{\omega^{\min} - 1} \cdot \frac{\omega^{\max} - 1}{\omega^{\max}} \cdot r^2 - \frac{1}{\omega^{\min} - 1}, \tag{29}$$

which is equivalent to

$$r^2 \leq \frac{\omega^{\max}}{\omega^{\max} - \omega^{\min}}. \tag{30}$$

From this, it follows with (22) that

$$\begin{aligned} r^2 - r^1 &\leq \left(1 - \frac{\omega^{\max} - 1}{\omega^{\max}} \right) r^2 \leq \left(1 - \frac{\omega^{\max} - 1}{\omega^{\max}} \right) \frac{\omega^{\max}}{\omega^{\max} - \omega^{\min}} \\ &= \frac{1}{\omega^{\max} - \omega^{\min}}. \end{aligned} \tag{31}$$

(iv) $\omega^{\min} < 0, \omega^{\max} > 1$: From (22) and (23) we get

$$r^2 - r^1 \leq \frac{r^1}{\omega^{\max} - 1}, \tag{32}$$

$$r^2 - r^1 \leq \frac{r^1 - 1}{\omega^{\min} - 1}. \tag{33}$$

If

$$r^1 \leq \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}}, \tag{34}$$

(32) yields

$$r^2 - r^1 \leq \frac{1}{\omega^{\max} - \omega^{\min}}. \tag{35}$$

Otherwise, if

$$r^1 \geq \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}}, \tag{36}$$

(33) yields the same.

(v) $\omega^{\min} < 0, \omega^{\max} = 1$: From (22), we get $r^1 \geq 0$, and, additionally using (23),

$$\begin{aligned} r^2 - r^1 &\leq \left(\frac{\omega^{\min}}{\omega^{\min} - 1} - 1 \right) r^1 - \frac{1}{\omega^{\min} - 1} \\ &\leq \frac{1}{1 - \omega^{\min}} = \frac{1}{\omega^{\max} - \omega^{\min}}. \end{aligned} \tag{37}$$

(vi) $\omega^{\min} < 0, \omega^{\max} < 1$: Applying first (23) and then (22), we get

$$r^1 \geq \frac{\left(\omega^{\min} - 1\right) r^2 + 1}{\omega^{\min}} \geq \frac{\omega^{\min} - 1}{\omega^{\min}} \cdot \frac{\omega^{\max}}{\omega^{\max} - 1} \cdot r^1 + \frac{1}{\omega^{\min}}, \tag{38}$$

which is equivalent to

$$r^1 \geq \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}}, \tag{39}$$

where we have used that the function $x \mapsto \frac{x-1}{x}$ is strictly monotonically increasing for $x < 0$. From this, it follows with (23) that

$$\begin{aligned} r^2 - r^1 &\leq \left(\frac{\omega^{\min}}{\omega^{\min} - 1} - 1 \right) r^1 - \frac{1}{\omega^{\min} - 1} \\ &\leq \left(\frac{\omega^{\min}}{\omega^{\min} - 1} - 1 \right) \frac{\omega^{\max} - 1}{\omega^{\max} - \omega^{\min}} - \frac{1}{\omega^{\min} - 1} \\ &= \frac{1}{\omega^{\max} - \omega^{\min}}. \end{aligned} \tag{40}$$

In each of these cases, the constant upper bound for $r^2 - r^1$ is attained if and only if $r^1 = r^{*,1}$ and $r^2 = r^{*,2}$. This shows that $\omega R^* \succcurlyeq \mathbf{0}_{P \times k}$ and, among all matrices $R \in \mathcal{D}_\omega$ with $r^1 < r^2$, R^* uniquely maximizes f_ω.

It can be shown analogously that, among all matrices $R \in \mathcal{D}_\omega$ with $r^1 > r^2$, the matrix which is obtained by exchanging the columns of R^* uniquely maximizes f_ω. As the maximal values in both cases coincide, this completes the proof. □

In fact, existence can be shown for any value of $k \in \mathbb{N}$:

Proposition 2. *The optimization problem (18) has a solution.*[1]

Proof. Define the function $g : \mathbb{R}^{P \times k} \to \mathbb{R}$ by

$$g(T) := \sum_{l=1}^{k} \frac{1}{P-1} \sum_{i=1}^{P} \left(T^{i,l} - \overline{T}^{\bullet,l} \right)^2. \tag{41}$$

[1] We thank an anonymous contributor to Mathematics Stackexchange for proposing the idea for the proof of this result.

Now, the maximization problem (18) is obviously equivalent to

$$\sup_{T \in \{\omega R \,:\, R \in \mathrm{GL}(k)\}} g(T) \tag{42}$$

subject to

$$\begin{aligned} T\mathbb{1}_k &= \mathbb{1}_P, \\ T &\succcurlyeq \mathbf{0}_{P\times k}. \end{aligned} \tag{43}$$

The feasible set of this optimization problem is bounded because the entries of every matrix T in the feasible set must fulfill $0 \leq t^{i,l} \leq 1$ for all $i \in \{1,\dots,P\}, l \in \{1,\dots,k\}$. It is also closed as the intersection of three closed sets (a finite-dimensional linear subspace, a set defined by equations and a set defined by non-strict inequalities). By the Heine-Borel theorem, the feasible set is compact, which implies that the continuous function g attains a maximum on this set. □

Remark (NNVM constraints for $k > 2$)**.** *Proposition 2 only shows existence, not uniqueness of the maximum. Unfortunately, we have not found a proof (or counterexample) for uniqueness in the case $k > 2$. Therefore, in the following, when we employ the constraints (NNVM) we restrict ourselves to $k = 2$, which we consider to be an important special case of our model. All other ingredients apart from the uniqueness of the solution of (18) are granted in the case $k > 2$ as well. However, the reader should be aware that we have no guarantee that β and ω are identifiable if we employ NNVM constraints in this case.*

We will need the following two lemmas, which are shown using basic linear algebra.

Lemma 1. *Let $\beta \in \mathbb{R}^{A\times k}$ and $\omega \in \mathbb{R}^{P\times k}$, and denote $r_1 := \mathrm{rank}(\beta), r_2 := \mathrm{rank}(\omega)$. If $\min(r_1, r_2) < k$, there exist $\tilde{k} \leq \min(r_1, r_2)$ and full-rank matrices $\tilde{\beta} \in \mathbb{R}^{A\times\tilde{k}}, \tilde{\omega} \in \mathbb{R}^{P\times\tilde{k}}$ such that*

$$\beta\omega^\top = \tilde{\beta}\tilde{\omega}^\top. \tag{44}$$

Proof. We set $\tilde{k} := \mathrm{rank}\left(\beta\omega^\top\right)$. Then, we choose a basis of the column space of $\beta\omega^\top$ and build up a matrix $\tilde{\beta} \in \mathbb{R}^{A\times\tilde{k}}$ column-wise out of the basis vectors. As the columns of $\beta\omega^\top$ lie in the span of the columns of $\tilde{\beta}$, we can find a matrix $\tilde{\omega} \in \mathbb{R}^{P\times\tilde{k}}$ such that the desired equality holds. Both $\tilde{\beta}$ and $\tilde{\omega}$ must have full rank due to

$$\tilde{k} = \mathrm{rank}\left(\beta\omega^\top\right) = \mathrm{rank}\left(\tilde{\beta}\tilde{\omega}^\top\right) \leq \min\left\{\mathrm{rank}\left(\tilde{\beta}\right), \mathrm{rank}\left(\tilde{\omega}^\top\right)\right\} \leq \tilde{k}. \tag{45}$$

□

Lemma 2. *Let $\beta, \tilde{\beta} \in \mathbb{R}^{A\times k}$ and $\omega, \tilde{\omega} \in \mathbb{R}^{P\times k}$ be full-rank matrices fulfilling*

$$\beta\omega^\top = \tilde{\beta}\tilde{\omega}^\top. \tag{46}$$

There is a matrix $R \in \mathrm{GL}(k)$ such that

$$\tilde{\omega} = \omega R \text{ and } \tilde{\beta} = \beta R^{-\top}. \tag{47}$$

Proof. Note that all the inverse matrices appearing in this proof exist because $\beta, \tilde{\beta}, \omega, \tilde{\omega}$ have full rank and $k \leq \min\{A, P\}$.

Multiplying (46) by $\left(\beta^\top \beta\right)^{-1} \beta^\top$ from the left, we get

$$\omega^\top = \left(\beta^\top \beta\right)^{-1} \beta^\top \tilde{\beta} \tilde{\omega}^\top, \tag{48}$$

which is equivalent to

$$\tilde{\omega} = \omega \left(\left(\beta^\top \beta\right)^{-1} \beta^\top \tilde{\beta} \right)^{-\top} =: \omega R. \tag{49}$$

Transposing (46) and multiplying by $\left(\omega^\top \omega\right)^{-1} \omega^\top$ from the left, we get

$$\beta^\top = \left(\omega^\top \omega\right)^{-1} \omega^\top \tilde{\omega} \tilde{\beta}^\top, \tag{50}$$

which is equivalent to

$$\tilde{\beta} = \beta \left(\left(\omega^\top \omega\right)^{-1} \omega^\top \tilde{\omega} \right)^{-\top} =: \beta S. \tag{51}$$

The observation that

$$R^\top S = \left(\beta^\top \tilde{\beta}\right)^{-1} \beta^\top \beta \omega^\top \omega \left(\tilde{\omega}^\top \omega\right)^{-1} \overset{(46)}{=} I_k \tag{52}$$

and, thus, $S = R^{-\top}$, completes the proof. □

Theorem 1. *Assume that the number of clusters k, is chosen based on the principle of parsimony. If the constraints (9) hold and*

(i) the constraints (IMI) are fulfilled, or

(ii) $k = 2$ and the constraints (NNVM) are fulfilled,

model (1) is identifiable in the sense of Definition 1.

Proof. First, note that if β or ω does not have rank k, we can by Lemma 1 reduce β and ω to full-rank matrices with column number smaller than k without changing the fitted death rates. In other words, we can achieve the same fit using a lower number of parameters. According to the principle of parsimony (which we implement by using the BIC as the model selection criterion), we would reduce k accordingly. So we can w.l.o.g. assume that ω and β have full rank k.

Now, assume that there are parameter vectors $\theta, \tilde{\theta} \in \Theta$ which fulfill (5), (9) and (i) (IMI) or (ii) (NNVM), where additionally $k = 2$. First, summation of (5) over t shows identifiability of α due to $\sum_{t=t_1}^{t_Y} \kappa_t^i = \sum_{t=t_1}^{t_Y} \tilde{\kappa}_t^i = 0$. Then,

summation over x shows identifiability of κ due to $\sum_{x=x_1}^{x_A} \beta_x^l = \sum_{x=x_1}^{x_A} \tilde{\beta}_x^l = 1$ and $\sum_{l=1}^{k} \omega^{i,l} = \sum_{l=1}^{k} \tilde{\omega}^{i,l} = 1$. We now get that

$$\sum_{l=1}^{k} \omega^{i,l} \beta_x^l = \sum_{l=1}^{k} \tilde{\omega}^{i,l} \tilde{\beta}_x^l \text{ for all } i \in \{1, \ldots, P\}, x \in \{x_1, \ldots, x_A\} \tag{53}$$

or, in matrix notation,

$$\beta \omega^\top = \tilde{\beta} \tilde{\omega}^\top. \tag{54}$$

By Lemma 2, we have

$$\tilde{\omega} = \omega R \text{ and } \tilde{\beta} = \beta R^{-\top} \tag{55}$$

for some $R \in \mathrm{GL}(k)$. Case distinction:

(IMI) The constraint on the first k rows of ω and $\tilde{\omega}$ implies

$$I_k = \tilde{\omega}^{1:k,1:k} = (\omega R)^{1:k,1:k} = I_k R = R, \tag{56}$$

which shows $\omega = \tilde{\omega}, \beta = \tilde{\beta}$.

(NNVM) The constraints on ω and $\tilde{\omega}$ imply $R \in \mathcal{D}_\omega$ and $R^{-1} \in \mathcal{D}_{\tilde{\omega}}$, so we have

$$f_{\tilde{\omega}}(I_k) = f_\omega(R) \leq f_\omega(I_k) = f_{\tilde{\omega}}\left(R^{-1}\right) \leq f_{\tilde{\omega}}(I_k). \tag{57}$$

This implies $f_\omega(R) = f_\omega(I_k)$, and it follows from Proposition 1 that R is a permutation matrix. Therefore, ω and $\tilde{\omega}$ as well as β and $\tilde{\beta}$ coincide up to the same (note that $R^{-\top} = R$) permutation of columns.

□

The following theorem shows that the constraints we have considered do not only uniquely determine the model parameters but also do not decrease the log-likelihood, which means they are in fact identifiability constraints. The proof of the theorem also shows a way to practically implement these identifiability constraints.

Theorem 2. *Assume that the number of clusters k, is chosen based on the principle of parsimony. The constraints (9) together with either (IMI) or, if we additionally have $k \leq 2$, (NNVM) do not change the fitted death rates, so in particular they do not decrease the log-likelihood, of the model (1).*

Proof. We first choose a particular solution, which we denote by $(\alpha, \beta, \kappa, \omega)$ for simplicity, of the unconstrained problem

$$\sup_\theta L(\theta). \tag{58}$$

As in the proof of Theorem 1, we can w.l.o.g. assume that β and ω have full rank k due to the principle of parsimony and Lemma 1.

The transformation

$$\begin{aligned} \beta_x^l &\to \frac{1}{c_{\beta,\omega}^l} \cdot \beta_x^l, \\ \omega^{i,l} &\to c_{\beta,\omega}^l \cdot \omega^{i,l} \end{aligned} \tag{59}$$

with $c_{\beta,\omega}^l \neq 0, l = 1, \ldots, k$ does not change the fit (or the rank of β, ω). With $c_{\beta,\omega}^l := \sum_{x=x_1}^{x_A} \beta_x^l$, we implement $\sum_{x=x_1}^{x_A} \beta_x^l = 1$ for all $l \in \{1, \ldots, k\}$.

The transformation

$$\begin{aligned} \omega^{i,l} &\to \frac{1}{c_{\omega,\kappa}^i} \cdot \omega^{i,l} \\ \kappa_t^i &\to c_{\omega,\kappa}^i \cdot \kappa_t^i \end{aligned} \tag{60}$$

with $c_{\omega,\kappa}^i \neq 0, i = 1, \ldots, P$ does not change the fit (or the rank of ω). With $c_{\omega,\kappa}^i := \sum_{l=1}^{k} \omega^{i,l}$, we implement $\sum_{l=1}^{k} \omega^{i,l} = 1$ for all $i \in \{1, \ldots, P\}$.

The transformation

$$\begin{aligned} \omega &\to \omega R, \\ \beta &\to \beta R^{-\top} \end{aligned} \tag{61}$$

with $R \in \mathrm{GL}(k)$ does not change the fit. Case distinction:

(IMI) As ω has full rank, we assume w.l.o.g. that $\omega^{1:k,1:k}$ is invertible (if there is no particular solution to (58) fulfilling this, the populations have to be renumbered accordingly). We now choose $R = \left(\omega^{1:k,1:k}\right)^{-1}$. This preserves the constraints that have already been implemented at previous steps and implements $\omega^{1:k,1:k} = I_k$.

(NNVM) Choose R such that it solves the optimization problem (18). As we require $k \leq 2$ in this case, R is uniquely determined according to Proposition 1 (if $k = 2$; for $k = 1$, it must obviously hold that $R = 1$). This preserves the constraints that have already been implemented at previous steps and implements $\omega \succcurlyeq \mathbf{0}_{P \times k}$ and $f_\omega(R) \leq f_\omega(I_k)$ for all $R \in \mathcal{D}_\omega$.

The transformation

$$\begin{aligned} \kappa_t^i &\to \kappa_t^i - c_{\alpha,\kappa}^i, \\ \alpha_x^i &\to \alpha_x^i + c_{\alpha,\kappa}^i \cdot \sum_{l=1}^{k} \omega^{i,l} \beta_x^l \end{aligned} \tag{62}$$

with $c_{\alpha,\kappa}^i \in \mathbb{R}, i = 1, \ldots, P$ does not change the fit. With $c_{\alpha,\kappa}^i := \frac{1}{Y} \sum_{t=t_1}^{t_Y} \kappa_t^i$, we implement $\sum_{t=t_1}^{t_Y} \kappa_t^i = 0$ for all $i \in \{1, \ldots, P\}$. □

The following corollary shows that the identifiability constraints (IMI) and (NNVM) are in some sense equivalent for $k = 2$.

Corollary 1. *For $k = 2$, the identifiability constraints (IMI) and (NNVM) induce identical parameters, provided that*

$$\omega^{1:2,1:2} = \begin{pmatrix} \omega^{min} & 1 - \omega^{min} \\ \omega^{max} & 1 - \omega^{max} \end{pmatrix} \tag{63}$$

holds for the particular solution of (58) in the proof of Theorem 2.

Proof. It is easily checked that, in this case, the inverse of $\omega^{1:2,1:2}$ equals R^* from Proposition 1. The proof of Theorem 2 shows that, consequently, both types of identifiability constraints induce the same values of ω and β. □

Further work on the model we have introduced should focus on the question whether the (NNVM) constraints also ensure identifiability for $k > 2$.