Electronic Journal of Statistics
Vol. 13 (2019) 67–93
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1515

Generalised additive dependency inflated models including aggregated covariates

Young K. Lee[*]

Kangwon National University
e-mail: youngklee@kangwon.ac.kr

Enno Mammen[†]

Heidelberg University
e-mail: mammen@math.uni-heidelberg.de

Jens P. Nielsen[‡]

City, University of London
e-mail: jens.nielsen.1@city.ac.uk

Byeong U. Park[§]

Seoul National University
e-mail: bupark@stats.snu.ac.kr

Abstract: Let us assume that X, Y and U are observed and that the conditional mean of U given X and Y can be expressed via an additive dependency of X, $\lambda(X)Y$ and $X+Y$ for some unspecified function λ. This structured regression model can be transferred to a hazard model or a density model when applied on some appropriate grid, and has important forecasting applications via structured marker dependent hazards models or structured density models including age-period-cohort relationships. The structured regression model is also important when the severity of the dependent variable has a complicated dependency on waiting times X, Y and the total waiting time $X+Y$. In case the conditional mean of U approximates a density, the regression model can be used to analyse the age-period-cohort model, also when exposure data are not available. In case the conditional mean of U approximates a marker dependent hazard, the regression model introduces new relevant age-period-cohort time scale interdependencies in understanding longevity. A direct use of the regression relationship introduced in this paper is the estimation of the severity of outstanding liabilities in non-life insurance companies. The technical approach taken is to use B-splines to capture the underlying one-dimensional

[*]Research of Young K. Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A2A01005039 and NRF-2018R1A2B6001068).

[†]Research of Enno Mammen was supported by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953.

[‡]Research of Jens P. Nielsen was supported by the Institute and Faculty of Actuaries, London, UK.

[§]Research of B. U. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001753).

unspecified functions. It is shown via finite sample simulation studies and an application for forecasting future asbestos related deaths in the UK that the B-spline approach works well in practice. Special consideration has been given to ensure identifiability of all models considered.

MSC 2010 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Structured nonparametric models, age-period-cohort model, identifiability, B-splines, UK mesothelioma mortality.

Received December 2017.

1. Introduction

Let us assume that X, Y and U are observed and that the conditional mean of U given X and Y can be expressed via an additive dependency of X, $\lambda(X)Y$ and $X + Y$ for some unspecified funtion λ, leading to the following mathematical definition of the "Generalised Additive Dependency Inflated Model including Aggregated Covariate" (GADIMAC): for a link function G

$$U = G\left(m_0 + m_1(X) + m_2(\lambda(X)Y) + m_3(X + Y)\right) + \varepsilon, \tag{1.1}$$

where the constant m_0 and the functions m_1, m_2, m_3, λ are unknown. Notice that this is a special case of the generalised structured regression model considered in Mammen and Nielsen (2003). In the case where U is the number of events within a suitable grid of X and Y, the conditional mean of U is essentially a density and one can use the GADIMAC model to identify and analyse the density version of the age-period-cohort model. Without the time acceleration $\lambda(X)$, the density age-period-cohort model is known to be hard to visualize, analyse and forecast, because the entering effects are only identifiable up to a line, see Kuang et al. (2008a,b, 2011) and Antonczyk et al. (2017). Also, one is left with complicated second order differences in the discrete case and complicated second order derivatives in our continuous case when working with canonical and well-defined parametrisation, see Nielsen and Nielsen (2014), O'Brien (2014), Riebler et al. (2012), Smith and Wakefield (2016) and Beutner et al. (2017).

The relevant age-period-cohort density version of GADIMAC is used to forecast the future asbestos-related deaths in the United Kingdom in the application in Section 4.3. Asbestos mortality data is characterized by its lack of proper exposure data and its complicated dependency structures on the entering time effects, see Peto et al. (1995) for an early approach and Hodgson et al. (2005), Rake et al. (2009), and Tan et al. (2010, 2011) for later approaches building various micro models to overcome the lack of exposure data. The recent approach by Martinez-Miranda et al. (2015, 2016) uses updated data and is simpler because exposure is directly modelled and estimated from the observed deaths. This paper adds to this latter approach by including further time scales, forecasting the peak of asbestos related deaths in the UK to be 2572 in the year 2018 and the total future UK asbestos-related deaths until the year 2032 to be forty eight thousands.

There are many potential applications of GADIMAC. One further potential application provides a solution to an omnipresent challenge in non-life insurance

when estimating the severity of outstanding liabilities. Here X is the waiting time from an insurance claim has happened till it is reported, Y is the waiting time from the claim being reported till its final settlement and U is the size of the claim. Another potential further application is within the current and important theme of longevity estimation, where X, Y and $X+Y$ represent cohort, age and period and U represents raw occurrence divided by exposure of some grid-points of discretised X's and Y's. In the longevity forecasting case the conditional mean of U is approximately equal to a two-dimensional hazard function as a function of cohort X and age Y. The GADIMAC model introduced by this paper provides a structured nonparametric representation of both past and future mortality when the calendar effect of $X+Y$ is extrapolated into the future. An authoritative exposition of current efforts on longevity forecasting could be the six models reviewed in Cairns et al. (2011), where a discrete time series is used to forecast calendar effects in all models as have become standard, see also some of the original proposals of Lee and Carter (1992), Lee and Miller (2001) and Renshaw and Haberman (2006). The GADIMAC model could introduce a welcome alternative modeling deterministic trends first before time series or other uncertainties complicate the visual and analytic impression of future mortality.

This paper develops theory identifying the GADIMAC model and introduces estimation techniques and asymptotic theory of GADIMAC models based on B-splines. Finite sample simulation studies show good performance of our new methodology and the usefulness of the new class of regression models is illustrated via a timely application to forecasting future asbestos related deaths in the UK. In Section 2 below the B-spline based estimation of the nonparametric structured model is constructed and the asymptotic theory is derived. Identifiability is discussed in Section 3. Practical implementation is considered in Section 4 with an implementation guide in Section 4.1 and finite sample simulation studies in Section 4.2 showing good performance of the estimation of the model. Finally the important practical forecast of future UK asbestos related deaths is given in Section 4.3.

2. Estimation and asymptotic properties of GADIMAC models

We assume that one observes independent real valued random variables U_i for $1 \leq i \leq n$ with mean $\mu(x_i, y_i)$ where $x_1, \ldots, x_n, y_1, \ldots, y_n$ are some deterministic design points on the real line. The regression function $\mu(x, y)$ satisfies

$$\mu(x, y) = G\left(m_0 + m_1(x) + m_2(\lambda(x)y) + m_3(x + y)\right) \tag{2.1}$$

for some unknown functions m_1, m_2, m_3, λ and for a known invertible link function G. Later, we will also apply this model to a random design case where one observes i.i.d. copies (X_i, Y_i, U_i) of (X, Y, U) such that ((2.1)) holds for the conditional mean $\mu(x, y) = E(U|X = x, Y = y)$. We assume that the tuples (x_i, y_i) lie in a connected subset $\mathcal{I}$ of a two-dimensional bounded rectangle.

Let I_1 denote the projection of $\mathcal{I}$ onto the x-axis, $I_2(\lambda) = \{\lambda(x)y : (x, y) \in \mathcal{I}\}$ and $I_3 = \{x + y : (x, y) \in \mathcal{I}\}$. In this section we discuss the estimation of the regression function μ with this structure using a set of observations U_i and design points $(X_i, Y_i) \in \mathcal{I}$. We will show that the function μ can be estimated with a one-dimensional nonparametric rate. Throughout this paper, we assume that $\mathcal{I}$ contains a rectangle $[\alpha, \beta] \times [0, \gamma]$ for some $\beta > \alpha > 0$ and $\gamma > 0$. Without loss of generality, we take $\alpha = 0$, since otherwise we may shift $\mathcal{I}$ along the x-axis and redefine the component functions m_1, m_3 and λ accordingly.

For $k \geq 2$ we consider the following estimator $\hat{m} = (\hat{m}_0, \hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{\lambda})$ that minimizes

$$n^{-1} \sum_{i=1}^{n} \left[U_i - G\left(\hat{m}_0 + \hat{m}_1(x_i) + \hat{m}_2(\hat{\lambda}(x_i) y_i) + \hat{m}_3(x_i + y_i) \right) \right]^2 + (\text{penalty}) \tag{2.2}$$

over $\hat{m} \in \mathcal{M} = \{(m_0, m_1, m_2, m_3, \lambda) : m_1(0) = m_2(0) = m_3(0) = 0, m_2'(0) = 1\}$. For the penalty, we consider

$$\begin{aligned} (\text{penalty}) = \rho_{1,n}\, T_1(\hat{m}_1) + \rho_{2,n}\, T_2(\hat{\lambda})^{(2k-1)/2} \int_{I_2(\hat{\lambda})} \hat{m}_2^{(k)}(z)^2\, dz \\ + \rho_{3,n}\, T_2(\hat{\lambda})^{1/2} \int_{I_2(\hat{\lambda})} \hat{m}_2'(z)^2\, dz + \rho_{4,n}\, T_3(\hat{m}_3), \end{aligned} \tag{2.3}$$

where $T_j(m) = \int_{I_j} m'(x)^2\, dx + \int_{I_j} m^{(k)}(x)^2\, dx$ for $j = 1, 3$, and

$$T_2(\lambda) = \int_{I_1} \lambda(x)^2\, dx + \int_{I_1} \lambda'(x)^2\, dx + \int_{I_1} \lambda^{(k)}(x)^2\, dx.$$

We present theory for the estimator $\hat{\mu}$ of the composite function μ, defined by

$$\hat{\mu}(x, y) = G\left(\hat{m}_0 + \hat{m}_1(x) + \hat{m}_2(\hat{\lambda}(x) y) + \hat{m}_3(x + y) \right).$$

Later we will use a simplified version of $\hat{\mu}$ in our numerical studies, where we replace $\mathcal{M}$ by a subspace containing only linear λ and spline functions m_j. In our theory and also in our numerical studies we will choose $\rho_n = \rho_{1,n} = \rho_{2,n} = \rho_{3,n} = \rho_{4,n}$. The following theorem shows that, under the assumption that the functions m_1, m_2, m_3 and λ allow derivatives up to order k, the choice $\rho_n \asymp n^{-2k/(2k+1)}$ leads to an estimator $\hat{\mu}$ that achieves a one-dimensional nonparametric rate.

Theorem 1. *For the components $(m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}$ of the underlying regression function suppose that m_1, m_2, m_3 and λ have derivatives of order k with bounded L_2-norm, where k is the constant in (2.3). Furthermore, we assume that G has an absolutely bounded derivative which is continuous at the point m_0 with $G'(m_0) > 0$, that the distributions of the error variables $\varepsilon_i = U_i - \mu(x_i, y_i)$ have subexponential tails,*

$$\sup_{1 \leq i \leq n} E\left(\exp(c^{-1} |\varepsilon_i|) \right) \leq c \quad \textit{for } c > 0 \textit{ large enough.}$$

Furthermore, we assume that there exists a sequence $\delta_n \to 0$ *with* $n^{k/(2k+1)}\delta_n \to \infty$ *such that the number of indices* i *with* $|x_i| \le \delta_n$ *and* $|y_i| \le \delta_n$ *can be bounded from below by* $C_\delta n\delta_n^2$ *for some constant* $C_\delta > 0$. *Then, it holds with* $\rho_n^{-1} = O_P(n^{2k/(2k+1)})$ *and* $\rho_n = O_P(n^{-2k/(2k+1)})$ *that*

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}(\hat{\mu}(x_i,y_i)-\mu(x_i,y_i))^2 \\ &= n^{-1}\sum_{i=1}^{n}\Big[G\left(\hat{m}_0+\hat{m}_1(x_i)+\hat{m}_2(\hat{\lambda}(x_i)y_i)+\hat{m}_3(x_i+y_i)\right) \\ &\qquad\qquad - G\Big(m_0+m_1(x_i)+m_2(\lambda(x_i)y_i)+m_3(x_i+y_i)\Big)\Big]^2 \\ &= O_P(n^{-2k/(2k+1)}). \end{aligned}$$

We note that this result does not imply that the functions m_1, m_2, m_3 and λ can be estimated with the rate $O_P(n^{-k/(2k+1)})$. A first step to such kind of results are identification results for our model that we discuss in the next section, which constitute the main contributions of this paper.

Now for the random design case, let $\hat{m} = (\hat{m}_0, \hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{\lambda})$ be defined as the minmizer of

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\Big[U_i - G\left(\hat{m}_0+\hat{m}_1(X_i)+\hat{m}_2(\hat{\lambda}(X_i)Y_i)+\hat{m}_3(X_i+Y_i)\right)\Big]^2 \\ & + (\text{penalty}) \end{aligned} \tag{2.4}$$

over $\hat{m} \in \mathcal{M} = \{(m_0, m_1, m_2, m_3, \lambda) : m_1(0) = m_2(0) = m_3(0) = 0, m_2'(0) = 1\}$. Then, we obtain the following analogue of Theorem 1.

Theorem 2. *Suppose that one observes i.i.d. copies* (X_i, Y_i, U_i) *of* (X, Y, U) *for* $1 \le i \le n$, *where* $\mu(x, y) = E(U|X = x, Y = y)$ *satisfies (2.1) and the distribution of* (X, Y) *is supported on* $\mathcal{I}$. *We assume that the conditions of Theorem 1 on* m_1, m_2, m_3, λ *and* G *hold, and that the conditional distributions of the error variables* $\varepsilon_i = U_i - \mu(X_i, Y_i)$ *have subexponential tails,*

$$\sup_{1\le i\le n} E\left(\exp(c^{-1}|\varepsilon_i|)|X_i, Y_i\right) \le c \quad a.s., \text{ for } c > 0 \text{ large enough.}$$

Furthermore, we assume that there exists a sequence $\delta_n \to 0$ *with* $n^{k/(2k+1)}\delta_n \to \infty$ *such that* $P(|X| \le \delta_n, |Y| \le \delta_n) \ge C_\delta\delta_n^2$ *for some constant* $C_\delta > 0$. *Then, it holds with* $\rho_n^{-1} = O_P(n^{2k/(2k+1)})$ *and* $\rho_n = O_P(n^{-2k/(2k+1)})$ *that*

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}(\hat{\mu}(X_i,Y_i)-\mu(X_i,Y_i))^2 \\ &= n^{-1}\sum_{i=1}^{n}\Big[G\left(\hat{m}_0+\hat{m}_1(X_i)+\hat{m}_2(\hat{\lambda}(X_i)Y_i)+\hat{m}_3(X_i+Y_i)\right) \\ &\qquad\qquad - G\Big(m_0+m_1(X_i)+m_2(\lambda(X_i)Y_i)+m_3(X_i+Y_i)\Big)\Big]^2 \\ &= O_P(n^{-2k/(2k+1)}). \end{aligned}$$

3. Identification of GADIMAC models

We discuss identification of the model. Suppose that there exists a consistent estimator $\hat{\mu}$ of the function μ defined by

$$\mu(x,y) = G\left(m_0 + m_1(x) + m_2(\lambda(x)y) + m_3(x+y)\right). \tag{3.1}$$

The question is if this implies that there exist consistent estimators of the functions m_1, m_2, m_3 and λ. We study the following question: Given a function $(x,y) \to \mu(x,y)$, does this function identify m_1, m_2, m_3 and λ up to a constant in the model (3.1)? Note that the same question arises in in-sample density forecasting where the model $f(x,y) = f_1(x)f_2(\lambda(x)y)f_3(x+y)$ can be transferred to this model by putting $\mu(x,y) = \log f(x,y)$, $m_1(x) = \log f_1(x)$, $m_2(z) = \log f_2(z)$ and $m_3(v) = \log f_3(v)$.

We say that m_1, m_2, m_3 and λ in (3.1) are identified up to a constant if $\mu(x,y) = G\left(\bar{m}_1(x) + \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x+y)\right)$ for some functions $\bar{m}_1, \bar{m}_2, \bar{m}_3$ and $\bar{\lambda}$ implies that

$$\begin{aligned} m_1(x) &= \bar{m}_1(x) - c_1 \text{ for } x \in I_1, \\ m_2(x) &= \bar{m}_2(x) - c_2 \text{ for } x \in I_2, \\ m_3(x) &= \bar{m}_3(x) - c_3 \text{ for } x \in I_3, \\ \lambda(x) &= \bar{\lambda}(x) \text{ for } x \in I_1 \end{aligned}$$

for some real numbers c_1, c_2, and c_3. We first discuss identification in case λ is known, and then in case λ is a linear function. We treat a more general case at the end.

3.1. The case of known λ

Let us assume that λ is known. The following theorem demonstrates that, in this case, m_j are identified up to a constant under some smoothness conditions and conditions on the shape of λ. This theorem serves as a basic building block for the discussion of the identification of m_j and λ in more general cases. For the formulation of the theorem we need the following additional notation. Denote by $\mathcal{I}^0$ the interior of $\mathcal{I}$. Put $I_1^0 = \{x : (x,y) \in \mathcal{I}^0 \text{ for some } y\}$, $I_2^0 = \{\lambda(x)y : (x,y) \in \mathcal{I}^0\}$, and $I_3^0 = \{x+y : (x,y) \in \mathcal{I}^0\}$.

Theorem 3. *Suppose that λ is a fixed continuously differentiable strictly positive function that is not equal to the function $x \to (\xi_1 + \xi_0 x)^{-1}$ for some $\xi_0, \xi_1 \in \mathbb{R}$. In particular, we do not allow that λ is a constant function. Assume that the closures of I_1^0, I_2^0 and I_3^0 are equal to I_1, $I_2 = I_2(\lambda)$ and I_3, respectively and that I_1^0 and I_3^0 are connected sets. Furthermore, we assume that m_1, m_2 and m_3 are twice continuously differentiable and that G is strictly monotone. Then in model (3.1), m_1, m_2, and m_3 are identified up to a constant.*

Thus for known λ we have a clear picture. Given some smoothness conditions, the model is identified up to constants if and only if λ is not equal to the function

$x \to (\xi_1 + \xi_0 x)^{-1}$ for some $\xi_0, \xi_1 \in \mathbb{R}$. Note that for constant λ, the functions m_1, m_2 and m_3 are only identified up to addition or substraction of linear functions, see Antonczyk et al. (2017). If λ is not equal to the function $x \to (\xi_1 + \xi_0 x)^{-1}$ for some $\xi_0, \xi_1 \in \mathbb{R}$ and if we put some constraints on m_1, m_2 and m_3 such as $m_1(0) = m_2(0) = m_3(0) = 0$, then m_1, m_2 and m_3 are completely identified. On the other side, if $\lambda(x)$ is equal to $(\xi_1 + \xi_0 x)^{-1}$ for some $\xi_0, \xi_1 \in \mathbb{R}$ then such a constraint does not lead to unique identification. To see this one can choose arbitrary choices of m_1, m_2 and m_3 and one can define $m_1^*(x) = m_1(x) + \xi \ln\left(1 + \frac{\xi_0}{\xi_1} x\right)$, $m_2^*(z) = m_2(z) + \xi \ln(1 + \xi_0 z)$, and $m_3^*(v) = m_3(v) - \xi \ln\left(1 + \frac{\xi_0}{\xi_1} v\right)$. Then one can easily verify that for all choices of $\xi \in \mathbb{R}$ it holds that

$$m_1(x) + m_2(\lambda(x)y) + m_3(x + y) = m_1^*(x) + m_2^*(\lambda(x)y) + m_3^*(x + y).$$

Thus we have no unique identification.

3.2. The case of unknown linear λ

Recall that, in our setting, $\mathcal{I}$ contains a nontrivial rectangle $[0, \beta] \times [0, \gamma]$ for some $\beta, \gamma > 0$. We note that this implies that $J_1(0) \equiv \{x : (x, 0) \in \mathcal{I}\}$ contains $[0, \beta]$ and $J_2(0) \equiv \{y : (0, y) \in \mathcal{I}\}$ contains $[0, \gamma]$. We consider the following constraints:

$$m_1(0) = m_2(0) = m_3(0) = 0,\ m_2'(0) = 1. \tag{3.2}$$

The first three conditions can be always achieved by redefining the functions m_1, m_2 and m_3. For the fourth condition this can be done if $m_2'(0) > 0$. If $m_2'(0) < 0$, we consider the link $\tilde{G}(z) = G(-z)$ so that the model (3.1) can be written as

$$U = \tilde{G}\left((-m_0) + (-m_1)(X) + (-m_2)(\lambda(X)Y) + (-m_3)(X + Y)\right) + \varepsilon.$$

We consider the case where λ is an unknown linear function $\lambda(x) = ax + b$ with $a \neq 0$. We note that, if a equals zero, then the functions m_j are not identifiable. The following theorem demonstrates that λ is identifiable under these conditions if m_2 is not linear, and that λ is identifiable only up to a constant if m_2 is linear. In the case where m_2 is linear, it follows that $m_2(z) = z$ from the constraints on m_2 in (3.2). This means that the model (3.1) reduces to $G^{-1}(\mu(x, y)) = m_0 + m_1(x) + (ax + b)y + m_3(x + y)$. In this case, we may have different sets of (m_1, m_3, b) that give the same function μ. For example, we have

$$m_1(x) + (ax + b)y + m_3(x + y) = \bar{m}_1(x) + (ax + \bar{b})y + \bar{m}_3(x + y)$$

with $\bar{m}_1(x) = m_1(x) + (\bar{b} - b)x$ and $\bar{m}_3(z) = m_3(z) + (b - \bar{b})z$.

Theorem 4. *Suppose that $\lambda(x) = ax + b$ for some nonzero constant a. Assume that $\mathcal{I}$ contains a nontrivial rectangle $[0, \beta] \times [0, \gamma]$ with $\beta, \gamma > 0$. Furthermore,*

we assume that m_1 is differentiable, m_2 is two times continuously differentiable and G is strictly monotone. Then in model (3.1) under the constraints (3.2), m_1, m_2, m_3 and λ are identified if m_2 is not linear. In case m_2 is linear, the function λ is identified up to a constant.

We now give a heuristic explanation why m_1, m_2, m_3 and λ can be consistently estimated under the assumptions of Theorem 4. For a rigorous proof of identifiability of the functions we refer to the appendix. We make the additional assumption that there exists an estimator $\hat{\nu}$ of $\nu(x,y) \equiv m_0 + m_1(x) + m_2(\lambda(x)y) + m_3(x+y)$ such that $\|\hat{\nu} - \nu\|_\infty = o_P(1)$, $\|\hat{\nu}_x - \nu_x\|_\infty = o_P(1)$ and $\|\hat{\nu}_y - \nu_y\|_\infty = o_P(1)$, where $f_x(x,y)$ and $f_y(x,y)$ for a bivariate function $f(x,y)$ denote its partial derivatives with respect to x or y, respectively. For the case where G is the identity function, such an estimator can be achieved by kernel smoothing of the estimator $\hat{\mu}$, i.e.

$$\hat{\nu}(x,y) = \frac{\int \hat{\mu}(u,v) K\left(\frac{u-x}{h}\right) K\left(\frac{v-y}{h}\right) du\, dv}{\int K\left(\frac{u-x}{h}\right) K\left(\frac{v-y}{h}\right) du\, dv}$$

for x and y in the interior of the support of μ and with boundary corrections of the kernel K at the boundary of the support. Let

$$q(x,y) = \frac{m_2((ax+b)y)}{ax+b} - y$$

for $ax + b \neq 0$ and $q(x,y) = 0$ for $ax + b = 0$. We have that uniformly in x and y

$$\begin{aligned}
\hat{\nu}(x,y) &= m_0 + m_1(x) + m_2((ax+b)y) + m_3(x+y) + o_p(1), && (3.3)\\
\hat{\nu}(x,0) &= m_0 + m_1(x) + m_3(x) + o_p(1), && (3.4)\\
\hat{\nu}_y(x,0) &= ax + b + m_3'(x) + o_p(1). && (3.5)
\end{aligned}$$

Furthermore, by more lengthy but straightforward calculations one can show that

$$\begin{aligned}
&\int_0^y \frac{1}{v}\left[\hat{\nu}_x(x,v) - \hat{\nu}_x(x,0) - \hat{\nu}_y(x+v,0) + \hat{\nu}_y(x,0)\right] dv\\
&\quad = a \cdot q(x,y) + o_p(1),\\
&\int_0^y \Bigg[\hat{\nu}_y(x,v) - \hat{\nu}_y(x+v,0) - \frac{x}{v}\big(\hat{\nu}_x(x,v) - \hat{\nu}_x(x,0) - \hat{\nu}_y(x+v,0)\\
&\qquad + \hat{\nu}_y(x,0)\big)\Bigg] dv\\
&\quad = b \cdot q(x,y) - \frac{1}{2}ay^2 + o_p(1).
\end{aligned} \qquad (3.6)$$

For linear functions m_2 we have $q(x,y) \equiv 0$ under our norming conditions. For nonlinear m_2, the functions $q(x,y)$ and y^2 are linearly independent. Thus (3.6) can be used to get a consistent estimator $\hat{a}$ of a in the case of a linear

function m_2, and consistent estimators $(\hat{a}, \hat{b})$ of (a, b) can be achieved in the case of a nonlinear function m_2. In the latter case we can replace a and b by $\hat{a}$ and $\hat{b}$ in (3.5) which gives a consistent estimator of $m_3'(x)$. Integration of this estimator results in a consistent estimator $\hat{m}_3$ of m_3. Using this estimator we get with the help of (3.4) a consistent estimator of m_1. Finally using all these estimators we can use (3.3) and we get a consistent estimator of m_2. Thus, we can consistently estimate all component functions m_1, m_2, m_3 and λ. Note that for this estimation we only need estimators of ν and of the integrals of partial derivatives of ν. In Theorem 1 we have argued that ν can be estimated with a one-dimensional nonparametric rate. We conjecture that, as in many other nonparametric models, the integrals of partial derivatives of ν can be estimated with the same rate as ν, see Lee et al. (2017) for example. This would imply that m_1, m_2, m_3 and λ can be estimated with a one-dimensional nonparametric rate.

3.3. More general cases

We now come to a more general case that includes nonlinear λ. The next lemma is a first step for analyzing identification of the component functions in the general case. We continue to make the norming constraints (3.2).

Lemma 1. *Suppose that the functions m_2 and $\bar{m}_2$ are continuously differentiable, that the functions m_1, m_2, m_3 and $\bar{m}_1$, $\bar{m}_2$, $\bar{m}_3$ fulfill the norming constraints (3.2), and that the functions λ and $\bar{\lambda}$ with m_j and $\bar{m}_j$ satisfy*

$$m_1(x) + m_2(\lambda(x)y) + m_3(x+y) = \bar{m}_1(x) + \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x+y). \tag{3.7}$$

Also, we assume that $\bar{\lambda}(0) \neq 0$. Then, for all $(x, y) \in \mathcal{I}$ with $x+y \in J_1(0) \cap J_2(0)$ and $\bar{\lambda}(x)y/\bar{\lambda}(0) \in J_1(0) \cap J_2(0)$, it holds that

$$\begin{aligned} 0 = \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)} \left[\bar{\lambda}\left(\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)} y \right) - \lambda\left(\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)} y \right) \right] - \left[\bar{\lambda}(x+y) - \lambda(x+y) \right] \\ + \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)} \lambda(0) m_2'\left(\lambda(0) \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)} y \right) - \lambda(x) m_2'(\lambda(x)y). \end{aligned} \tag{3.8}$$

We make use of Lemma 1 for identification of unknown λ under the assumption that λ is linear in a small neighborhood of zero. Specifically, we assume that $\lambda(0) \neq 0$ and there exists $\varepsilon > 0$ such that

$$\lambda(x) = \lambda(0) + \lambda'(0)x, \quad 0 \leq x \leq \varepsilon. \tag{3.9}$$

The assumption is clearly more general than the one in Theorem 4. It only requires linearity in an arbitrarily small neighborhood of zero and thus allows a general class of nonlinear functions λ. The following theorem demonstrates that λ is identifiable within this wider class if m_2 is not linear in any neighborhood of zero and if $J_1(0) \cap J_2(0)$ equals I_1, the domain of λ. The theorem is based on the same set of conditions for m_j in Theorem 4. The condition that m_2 is not linear in any neighborhood of zero is implied by the condition that $m_2''(0) \neq 0$ if m_2'' is continuous.

Theorem 5. *Assume that λ is differentiable, $\lambda(0) \neq 0$ and $\lambda'(0) \neq 0$, and that $\mathcal{I}$ contains a nontrivial rectangle $[0,\beta] \times [0,\gamma]$ with $\beta, \gamma > 0$. Assume also that (3.9) holds for some $\varepsilon > 0$, and that m_j and G fulfill the conditions in Theorem 4. Then in model (3.1) under the constraints (3.2), m_1, m_2, m_3 and λ are identified if $J_1(0) \cap J_2(0) = I_1$ and m_2 is not linear in any neighborhood of zero.*

4. Finite sample studies and an application forecasting asbestos related deaths in UK

This section first introduces in Section 4.1 the necessary considerations when implementing the method in practice, and Section 4.2 presents finite sample simulation studies showing good performance of our B-spline estimation approach. The important forecasting result on future asbestos related deaths are presented in Section 4.3 with some new and interesting forecasts that might be of interest for some policy makers, health economists or non-life insurers. Note that the methods considered here do not strictly belong to the recently defined class of "in-sample forecasters". To qualify as an in-sample forecaster, the forecast should be a function of one-dimensional functions that are fully estimated in-sample, see Martinez-Miranda et al. (2015), Lee et al. (2015, 2017) and Hiabu et al. (2016). However, the calendar effect has to be extrapolated when applying GADIMAC to the modeling of the age-period-cohort relationship of asbestos related deaths, disqualifying the approach as in-sample forecasting. The necessary extrapolation of the calendar effect does compromise the otherwise simple interpretation of the GADIMAC model making the forecasting exercise non-trivial and non-automatic. There will be more comments on this in Section 4.3 below.

4.1. Practical implementation of GADIMAC models

Before we present the results of our simulation study and real data example, we describe briefly how we get the estimators that minimize the objective function (2.4). For simplicity, we choose λ to be linear and thus assume $\lambda(x) = \theta_0 + \theta_1 x$ for some unknown θ_0 and θ_1. The procedure may be generalized to any parametric function or even to a nonparametric model for λ.

We are to minimize the objective function at (2.4) over $(m_0, \theta_0, \theta_1) \in \mathbb{R}^3$ and (m_1, m_2, m_3) with each m_j in the space of cubic B-splines ($k = 2$) satisfying the constraints

$$m_1(0) = m_2(0) = m_3(0) = 0, m_2'(0) = 1. \tag{4.1}$$

Instead of the penalty at (2.3) we use a simpler version given by

$$\begin{aligned}(\text{penalty}) = {} & \rho_{1,n}\, \tilde{T}_1(\hat{m}_1) + \rho_{2,n}\, \tilde{T}_2(\hat{\lambda})^{(2k-1)/2} \int_{I_2(\hat{\lambda})} \hat{m}_2^{(k)}(z)^2\, dz \\ & + \rho_{3,n}\, \tilde{T}_3(\hat{m}_3),\end{aligned} \tag{4.2}$$

where $\tilde{T}_j(m) = \int_{I_j} m^{(k)}(x)^2\,dx$ for $j = 1, 3$, and $\tilde{T}_2(\lambda) = \int_{I_1} \lambda(x)^2\,dx$. For the penalty at (4.2) we omit $\int \hat{m}_j'(u)^2\,du$ for $1 \le j \le 3$ in (2.3) since we may show that Theorem 1 remains to hold under this change at the cost of much more involved arguments based on a decomposition of the functions $\hat{m}_j$ as sums of a polynomial and a smooth function. In an additional technical step we have to show that the polynomials can be estimated with the parametric rate $n^{-1/2}$. This would be more involved than the discussion of smoothing splines in van de Geer (2000) because of the nonlinear nature of our model. Furthermore, we omit $\int \lambda'(u)^2\,du + \int \lambda^{(k)}(u)^2\,du$ in $T_2(\lambda)$ because we use parametric linear fits for λ.

The minimisation problem is nonlinear since we have a link function G. Even if we choose the identity link, it is still a nonlinear optimization problem since θ_0 and θ_1 enter the objective function in the arguments of the basis functions for cubic B-splines. We suggest an iteration scheme, which we actually employed in simulation study and real data example. The procedure starts by initializing $\hat{\theta}_0$ and $\hat{\theta}_1$. Then, (i) find $\hat{m}_0$ and the B-spline coefficients of $\hat{m}_j$ under the constraints (4.1) that minimize (2.4); (ii) after we get the estimators $\hat{m}_0$ and $\hat{m}_j$, we update $\hat{\theta}_0$ and $\hat{\theta}_1$ by minimizing (2.4) with the estimated $\hat{m}_j$ being plugged into (2.4). The minimization problem in (i) requires another iteration if G is not the identity link, while the updating task in (ii) is nonlinear and thus needs an iteration even if G is the identity link.

We iterate the above procedures (i) and (ii) until convergence. We stop the iteration when the changes in $\hat{m}_j$ are sufficiently small. To be more specific, let $\hat{m}_j^{[\ell]}$ for $0 \le j \le 3$ and $\hat{\lambda}^{[\ell]}$ denote the component estimators at the ℓth iteration step. When

$$(\hat{m}_0^{[\ell]} - \hat{m}_0^{[\ell-1]})^2 + \sum_{j=1}^{3} \int (\hat{m}_j^{[\ell]}(u) - \hat{m}_j^{[\ell-1]}(u))^2 du + \int (\hat{\lambda}^{[\ell]}(u) - \hat{\lambda}^{[\ell-1]}(u))^2 du$$

falls below a threshold value, we stop the iteration for $\hat{m}_j$ taking $\hat{m}_j = \hat{m}_j^{[\ell]}$ and then find the final update $\hat{\lambda}^{[\ell]}$ for $\hat{\lambda}$ by minimizing (2.4). This iteration scheme worked very well in our numerical studies. As the threshold value in the stopping criterion, we chose 10^{-4}.

4.2. Finite sample simulation study

We generated (X, Y) from the uniform distribution over $[0, 1]^2$ and U from the model $U = m_0 + m_1(X) + m_2(\lambda(X)Y) + m_3(X + Y) + \varepsilon$, where $\lambda(z) = \theta_0 + \theta_1 z$ and $\varepsilon \sim N(0, \sigma^2)$. We set $m_0 = 1$, $m_1(z) = x^2$, $m_2(z) = z^2/10 + z$, $m_3(z) = z^3/8$, $\theta_0 = \theta_1 = 1$. We chose three noise levels, $\sigma^2 = 0.01, 0.1, 0.3$, and two sample sizes $n = 400$ and $1,000$. The penalty constants $\rho_{j,n}$ were set to $\rho_n = 0.12 \times n^{-2k/(2k+1)} = 0.12 \times n^{-4/5}$.

We found that the MISE (Mean Integrated Squared Error) properties of the proposed estimators do not depend much on the choice of the initial values of $\hat{\theta}_0$

and $\hat{\theta}_1$. The results presented in Table 1 are for the choice $\hat{\theta}_0 = 1.3$ and $\hat{\theta}_1 = 0.5$. For these results, we also used equally spaced knots and chose the number of knots that minimizes the mean integrated squared error

$$\text{MISE} = E \int_0^1 \int_0^1 (\hat{\mu}(x,y) - \mu(x,y))^2 \, dx \, dy,$$

where $\hat{\mu}(x,y) = \hat{m}_0 + \hat{m}_1(x) + \hat{m}_2(\hat{\lambda}(x)y) + \hat{m}_3(x+y)$. The table includes the values of MISE, as defined above, IV (Integrated Variance) and ISB (Integrated Squared Bias) defined by

$$\text{IV} = E \int_0^1 \int_0^1 (\hat{\mu}(x,y) - E\hat{\mu}(x,y))^2 \, dx \, dy,$$

$$\text{ISB} = \int_0^1 \int_0^1 (E\hat{\mu}(x,y) - \mu(x,y))^2 \, dx \, dy.$$

It also reports those values for each component estimator $\hat{m}_j$, where for $\hat{m}_0$ they are the values of $E(\hat{m}_0 - m_0)^2$, $\text{var}(\hat{m}_0)$ and $(E\hat{m}_0 - m_0)^2$, respectively.

The results in Table 1 support that our proposed method works very well for finite sample sizes. Overall, the values of MISE, IV and ISB of the regression function estimator $\hat{\mu}(x,y)$ decrease very fast as the sample size n increases, or as the noise level σ^2 gets smaller. For the component function estimators $\hat{m}_j$ the values of MISE also decreases rather fast as n increases, except $\hat{m}_2$. For the latter component the MISE goes down slowly. If we investigate the numbers more closely, we find that the bias of $\hat{m}_2$ stays unchanged or even gets larger slightly as n increases, which results in the slow decline in MISE. We think this is partly owing to the structural complexity that the function m_2 enters our model in the form of $m_2(\lambda(x)y)$ with another unknown function λ. The bias of the estimator comes from the inaccuracy of the spline approximation, which may be hard to reduce, for this particular component, by choosing the tuning parameters depending on the sample size and noise level. Indeed, we also find in the table that the bias of $\hat{m}_2$ does not improve as the noise level σ^2 decreases. The biases of other component estimators $\hat{m}_j$ do not change much as the sample size or the noise level changes. However, the variances of all component estimators $\hat{m}_j$ decrease as n increases or σ^2 decreases.

4.3. Forecasting asbestos related deaths in the UK

As an example of implementing our method, we considered the UK mesothelioma mortality dataset. It consists of the counts of deaths caused by exposure to asbestos, given by year (1980–2012) and age (25–94) at the time of death. The total number of the deaths during the period and in the range of age is 46,348. Basically, for this dataset one may take the variable x to be the cohort and y the age of death. Thus $x =$ (year of death) $- y$. To put the support of

TABLE 1
Mean integrated squared error (MISE), integrated square bias (ISB) and integrated variance (IV) of the whole regression function estimator $\hat{m}$ and of the component function estimators $\hat{m}_j$ and $\hat{\lambda}$, based on 100 MC samples of sizes $n = 400$ and $n = 1,000$.

n	σ^2	Criterion	$\hat{m}_0$	$\hat{m}_1$	$\hat{m}_2$	$\hat{m}_3$	$\hat{\lambda}$	$\hat{m}$
400	0.01	MISE	0.0048	0.0041	0.0076	0.0090	0.0024	0.0043
		IV	0.0005	0.0019	0.0004	0.0028	0.0009	0.0004
		ISB	0.0043	0.0022	0.0072	0.0062	0.0015	0.0039
	0.1	MISE	0.0095	0.0178	0.0104	0.0278	0.0084	0.0246
		IV	0.0052	0.0156	0.0030	0.0217	0.0068	0.0037
		ISB	0.0043	0.0022	0.0074	0.0061	0.0016	0.0209
	0.3	MISE	0.0198	0.0461	0.0169	0.0725	0.0226	0.0663
		IV	0.0156	0.0445	0.0090	0.0650	0.0207	0.0112
		ISB	0.0042	0.0016	0.0079	0.0075	0.0019	0.0551
1,000	0.01	MISE	0.0029	0.0030	0.0072	0.0056	0.0018	0.0012
		IV	0.0004	0.0006	0.0002	0.0014	0.0002	0.0002
		ISB	0.0025	0.0024	0.0070	0.0042	0.0016	0.0010
	0.1	MISE	0.0060	0.0077	0.0096	0.0189	0.0036	0.0055
		IV	0.0038	0.0056	0.0018	0.0140	0.0021	0.0014
		ISB	0.0022	0.0021	0.0078	0.0049	0.0015	0.0041
	0.3	MISE	0.0130	0.0182	0.0139	0.0468	0.0076	0.0156
		IV	0.0113	0.0164	0.0054	0.0417	0.0061	0.0042
		ISB	0.0017	0.0018	0.0085	0.0051	0.0015	0.0114

the data as a subset of $[0, a] \times [0, b]$ for some $a, b > 0$, we made the following transformation:

$$x = (\text{Year of Death}) - 1980 + 69 - y, \quad y = (\text{Age of Death}) - 25. \tag{4.3}$$

We note that $a = 101$ and $b = 69$ with the above transformation. The lowest possible year of birth, $1980 - 94 = 1886$ for those who died in the year 1980 at the age 94, is transformed to the cohort value $x = 0$ and the highest, $2012 - 25 = 1987$ for those who died in 2012 at the age 25, to $x = 101$. The support set $\mathcal{I}$ of the transformed (x, y) is a parallelogram given by

$$\mathcal{I} = \{(x, y) : 0 \le y \le 69,\ 69 - y \le x \le 101 - y\}.$$

Let $U(k, l)$ denote the death count at age k in year l. For each age $k \in \{25, 26, \ldots, 94\}$ and calendar year $j \in \{1980, 1981, \ldots, 2012\}$, let $x(k, l) = l - 1980 + 69 - y(k, l)$ and $y(k, l) = k - 25$. We considered the model

$$\begin{aligned} U(k,l) = \exp\Big[& m_0 + m_1\big(x(k,l)\big) + m_2\big(\lambda(x(k,l))y(k,l)\big) \\ & + m_3\big(x(k,l) + y(k,l)\big)\Big] + \varepsilon(k,l), \end{aligned} \tag{4.4}$$

where λ is a linear function. In general, it is not easy to check if the dataset comes from a distribution that satisfies the subexponentiality condition in Theorem 1. However, the condition holds if $U(k, l)$ for each given (k, l) follows a Poisson distribution, which one typically assumes for count data such as ours. To choose

K_j, the numbers of knots for the cubic B-splines in the approximation of m_j, we employed a cross validatory (CV) criterion. For the CV criterion, we chose $100 \times \alpha\,\%$ among those ages in $\{25, 26, \ldots, 94\}$ for each calendar year l. Call the set of chosen integers $\mathcal{V}_l$. We then fitted the model at (4.4) with those remaining $100 \times (1-\alpha)\,\%$ of the data on the grid

$$\mathcal{T} \equiv \big\{(k,l) : k \in \{25, 26, \ldots, 94\} \setminus \mathcal{V}_l,\ 1980 \le l \le 2012\big\}.$$

Denote the estimated constant and component functions by $\hat{m}_j^{\mathcal{T}}$ and $\hat{\lambda}^{\mathcal{T}}$. We computed

$$(\mathrm{CV}) = \sum_{(k,l)\in\mathcal{T}} \Big(U(k,l) - \exp\Big[\hat{m}_0^{\mathcal{T}} + \hat{m}_1^{\mathcal{T}}\big(x(k,l)\big) + \hat{m}_2^{\mathcal{T}}\big(\hat{\lambda}^{\mathcal{T}}(x(k,l))y(k,l)\big) + \hat{m}_3^{\mathcal{T}}\big(x(k,l)+y(k,l)\big)\Big]\Big)^2.$$

In our application we chose $\alpha = 0.1$ in the above CV criterion. The penalty constants $\rho_{j,n}$ were set to $\rho_n = 0.0022$. Actually, we found that the associated Hessian matrix of the quadratic objective function at (2.4) was not invertible for too small values of ρ_n. The results of the application of our method to the mortality data are shown in Figures 1. We used the estimated model to forecast the death counts in the future years $2012 + \delta$ for $\delta \ge 1$. For this we considered only the cohort group who were born during the years 1886–1987, among whom the numbers of deaths in the future years were our target for forecasting. The set of transformed (x, y) for those who will die in the year $2012 + \delta$ is

$$\begin{aligned} G_\delta &= \{(x,y) : x + y + 1980 - 69 = 2012 + \delta,\ 0 \le x \le 101,\ 0 \le y \le 69\} \\ &= \{(x,y) : x + y = 101 + \delta,\ 0 \le x \le 101,\ 0 \le y \le 69\} \end{aligned}$$

We computed the forecasted number for the year $2012 + \delta$ by the formula

$$N_\delta = \sum_{(x,y)\in G_\delta} \exp\left[\hat{m}_0 + \hat{m}_1(x) + \hat{m}_2(\lambda(x)y) + \hat{m}_3(x+y)\right]. \tag{4.5}$$

Note that our estimate $\hat{m}_3$ is nonparametric and thus it is defined only on the range of $x + y$, which is the interval $[0, 101]$. We used a quadratic extrapolation of $\hat{m}_3$ for future times $x+y > 101$ in the forecasting formula. We briefly compare this forecasting methodology to the three forecast options I_0, I_1 and I_2 from the discrete age-period-cohort model considered in the applied claims study of Kuang et al. (2011). Had we chosen to use a line instead of a quadratic extrapolation then it would correspond to the I_0 forecast as defined in the latter paper. Had we used a minimum of recent information to estimate our quadratic extrapolation, then it would have corresponded to the I_2 forecast of the latter paper that defines I_1 as the extrapolated line based on a minimum of recent information. In our method we used all available information on the calendar time to estimate a quadratic extrapolation.

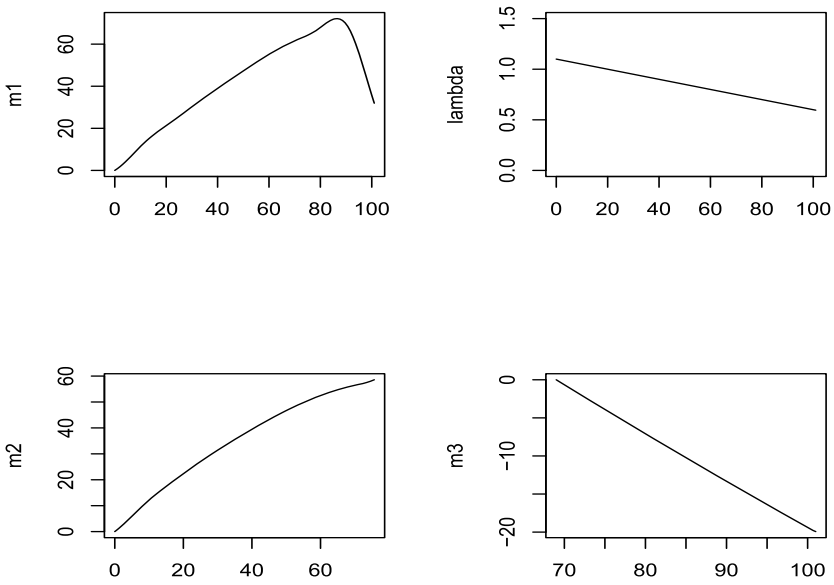

FIG 1. *Estimates of the three component functions m_j and the time transformation λ obtained by applying the model (4.4) to the asbestos mortality data.*

The forecasting result is shown in Figure 2. The peak year is 2018 and the peak number of deaths is 2,572. The total number of deaths during the period 2013–2032 is predicted to be 48,007. Martinez-Miranda et al. (2016) worked on the male-subset of the same data set and estimated male asbestos related deaths to peak in the year 2017 and to total around 2,100 deaths that year. Our predictions are at a higher level because both males and females are considered. However, the two studies do seem to be in a reasonable relationship to each other given that more males than woman seemed to have been exposed to asbestos at life threatening levels. While it is beyond the scope of this paper, it would be interesting to undertake a detailed applied statistical study, where male, female and joint male-female mortality rates of asbestos related deaths in the UK are closely investigated and where the methodology of the current paper and that of Martinez-Miranda et al. (2016) are implemented, compared and discussed. Also, it would be interesting to collect data up till and including 2017 to see whether short term forecasting of the two alternative methods do indeed work around the peak year of number of deaths.

In Figure 1 we notice that time acceleration λ is decreasing indicating that time goes slower implying increasing lifetimes in later calendar years. There is a complicated interaction between the mortality with age component m_2 that has a reasonable mortality shape and then complicated m_1 and m_3 functions that rapidly increase and decrease respectively and in this way to some extent leveling each other out over time. The intuition of the four one-dimensional functions and their interplay is not easy and it is the resulting forecasts, as

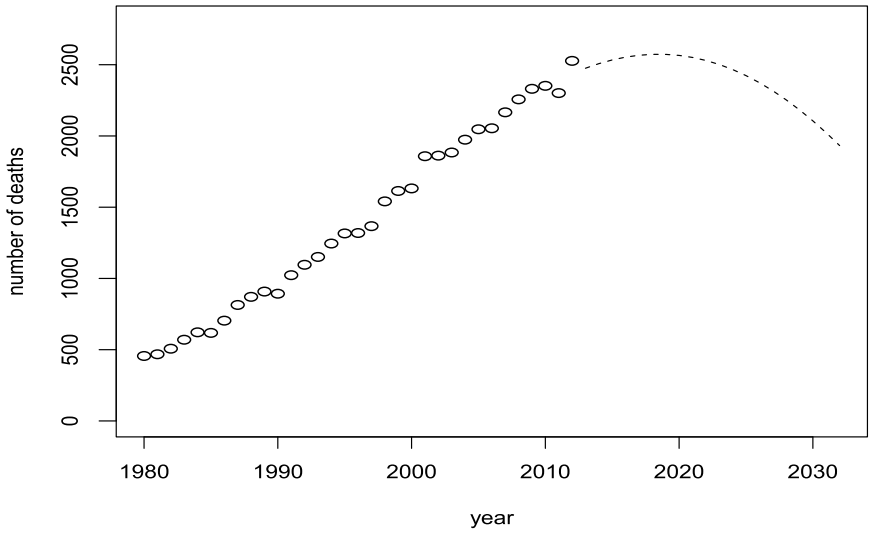

FIG 2. *Small circles indicate the observed numbers of asbestos deaths in the UK until the year 2012. Forecasted numbers of deaths are depicted as a dashed curve.*

shown in Figure 2 with future number of deaths over the years, that perhaps is the best driver of our intuition. However, the forecasts are based on just four functions and that is after all easier to understand than hundreds of discretely estimated parameters as in Martinez-Miranda et al. (2015, 2016). But it still does take a trained applied statistician to do these forecasts well. It will most likely never be a fully automatic exercise.

Appendix. Proofs

Proof of Theorem 1. The main idea of the proof is to apply uniform bounds from empirical process theory and to proceed as e.g. in the proof of Theorem 10.2 in van de Geer (2000). For this purpose we have to check entropy conditions for function classes defined by constraints of our penalty terms. For a related proof in a neural network model, see also Horowitz and Mammen (2007). For a constant $c > |m_0^*|$, where m_0^* denotes the true value of m_0, we define

$$\begin{aligned}\mathcal{F}_c &= \Big\{ f : f(x,y) = G(m_0 + m_1(x) + m_2(\lambda(x)y) + m_3(x+y)) : \\ &\qquad (m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c \Big\}, \\ \mathcal{F}_c^* &= \Big\{ f : f(x,y) = G(m_0 + m_1(x) + m_2(\lambda(x)y) + m_3(x+y)) : \\ &\qquad (m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c^* \Big\},\end{aligned}$$

$$\begin{aligned}\mathcal{M}_c &= \Big\{(m_0,m_1,m_2,m_3,\lambda) : |m_0| \le c, m_1(0) = m_2(0) = m_3(0) = 0,\\ &\qquad T_2(\lambda) = 1\Big\},\\ \mathcal{M}_c^* &= \Big\{(m_0,m_1,m_2,m_3,\lambda) \in \mathcal{M}_c : J(m_1,m_2,m_3,\lambda) \le 1\Big\},\end{aligned}$$

where

$$\begin{aligned}J(m_1,m_2,m_3,\lambda) &= T_1(m_1) + T_2(\lambda)^{(2k-1)/2}\int_{I_2(\lambda)} m_2^{(k)}(z)^2\,dz\\ &\quad + T_2(\lambda)^{1/2}\int_{I_2(\lambda)} m_2'(z)^2\,dz + T_3(m_3).\end{aligned}$$

Note that we replace the norming condition $m_2'(0) = 1$ by $T_2(\lambda) = 1$. This can be done because the penalty function J is defined so that $J(m_1,m_2,m_3,\lambda) = J(m_1,m_2^d,m_3,\lambda^d)$ for all $d > 0$ where $m_2^d(z) = m_2(d^{-1}z)$ and $\lambda^d(x) = d\cdot\lambda(x)$. Note also that we added an additional constraint $|m_0| \le c$ for some constant $c > |m_0^*|$. We will get rid of this constraint below.

We will show below that for $c > 0$ there exists a constant $C > 0$ with

$$H_B(\delta, \mathcal{F}_c^*, \|\cdot\|_\infty) \le C\delta^{-1/k} \tag{A.1}$$

for $\delta > 0$. Here, $H_B(\delta, \mathcal{F}_c^*, \|\cdot\|_\infty)$ denotes the δ-entropy with bracketing for the class $\mathcal{F}_c^*$ with respect to the supnorm $\|\cdot\|_\infty$. We now argue that (A.1) implies the statement of the theorem. For seeing this one can proceed as in the classical empirical process results for penalized least squares, see van de Geer (2000). We shortly outline this now. One first makes use of the basic inequality

$$\|G\hat{m}^c - Gm^*\|_n^2 + \rho_n J(\hat{m}^c) \le 2|\langle\varepsilon, G\hat{m}^c - Gm^*\rangle_n| + \rho_n J(m^*). \tag{A.2}$$

Here, ε, $G\hat{m}^c$ and Gm^* denote the n-dimensional vectors with elements ε_i, $G(\hat{m}_0^c+\hat{m}_1^c(x_i)+\hat{m}_2^c(\hat{\lambda}^c(x_i)y_i)+\hat{m}_3^c(x_i+y_i))$ and $G(m_0^*+m_1^*(x_i)+m_2^*(\lambda^*(x_i)y_i)+m_3^*(x_i+y_i))$, respectively. Also, by $\|x\|_n$ we denote the empirical norm $\|x\|_n^2 = n^{-1}\sum_{i=1}^n x_i^2$ for $x \in \mathbb{R}$. The estimators $\hat{m}_0^c, \hat{m}_1^c, \ldots$ are the penalized least squares estimators in the model $\mathcal{M}_c$ where the additional constraint $|m_0| \le c$ has been put. By $m_0^*, m_1^*, \ldots$ we denote the components of the true regression function. Note that we have assumed that c is large enough such that the true regression function lies in $\mathcal{F}_c$.

To get a bound for the first term on the right hand side of the basic inequality (A.2) one can apply the following result:

For deterministic values $v_1, \ldots v_n$ in $\mathbb{R}^d$ for some $d \geq 1$ and for independent random variables ε_i in $\mathbb{R}$ $(1 \leq i \leq n)$ suppose that $E(\varepsilon_i) = 0$ and that $E[\exp(C^{-1}|\varepsilon_i|)] \leq C$ for C large enough. Then for a bounded class $\mathcal{A}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$ with

$$\sup_{\delta>0} \delta^{\nu} H_B(\delta, \mathcal{A}, \|\cdot\|_{\infty}) < \infty$$

for some $0 < \nu < 2$, it holds that

$$\sup_{a \in \mathcal{A}} \frac{n^{-1} \sum_{i=1}^n \varepsilon_i a(v_i)}{(\max\{\|a\|_n, n^{-1/(2+\nu)}\})^{1-\nu/2}} = O_P(n^{-1/2}),$$

where $\|a\|_n$ is the empirical norm such that $\|a\|_n^2 = n^{-1} \sum_{i=1}^n a(v_i)^2$.

This result can be achieved from Corollary 8.8 in van de Geer (2000). See also the remark after the statement of this corollary.

One can easily check that the elements of $G\hat{m}^c$ are absolutely bounded by a constant times $\sqrt{J(\hat{m}^c)}$. To see this, note that for a function $f : [0,1] \to \mathbb{R}$ with $\int_0^1 f(x)^2 \mathrm{d}x \leq J$ and $\int_0^1 f'(x)^2 \mathrm{d}x \leq J$ with a constant J one can find an element $x^* \in [0,1]$ with $|f(x^*)| \leq \sqrt{J}$. This gives for all $x \in [0,1]$ that $|f(x)| \leq |f(x^*)| + |\int_{x^*}^x f'(u)\mathrm{d}u| \leq |f(x^*)| + (\int_0^1 f'(u)^2 \mathrm{d}u)^{1/2} \leq 2\sqrt{J}$.

We apply the empirical process result with $a = (G\hat{m}^c - Gm^*)/(1 + \sqrt{J(\hat{m}^c)})$. Put $\hat{J}^{1/2} = 1 + \sqrt{J(\hat{m}^c)}$ and $\Delta = \|G\hat{m}^c - Gm^*\|_n$. Then we get that

$$\begin{aligned} |\langle \varepsilon, G\hat{m}^c - Gm^* \rangle_n| &= O_P\left(n^{-1/2} \left(\frac{\Delta}{\hat{J}^{1/2}} \right)^{1-\nu/2} \hat{J}^{1/2} \right) \\ &= O_P\left(n^{-1/2} \Delta^{(2k-1)/(2k)} \hat{J}^{1/(4k)} \right) \end{aligned}$$

on the event that $\Delta \hat{J}^{-1/2} \geq n^{-1/(2+\nu)} = n^{-k/(2k+1)}$ and

$$|\langle \varepsilon, G\hat{m}^c - Gm^* \rangle_n| = O_P\left(n^{-1/2} n^{-\frac{1-\nu/2}{2+\nu}} \hat{J}^{1/2} \right) = O_P\left(n^{-2k/(2k+1)} \hat{J}^{1/2} \right)$$

on the event that $\Delta \hat{J}^{-1/2} \leq n^{-1/(2+\nu)} = n^{-k/(2k+1)}$. For the first event we get from the basic inequality (A.2) and $\hat{J} \leq \Delta^2 n^{2k/(2k+1)}$ that

$$\begin{aligned} \Delta^2 &= O_P\left(n^{-1/2} \Delta^{(2k-1)/(2k)} \hat{J}^{1/(4k)} \right) + O_P(\rho_n) \\ &= O_P\left(n^{-1/2} \Delta n^{1/(2(2k+1))} \right) + O_P(\rho_n) \\ &= O_P\left(n^{-k/(2k+1)} \Delta \right) + O_P(\rho_n) \\ &= O_P\left(\rho_n^{1/2} \Delta + \rho_n \right). \end{aligned}$$

Thus we have $\Delta^2 = O_P(\rho_n)$ on the first event. Note that because of $\hat{J} \leq \Delta^2 n^{2k/(2k+1)}$ this also implies that $\hat{J} = O_P(1)$ on the first event. On the second

event, we get from the basic inequality (A.2)

$$\begin{aligned}\rho_n \hat{J} &= O_P\left(n^{-2k/(2k+1)}\hat{J}^{1/2} + \rho_n\right) \\ &= O_P\left(\rho_n \hat{J}^{1/2} + \rho_n\right).\end{aligned}$$

Thus on this event we have $\hat{J} = O_P(1)$ and because of $\Delta \hat{J}^{-1/2} \leq n^{-k/(2k+1)}$ we get that $\Delta = O_P(n^{-k/(2k+1)})$. This shows that

$$\|G\hat{m}^c - Gm^*\|_n^2 = O_P(n^{-2k/(2k+1)}). \tag{A.3}$$

Compare also Mammen and van de Geer (1997) for a related application of the above empirical process bound.

We now argue that $G\hat{m}^c = G\hat{n}$ with probability tending to one. For this claim we make use of the result $J(\hat{m}^c) = O_P(1)$ that we have just proved. We now argue that this implies that the derivatives of $\hat{m}_1^c$, $\hat{m}_2^c$, $\hat{\lambda}^c$, and $\hat{m}_3^c$ are uniformly bounded by a random variable that is of order $O_P(1)$. For a proof of this claim we argue first that the L_2 norms of the first order and second order derivatives of these functions are of order $O_P(1)$. This gives a bound of order $O_P(1)$ for the supnorm of the first order derivatives where in this step we make use of the same argument used above for showing that the elements $G\hat{m}^c$ are absolutely bounded by a constant times $\sqrt{J(\hat{m}^c)}$. For bounding the L_2 norms of the first order and second order derivatives one can make use of interpolation inequalities: it holds that $\int(\varphi^{(j)}(x))^2 \mathrm{d}x \leq C\gamma^{-2j}\int(\varphi(x))^2\mathrm{d}x + C\gamma^{2(l-j)}\int(\varphi^{(l)}(x))^2\mathrm{d}x$ for functions φ with $\int(\varphi^{(l)}(x))^2\mathrm{d}x < \infty$, for $1 \leq j \leq l$ and for $\gamma > 0$ with a constant C depending only on the integration region, see Agmon (1965).

By applying these bounds for the first-order derivatives we obtain

$$|\hat{m}_1^c(x)| \leq R_n\delta_n, \quad |\hat{m}_2^c(\hat{\lambda}^c(x)y)| \leq R_n\delta_n, \quad |\hat{m}_3^c(x+y)| \leq R_n\delta_n, \quad 0 \leq x, y \leq \delta_n$$

for a random variable $R_n = O_P(1)$, where δ_n is defined in the statement of the theorem and fulfills $n^{k/(2k+1)}\delta_n \to \infty$ and $\delta_n \to 0$. Choose $\delta > 0$. We get with some further constants $C_1, C_2, \ldots > 0$:

$$\begin{aligned}&O_P(\delta_n^{-2}n^{-2k/(2k+1)}) \\ &\quad\geq \frac{1}{n}\delta_n^{-2}\sum_{i=1}^n \Big(G(\hat{m}_0^c + \hat{m}_1^c(x_i) + m_2(\hat{\lambda}^c(x_i)y_i) + \hat{m}_3^c(x_i+y_i)) \\ &\qquad -G(m_0^* + m_1^*(x_i) + m_2(\lambda^*(x_i)y_i) + m_3^*(x_i+y_i))\Big)^2 \\ &\qquad \times I_{[|x_i|\leq\delta_n, |y_i|\leq\delta_n, |\hat{m}_0^c - m_0^*|\geq\delta]} \\ &\quad\geq n^{-1}\delta_n^{-2}C_1\sum_{i=1}^n(|\hat{m}_0^c - m_0^*| - C_2R_n\delta_n)^2 I_{[|x_i|\leq\delta_n, |y_i|\leq\delta_n, |\hat{m}_0^c - m_0^*|\geq\delta]} \\ &\quad\geq C_3\delta^2 I_{[|\hat{m}_0^c - m_0^*|\geq\delta]} + o_P(1).\end{aligned}$$

Thus we have that the probability of the event $\{|\hat{m}_0^c - m_0^*| \geq \delta\}$ converges to 0 and $\hat{m}_0^c = m_0^* + o_P(1)$. This shows that with probability tending to one the

constraint $|m_0| \le c$ is not active in the calculation of $\hat{m}_0^c, \hat{m}_1^c, \ldots$. This implies that $G\hat{m}^c = G\hat{m}$ with probability tending to one.

We now come to the proof of the entropy bound (A.1). For the proof we make use of the entropy bound for Sobolev classes,

$$H_B\left(\delta, \left\{g : [0,1] \to \mathbb{R} : \|g\|_\infty \le 1, \int_0^1 g^{(k)}(x)^2\, dx \le 1\right\}, \|\cdot\|_\infty\right) \le C_1 \delta^{-1/k} \tag{A.4}$$

for some constant $C_1 > 0$. For (A.4) the reader is referred to Birman and Solomjak (1967). Using similar arguments as above we now argue that λ, m_1, m_2 and m_3 are uniformly absolutely bounded. Note that for $(m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c^*$ we have, because of $T_2(\lambda) = 1$, that $|\lambda(x^\lambda)| < C_2$ for an element x^λ of $\mathcal{I}_1$ depending on λ with some constant $C_2 > 0$ (not depending on λ). Thus $\|\lambda\|_\infty \le C_2 + \sup_{x^* \in \mathcal{I}_1} \int |\lambda'(x)| I_{[x^* \le x \le x^\lambda \text{ or } x^\lambda \le x \le x^*]} \mathrm{d}x \le C_2 + C_3 \left(\int_{\mathcal{I}_1} |\lambda'(x)|^2 \mathrm{d}x\right)^{1/2} \le C_2 + C_3$ for some constant $C_3 > 0$. In particular, the length of the intervals $I_2(\lambda)$ can be uniformly bounded. We have $T_1(m_1) \le 1$, $T_3(m_3) \le 1$ and

$$\int_{I_2(\lambda)} m_2^{(k)}(x)^2\, dx + \int_{I_2(\lambda)} m_2'(x)^2\, dx \le 1. \tag{A.5}$$

From all these inequalities we get that $\|m_j\|_\infty \le C_4$ for $1 \le j \le 3$ for some constant $C_4 > 0$. Thus, we can apply the entropy bound (A.4) for $1 \le j \le 3$ to all functions m_j with $(m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c^*$ for some $m_k (k \ne j)$ and λ, and to the Sobolev class of all functions λ with $(m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c^*$ for some m_0, m_1, m_2, m_3. Thus, the bound (A.1) follows if we prove that $\|m_2'\|_\infty \le C_5$ for some constant $C_5 > 0$ as long as $(m_0, m_1, m_2, m_3, \lambda) \in \mathcal{M}_c^*$ for some m_0, m_1, m_3 and λ. Such a bound for $\|m_2'\|_\infty$ can be achieved by noting that $\int_{I_2(\lambda)} m_2''(x)^2\, dx < C_6$ for some constant $C_6 > 0$. This inequality follows with $j = 1$, $l = k - 1$, $\varphi = m_2'$ from the interpolation inequality that we already used above: $\int (\varphi^{(j)}(x))^2 \mathrm{d}x \le C\gamma^{-2j} \int (\varphi(x))^2 \mathrm{d}x + C\gamma^{2(l-j)} \int (\varphi^{(l)}(x))^2 \mathrm{d}x$ for functions φ with $\int (\varphi^{(l)}(x))^2 \mathrm{d}x < \infty$, for $1 \le j \le l$ and for $\gamma > 0$ with a constant C depending only on the integration region, see Agmon (1965). This concludes the proof of the theorem. □

Proof of Theorem 2. The theorem follows by a simple application of Theorem 1. Choose $\kappa_n \to \infty$. Then, Theorem 1 implies that

$$P(n^{-1} \sum_{i=1}^n (\hat{\mu}(X_i, Y_i) - \mu(X_i, Y_i))^2 > \kappa_n n^{-2k/(2k+1)} \mid X_1, Y_1, \ldots, X_n, Y_n) \to 0,$$

almost surely. This implies that the expectation of this conditional probability converges to zero. Because this holds for all $\kappa_n \to \infty$ we get the statement of the theorem. □

Proof of Theorem 3. Choose $\bar{m}_1$, $\bar{m}_2$, and $\bar{m}_3$ with

$$m_1(x) + m_2(\lambda(x)y) + m_3(x + y) = \bar{m}_1(x) + \bar{m}_2(\lambda(x)y) + \bar{m}_3(x + y)$$

for $(x, y) \in \mathcal{I}^0$. Put $\delta_j = \bar{m}_j - m_j$ for $j \in \{1, 2, 3\}$. Then

$$\delta_1(x) + \delta_2(\lambda(x)y) + \delta_3(x + y) = 0. \tag{A.6}$$

We have to show that this equality implies

$$\delta_j(x) = c_j \text{ for } x \in I_j \tag{A.7}$$

for some real numbers c_j for $j \in \{1, 2, 3\}$. From (A.6) we get

$$\begin{aligned} \delta_1'(x) + \lambda'(x)y\delta_2'(\lambda(x)y) + \delta_3'(x + y) = 0, \\ \lambda(x)\delta_2'(\lambda(x)y) + \delta_3'(x + y) = 0, \end{aligned} \tag{A.8}$$

and thus

$$\begin{aligned} \lambda'(x)\delta_2'(\lambda(x)y) + \lambda'(x)\lambda(x)y\delta_2''(\lambda(x)y) + \delta_3''(x + y) = 0, \\ \lambda^2(x)\delta_2''(\lambda(x)y) + \delta_3''(x + y) = 0. \end{aligned} \tag{A.9}$$

By taking the difference of the last two equations we get

$$\lambda'(x)\delta_2'(\lambda(x)y) + [\lambda'(x)\lambda(x)y - \lambda^2(x)]\delta_2''(\lambda(x)y) = 0. \tag{A.10}$$

Choose $(x_1, y_1) \in \mathcal{I}^0$. We consider the following three cases of (x_1, y_1).

Case 1: There exists a tuple $(x_2, y_2) \in \mathcal{I}^0$ with $\lambda(x_2)y_2 = \lambda(x_1)y_1$ and $\lambda'(x_2)/\lambda^2(x_2) \neq \lambda'(x_1)/\lambda^2(x_1)$.

Case 2: For all $\epsilon > 0$ there exist $x \in I_1^0, |x - x_1| \leq \epsilon$ with $\lambda(x) \neq \lambda(x_1)$ and for all tuples $(x_2, y_2) \in \mathcal{I}^0$ with $\lambda(x_2)y_2 = \lambda(x_1)y_1$ it holds that $\lambda'(x_2)/\lambda^2(x_2) = \lambda'(x_1)/\lambda^2(x_1)$.

Case 3: There exists an $\epsilon > 0$ such that $\lambda(x) = \lambda(x_1)$ for all $x \in I_1^0, |x - x_1| \leq \epsilon$.

In Case 1 we put $z = \lambda(x_1)y_1 = \lambda(x_2)y_2$. Then, it holds that

$$a_j\delta_2'(z) + b_j\delta_2''(z) = 0,$$

where $a_j = \lambda'(x_j)$ and $b_j = \lambda'(x_j)z - \lambda^2(x_j)$ for $j \in \{1, 2\}$. Because of $\lambda'(x_1)/\lambda^2(x_1) \neq \lambda'(x_2)/\lambda^2(x_2)$ the matrix

$$\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}$$

has full rank. Thus, we have that

$$\delta_2'(z) = \delta_2''(z) = 0. \tag{A.11}$$

From (A.8) we also get that $\delta_3'(x_1 + y_1) = 0$ and $\delta_1'(x_1) = 0$.

Consider now Case 2. In this case we have for x in a neighborhood of x_1 that $\lambda'(x)/\lambda^2(x) = \rho_0$ with $\rho_0 = \lambda'(x_1)/\lambda^2(x_1)$. Solutions of this differential equation fulfill that $\lambda(x) = -(\rho_1 + \rho_0 x)^{-1}$ for x in a neighborhood of x_1 with some constant ρ_1. From (A.10) we get for $\delta_2'(z)$ for z in a neighborhood of

$y_1\lambda(x_1)$ the differential equation $\rho_0\delta_2'(z) + (\rho_0 z - 1)\delta_2''(z) = 0$. Solutions of this equation fulfill with a constant ρ_2 for z in a neighborhood of $y_1\lambda(x_1)$ that $\delta_2'(z) = \rho_2(\rho_0 z - 1)^{-1}$. From (A.8) we get that

$$\delta_3'(x+y) = -\rho_2\big(\rho_0(x+y) + \rho_1\big)^{-1}.$$

Putting these expressions of λ, δ_2' and δ_3' into the first equation of (A.8) gives $\delta_1'(x) = \rho_2(\rho_1 + \rho_0 x)^{-1}$.

We now come to Case 3. Now we have that $\lambda(x) = \lambda(x_1)$ for x in a neighborhood of x_1. From (A.9) we get that $\delta_3''(x+y) = 0$ and $\delta_2''(\lambda(x)y) = 0$ for all (x, y) in a neighborhood of (x_1, y_1). Thus, $\delta_1'(x)$, $\delta_2'(z)$, and $\delta_3'(x+y)$ are constant for all (x, y) in a neighborhood of (x_1, y_1) and z in a neighborhood of $\lambda(x_1)y_1$ and satisfy $\delta_1'(x) + \delta_3'(x+y) = 0$ and $\lambda\delta_2'(z) + \delta_3'(x+y) = 0$. In particular, we have that $\delta_1'(x) = \gamma_1$ for x in a neighborhood of x_1 with $\gamma_1 = \delta_1'(x_1)$.

Thus, with some constants $\gamma_0 > 0$ and $\gamma_1, \rho_0, \rho_1, \rho_2$, we get for $(x_1, y_1) \in \mathcal{I}^0$ that there exist three possibilities: (1) $\delta_1'(x_1) = 0$, $\delta_1''(x_1) = 0$, $\delta_2'(\lambda(x_1)y_1) = 0$ and $\delta_3'(x_1 + y_1) = 0$; (2) $\delta_1'(x) = \rho_2(\rho_1 + \rho_0 x)^{-1}$, $\delta_1''(x) = -\rho_0\rho_2(\rho_1 + \rho_0 x)^{-2}$ and $\lambda(x) = -(\rho_1 + \rho_0 x)^{-1}$ for x in an interval that contains x_1 and $\delta_2'(\lambda(x_1)y_1) = \rho_2(\rho_0\lambda(x_1)y_1 - 1)^{-1}$and $\delta_3'(x_1 + y_1) = -\rho_2\big(\rho_0(x_1 + y_1) + \rho_1\big)^{-1}$; (3) $\delta_1'(x) = \gamma_1$, $\delta_1''(x) = 0$ and $\lambda(x) = \gamma_0$ for x in an interval that contains x_1 and $\delta_2'(\lambda(x_1)y_1) = \gamma_1/\lambda(x_1)$and $\delta_3'(x_1 + y_1) = -\gamma_1$. We can assume that $\gamma_1 > 0$ and $\rho_2 > 0$ because for (x_1, y_1) in (2) with $\rho_2 = 0$ or in (3) with $\gamma_1 = 0$, respectively, we have that (x_1, y_1) fulfills (1).

We now make use of the continuity of δ_1', δ_1'' and λ. This implies that intervals of x in (2) can only overlap if they have the same constants ρ_0, ρ_1, ρ_2. Furthermore, intervals of x in (3) can only overlap if they have the same constant γ_1. And, values of x_1 with (1) cannot lie in intervals of (2) or in intervals of (3). We conclude that for all $x_1 \in I_1^0$ the same case (1), (2) or (3) holds. By assumption, we have excluded that $\lambda(x) = -(\rho_1 + \rho_0 x)^{-1}$ with some constants ρ_0, ρ_1 or that λ is constant. Thus for all $x_1 \in I_1^0$ case (1) applies and we have that $\delta_1'(x_1) = 0$, $\delta_1''(x_1) = 0$, $\delta_2'(\lambda(x_1)y_1) = 0$ and $\delta_3'(x_1 + y_1) = 0$ for all $(x_1, y_1) \in \mathcal{I}^0$. □

Proof of Theorem 4 . Let the functions m_1, m_2, m_3, λ, $\bar{m}_1$, $\bar{m}_2$, $\bar{m}_3$, $\bar{\lambda}$ satisfy

$$\begin{aligned}
&m_1(x) + m_2(\lambda(x)y) + m_3(x+y)\\
&\quad = \bar{m}_1(x) + \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x+y),\ (x,y) \in \mathcal{I},\\
&\lambda(x) = ax + b, \quad \bar{\lambda}(x) = \bar{a}x + \bar{b},\\
&m_2(0) = \bar{m}_2(0) = m_3(0) = \bar{m}_3(0) = 0, \quad m_2'(0) = \bar{m}_2'(0) = 1
\end{aligned}$$

for some nonzero constants $a, b, \bar{a}$ and $\bar{b}$. Assume that m_1 and $\bar{m}_1$ are differentiable and that m_2 and $\bar{m}_2$ are two times continuously differentiable. With $\delta_1(x) = \bar{m}_1(x) - m_1(x)$ and $\delta_3(x) = \bar{m}_3(x) - m_3(x)$ we have

$$\delta_1(x) + \bar{m}_2((\bar{a}x + \bar{b})y) - m_2((ax + b)y) + \delta_3(x+y) = 0, \quad (x,y) \in \mathcal{I}. \quad \text{(A.12)}$$

By putting $y = 0$ into (A.12), it gives $\delta_1(x) = -\delta_3(x)$ for all $x \in J_1(0)$ and we get

$$\delta_1(x) + \bar{m}_2((\bar{a}x + \bar{b})y) - m_2((ax + b)y) - \delta_1(x+y) = 0 \quad \text{(A.13)}$$

for all x and y with $(x, y) \in \mathcal{I}$ and $x + y \in J_1(0)$. If we differentiate both sides of the equation (A.13) with respect to y and if we put $y = 0$ afterwards, then we obtain

$$\delta_1'(x) = (\bar{a}x + \bar{b}) - (ax + b) = (\bar{a} - a)x + \bar{b} - b, \quad x \in J_1(0).$$

Because of $\delta_1(0) = 0$ this shows

$$\delta_1(x) = \frac{1}{2}(\bar{a} - a)x^2 + (\bar{b} - b)x, \quad x \in J_1(0). \tag{A.14}$$

By plugging this back into (A.13) we establish

$$\bar{m}_2((\bar{a}x + \bar{b})y) - m_2((ax + b)y) = \frac{1}{2}(\bar{a} - a)(2xy + y^2) + (\bar{b} - b)y \tag{A.15}$$

for all x and y with $(x, y) \in \mathcal{I}$ and $x + y \in J_1(0)$. By taking second-order derivatives of both sides of the equation (A.15), we get

$$\begin{aligned} \bar{a}^2\bar{m}_2''((\bar{a}x + \bar{b})y) - a^2 m_2''((ax + b)y) = 0, \\ \bar{a}(\bar{a}x + \bar{b})\bar{m}_2''((\bar{a}x + \bar{b})y) - a(ax + b)m_2''((ax + b)y) = 0 \end{aligned} \tag{A.16}$$

for all x and y with $(x, y) \in \mathcal{I}$ and $x + y \in J_1(0)$.

Now let $v_1(x, y) = \bar{m}_2''((\bar{a}x + \bar{b})y)$ and $v_2(x, y) = m_2''((ax + b)y)$. Then, $\mathbf{v} = (v_1, v_2)^\top$ solves the linear equations

$$\begin{aligned} \bar{a}^2 v_1(x, y) - a^2 v_2(x, y) = 0, \\ \bar{a}(\bar{a}x + \bar{b})v_1(x, y) - a(ax + b)v_2(x, y) = 0. \end{aligned}$$

These equations have the unique solution $\mathbf{v}(x, y) = \mathbf{0}$ if

$$0 \neq \bar{a}^2 a(ax + b) - \bar{a}a^2(\bar{a}x + \bar{b}) = \bar{a}a[\bar{a}b - a\bar{b}].$$

We have two cases $a\bar{b} \neq \bar{a}b$ and $a\bar{b} = \bar{a}b$. In the first case, $\bar{m}_2''((\bar{a}x + \bar{b})y)$ and $m_2''((ax + b)y)$ are zero for all x and y with $(x, y) \in \mathcal{I}$ and $x + y \in J_1(0)$, so that $\bar{m}_2$ and m_2 are linear on a nontrivial interval that includes $\{0\}$. Due to the constraints that $\bar{m}_2'(0) = m_2'(0) = 1$ and $\bar{m}_2(0) = m_2(0) = 0$, we must have

$$\bar{m}_2(z) = z = m_2(z)$$

for all z in the nontrivial interval. Putting this back to (A.15) we get

$$\frac{1}{2}(\bar{a} - a)y^2 = 0. \tag{A.17}$$

for infinitely many y. Thus we conclude $\bar{a} = a$.

Now we consider the second case where $a\bar{b} = \bar{a}b$. Since $a \neq 0$ and $\bar{a} \neq 0$, the latter happens either (i) when $b = \bar{b} = 0$ or (ii) when $b \neq 0$ and $\bar{b} \neq 0$. In the case (i), differentiating (A.15) with respect to y and putting $x = 0$ gives $(\bar{a} - a)y = 0$ for all $y \in J_1(0) \cap J_2(0)$. This gives $a = \bar{a}$. In the case (ii), let

$c = \bar{a}/a = \bar{b}/b$. Differentiating both sides of (A.15) with respect to x and then putting $x = 0$ gives

$$\bar{a} \cdot \bar{m}_2'(\bar{b}y) - a \cdot m_2'(by) = \bar{a} - a, \quad y \in J_1(0) \cap J_2(0).$$

Using the identity $a\bar{b} = \bar{a}b$, we get $\bar{a} \cdot \bar{m}_2'(cz) - a \cdot m_2'(z) = \bar{a} - a$. By integrating both sides of the resulting identity and utilizing the constraints $\bar{m}_2(0) = m_2(0) = 0$, we obtain

$$\bar{m}_2(cz) - m_2(z) = (c-1)z$$

for all z in an interval $[0, \alpha]$ with some $\alpha > 0$. This and the identity (A.15) entail

$$\begin{aligned} \frac{1}{2}(\bar{a} - a)(2xy + y^2) + (\bar{b} - b)y &= \bar{m}_2((\bar{a}x + \bar{b})y) - m_2((ax + b)y) \\ &= \bar{m}_2(c(ax + b)y) - m_2((ax + b)y) \\ &= (c-1)(ax + b)y \end{aligned} \tag{A.18}$$

for infinitely many pairs (x, y). Comparing the coefficients of both sides of the identity (A.18), we establish (A.17) again for infinitely many y. This implies $\bar{a} = a$ so that $c = 1$ and thus $\bar{b} = b$. This completes the proof. □

Proof of Lemma 1. From (3.2) and (3.7) with $y = 0$ we get

$$\bar{m}_1(x) + \bar{m}_3(x) = m_1(x) + m_3(x), \quad x \in J_1(0).$$

Together with (3.7) this gives

$$m_2(\lambda(x)y) + m_3(x + y) - m_3(x) = \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x + y) - \bar{m}_3(x) \tag{A.19}$$

for all $(x, y) \in \mathcal{I}$ with $x \in J_1(0)$. For $x = 0$ we have that $m_2(\lambda(0)y) + m_3(y) = \bar{m}_2(\bar{\lambda}(0)y) + \bar{m}_3(y)$ for all $y \in J_2(0)$. This with (A.19) entails that

$$\begin{aligned} &m_2(\lambda(x)y) + m_3(x + y) + m_2(\lambda(0)x) \\ &\qquad = \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x + y) + \bar{m}_2(\lambda(0)x) \end{aligned} \tag{A.20}$$

for all $(x, y) \in \mathcal{I}$ with $x \in J_1(0) \cap J_2(0)$. Furthermore, it holds that

$$\begin{aligned} m_2(\lambda(0)(x + y)) + m_3(x + y) = \bar{m}_2(\bar{\lambda}(0)(x + y) + \bar{m}_3(x + y), \\ x + y \in J_2(0). \end{aligned} \tag{A.21}$$

Thus, by taking the difference the two equations (A.20) and (A.21) we obtain

$$\begin{aligned} &m_2(\lambda(x)y) - m_2(\lambda(0)(x + y)) + m_2(\lambda(0)x) \\ &\qquad = \bar{m}_2(\bar{\lambda}(x)y) - \bar{m}_2(\bar{\lambda}(0)(x + y)) + \bar{m}_2(\bar{\lambda}(0)x) \end{aligned} \tag{A.22}$$

for all $(x, y) \in \mathcal{I}$ with $x \in J_1(0) \cap J_2(0)$ and $x + y \in J_2(0)$. By taking the derivatives of both sides of (A.22) we get

$$\begin{aligned} &\lambda(x)m_2'(\lambda(x)y) - \lambda(0)m_2'(\lambda(0)(x + y)) \\ &\qquad = \bar{\lambda}(x)\bar{m}_2'(\bar{\lambda}(x)y) - \bar{\lambda}(0)\bar{m}_2'(\bar{\lambda}(0)(x + y)) \end{aligned} \tag{A.23}$$

for all $(x,y) \in \mathcal{I}$ with $x \in J_1(0) \cap J_2(0)$ and $x+y \in J_2(0)$. In particular for $y=0$, this writes as

$$\lambda(x) - \lambda(0) m_2'(\lambda(0)x) = \bar{\lambda}(x) - \bar{\lambda}(0)\bar{m}_2'(\bar{\lambda}(0)x), \quad x \in J_1(0) \cap J_2(0). \tag{A.24}$$

By plugging (A.24) two times into (A.23) we get that, for all $(x,y) \in \mathcal{I}$ with $x+y \in J_1(0) \cap J_2(0)$ and $\bar{\lambda}(x)y/\bar{\lambda}(0) \in J_1(0) \cap J_2(0)$,

$$\begin{aligned}
0 &= \bar{\lambda}(x)\bar{m}_2'(\bar{\lambda}(x)y) - \bar{\lambda}(0)\bar{m}_2'(\bar{\lambda}(0)(x+y)) \\
&\qquad - \lambda(x)m_2'(\lambda(x)y) + \lambda(0)m_2'(\lambda(0)(x+y)) \\
&= \bar{\lambda}(x)\bar{m}_2'(\bar{\lambda}(x)y) - \bar{\lambda}(x+y) - \lambda(x)m_2'(\lambda(x)y) + \lambda(x+y) \\
&= \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}\bar{\lambda}(0)\bar{m}_2'\left(\bar{\lambda}(0)\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}y\right) - \lambda(x)m_2'(\lambda(x)y) - \bar{\lambda}(x+y) + \lambda(x+y) \\
&= \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}\bar{\lambda}\left(\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}y\right) - \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}\lambda\left(\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}y\right) + \frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}\lambda(0)m_2'\left(\lambda(0)\frac{\bar{\lambda}(x)}{\bar{\lambda}(0)}y\right) \\
&\qquad - \lambda(x)m_2'(\lambda(x)y) - \bar{\lambda}(x+y) + \lambda(x+y).
\end{aligned}$$

This shows the claim of the lemma. □

Proof of Theorem 5. Let the functions m_1, m_2, m_3, λ, $\bar{m}_1$, $\bar{m}_2$, $\bar{m}_3$, $\bar{\lambda}$ satisfy

$$\begin{aligned}
&m_1(x) + m_2(\lambda(x)y) + m_3(x+y) = \bar{m}_1(x) + \bar{m}_2(\bar{\lambda}(x)y) + \bar{m}_3(x+y), \\
&\lambda(x) = \lambda'(0)x + \lambda(0), \quad \bar{\lambda}(x) = \bar{\lambda}'(0)x + \bar{\lambda}(0), \quad x \in [0,\varepsilon] \\
&m_2(0) = \bar{m}_2(0) = m_3(0) = \bar{m}_3(0) = 0, \quad m_2'(0) = \bar{m}_2'(0) = 1
\end{aligned}$$

for some $\varepsilon > 0$. We first note that, with the local linearity of λ and $\bar{\lambda}$ at zero, we may follow the arguments in the proof of Theorem 4 for sufficiently small $x, y \geq 0$. Thus, if $\lambda'(0)\bar{\lambda}(0) \neq \bar{\lambda}'(0)\lambda(0)$, then m_2 and $\bar{m}_2$ are linear in a small neighborhood of zero. This means that, if both m_2 and $\bar{m}_2$ are not linear in any small neighborhood of zero, then we must have $\lambda'(0)\bar{\lambda}(0) = \bar{\lambda}'(0)\lambda(0)$, so that we may conclude $\lambda(0) = \bar{\lambda}(0)$ and $\lambda'(0) = \bar{\lambda}'(0)$ as in the proof of Theorem 4.

We now use Lemma 1. If we differentiate both sides of (3.8) with respect to x and put $x = 0$ afterwards, then we get

$$\begin{aligned}
0 = {} & \frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)}\Big(\bar{\lambda}(y) - \lambda(y)\Big) + \left(\frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)}y - 1\right)\Big(\bar{\lambda}'(y) - \lambda'(y)\Big) \\
& + \lambda(0)\left(\frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)} - \frac{\lambda'(0)}{\lambda(0)}\right) m_2'(\lambda(0)y) \\
& + \lambda(0)^2\left(\frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)} - \frac{\lambda'(0)}{\lambda(0)}\right) y\, m_2''(\lambda(0)y), \quad y \in J_1(0) \cap J_2(0).
\end{aligned} \tag{A.25}$$

By integrating both sides of (A.25) we obtain

$$\begin{aligned}
&\left(\frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)}y - 1\right)\Big(\bar{\lambda}(y) - \lambda(y)\Big) + \lambda(0)\left(\frac{\bar{\lambda}'(0)}{\bar{\lambda}(0)} - \frac{\lambda'(0)}{\lambda(0)}\right) y\, m_2'(\lambda(0)y) \\
&\qquad = \lambda(0) - \bar{\lambda}(0), \quad y \in J_1(0) \cap J_2(0).
\end{aligned} \tag{A.26}$$

Since $\lambda(0) = \bar{\lambda}(0)$ and $\lambda'(0) = \bar{\lambda}'(0)$, the identity (A.26) gives that $\bar{\lambda}(y) - \lambda(y) = 0$ for all $y \in J_1(0) \cap J_2(0)$ with $y \cdot \bar{\lambda}'(0)/\bar{\lambda}(0) \neq 1$. Since λ and $\bar{\lambda}$ are continuous, this implies $\bar{\lambda} \equiv \lambda$ on $J_1(0) \cap J_2(0)$. □

References

Agmon, S.(1965). *Lectures on Elliptic Boundary Value Problems*. Van Nostrand, Princeton, NJ. MR0178246

Antonczyk, D., Fitzenberger, B., Mammen, E. and Yu, K. (2017). A nonparametric approach to identify age, time and cohort effects. Preprint.

Beutner, E. A., Reese, S. B. and Urbain, J. P. (2017). Identifiability issues of age-period and age-period-cohort models of the Lee-Carter type. *Insurance: Mathematics and Economics* **75**, 117–125. MR3670066

Birman, M. Š., Solomjak, M. J. (1967). Piecewise polynomial approximations of functions of classes W_p^α (Russian). *Mat. Sb. (N.S.)* **73**, 331–355. MR0217487

Cairns, A. J. G., Blake, D., Dowdb, K., Coughlan, G. D. and Epstein, D. (2011). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics* **48**, 355–367. MR2820048

Hiabu, M., Mammen, E., Martinez-Miranda, M. D. and Nielsen, J. P. (2016). In-sample forecasting with local linear survival densities. *Biometrika* **103**, 843–859. MR3620443

Hodgson, J. T., McElvenny, D. M. and Darnton, A. J. (2005). The expected burden of mesothelioma mortality in Great Britain from 2002 to 2050. *British Journal of Cancer* **92**, 587–593.

Horowitz, J. L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link function. *Annals of Statistics* **35**, 2589–2619. MR2382659

Kuang, D., Nielsen, B. and Nielsen, J. P. (2008a). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 979–986. MR2461224

Kuang, D., Nielsen, B. and Nielsen, J. P. (2008b). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 987–991. MR2461225

Kuang, D., Nielsen, B. and Nielsen, J. P. (2011). Forecasting in an extended chain-ladder-type model. *Journal of Risk and Insurance* **78**, 345–359.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of American Statistical Association* **87**, 659–675

Lee, R. D. and Miller, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography* **38**, 537–549 MR2011262

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B. U. ((2015). Asymptotics for In-Sample Density Forecasting. *Annals of Statistics* **43**, 620–651. MR3319138

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B. U. (2017). Operational time and in-sample density forecasting. *Annals of Statistics* **45**, 1312–1341. MR3662456

Mammen, E. and Nielsen, J. P. (2003). Generalised structured models. *Biometrika* **90**, 551–566. MR2006834

Mammen, E., and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics* **25**, 1014–1035. MR1447739

Martinez-Miranda, M. D., Nielsen, B. and Nielsen, J. P. (2015). Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. *Journal of Royal Statistical Society Series A* **178**, 29–55. MR3291760

Martinez-Miranda, M. D., Nielsen, B. and Nielsen, J. P. (2016). Simple benchmark for mesothelioma projection for Great Britain. *Occupational and Environmental Medicine* **73**, 561–563.

Nielsen, B. and Nielsen, J. P. (2014). Identification and forecasting in mortality models. *The Scientific World Journal*, 347043.

O-Brien, R.(2014). *Age-Period-Cohort Models: Approaches and Analyses with Aggregate Data*. Chapman & Hall/CRC Press, London.

Peto, J., Hodgson, J. T. and Matthews, F. E. (1995). Continuing increase in mesothelioma mortality in Britain. *Lancet* **345**, 535–539.

Rake, C., Gilham, C. and Hatch, J. (2009). Occupational, domestic and environmental mesothelioma risks in the British population: a case-control study. *British Journal of Cancer* **100**, 1175–1183.

Renshaw, A. J. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* **38**, 556–570.

Riebler, A., Held, L. and Rue, H. (2012). Estimation and extrapolation of time trends in registry data-Borrowing strength from related populations. *Annals of Applied Statistics* **6**, 304–333. MR2951539

Smith, T. R. and Wakefield, J. (2016). A review and comparison of age-period-cohort models for cancer incidence. *Statistical Science* **31**, 591–610. MR3598741

Tan, E., Warren, N. and Darnton, A. J. (2010). Projection of mesothelioma mortality in Britain using Bayesian methods. *British Journal of Cancer* **103**, 430–436.

Tan, E., Warren, N., Darnton, A. J. (2011). Modelling mesothelioma mortality in Great Britain using the two-stage clonal expansion model. *Occupational and Environmental Medicine* **68**, A60. doi:10.1136/oemed-2011-100382.194

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.