Calendar effect and in-sample forecasting

Enno Mammen
Heidelberg University, Germany
mammen@math.uni-heidelberg.de
María Dolores Martínez-Miranda
University of Granada, Spain
mmiranda@ugr.es
Jens Perch Nielsen
Cass Business School, City, University of London, U.K.
Jens.Nielsen.1@city.ac.uk
Michael Vogt
University of Bonn, Germany
michael.vogt@uni-bonn.de

March 14, 2020

A very popular forecasting tool in the actuarial sciences is the so-called chain ladder. Mammen et al. (2015) recently introduced in-sample forecasting - generalizing continuous chain ladder of Martínez-Miranda et al. (2013) - as a general forecasting technique applicable in many fields. The main aim of this paper is to develop an extended version of the continuous chain ladder which is of interest not only for actuaries but which has many potential applications in economics and other fields. The statistical problem underlying the extended continuous chain ladder is to estimate and forecast a structured nonparametric density. In the theoretical part of the paper, we develop methodology to approach this problem. The usefulness of the methods is illustrated by empirical examples from economics and the actuarial sciences.

Keywords: nonparametric density estimation, kernel smoothing, backfitting.
JEL classifications: C14, C53.

1 Introduction

One of the main forecasting tools in the actuarial sciences is the so-called chain ladder methodology. Recently, Martínez-Miranda et al. (2013) and Mammen et al. (2015) introduced and generalized the continuous chain ladder approach to improve on the classic technique. In this paper, we develop an extension of the continuous chain ladder which is not only useful for forecasting problems in the actuarial sciences but which is of much broader interest with many potential applications in economics.

The forecasting problem underlying the continuous chain ladder is as follows: Suppose we observe a data sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$, where (X_i, Y_i) are i.i.d. copies of a two-dimensional random variable (X, Y). Moreover, assume that (X, Y) has a multiplicative density of the form $f(x, y) = f_1(x)f_2(y)$ which is supported on a proper subset $\mathcal{I}$ of the unit square $[0, 1]^2$. To fix ideas, let $\mathcal{I}$ be the triangle $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x + y \leq 1\}$. The aim is to estimate the multiplicative components f_1 and f_2 and to forecast the density f to the region $\mathcal{I}^c := [0, 1]^2 \setminus \mathcal{I}$ with the help of the produced estimates. As shown in Mammen et al. (2015) and Lee et al. (2015), the density components $f_j : [0, 1] \to \mathbb{R}$ can be estimated by an iterative backfitting algorithm which yields estimates $\hat{f}_j : [0, 1] \to \mathbb{R}$ for $j = 1, 2$. This allows to define an estimator $\hat{f}(x, y) = \hat{f}_1(x)\hat{f}_2(y)$ of the density $f(x, y)$ for all points $(x, y) \in [0, 1]^2$. Hence, it is possible to estimate the density $f(x, y)$ not only on the support $\mathcal{I}$ but on the whole unit square $[0, 1]^2$. In particular, it is possible to predict the density $f(x, y)$ at points $(x, y) \in \mathcal{I}^c$ without any extrapolation. Martínez-Miranda et al. (2013) introduced the term *in-sample forecasting* to describe this phenomenon.

On first sight, the forecasting problem underlying the continuous chain ladder appears to be rather specific. However, it turns out to be a suitable framework for a wide range of applications. We give some examples to illustrate this.

Example 1. We first revisit the original application of the chain ladder methodology: claims reserving of outstanding liabilities in non-life insurance. Suppose we have data on n different insurance claims $i = 1, \ldots, n$. For each claim i, we observe (X_i, Y_i), where X_i is the time point when claim i incurred and Y_i is the time delay with which the claim was reported to the insurance. The value $X_i + Y_i$ specifies the actual time point when claim i was reported and is usually called *calendar time* in the literature. For simplicity of exposition, we normalize X_i and Y_i to take values in the unit interval $[0, 1]$, implying that the data points (X_i, Y_i) are supported on the triangle $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x + y \leq 1\}$. The diagonals of the triangle $\mathcal{I}$ correspond to calendar time. In particular, for each $t \in [0, 1]$, the diagonal $D(t) = \{(x, y) \in \mathcal{I} : x+y = t\}$ specifies the points (x, y) with calendar time $x+y = t$. One of the most important problems in non-life insurance is to forecast the density f of the observations (X_i, Y_i) to the upper triangle $\mathcal{I}^c := [0, 1]^2 \setminus \mathcal{I}$. More precisely speaking, actuaries are mainly interested in approximating the quantity $\mathcal{L} = \int_{(x,y)\in\mathcal{I}^c} f(x, y)dxdy$, which gives the outstanding number of insurance claims and thus represents the future liabilities for the insurance company. We come back to this example in Section 6.1.

Example 2. The next example is concerned with the analysis of fertility, which is an important subject in economics; cp. Aaronson et al. (2014), Baudin et al. (2015),

Momota (2016), Cooley and Henriksen (2018) and Cooley et al. (2019) among many others for economic studies of fertility. Fertility trends have profound societal and economic implications. In many developed countries, fertility rates have dropped dramatically over the last few decades, which has strong impacts on key economic variables such as growth, tax returns and social security contributions. Paired with longer life expectancy, the sharp decrease in fertility poses serious problems to social welfare and health care systems. A better understanding of the dynamics of fertility and reliable projections into the future are of vital importance to deal with these problems.

Consider the following forecasting problem: Suppose we observe data (X_i, Y_i) on a large number of births $i = 1, \dots, n$, where X_i is the birth cohort of the mother (i.e. her birth date) and Y_i is the age of the mother at the time of birth i. For simplicity of exposition, we normalize X_i and Y_i to take values in the unit interval $[0, 1]$. The data points (X_i, Y_i) are thus supported on the trapezium $\mathcal{I} = \{(x, y) \in [0, 1]^2 : c \leq x + y \leq 1\}$ for some $c > 0$, where $[c, 1]$ is the interval of calendar times for which we have observations. Our aim is to model and forecast the fertility density f of the observations (X_i, Y_i). A simple model is $f(x, y) = f_1(x)f_2(y)$, where f_1 represents the child-birth density w.r.t. cohort of the mother and f_2 the child-birth density w.r.t. age of the mother. This simple model decomposes the fertility density f into a cohort effect f_1 and an age effect f_2. With the help of the continuous chain ladder approach, we can estimate the density components f_1 and f_2 and forecast the fertility density f into the future, i.e., to the upper triangle $\{(x, y) \in [0, 1]^2 : x + y > 1\}$. In Section 6.2, we analyze this application example in detail.

Example 3. A further example comes from labour economics and is concerned with unemployment forecasts. Suppose we observe data (X_i, Y_i) for a large number of unemployment benefit receivers $i = 1, \dots, n$, where X_i is the time point where individual i started to receive benefits and Y_i is the benefit duration. As in the previous examples, we normalize X_i and Y_i to take values in the unit interval $[0, 1]$, implying that the data points (X_i, Y_i) are supported on the triangle $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x+y \leq 1\}$. Our goal is to model and forecast the density f of the observations (X_i, Y_i). Imposing the simple model $f(x, y) = f_1(x)f_2(y)$, the density f can be projected into the future by in-sample forecasting and the resulting density forecast can be used to make unemployment predictions. For instance, we can compute forecasts of the quantity $\mathcal{L}(t) = \int_{(x,y)\in D(t)} f(x, y)dxdy$ for $t > 1$ with $D(t) = \{(x, y) \in [0, 1]^2 : x + y = t\}$, which (roughly speaking) specifies the number of benefit leavers at the future time point $t > 1$. This application example has been studied in detail in Wilke (2018), where the classic chain ladder methodology was compared with several other forecasting methods.

The list of application examples could be easily continued. A further interesting field of application are mortality studies; see e.g. Mammen et al. (2015), Martínez-Miranda et al. (2015) and Martínez-Miranda et al. (2016) for an application to asbestos mortality. Even though mortality studies are mainly conducted in epidemiology and public health, they are also relevant for questions in health economics. We finally mention that the continuous chain ladder methodology can also be used to analyze obesity rates. Recently, the issue of obesity and its economic implications have received growing attention in the economics literature; see e.g. the studies in Baum and Ruhm (2009), An and Xiang (2016) and Fannon et al. (2018).

Even though the density model $f(x, y) = f_1(x)f_2(y)$ is a good baseline to approach a number of forecasting problems such as those in Examples 1–3, one may argue that it is too simplistic. In many applications including those of Examples 1–3, one can expect the density $f(x, y)$ not to be a simple product of two components $f_1(x)$ and $f_2(y)$. The fertility density $f(x, y)$ in Example 2, for instance, is presumably not only influenced by a cohort-of-mother effect $f_1(x)$ and an age-of-mother effect $f_2(y)$. Quite arguably, it is also strongly affected by the societal and economic conditions of the time, that is, an additional calendar time effect is likely to be present. A more plausible model for the fertility density f thus has the form

$$f(x, y) = f_1(x)f_2(y)f_3(x + y), \tag{1.1}$$

where $f_3(x + y)$ is an additional density component that depends on calendar time $x + y$. In the sequel, we call (1.1) the extended continuous chain ladder model and f_3 the calendar effect of the model. Similarly, a calendar effect can be expected to be present in Example 3 since the unemployment density $f(x, y)$ is very likely to be influenced by the macroeconomic conditions at calendar time $x + y$, or put differently, by business cycle fluctuations.

The extended continuous chain ladder model (1.1) is much more difficult to handle than the simple version $f(x, y) = f_1(x)f_2(y)$ without a calendar effect. First of all, the estimation theory for the simple model developed in Mammen et al. (2015) and Lee et al. (2015) does not carry over to the extended model in a straightforward way. Moreover, in the extended model, in-sample forecasting is not possible any more. The problem is that the calendar effect $f_3(x + y)$ can only be estimated at time points $x + y \leq 1$ (given that $\mathcal{I}$ is the triangle or trapezium support). Hence, to predict the density $f(x, y)$ at future time points $x + y > 1$, the function f_3 needs to be extrapolated. The main contribution of our paper is twofold: (i) We provide new methodology and theory for estimating the density components f_1, f_2 and f_3 in the extended continuous chain ladder model (1.1). (ii) We develop a novel forecasting strategy to forecast the density f to the region $\mathcal{I}^c = [0, 1] \setminus \mathcal{I}$. Our estimation and

forecasting methodology is developed in Sections 2–5. The practical relevance of our methods is demonstrated by an actuarial application (Example 1) and an economic application (Example 2) in Section 6. Moreover, we investigate the finite sample performance of our methods by a simulation study in Section 7.

The extended continuous chain ladder model (1.1) is closely related to age-period-cohort models which have a long tradition in econometrics, biostatistics, actuarial science and other fields; see e.g. Heckman and Robb (1985), Carstensen (2007) and Kuang et al. (2011). Specifically, model (1.1) can be regarded as a continuous version of the discrete age-period-cohort framework of Kuang et al. (2011). Despite this close connection, our estimation and forecasting methods are very different from those used in the context of age-period-cohort models. In particular, our forecasting strategy differs strongly from the forecasting methods developed in Kuang et al. (2011). In Section 4, we discuss these differences in detail. Moreover, in the supplementary material, we give a new interpretation of the forecasts constructed in Kuang et al. (2011) in the light of our forecasting strategy. We believe that the forecasting approach of Kuang et al. (2008a,b, 2011) and its further developments in Nielsen and Nielsen (2014), Nielsen (2015, 2018), Harnau (2018a,b), Fannon et al. (2018) and Harnau and Nielsen (2018) could benefit from the new insight of this paper and the provided supplementary material.

We finally note that while in-sample forecasting originally grew out of actuarial reserving techniques, only recently, see Bischofberger et al. (2019a), it has been proved that the methodology generalizes to the case where each future event has a marker tight to it (the size of a claim in reserving). The only requirement is that this marker also obeys some structure, which is multiplicativity in the continuous chain ladder case.

2 The extended continuous chain ladder model

In this section, we describe the extended continuous chain ladder model in detail which underlies our analysis. Let $\{(X_i, Y_i) : i = 1, \dots, n\}$ be a sample of data, where (X_i, Y_i) are i.i.d. copies of a two-dimensional random variable (X, Y). The variable (X, Y) is assumed to have a multiplicative density of the form

$$f(x, y) = f_1(x) f_2(y) f_3(x + y) \tag{2.1}$$

which is supported on a proper subset $\mathcal{I}$ of the unit square $[0, 1]^2$, that is, $P((X, Y) \in \mathcal{I}) = 1$. In most applications, including those discussed in the examples of the introduction, $\mathcal{I}$ is either a triangle of the form $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x + y \leq 1\}$ or a trapezium of the form $\mathcal{I} = \{(x, y) \in [0, 1]^2 : c \leq x + y \leq 1\}$ for some $0 < c < 1$. In

what follows, we thus restrict attention to the triangle and the trapezium support. It is however not difficult to extend our methods and theory to other types of support $\mathcal{I}$. Note that Lee et al. (2015) also include a calendar component f_3 in their version of the continuous chain ladder model. However, they impose very severe structural constraints on this component. In particular, they assume that f_3 is a periodic function, which strongly restricts its usefulness in applications. Our model, in contrast, allows for a general component function f_3 which is flexible enough to capture calendar effects present in the data. The precise conditions on the density components f_1, f_2 and f_3 in model (2.1) are laid out below. Our aim is to estimate the density components f_1, f_2 and f_3 from the data sample $\{(X_i, Y_i) : i = 1, \dots, n\}$ and to forecast the density f to the region $\mathcal{I}^c := [0,1]^2 \setminus \mathcal{I}$ with the help of the estimated components.

Importantly, the model formulated in equation (2.1) is not identified, that is, the density components f_1, f_2 and f_3 are not uniquely determined. Specifically, $f(x,y) = f_1(x)f_2(y)f_3(x+y)$ can be rewritten as $f(x,y) = g_1(x)g_2(y)g_3(x+y)$, where

$$g_1(x) = e^{-a_1}e^{-bx}f_1(x) \tag{2.2}$$

$$g_2(y) = e^{-a_2}e^{-by}f_2(y) \tag{2.3}$$

$$g_3(x+y) = e^{a_1+a_2}e^{b(x+y)}f_3(x+y) \tag{2.4}$$

with arbitrary real-valued constants a_1, a_2 and b. For identification, we impose the following conditions: First of all, we normalize the component functions f_1 and f_2 to integrate to 1, that is,

$$\int_0^1 f_1(x)dx = 1 \tag{IC1}$$

$$\int_0^1 f_2(y)dy = 1. \tag{IC2}$$

Conditions (IC1) and (IC2) make sure that f_1 and f_2 can be interpreted as proper densities. In addition, we assume the following: there exists a constant $\kappa^* > 0$ such that f_3 is a constant function on the interval $[1-\kappa^*, 1]$. More formally speaking, we suppose that

$$f_3(z) = \text{const.} \quad \text{for all } z \in [1-\kappa^*, 1], \tag{IC3}$$

where $\kappa^* > 0$ is the largest real number such that (IC3) is satisfied.

The heuristic idea behind the restriction (IC3) is as follows: On a logarithmic scale, the multiplicative density $f(x,y) = f_1(x)f_2(y)f_3(x+y)$ becomes $\log f(x,y) = \log f_1(x) + \log f_2(y) + \log f_3(x+y)$. Suppose that the logarithmic calendar effect $\log f_3$ is differentiable. By definition of differentiability, this means that it can locally

be well approximated by a linear function. Hence, we may in particular assume that $\log f_3$ is approximately linear locally around the present time point $z = 1$, that is,

$$\log f_3(z) = a + bz + \psi(z) \quad \text{for all } z \in [1 - \kappa^*, 1], \tag{2.5}$$

where κ^* is a small positive constant and the function ψ is approximately equal to zero. In terms of the multiplicative model (2.1), this means that

$$f_3(z) = e^{a+bz} e^{\psi(z)} \quad \text{for all } z \in [1 - \kappa^*, 1].$$

As indicated by (2.2)–(2.4), we can renormalize the density components f_1, f_2 and f_3 such that the exponential e^{a+bz} is eliminated from f_3 and shifted to the other two components f_1 and f_2. On a logarithmic scale, this means that the linear part $a + bz$ is subtracted from f_3 and added to the other two component functions. After renormalization, we obtain that $f_3(z) = e^{\psi(z)}$, or equivalently, $\log f_3(z) = \psi(z)$ for all $z \in [1 - \kappa^*, 1]$. Assuming that $\log f_3(z)$ is not only approximately but exactly linear on $[1 - \kappa^*, 1]$, i.e., assuming that $\psi \equiv 0$, we in particular get that $\log f_3(z)$ and thus the calendar effect $f_3(z)$ itself is constant for $z \in [1 - \kappa^*, 1]$. This is exactly the restriction imposed by (IC3). To sum up, our heuristic discussion has shown the following: (IC3) is heuristically motivated by smoothness considerations. In particular, it is approximately fulfilled by any smooth calendar effect f_3 and κ^* sufficiently small. However, it is not satisfied exactly in general. By assuming it, we impose some shape constraint on the calendar effect f_3. From a practical perspective, however, such a constraint is very natural (provided that f_3 is smooth). As we will see in Section 4, a model with the restriction (IC3) on the calendar effect is particularly suited to forecasting purposes.

Under (IC1)–(IC3) together with some smoothness and boundedness conditions, the component functions f_1, f_2 and f_3 are identified in model (2.1). More specifically, we have the following result:

Proposition 1. *Let* (IC1)–(IC3) *be satisfied and assume the following:*

(i) *f_1, f_2 and f_3 are bounded away from zero and infinity on their supports $\mathcal{I}_1 = [0, 1]$, $\mathcal{I}_2 = [0, 1]$ and $\mathcal{I}_3 = \{z \in [0, 1] : z = x + y \text{ for some } (x, y) \in \mathcal{I}\}$, respectively.*

(ii) *f_1 and f_2 are differentiable on $\mathcal{I}_1 = \mathcal{I}_2 = [0, 1]$.*

Then f_1, f_2 and f_3 are identified. More precisely, let g_1, g_2 and g_3 be functions such that $f(x, y) = g_1(x)g_2(y)g_3(x+y)$ for any $(x, y) \in \mathcal{I}$ and let them satisfy (IC1)–(IC3) *along with (i) and (ii). Then $g_j(w) = f_j(w)$ for all $w \in \mathcal{I}_j$ and $j = 1, 2, 3$.*

Proof. Consider the region $\mathcal{I}_{\kappa^*} = \{(x, y) \in \mathcal{I} : 1 - \kappa^* \leq x + y \leq 1\}$. In this region, it holds that $f(x, y) = c_3 f_1(x) f_2(y)$ with $f_3(x + y) = c_3$ and some constant $c_3 \in \mathbb{R}$. Applying Theorem 1 from Lee et al. (2015) to the model $f(x, y) = c_3 f_1(x) f_2(y)$ on the region $\mathcal{I}_{\kappa^*}$, we obtain that the functions f_1 and f_2 are identified on their support $\mathcal{I}_1 = \mathcal{I}_2 = [0, 1]$. From this, it trivially follows that f_3 is identified on $\mathcal{I}_3$ as well. □

3 Estimation method

Lee et al. (2015) studied the problem of estimating the density in model (2.1) when the function f_3 is constant (or more generally, when f_3 is periodic). They proposed a backfitting algorithm to estimate the two density components f_1 and f_2. Among a number of in-sample forecasting techniques Bischofberger et al. (2019b) found that projecting the data down on the multiplicative space of interest (equivalent to the backfitting approach) seems to be the best thing to do in practice. Even better than the dimensional-reducing time reversion trick exploited by Hiabu et al. (2016) and Bischofberger (2020). Backfitting methods are not only used for density estimation. They are also very popular for estimating generalized additive regression models, see Opsomer and Ruppert (1997), Mammen et al. (1999), Yu et al. (2008), Mammen et al. (2009), Fengler et al. (2015) among many others. Estimation procedures closely related to backfitting were for example proposed in Linton et al. (2001), Linton and Mammen (2008) and Connor et al. (2012). In what follows, we generalize the backfitting approach of Lee et al. (2015) to the extended model (2.1) with a general calendar effect f_3 that satisfies (IC3). To keep the exposition simple, we assume throughout the section that the constant κ^* in (IC3) is known. In Section 4, we discuss a cross-validation approach to choose κ^*.

We estimate the density $f(x, y) = f_1(x) f_2(y) f_3(x + y)$ in a region $\mathcal{S} \subseteq \mathcal{I}$ where we have sufficiently many data points. When $\mathcal{I}$ is the triangle support $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x+y \leq 1\}$ or the trapezium support $\mathcal{I} = \{(x, y) \in [0, 1]^2 : c \leq x+y \leq 1\}$, the data tend to be sparse around the two corner points $(1, 0)$ and $(0, 1)$. For this reason, we exclude small neighbourhoods around these two points from the estimation. More formally, we estimate f on the subset

$$\mathcal{S} = \big\{(x, y) \in \mathcal{I} : x \leq 1 - \delta \text{ and } y \leq 1 - \delta\big\}$$

for some small $\delta > 0$.[1] For convenience, we impose slightly different identification

[1]The parameter δ is mainly of theoretical importance. In practice, we may set δ to any small positive number. As the precise choice of δ has little effect on the estimation procedure in practice, we suggest to simply set δ equal to 0 when implementing our approach (even though this is not fully correct in terms of the theory).

conditions on f_1 and f_2. In particular, we replace (IC1) and (IC2) by the constraints

$$\int_{\mathcal{S}_1} f_1(x)dx = 1 \quad \text{and} \quad \int_{\mathcal{S}_2} f_2(y)dy = 1, \tag{3.1}$$

where

$$\mathcal{S}_1 = \big\{x \in [0,1] : (x,y) \in \mathcal{S} \text{ for some } y \in [0,1]\big\}$$
$$\mathcal{S}_2 = \big\{y \in [0,1] : (x,y) \in \mathcal{S} \text{ for some } x \in [0,1]\big\}.$$

Note that $\mathcal{S}_1 = \mathcal{S}_2 = [0, 1-\delta]$ both in the triangle and the trapezium support case. For later reference, we additionally define $\mathcal{S}_3 = \{z \in [0,1] : z = x + y \text{ for some } (x,y) \in \mathcal{S}\}$. Throughout the section, we assume that the functions f_1, f_2 and f_3 are normalized to satisfy (3.1) and (IC3).

The density components f_1, f_2 and f_3 fulfill the integral equations

$$f_1(x) = \frac{f_{w,1}(x)}{\int_{\mathcal{J}_2(x)} f_2(y) f_3(x+y)dy} \tag{3.2}$$

$$f_2(y) = \frac{f_{w,2}(y)}{\int_{\mathcal{J}_1(y)} f_1(x) f_3(x+y)dx} \tag{3.3}$$

$$f_3(z) = \frac{1_{[0,1-\kappa^*)}(z) f_{w,3}(z)}{\int_{\mathcal{J}_3(z)} f_1(x) f_2(z-x)dx} + \frac{1_{[1-\kappa^*,1]}(z) \int_{1-\kappa^*}^1 f_{w,3}(v)dv}{\int_{1-\kappa^*}^1 \int_{\mathcal{J}_3(v)} f_1(x) f_2(v-x)dxdv}, \tag{3.4}$$

where $1_A(x)$ is the indicator function defined by $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise and we use the notation

$$f_{w,1}(x) = \int_{\mathcal{J}_2(x)} f(x,y)dy$$
$$f_{w,2}(y) = \int_{\mathcal{J}_1(y)} f(x,y)dx$$
$$f_{w,3}(z) = \int_{\mathcal{J}_3(z)} f(x,z-x)dx$$

together with

$$\mathcal{J}_2(x) = \{y \in [0,1] : (x,y) \in \mathcal{S}\}$$
$$\mathcal{J}_1(y) = \{x \in [0,1] : (x,y) \in \mathcal{S}\}$$
$$\mathcal{J}_3(z) = \{x \in [0,1] : (x,z-x) \in \mathcal{S}\}.$$

To estimate the density components f_1, f_2 and f_3, we construct empirical versions of the integral equations (3.2)–(3.4). Our estimators of f_1, f_2 and f_3 are defined as the solution to these empirical integral equations.

To define empirical versions of (3.2)–(3.4), we let $\hat{f}$ be a two-dimensional estimator of f (e.g. the local linear estimator of Nielsen (1999) which is introduced below) and let

$$\hat{f}_{w,1}(x) = \int_{\mathcal{J}_2(x)} \hat{f}(x,y)dy$$
$$\hat{f}_{w,2}(y) = \int_{\mathcal{J}_1(y)} \hat{f}(x,y)dx$$
$$\hat{f}_{w,3}(z) = \int_{\mathcal{J}_3(z)} \hat{f}(x,z-x)dx$$

be estimators of $f_{w,1}$, $f_{w,2}$ and $f_{w,3}$, respectively. With this notation at hand, we define estimators $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ of the functions f_1, f_2, f_3 as the solutions to the empirical integral equations

$$\hat{f}_1(x) = \hat{\phi}_1 \frac{\hat{f}_{w,1}(x)}{\int_{\mathcal{J}_2(x)} \hat{f}_2(y)\hat{f}_3(x+y)\ dy} \tag{3.5}$$

$$\hat{f}_2(y) = \hat{\phi}_2 \frac{\hat{f}_{w,2}(y)}{\int_{\mathcal{J}_1(y)} \hat{f}_1(x)\hat{f}_3(x+y)\ dx} \tag{3.6}$$

$$\hat{f}_3(z) = \hat{\phi}_3 \frac{1_{[0,1-\kappa^*)}(z)\hat{f}_{w,3}(z)}{\int_{\mathcal{J}_3(z)} \hat{f}_1(x)\hat{f}_2(z-x)dx} + \hat{\phi}_3 \frac{1_{[1-\kappa^*,1]}(z)\int_{1-\kappa^*}^1 \hat{f}_{w,3}(v)dv}{\int_{1-\kappa^*}^1 \int_{\mathcal{J}_3(v)} \hat{f}_1(x)\hat{f}_2(v-x)dxdv} \tag{3.7}$$

under the constraints

$$\int_{\mathcal{S}} \hat{f}_1(x)dx = 1, \quad \int_{\mathcal{S}} \hat{f}_2(y)dy = 1 \quad \text{and} \quad \int_{\mathcal{S}} \hat{f}_1(x)\hat{f}_2(y)\hat{f}_3(x+y)dxdy = \hat{\vartheta}, \tag{3.8}$$

where $\hat{\vartheta} = n^{-1}\sum_{i=1}^n 1((X_i,Y_i) \in \mathcal{S})$ is an estimator of $\vartheta = \int_{\mathcal{S}} f(x,y)dxdy$. The coefficients $\hat{\phi}_j$ $(j = 1,2,3)$ in (3.5)–(3.7) are chosen such that the constraints in (3.8) are satisfied.

The solutions $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ to the integral equations (3.5)–(3.7) cannot be computed explicitly in general. They can however be approximated by the following backfitting algorithm:

Step 0. Let $\hat{f}_1^{[0]}$, $\hat{f}_2^{[0]}$ be starting values for estimating f_1, f_2 which satisfy the first two constraints in (3.8). Calculate

$$\tilde{f}_3^{[0]}(z) = \begin{cases} \dfrac{\hat{f}_{w,3}(z)}{\int_{\mathcal{J}_3(z)} \hat{f}_1^{[0]}(x)\hat{f}_2^{[0]}(z-x)dx} & \text{for } z \in [0, 1-\kappa^*) \\ \dfrac{\int_{1-\kappa^*}^1 \hat{f}_{w,3}(v)dv}{\int_{1-\kappa^*}^1 \int_{\mathcal{J}_3(v)} \hat{f}_1^{[0]}(x)\hat{f}_2^{[0]}(v-x)dxdv} & \text{for } z \in [1-\kappa^*, 1] \end{cases}$$

and set $\hat{f}_3^{[0]}(z) = \hat{\phi}_3^{[0]}\tilde{f}_3^{[0]}(z)$, where $\hat{\phi}_3^{[0]}$ is chosen such that the third constraint in (3.8) is satisfied.

Step r. Let $\hat{f}_1^{[r-1]}$, $\hat{f}_2^{[r-1]}$ and $\hat{f}_3^{[r-1]}$ be the backfitting estimates from the previous iteration step. Compute updates as follows:

(a) Calculate

$$\tilde{f}_1^{[r]}(x) = \frac{\hat{f}_{w,1}(x)}{\int_{\mathcal{J}_2(x)} \hat{f}_2^{[r-1]}(y)\hat{f}_3^{[r-1]}(x+y)dy}$$

and set $\hat{f}_1^{[r]}(x) = \hat{\phi}_1^{[r]}\tilde{f}_1^{[r]}(x)$, where $\hat{\phi}_1^{[r]}$ is chosen such that the first constraint of (3.8) is fulfilled.

(b) Calculate

$$\tilde{f}_2^{[r]}(y) = \frac{\hat{f}_{w,2}(y)}{\int_{\mathcal{J}_1(y)} \hat{f}_1^{[r]}(x)\hat{f}_3^{[r-1]}(x+y)dx}$$

and set $\hat{f}_2^{[r]}(y) = \hat{\phi}_2^{[r]}\tilde{f}_2^{[r]}(y)$, where $\hat{\phi}_2^{[r]}$ is chosen such that the second constraint of (3.8) is satisfied.

(c) Compute $\hat{f}_3^{[r]}$ analogous to $\hat{f}_3^{[0]}$ in Step 0.

Iterate this procedure until some convergence criterion is satisfied.

To run the backfitting algorithm described above, we require an estimator $\hat{f}$ of the two-dimensional density f. Since the standard two-dimensional kernel density estimator is in general not consistent at the boundary of the support $\mathcal{S}$, we work with a local linear estimator which does not suffer from boundary problems. In particular, we let $\hat{f}$ be the local linear estimator of Nielsen (1999) which is defined as follows: Let

$$\tilde{f}_{b_1,b_2}(x,y) = \frac{1}{nb_1b_2}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{b_1}\Big)K\Big(\frac{Y_i - y}{b_2}\Big)W_i$$

be a standard kernel density estimator of f, where $W_i = 1((X_i, Y_i) \in \mathcal{S})$, K is a kernel function and (b_1, b_2) is the bandwidth vector. Throughout the paper, we assume that the kernel K is a symmetric probability density which is Lipschitz continuous and has bounded support. Moreover, let $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2)$ be the solution to the minimization problem

$$\hat{\boldsymbol{\eta}}(x,y) = \arg\min_{\boldsymbol{\eta}=(\eta_0,\eta_1,\eta_2)} \lim_{b_1,b_2\to 0} \int_{\mathcal{S}} \big[\tilde{f}_{b_1,b_2}(v,w) - \boldsymbol{a}(v,w;x,y)^\top \boldsymbol{\eta}(x,y)\big]^2 \times K\Big(\frac{v-x}{h_1}\Big)K\Big(\frac{w-y}{h_2}\Big)dvdw, \tag{3.9}$$

where $\boldsymbol{a}(v,w;x,y) = (1, (v-x)/h_1, (w-y)/h_2)^\top$. It can be shown that

$$\hat{\boldsymbol{\eta}}(x,y) = \boldsymbol{A}(x,y)^{-1}\boldsymbol{b}(x,y), \tag{3.10}$$

where

$$\boldsymbol{A}(x,y) = \int_{\mathcal{S}} \boldsymbol{a}(v,w;x,y)\boldsymbol{a}(v,w;x,y)^\top h_1^{-1}h_2^{-1}K\Big(\frac{v-x}{h_1}\Big)K\Big(\frac{w-y}{h_2}\Big)dvdw \quad (3.11)$$

$$\boldsymbol{b}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}(X_i,Y_i;x,y)h_1^{-1}h_2^{-1}K\Big(\frac{X_i-x}{h_1}\Big)K\Big(\frac{Y_i-y}{h_2}\Big)W_i. \quad (3.12)$$

We provide some details on the derivation of formula (3.10) in the Appendix. The local linear estimator $\hat{f}$ is defined as the first component of the vector $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2)$, that is, $\hat{f} = \hat{\eta}_0$.

4 Forecasting

4.1 The forecasting method

The backfitting algorithm described in Section 3 yields an estimate of the calendar effect $f_3(z)$ up to the present time point $z = 1$. To obtain an estimate of the two-dimensional density f on the whole unit square $[0,1]^2$, we need to extrapolate the estimate of $f_3(z)$ into the future, that is, to time points $z \in (1,2]$. Our extrapolation strategy is closely related to the identification constraints (IC1)–(IC3) imposed on the functions f_1, f_2 and f_3. These constraints normalize the calendar effect in such a way that it makes sense to simply extrapolate it constantly into the (near) future. The heuristic reason for this is as follows: As already discussed in Section 2, if the calendar effect f_3 is smooth, it is approximately linear around the present time point $z = 1$ on a logarithmic scale, that is,

$$\log f_3(z) = a + bz + \psi(z) \quad \text{for all } z \in [1-\kappa^*, 1+\lambda],$$

where κ^* and λ are small positive constants and the function ψ is approximately equal to zero. For our heuristic discussion, we neglect the function ψ and assume that $\log f_3(z)$ is linear for $z \in [1-\kappa^*, 1+\lambda]$, that is,

$$\log f_3(z) = a + bz \quad \text{for all } z \in [1-\kappa^*, 1+\lambda].$$

Transforming back to the multiplicative model, this means that

$$f_3(z) = e^{a+bz} \quad \text{for all } z \in [1-\kappa^*, 1+\lambda]. \quad (4.1)$$

As discussed in Section 2, our identification strategy normalizes the functions f_1, f_2 and f_3 such that the exponential e^{a+bz} is eliminated from f_3 and shifted to the other components. After this normalization, f_3 is constant on the interval $[1-\kappa^*, 1+\lambda]$. In particular, by imposing the constraints (IC1)–(IC3), we obtain that

$$f_3(z) = c_3 \quad \text{for all } z \in [1-\kappa^*, 1+\lambda]$$

and some constant $c_3 \in \mathbb{R}$. Under this normalization, it is most natural to extrapolate the calendar effect constantly into the future, that is, we set $\hat{f}^{\text{fc}}(z) = \hat{c}_3$ for $z \in (1, 1+\lambda]$, where $\hat{c}_3$ is an estimate of c_3. More formally, our extrapolation strategy is as follows:

Step 1. Estimate f_1, f_2, f_3 by the backfitting algorithm from Section 3, where κ^* is assumed to be known. We discuss a cross-validation procedure to choose κ^* below. Denote the resulting estimates by $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$.

Step 2. Extrapolate $\hat{f}_3$ constantly into the future, that is, set $\hat{f}_3^{\text{fc}}(z) = \hat{f}_3(1)$ for $z > 1$ and forecast $f(x, y)$ at points $x + y > 1$ by

$$\hat{f}^{\text{fc}}(x, y) = \hat{f}_1(x)\hat{f}_2(y)\hat{f}_3^{\text{fc}}(x + y).$$

The constant κ^* in the constraint (IC3) is usually not known in practice. We propose the following cross-validation procedure for the choice of κ^*: Pick some small $\lambda > 0$ (e.g. corresponding to the forecast horizon $(1, 1+\lambda]$) and define

$$\mathcal{S}_\lambda^< = \{(x, y) \in \mathcal{S} : x + y \le 1 - \lambda\}$$
$$\mathcal{S}_\lambda^> = \{(x, y) \in \mathcal{S} : x + y > 1 - \lambda,\, x \le 1 - \lambda,\, y \le 1 - \lambda\}.$$

Let $\mathcal{D}$ be the set of data points (X_i, Y_i) that lie in $\mathcal{S}_\lambda^<$, that is, $(X_i, Y_i) \in \mathcal{S}_\lambda^<$. For any $\kappa \in (\lambda, 1]$, compute the backfitting estimators $\hat{f}_1^\kappa$, $\hat{f}_2^\kappa$, $\hat{f}_3^\kappa$ from the data sample $\mathcal{D}$ as described in Section 3, where the set $\mathcal{S}$ is replaced by $\mathcal{S}_\lambda^<$. Next define

$$\hat{f}_3^\kappa(z) = \hat{f}_3^\kappa(1 - \lambda) \quad \text{for } z \in (1 - \lambda, 1],$$

thus constantly extrapolating the estimated calendar effect into the region $(1-\lambda, 1]$. Set

$$\hat{f}^\kappa(x, y) = \hat{f}_1^\kappa(x)\hat{f}_2^\kappa(y)\hat{f}_3^\kappa(x + y) \quad \text{for } (x, y) \in \mathcal{S}_\lambda^>$$

and consider the MISE criterion

$$\text{MISE}(\kappa) = \int_{\mathcal{S}_\lambda^>} \left\{\hat{f}^\kappa(x, y) - f(x, y)\right\}^2 dxdy.$$

Minimizing MISE(κ) with respect to κ is equivalent to minimizing

$$\int_{\mathcal{S}_\lambda^>} \{\hat{f}^\kappa(x, y)\}^2 dxdy - 2\int_{\mathcal{S}_\lambda^>} \hat{f}^\kappa(x, y) f(x, y) dxdy,$$

which can be estimated by

$$\text{CV}(\kappa) = \int_{\mathcal{S}_\lambda^>} \{\hat{f}^\kappa(x, y)\}^2 dxdy - \frac{2}{n}\sum_{i=1}^n 1((X_i, Y_i) \in \mathcal{S}_\lambda^>)\hat{f}^\kappa(X_i, Y_i).$$

We finally define our estimator of κ by

$$\hat{\kappa} = \arg\min_{\kappa \in (\lambda,1]} \mathrm{CV}(\kappa). \tag{4.2}$$

This cross-validation procedure to choose κ is similar in spirit to methods developed in Pesaran and Timmermann (2007) in a completely different context. The econometric problem considered there is to select the optimal estimation window for prediction in a linear regression model with a structural break. To approach this problem, Pesaran and Timmermann (2007) propose a cross-validation procedure similar to ours: Pseudo out-of-sample forecasts based on different estimation windows are evaluated by a cross-validation criterion. The estimation window is then selected by minimizing this criterion.

4.2 Comparison with other forecasting approaches

As already mentioned in the introduction, the extended continuous chain ladder model (2.1) can be regarded as a continuous version of the discrete age-period-cohort model of Kuang et al. (2011). Nevertheless, our forecasting strategy is very different from standard forecasting procedures employed in the context of age-period-cohort models. To highlight the main differences, we compare our forecasting strategy to the procedures developed in Kuang et al. (2011).

In Kuang et al. (2011), three different forecasters are defined which are named as $I(0)$ (zero-times), $I(1)$ (one-time) and $I(2)$ (two-times) integrators. An overview over the discrete age-period-cohort model of Kuang et al. (2011) and a precise definition of the $I(0)$, $I(1)$ and $I(2)$ forecasters is provided in the supplementary material. All three forecasters extrapolate the (logarithmic) calendar effect linearly into the future, where the slope for extrapolation is estimated in different ways. In particular, the three methods use the estimated calendar effect up to today in different ways to determine the slope for extrapolation. The $I(2)$ method only uses the most recent past of the estimated calendar effect to determine the slope, while the $I(0)$ and $I(1)$ approaches use the complete past. Whereas the forecasting methods of Kuang et al. (2011) focus on estimating the slope of the calendar effect in a suitable way, we essentially eliminate the calendar effect from the model by our identification strategy, that is, by normalizing it to have zero slope in the recent past from the time point $1-\kappa$ onwards. This allows us to employ the simplest possible forecasting strategy: constant extrapolation.

In the supplementary material, we show that the $I(0)$, $I(1)$ and $I(2)$ forecasters of Kuang et al. (2011) can be re-interpreted in terms of our forecasting strategy. In particular, we show that the forecasters can be re-produced by imposing specific normalization constraints on the parameters of the discrete age-period-cohort model

Method	Normalization constraints on f_3
$I(0)$	$\int_0^1 \log(f_3(z))dz = 0$ and $\int_0^1 z \log(f_3(z))dz = 0$
$I(1)$	$f_3(0) = f_3(1)$
$I(2)$	$f_3(1-\eta) = f_3(1)$ with some small $\eta > 0$

Table 1: Normalization constraints on f_3 for the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasters. Note that unlike κ, η is not a tuning parameter which is selected in a data-driven way. It is rather chosen adhoc and set to a very small positive value.

and by extrapolating the estimated calendar effect constantly into the future under these constraints. Continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasters can be obtained by imposing analogous normalization constraints on the densities in our continuous model $f(x,y) = f_1(x)f_2(y)f_3(x+y)$. Specifically, to obtain continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasts, we re-normalize the calendar effect f_3 according to the constraints listed in Table 1. The main difference between our approach and the $I(0)$, $I(1)$ and $I(2)$ forecasts thus lies in the way the calendar effect f_3 is normalized before it is extrapolated constantly.

The normalization constraints imposed on f_3 by our approach and by the $I(0)$, $I(1)$ and $I(2)$ forecasts are related as follows to each other:

(i) The constraints of the $I(0)$ and $I(1)$ methods can be interpreted as eliminating the linear part of the logarithmic calendar effect $\log f_3$ in the entire past, that is, on the interval $[0,1]$. To see this, suppose that $\log f_3$ is a linear function on $[0,1]$ of the form $\log f_3(z) = a + bz$. The constraint $f_3(0) = f_3(1)$ of the $I(1)$ method obviously implies that $b = 0$, that is, it normalizes $\log f_3(z)$ to be a constant function. Similarly, the two constraints $\int_0^1 \log(f_3(z))dz = 0$ and $\int_0^1 z \log(f_3(z))dz = 0$ of the $I(0)$ method imply that $a = b = 0$.

(ii) Analogously, the constraint $f_3(1-\eta) = f_3(1)$ of the $I(2)$ method can be interpreted as eliminating the linear part of $\log f_3$ in the most recent past, in particular, on the interval $[1-\eta, 1]$. Note that unlike κ, the value η is not a parameter that is chosen in a data-driven way. It is rather a fixed constant which is set adhoc to a very small positive value.

(iii) Our approach normalizes the logarithmic calendar effect $\log f_3$ such that the linear part on the interval $[1-\kappa, 1]$ gets eliminated. Importantly, we do not pick the parameter κ adhoc. We rather select it in a data-driven way (by our cross-validation procedure) to estimate the largest interval $[1-\kappa^*, 1]$ where f_3 is (approximately) linear. We then use the estimate of the interval $[1-\kappa^*, 1]$ to normalize f_3 appropriately.

According to the remarks (i)–(iii), the $I(0)$, $I(1)$ and $I(2)$ forecasters can be regarded as extreme cases of our approach: The $I(0)$ and $I(1)$ forecasters eliminate the linear part of $\log f_3$ in the entire past $[0,1]$. However, this only makes sense if $\log f_3$ is indeed approximately linear on $[0,1]$. The $I(2)$ method, in contrast, takes into account only the most recent past $[1-\eta,1]$ to eliminate the linear part from $\log f_3$. If $\log f_3$ is approximately linear on a much larger interval than $[1-\eta,1]$, this is suboptimal as the shape constraint of $\log f_3$ (its linearity) is only exploited on the small subinterval $[1-\eta,1]$. In contrast to the $I(0)$, $I(1)$ and $I(2)$ forecasters, our method adapts to the unknown form of the calendar effect f_3 by approximating the largest interval $[1-\kappa^*,1]$ where $\log f_3$ is (approximately) linear. For this reason, we expect our method to produce better forecasts than the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ methods. We demonstrate this in our simulation study in Section 7.

5 Theoretical results

In this section, we derive some theoretical properties of the estimators $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ introduced in Section 3. We use the following notation: $L^2(\mathcal{S}_j)$ is the space of square integrable functions $q:\mathcal{S}_j\to\mathbb{R}$ for $j=1,2$ and $L^2_{\kappa^*}(\mathcal{S}_3)$ denotes the space of square integrable functions $q:\mathcal{S}_3\to\mathbb{R}$ which are constant on the interval $[1-\kappa^*,1]$. We let $\mathcal{L}=L^2(\mathcal{S}_1)\times L^2(\mathcal{S}_2)\times L^2_{\kappa^*}(\mathcal{S}_3)$ and write $\boldsymbol{g}=(g_1,g_2,g_3)^\top\in\mathcal{L}$ along with $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)^\top\in\mathbb{R}^3$. With this notation at hand, we define the operator

$$\mathcal{F}:\mathbb{R}^3\times\mathcal{L}\to\mathbb{R}^3\times\mathcal{L}$$

by setting $\mathcal{F}(\boldsymbol{\theta},\boldsymbol{g})(x,y,z)=\big(\mathcal{F}_1(\boldsymbol{\theta},\boldsymbol{g}),\,\mathcal{F}_2(\boldsymbol{\theta},\boldsymbol{g}),\,\mathcal{F}_3(\boldsymbol{\theta},\boldsymbol{g}),\,\mathcal{F}_4(\boldsymbol{\theta},\boldsymbol{g})(x),\,\mathcal{F}_5(\boldsymbol{\theta},\boldsymbol{g})(y),\,\mathcal{F}_6(\boldsymbol{\theta},\boldsymbol{g})(z)\big)^\top$, where

$$\begin{aligned}
\mathcal{F}_1(\boldsymbol{\theta},\boldsymbol{g}) &= 1-\int_{\mathcal{S}_1} g_1(x)dx\\
\mathcal{F}_2(\boldsymbol{\theta},\boldsymbol{g}) &= 1-\int_{\mathcal{S}_2} g_2(y)dy\\
\mathcal{F}_3(\boldsymbol{\theta},\boldsymbol{g}) &= \vartheta-\int_{\mathcal{S}} g_1(x)g_2(y)g_3(x+y)dxdy
\end{aligned}$$

with $\vartheta=\int_{\mathcal{S}} f(x,y)dxdy$ and

$$\begin{aligned}
\mathcal{F}_4(\boldsymbol{\theta},\boldsymbol{g})(x) &= \int_{\mathcal{J}_2(x)}\big\{\theta_1 f(x,y)-g_1(x)g_2(y)g_3(x+y)\big\}dy\\
\mathcal{F}_5(\boldsymbol{\theta},\boldsymbol{g})(y) &= \int_{\mathcal{J}_1(y)}\big\{\theta_2 f(x,y)-g_1(x)g_2(y)g_3(x+y)\big\}dx
\end{aligned}$$

$$\mathcal{F}_6(\boldsymbol{\theta},\boldsymbol{g})(z) = 1_{[0,1-\kappa^*)}(z)\int_{\mathcal{J}_3(z)}\big\{\theta_3 f(x,z-x) - g_1(x)g_2(z-x)g_3(z)\big\}dx$$
$$+\, 1_{[1-\kappa^*,1]}(z)\,\frac{1}{\kappa^*}\int_{1-\kappa^*}^{1}\int_{\mathcal{J}_3(v)}\big\{\theta_3 f(x,v-x) - g_1(x)g_2(v-x)g_3(v)\big\}dxdv.$$

The true density components $\boldsymbol{f} = (f_1, f_2, f_3)^\top$ are characterized by the equation

$$\mathcal{F}(\boldsymbol{\phi}, \boldsymbol{f}) = \mathbf{0}, \tag{5.1}$$

where $\boldsymbol{\phi} = (1,1,1)^\top$. This is equivalent to saying that the component functions f_1, f_2 and f_3 satisfy the integral equations (3.2)–(3.4).

The estimator $\hat{\boldsymbol{f}} = (\hat{f}_1, \hat{f}_2, \hat{f}_3)^\top$ of the density components $\boldsymbol{f} = (f_1, f_2, f_3)^\top$ defined in Section 3 can be characterized as the solution to an empirical version of (5.1): Let $\hat{\mathcal{F}}_j$ be operators which are defined analogously as $\mathcal{F}_j$ for $1 \le j \le 6$ with the density f and the parameter ϑ replaced by the estimators $\hat{f}$ and $\hat{\vartheta}$ from Section 3. The estimators $\hat{\boldsymbol{f}} = (\hat{f}_1, \hat{f}_2, \hat{f}_3)^\top$ and $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3)^\top$ are given as the solution to the equation

$$\hat{\mathcal{F}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{f}}) = \mathbf{0}, \tag{5.2}$$

where $\hat{\mathcal{F}}(\boldsymbol{\theta},\boldsymbol{g})(x,y,z) = (\hat{\mathcal{F}}_1(\boldsymbol{\theta},\boldsymbol{g}),\ \hat{\mathcal{F}}_2(\boldsymbol{\theta},\boldsymbol{g}),\ \hat{\mathcal{F}}_3(\boldsymbol{\theta},\boldsymbol{g}),\ \hat{\mathcal{F}}_4(\boldsymbol{\theta},\boldsymbol{g})(x),\ \hat{\mathcal{F}}_5(\boldsymbol{\theta},\boldsymbol{g})(y),$ $\hat{\mathcal{F}}_6(\boldsymbol{\theta},\boldsymbol{g})(z))^\top$.

We now construct and examine a theoretical approximation of the estimator $\hat{\boldsymbol{f}}$. To do so, we define the operator

$$\mathcal{G}(\boldsymbol{\theta},\boldsymbol{g}) = \mathcal{F}(\mathbf{1}+\boldsymbol{\theta}, \boldsymbol{f}\circ(\mathbf{1}+\boldsymbol{g})),$$

where $\boldsymbol{f}\circ(\mathbf{1}+\boldsymbol{g})$ with $\mathbf{1} = (1,1,1)^\top$ denotes the componentwise multiplication of the two function vectors $\boldsymbol{f}$ and $(\mathbf{1}+\boldsymbol{g})$. By construction, $\mathcal{G}(\mathbf{0},\mathbf{0}) = \mathbf{0}$. The Fréchet derivative $\mathcal{G}'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})$ of $\mathcal{G}(\boldsymbol{\theta},\boldsymbol{g})$ at $(\boldsymbol{\theta},\boldsymbol{g}) = (\mathbf{0},\mathbf{0})$ in the direction $(\boldsymbol{d},\boldsymbol{\delta})$ is given by

$$\mathcal{G}'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(x,y,z) = \big(\mathcal{G}_1'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}), \mathcal{G}_2'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}), \mathcal{G}_3'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}),$$
$$\mathcal{G}_4'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(x), \mathcal{G}_5'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(y), \mathcal{G}_6'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(z)\big)^\top,$$

where

$$\mathcal{G}_1'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}) = -\int_{\mathcal{S}_1} f_1(x)\delta_1(x)dx$$
$$\mathcal{G}_2'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}) = -\int_{\mathcal{S}_2} f_2(y)\delta_2(y)dy$$
$$\mathcal{G}_3'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}) = -\int_{\mathcal{S}} f(x,y)\delta_+(x,y)dxdy$$
$$\mathcal{G}_4'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(x) = \int_{\mathcal{J}_2(x)} f(x,y)\big\{d_1 - \delta_+(x,y)\big\}dy$$

$$\mathcal{G}_5'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(y) = \int_{\mathcal{J}_1(y)} f(x,y)\big\{d_2 - \delta_+(x,y)\big\}dx$$

$$\begin{aligned}\mathcal{G}_6'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta})(z) &= 1_{[0,1-\kappa^*)}(z)\int_{\mathcal{J}_3(z)} f(x,z-x)\big\{d_3 - \delta_+(x,z-x)\big\}dx \\ &\quad + 1_{[1-\kappa^*,1]}(z)\,\frac{1}{\kappa^*}\int_{1-\kappa^*}^1\int_{\mathcal{J}_3(v)} f(x,v-x)\big\{d_3 - \delta_+(x,v-x)\big\}dxdv\end{aligned}$$

with $\delta_+(x,y) = \delta_1(x) + \delta_2(y) + \delta_3(x+y)$. To make the notation more concise, we write $\mathcal{G}'(\mathbf{0},\mathbf{0})(\boldsymbol{d},\boldsymbol{\delta}) = \mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})$ in what follows. Analogous to the definition of $\mathcal{G}$, we let $\hat{\mathcal{G}}(\boldsymbol{\theta},\boldsymbol{g}) = \hat{\mathcal{F}}(\mathbf{1}+\boldsymbol{\theta}, \boldsymbol{f}\circ(\mathbf{1}+\boldsymbol{g}))$ and note that

$$\hat{\mathcal{G}}(\mathbf{0},\mathbf{0}) = \hat{\mathcal{F}}(\boldsymbol{\phi},\boldsymbol{f}) = \begin{pmatrix} 0 \\ 0 \\ \hat{\vartheta} - \vartheta \\ \boldsymbol{f}_w \circ \hat{\boldsymbol{\mu}} \end{pmatrix},$$

where

$$\begin{aligned}\boldsymbol{f}_w(x,y,z) &= \big(f_{w,1}(x), f_{w,2}(y), 1_{[0,1-\kappa^*)}(z)\, f_{w,3}(z) + 1_{[1-\kappa^*,1]}(z)\, c_{w,3}\big)^\top \\ \hat{\boldsymbol{\mu}}(x,y,z) &= \big(\hat{\mu}_1(x), \hat{\mu}_2(y), \hat{\mu}_3(z)\big)^\top\end{aligned}$$

with $c_{w,3} = (\kappa^*)^{-1}\int_{1-\kappa^*}^1 f_{w,3}(v)dv$ and

$$\hat{\mu}_1(x) = f_{w,1}^{-1}(x)\int_{\mathcal{J}_2(x)}\big[\hat{f}(x,y) - f(x,y)\big]dy \tag{5.3}$$

$$\hat{\mu}_2(y) = f_{w,2}^{-1}(y)\int_{\mathcal{J}_1(y)}\big[\hat{f}(x,y) - f(x,y)\big]dx \tag{5.4}$$

$$\begin{aligned}\hat{\mu}_3(z) &= 1_{[0,1-\kappa^*)}(z)\, f_{w,3}^{-1}(z)\int_{\mathcal{J}_3(z)}\big[\hat{f}(x,z-x) - f(x,z-x)\big]dx \\ &\quad + 1_{[1-\kappa^*,1]}(z)\, c_{w,3}^{-1}\,\frac{1}{\kappa^*}\int_{1-\kappa^*}^1\int_{\mathcal{J}_3(v)}\big[\hat{f}(x,v-x) - f(x,v-x)\big]dxdv.\end{aligned} \tag{5.5}$$

Letting $\mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}$ be the inverse of $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ (which exists by Lemma 1 in the Appendix), we define $\bar{\boldsymbol{f}} = (\bar{f}_1, \bar{f}_2, \bar{f}_3)^\top$ and $\bar{\boldsymbol{\phi}} = (\bar{\phi}_1, \bar{\phi}_2, \bar{\phi}_3)^\top$ by the equation

$$\begin{pmatrix} \bar{\boldsymbol{\phi}} - \boldsymbol{\phi} \\ (\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f} \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}\begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ \hat{\boldsymbol{\mu}} \end{pmatrix}. \tag{5.6}$$

The quantities $\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}$ and $(\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f}$ can be interpreted as follows: Let $\hat{\mathcal{G}}'_{(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})$ be the derivative of $\hat{\mathcal{G}}(\boldsymbol{\theta},\boldsymbol{g})$ in the direction $(\boldsymbol{d},\boldsymbol{\delta})$ and let $\hat{\mathcal{G}}'^{-1}_{(\boldsymbol{\theta},\boldsymbol{g})}$ be the inverse of $\hat{\mathcal{G}}'_{(\boldsymbol{\theta},\boldsymbol{g})}$. By Newton's method, we can approximate the root of the operator $\hat{\mathcal{G}}$ by the iteration

$$(\boldsymbol{\theta},\boldsymbol{g})_{\ell+1} = (\boldsymbol{\theta},\boldsymbol{g})_\ell - \hat{\mathcal{G}}'^{-1}_{((\boldsymbol{\theta},\boldsymbol{g})_\ell)}\hat{\mathcal{G}}\big((\boldsymbol{\theta},\boldsymbol{g})_\ell\big)$$

for $\ell = 0, 1, 2 \ldots$ with some starting value $(\boldsymbol{\theta}, \boldsymbol{g})_0$. Setting $(\boldsymbol{\theta}, \boldsymbol{g})_0 = (\mathbf{0}, \mathbf{0})$ and performing one Newton step, we get that

$$(\boldsymbol{\theta}, \boldsymbol{g})_1 = -\hat{\mathcal{G}}'^{-1}_{(\mathbf{0},\mathbf{0})}\, \hat{\mathcal{G}}(\mathbf{0}, \mathbf{0}) \approx \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})} \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ \hat{\boldsymbol{\mu}} \end{pmatrix}.$$

Hence, we obtain $(\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, (\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f})$ by performing one approximate Newton step from $(\mathbf{0}, \mathbf{0})$ into the direction of the root $(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}, (\hat{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f})$ of $\hat{\mathcal{G}}$.

Our first theoretical result shows that $\bar{\boldsymbol{\phi}}$ and $\bar{\boldsymbol{f}}$ give a good approximation to the estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{f}}$.

Theorem 1. *Let the conditions of Proposition 1 be satisfied. Moreover, let $\hat{f}$ be any estimator with the property that $\hat{f}(x, y) = f(x, y) + O_p(\varepsilon_n)$ uniformly for $(x, y) \in \mathcal{S}$. Then with probability tending to 1, there exists a solution $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{f}})$ of the equation $\hat{\mathcal{F}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{f}}) = \mathbf{0}$ and it holds that*

$$|\hat{\phi}_j - \bar{\phi}_j| = O_p(\varepsilon_n^2 + n^{-1/2})$$
$$\sup_{w \in \mathcal{S}_j} |\hat{f}_j(w) - \bar{f}_j(w)| = O_p(\varepsilon_n^2 + n^{-1/2})$$

for $j = 1, 2, 3$.

Standard theory for kernel smoothing yields that $\varepsilon_n = n^{-3/10}\sqrt{\log n}$ for the local linear estimator $\hat{f}$ with $h_1 \sim h_2 \sim n^{-1/5}$. According to Theorem 1, $\hat{f}_j(w) - \bar{f}_j(w) = O_p(n^{-1/2})$ uniformly on $\mathcal{S}_j$ in this case. In addition to this, we can show that $\bar{f}_j(w) - f_j(w) = O_p(n^{-2/5}\sqrt{\log n})$ uniformly over $\mathcal{S}_j$. As a consequence, the first-order asymptotic properties of $\hat{f}_j$ are identical to those of $\bar{f}_j$. These asymptotic properties are summarized by the following two theorems. The first result specifies the uniform convergence rate of $\hat{f}_j$.

Theorem 2. *Let the conditions of Proposition 1 be satisfied and suppose that the density f is twice continuously differentiable on $\mathcal{S}$. Moreover, let the kernel K be supported on $[-1, 1]$, symmetric and Lipschitz continuous. Finally, let the bandwidths h_j for $j = 1, 2$ be such that $n^{1/5} h_j \to c_j$ for some constants $c_j > 0$. Then it holds that*

$$\sup_{w \in \mathcal{S}_j} |\hat{f}_j(w) - f_j(w)| = O_p(n^{-2/5}\sqrt{\log n})$$

for $j = 1, 2, 3$.

The next result specifies the asymptotic distribution of the estimators $\hat{f}_j$. To formulate it, we introduce some additional notation. As in Theorem 2, we suppose that $n^{1/5} h_j \to c_j > 0$ for $j = 1, 2$ and let

$$\tilde{f}^B(x, y) = \frac{1}{2} \int u^2 K(u) du \Big[c_1^2 \frac{\partial^2 f(x, y)}{\partial^2 x} + c_2^2 \frac{\partial^2 f(x, y)}{\partial^2 y} \Big]. \tag{5.7}$$

For $j = 1, 2, 3$, we define $\tilde{\mu}_j^B$ analogously as $\hat{\mu}_j$ in (5.3)–(5.5) with $\hat{f} - f$ replaced by $\tilde{f}^B$. Writing $\tilde{\boldsymbol{\mu}}^B = (\tilde{\mu}_1^B, \tilde{\mu}_2^B, \tilde{\mu}_3^B)$, we let $\boldsymbol{d} = (d_1, d_2, d_3) \in \mathbb{R}^3$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) \in \boldsymbol{\mathcal{L}}$ be the solution of the equation

$$\begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{\beta} \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})} \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ n^{-2/5} \tilde{\boldsymbol{\mu}}^B \end{pmatrix}. \tag{5.8}$$

Equivalently, $\boldsymbol{d}$ and $\boldsymbol{\beta}$ can be defined as the solution to the backfitting equation (A.26) in the Appendix. As we will see, β_j plays the role of the asymptotic bias of $\hat{f}_j$. We finally introduce the terms

$$\begin{aligned}
\sigma_1^2(x) &= c_1^{-1} f_{w,1}(x)^{-1} \int K^2(u) du \\
\sigma_2^2(y) &= c_2^{-1} f_{w,2}(y)^{-1} \int K^2(u) du \\
\sigma_3^2(z) &= c_2^{-1} f_{w,3}(z)^{-1} 1_{[0,1-\kappa^*)}(z) \int [K * K(u)][K * K(-c_1 u / c_2)] du
\end{aligned}$$

with $K * K(u) = \int K(\varphi) K(\varphi - u) d\varphi$, which turn out to be the asymptotic variances of $\hat{f}_j$ for $j = 1, 2, 3$. We are now in a position to specify the asymptotic distribution of the estimators $\hat{f}_j$.

Theorem 3. *Let the conditions of Theorem 2 be satisfied. Then for any $j = 1, 2, 3$ and any fixed point w in the interior of $\mathcal{S}_j$, it holds that*

$$n^{2/5} \frac{\hat{f}_j(w) - f_j(w)}{f_j(w)} \xrightarrow{d} N\big(\beta_j(w), \sigma_j^2(w)\big).$$

The proofs of Theorems 1–3 are given in the Appendix.

6 Case studies

In this section, we consider two applications of the density forecasting methods described above. The first application comes from the actuarial sciences and is on claims reserving forecasting in non-life insurance (cp. Example 1 from the introduction). Similar case studies were conducted by Mammen et al. (2015) and Hiabu et al. (2016) to illustrate a simpler multiplicative density model where f_3 is a constant function. The second application example comes from economics and deals with forecasting fertility rates in Italy and the US (cp. Example 2 from the introduction).

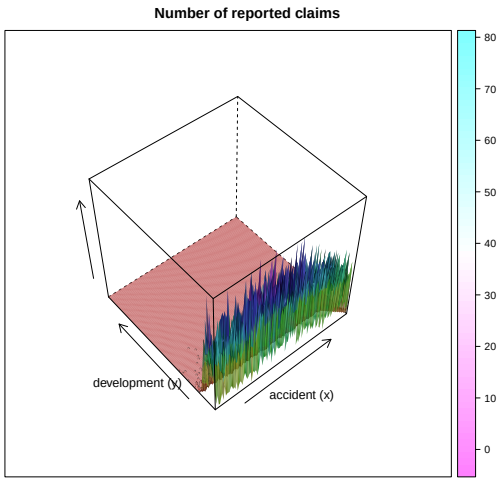

Figure 1: Histogram of the reserving data. The number of reported claims during 10 years is shown according to the accident time (x) and the reporting delay (y).

6.1 Claims reserving in non-life insurance

In this example, we observe data (X_i, Y_i) on n different insurance claims $i = 1, \ldots, n$, where X_i specifies the time point when claim i incurred and Y_i is the time delay until claim i was reported to the insurance. The data points (X_i, Y_i) take values in the triangle $\mathcal{I} = \{(x, y) : 0 \leq x, y \leq T,\, x + y \leq T\}$, where x is the accident time (i.e. the time when the claim incurs), y is the claims development time (i.e. the time delay until the claim is reported) and $[0, T]$ (with $T > 0$) is the time observation window.[2] Our aim is to estimate and forecast the density f of (X_i, Y_i). In particular, we are interested in forecasting the quantity $\mathcal{L} = \int_{(x,y)\in\mathcal{I}^c} f(x, y)dxdy$, which gives the outstanding number of insurance claims in the upper triangle $\mathcal{I}^c = [0, T]^2 \setminus \mathcal{I}$. This number represents the future liabilities for the company.

Traditionally, actuaries work with the data aggregated in so-called run-off triangles. A run-off triangle can be written as $\aleph_m = \{N_{st} : (s, t) \in \mathcal{I}_m\}$, where $\mathcal{I}_m = \{(s, t) : s = 1, \ldots, m;\, t = 1, \ldots, m;\, s+t-1 \leq m\}$ and N_{st} is the total number of claims incurred in period (week, month, quarter or year) s and reported in period $s + t - 1$, that is, with $t - 1$ periods delay. The quantities N_{st} are usually referred to as frequencies in the literature and can be computed from the individual claims data (X_i, Y_i) as $N_{st} = \sum_{i=1}^{n} 1_{st}(X_i, Y_i)$, where $1_{st}(X_i, Y_i)$ is the indicator function which equals 1 iff X_i is a time point in period s and Y_i a time point in period t. The quantities N_{st} are thus nothing else than the values of the histogram of the individual claims data computed with bin width equal to the considered period (which is usually a week, a month, a quarter or a year). Figure 1 shows the data set we

[2]When deriving the estimation methods and theory in the previous sections, we have used the normalization $T = 1$ for convenience.

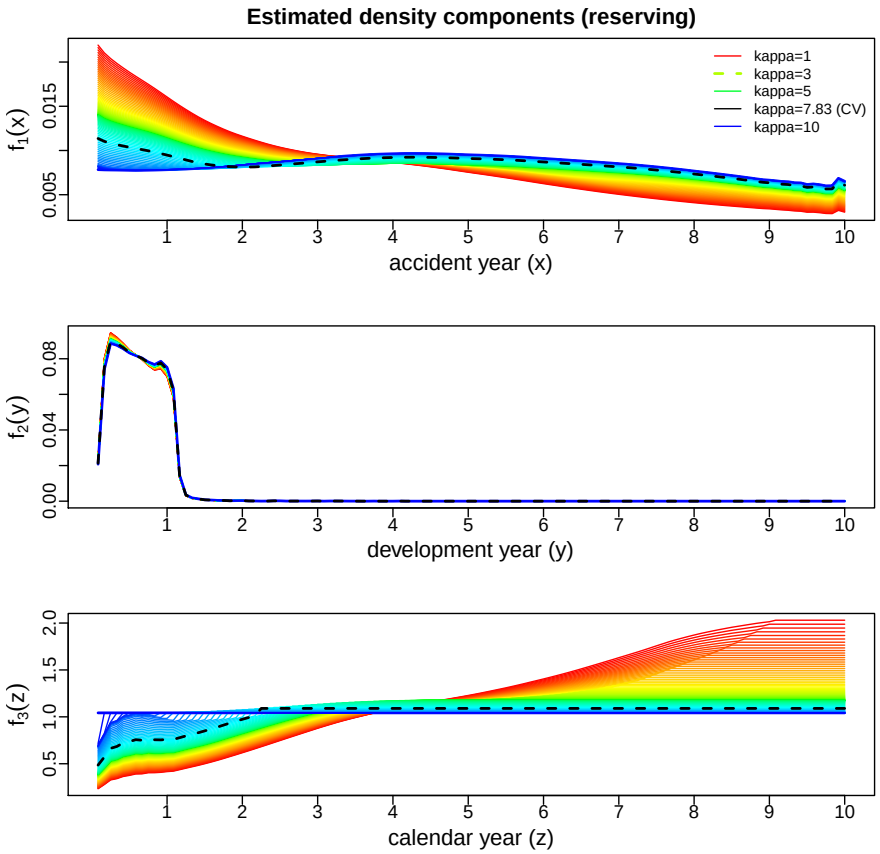

Figure 2: Estimated density components for the reserving data considering different κ values (given in years). The dashed black lines show the estimated density components for the cross-validated choice of κ, which is $\kappa = 7.83$ years.

analyze in this section. The data consist of monthly frequencies N_{st} of reported claims from a motor business in Cyprus. The sample size amounts to $n = 55384$ claims which were reported between 2004 and 2013. Figure 1 presents the histogram of the frequencies N_{st} (with bin width equal to one month).

We derive forecasts on the basis of the multiplicative density model $f(x, y) = f_1(x)f_2(y)f_3(x + y)$ from (2.1), where f_1 and f_2 are the density components corresponding to accident and reporting (development) time, respectively, and f_3 is a function which describes the calendar effect. In our theoretical framework, the estimators of the density components are directly computed from the sample of individual data (X_i, Y_i). In particular, we first estimate the two-dimensional density f by a local linear kernel estimator $\hat{f}$ from the data sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ and then estimate the density components f_1, f_2 and f_3 by running the backfitting procedure from Section 3 with the pilot estimator $\hat{f}$. In practice, an estimator of f can also be computed from the histogram values N_{st}. (It may even be necessary to do so as the individual claims data are not always available.) Estimating f from the

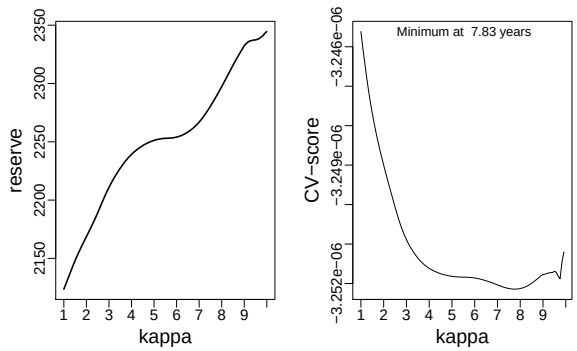

Figure 3: The effect of the parameter κ on the predicted reserves, i.e., on the predicted number of outstanding liabilities $\mathcal{L}$. The left-hand panel shows the predicted reserves $\mathcal{L}$ for different values of κ, the right-hand panel gives the cross-validation score minimized to choose κ.

values N_{st} by our kernel methods essentially amounts to smoothing the histogram of the individual claims data, where the amount of smoothing is determined by the chosen bandwidths. In the data example at hand, we compute the local linear kernel estimator of f from the values N_{st} with the bandwidths $\hat{h}_1$ and $\hat{h}_2$ chosen by cross-validation ($\hat{h}_1 = 1.77$ years, $\hat{h}_2 = 0.08$ years).

The estimated density components f_1, f_2 and f_3 produced by our backfitting algorithm for different values of the parameter κ (given in years) are shown in Figure 2. The biggest value, $\kappa = 10$ years, corresponds to a model with constant calendar effect. In the graphs, we have highlighted the results for the parameter value chosen by the proposed cross-validation method (with $\lambda = 1$ year), which is $\kappa = 7.83$ years. Figure 3 shows the effect of κ on the predicted number of outstanding liabilities $\mathcal{L} = \int_{(x,y)\in\mathcal{I}^c} f(x,y)dxdy$ as well as the cross-validation score minimized to choose κ. The predicted number of outstanding liabilities $\mathcal{L}$ for the cross-validated choice of κ is 2324 compared to the bigger number 2344 obtained by using a model with constant calendar effect.

6.2 Forecasting of fertility rates

Fertility trends have important societal and economic implications. For this reason, the analysis of fertility has received a lot of attention in economics; see Aaronson et al. (2014), Baudin et al. (2015), Momota (2016), Cooley and Henriksen (2018) and Cooley et al. (2019) among many others. As the statistical analysis of fertility data is quite demanding, the subject is also well-studied in the statistics literature; see for example Lee (1993), Hyndman and Ullah (2007) and Shang (2019).

In this section, we analyze two samples of fertility rates data from Italy and

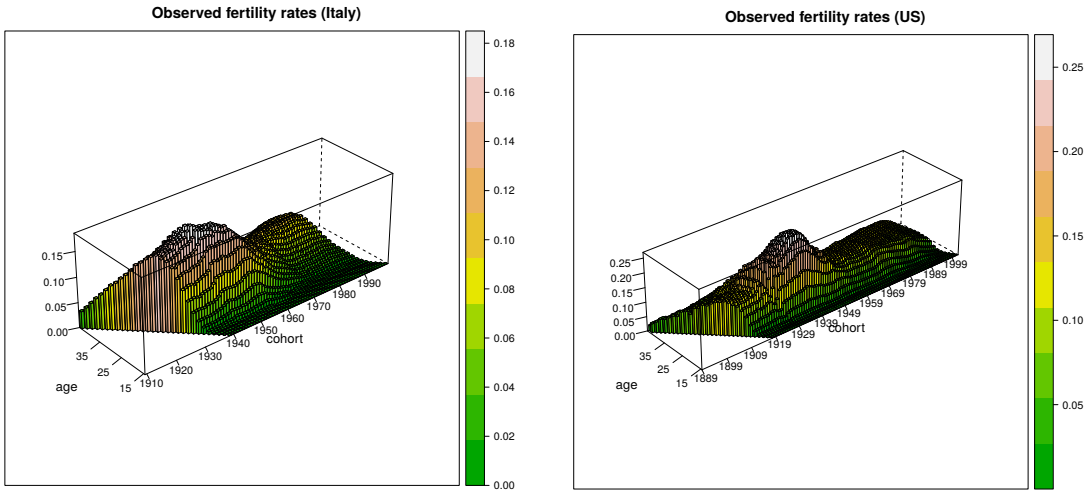

Figure 4: Observed fertility rates for Italy and the US.

the US available on the Human Fertility Database.[3] For both countries, the data sample consists of observations R_{st} which specify the fertility rate by cohort s (in years) and age t (in years). More precisely, $R_{st} = B_{st}/N_{st}$, where B_{st} is the number of births for women of age t born in year s and N_{st} is the total number of women of age t born in year s. Figure 4 shows the observed fertility rates R_{st} for both Italy and the US. For both countries, the age t ranges from 15 to 44 years, while the cohort s covers the years 1939–1999 for Italy and the years 1918–2002 for the US. Inspecting Figure 4, the data can be seen to have a trapezium support of the form $\mathcal{I}_{\underline{m},\overline{m}} = \{(s,t) : s = 1,\dots,\overline{m};\, t = 1,\dots,\overline{m};\, \underline{m} \le s+t-1 \le \overline{m}\}$, where $\underline{m}$ and $\overline{m}$ take different values for Italy and the US.

The birth counts B_{st} and the fertility rates R_{st} can be related as follows to our statistical model introduced in Section 2. The birth counts B_{st} are computed from individual data (X_i, Y_i) on a large number of births $i = 1,\dots,n$ (which are however not available on the Human Fertility Database). For each birth i, we let X_i denote the birth date of the mother and Y_i her age. With these individual data, the birth counts B_{st} are given by $B_{st} = \sum_{i=1}^{n} 1_{st}(X_i, Y_i)$, where $1_{st}(X_i, Y_i)$ is the indicator function which equals 1 iff $X_i \in (s-1, s]$ and $Y_i \in (t-1, t]$. Letting f^B be the density of the individual birth data (X_i, Y_i), the birth counts B_{st} can be regarded as histogram values which estimate the quantities $\int_{s-1}^{s}\int_{t-1}^{t} f^B(x,y)dxdy$. Similarly, the rates data R_{st} can be interpreted as approximations of the quantities $\int_{s-1}^{s}\int_{t-1}^{t} f^R(x,y)dxdy$, where f^R is some underlying intensity function. Note that unlike f^B, the function f^R is not a proper density in general which integrates up to 1. This is however not an issue at all since f^R can be easily re-normalized to integrate to 1.

[3]The data can be requested on the webpage `www.humanfertility.org/cgi-bin/main.php`.

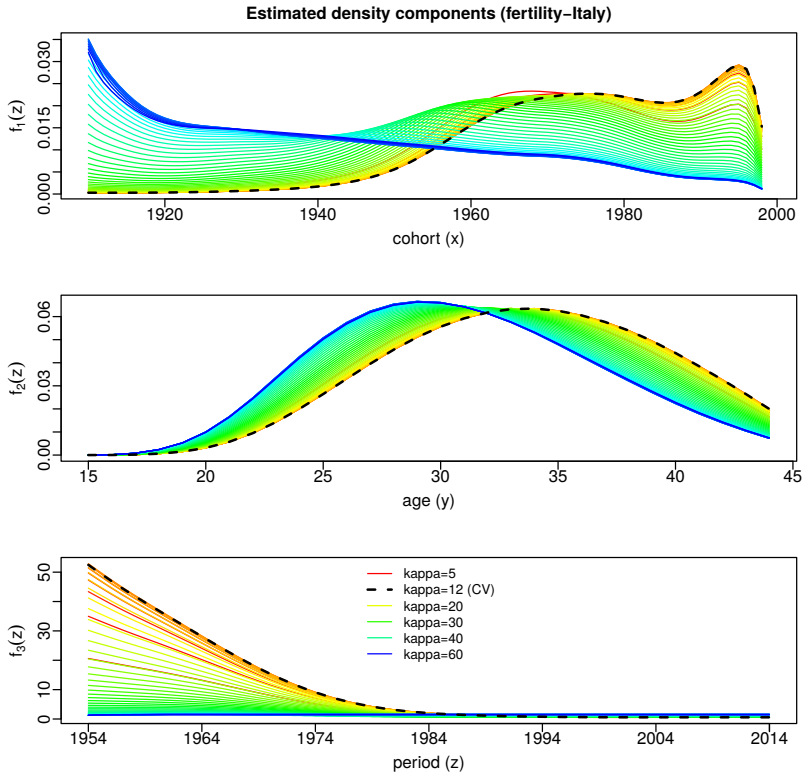

Figure 5: Estimated density components for Italy.

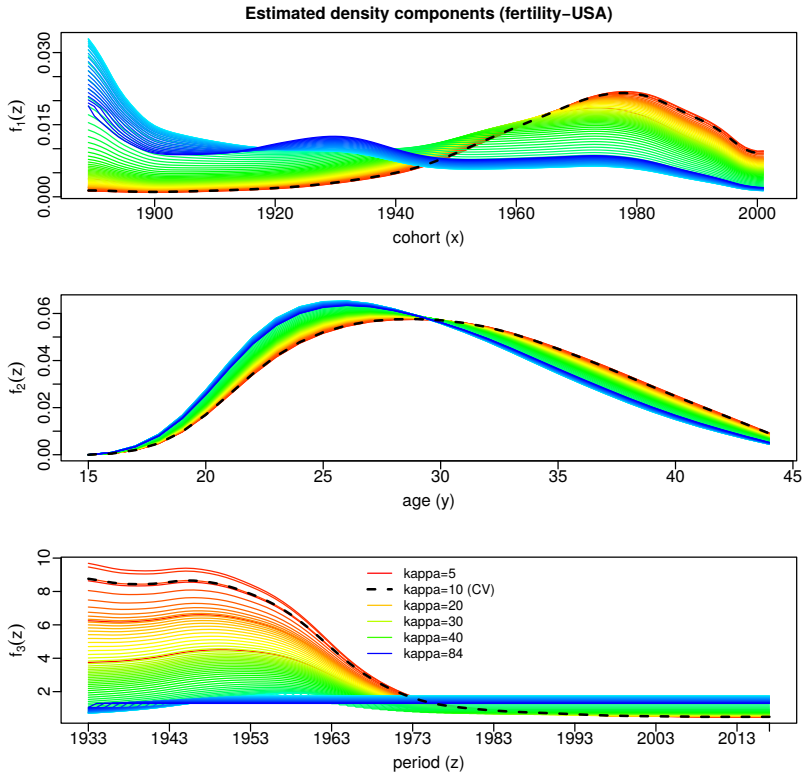

Figure 6: Estimated density components for the US.

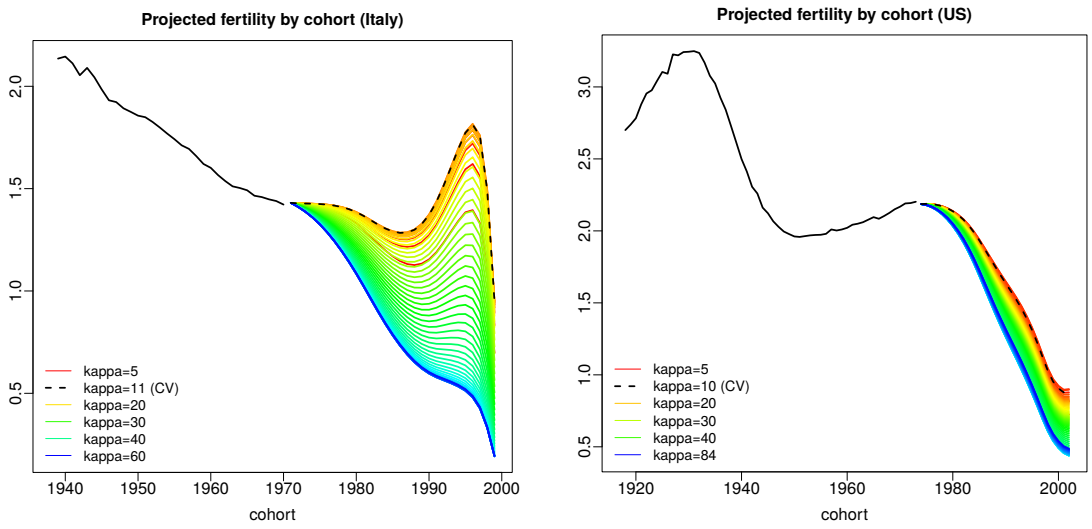

Figure 7: Observed plus projected total fertility rates by cohort for Italy and the US.

In what follows, we analyze the fertility rates R_{st} rather than the absolute fertility numbers B_{st} because the rates are the quantities of primary interest both for research and policy purposes. In particular, we estimate and forecast the density f^R which underlies the fertility rate values R_{st} rather than the density f^B which underlies the birth counts B_{st}. We impose the multiplicative structure $f(x,y) = f_1(x)f_2(y)f_3(x+y)$ from (2.1) on the fertility rate density $f = f^R$ and estimate the density components f_1, f_2 and f_3 by our methods from Section 3. More precisely speaking, we first compute a local linear kernel density estimator of f with cross-validated bandwidths ($\hat{h}_1 = 6.8$ and $\hat{h}_2 = 2.5$ years for Italy, $\hat{h}_1 = 5.3$ and $\hat{h}_2 = 2.5$ years for the US) from the rates data R_{st} and then apply our backfitting algorithm to obtain estimates of the density components f_1, f_2 and f_3. The estimated density components produced by the backfitting algorithm for different values of the parameter κ (given in years) are shown in Figures 5 and 6. The biggest value of κ corresponds to a model with constant calendar effect. In the graphs, we have highlighted the results for the value of κ chosen by the proposed cross-validation method, which was implemented with $\lambda = 1$ year. As can be clearly seen, our estimation results suggest that there is a strongly decreasing calendar effect present in the data both for Italy and the US.

Figure 7 shows our estimates of the (observed plus projected) total fertility rate by cohort (TFR_c) for both Italy and the US. More specifically, it reports our estimates of the quantity $\mathrm{TFR}_c(s) = \int_0^{\overline{m}} \{\int_{s-1}^{s} f(x,y)dx\}dy$ for the cohorts s available in the two data sets. Roughly speaking, the quantity $\mathrm{TFR}_c(s)$ gives the average number of children born to a woman of cohort s who survives until the end of her reproductive life. As argued in academic studies such as Hvidtfeldt et al. (2010),

the quantity TFR_c which is based on a full age-period-cohort decomposition of the data is an accurate measure of fertility. Other common measures such as the total fertility rate by period (TFR_p), in contrast, are shown to be less accurate, resulting in significant underestimation of fertility. Hence, the values of the TFR_p measure, which are for example reported by the World Bank, should be treated with caution. Our methodology allows to produce predictions of the more accurate TFR_c measure. In the case of Italy and the US, the estimated and predicted TFR_c values can be seen to have a falling tendency in both countries, the tendency being somewhat stronger in the US. To be more specific, we have a closer look at the graph for the US. The TFR_c values are fairly stable from the 1950s up to the late 1970s but are predicted to drop strongly from this point onwards. In particular, whereas the cohort of US women born in 1970 had approximately 2 children on average, the cohort of US women born in 2000 (that is, the cohort of women who are approx. 19 years old today) is predicted to have less than 1 child on average. Hence, our new continuous age-period-cohort model predicts alarmingly low future fertility for the US (as well as for Italy).

7 Simulations

In what follows, we examine the finite sample properties of our methods by Monte Carlo experiments.

7.1 Simulation design

We consider a two-dimensional density of the type $f(x, y) = f_1(x)f_2(y)f_3(x + y)$, supported on the triangle $\mathcal{I} = \{(x, y) \in [0, 1]^2 : x+y \leq 1\}$. The density components f_1, f_2 and f_3 are normalized to satisfy the identification constraints (IC1)–(IC3). The density f is extended to the full unit square, $[0, 1]^2$, by defining $f_3(z) = f_3(1)$ for $1 < z \leq 2$. The observation region is $\mathcal{I}$ and the forecasting region is $[0, 1]^2 \setminus \mathcal{I}$. We consider two scenarios, one with constant calendar effect and one with non-constant effect. The first scenario is the correct model for the continuous chain ladder approach of Martínez-Miranda et al. (2013) and the case of $\kappa^* = 1$ in our approach. The density components for each scenario are defined below. In all cases, we simulate pseudo data in the triangle $\mathcal{I}$ in an aggregated form (as counts): We define a two-dimensional grid of 100×100 equally spaced points in $[0, 1]^2$. At each grid point (x, y) in the observation region $\mathcal{I}$, we simulate counts from a Binomial with size n and probability $nf(x, y)/100^2$. We consider the sample sizes $n = 10^4, 10^5$ and 10^6, and simulate 500 pseudo samples for each case.

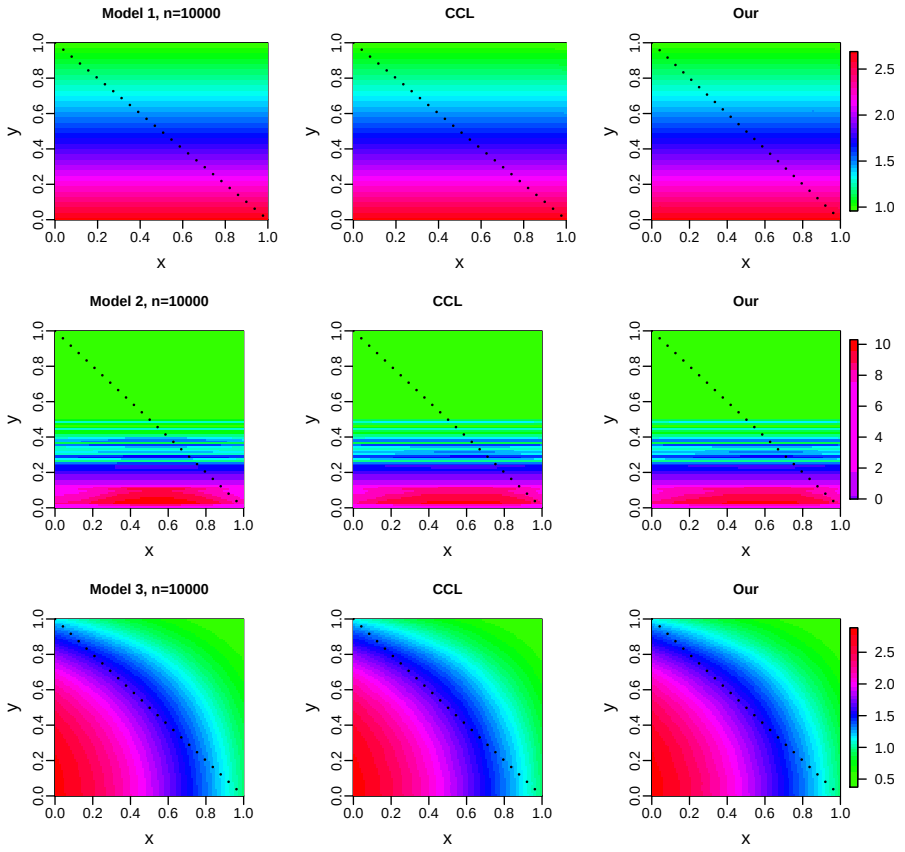

Figure 8: Models 1 to 3 with constant calendar effect. True two-dimensional densities (first column), CCL forecasts (second column) and our forecasts (third column).

We compare our approach with the following procedures: (i) the continuous chain ladder (CCL) approach of Martínez-Miranda et al. (2013) which does not take into account a calendar effect, (ii) a benchmark approach which is identical to our method with κ^* assumed to be known, (iii) the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasters of Kuang et al. (2011) which were introduced in Section 4.2. Our approach is implemented as follows: In a first step, we compute estimates $\hat{f}_1$, $\hat{f}_2$ and $\hat{f}_3$ of the density components f_1, f_2 and f_3 by carrying out the backfitting algorithm described in Section 3 with the same implementation choices as in the two case studies from Section 6. We in particular compute data-driven bandwidths $\hat{h}_1$ and $\hat{h}_2$ by rescaling (by a factor $n^{-1/5}/n^{-1/6}$) the cross-validated bandwidths of the unstructured local linear density estimator. To compute the cross-validation choice of κ defined in (4.2), we set $\lambda = 0.01$, which corresponds to the smallest forecast horizon according to the generation of the data in the triangle (as the bin size of the generated data is 0.01). In the second step, we compute density forecasts by constantly extrapolating the estimated calendar effect into the future as described in Section 4. The CCL and benchmark approaches are implemented in exactly the

Model	n	Our	CCL	I(0)	I(1)	I(2)
1	1e+04	1.1810	1.1034	1.2968	1.3006	3.1544
	1e+05	0.1185	0.1118	0.1303	0.1297	0.3127
	1e+06	0.0129	0.0124	0.0142	0.0139	0.0335
2	1e+04	2.5854	2.5433	2.9370	2.6679	2.6765
	1e+05	0.8282	0.8247	0.8036	0.8585	0.8954
	1e+06	0.1107	0.1101	0.1121	0.1124	0.1165
3	1e+04	0.8369	0.7738	0.8947	0.9245	2.5450
	1e+05	0.0826	0.0773	0.0865	0.0884	0.2235
	1e+06	0.0088	0.0083	0.0088	0.0091	0.0262

Table 2: MISE values of the estimated density $\hat{f}$ (averaged over 500 simulation runs and multiplied by 10^6) for the models with constant calendar effect.

same way, the only difference being that we set $\kappa = 1$ in the CCL case (which means that there is no calendar effect) and $\kappa = \kappa^*$ in the benchmark case (which means that the true κ^* is known). Finally, the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasts are computed as follows: We estimate the density components f_1, f_2 and f_3 by the backfitting algorithm of Section 3, where the normalization constraint (IC3) is replaced by those in Table 1 and the same bandwidths $\hat{h}_1$ and $\hat{h}_2$ as for our approach are used. We then produce density forecasts by constantly extrapolating the estimated calendar effect. The parameter η in the normalization constraint of the $I(2)$ forecast is set to the smallest possible value, which is $\eta = 0.01$ (as the bin size of the generated data is 0.01).

To evaluate the quality of the density forecasts produced by our approach and the competing methods, we proceed as follows: We apply each method to estimate the density $f(x, y)$ on the triangle $\mathcal{I}$ and to forecast it to the full unit square $[0, 1]^2$. For each method, we thus obtain a density forecast $\hat{f}(x, y)$ at all $(x, y) \in [0, 1]^2$. The performance of the density forecast $\hat{f}$ is measured by the MISE criterion $\text{MISE}(\hat{f}) = \int_0^1 \int_0^1 \{\hat{f}(x, y) - f(x, y)\}^2 dxdy$.

7.2 Scenario 1: constant calendar effect

To start with, we consider a simulation scenario where there is no calendar effect. This is the correct model for the CCL method of Martínez-Miranda et al. (2013). We let f_3 be constant and define three theoretical models. Model 1 has $f_1(x) \equiv 1$ and $f_2(y) = (1 - e^{-1})e^{-y}$. Model 2 has the two underlying density components estimated from the reserving data in the first case study above (see Figure 2, case of $\kappa = 10$ years). Model 3 has $f_1(x) = 3/2 - x$ and $f_2(y) = 5/4 - 3y^2/4$ and was previously considered by Lee et al. (2015). Note that in these three models, the

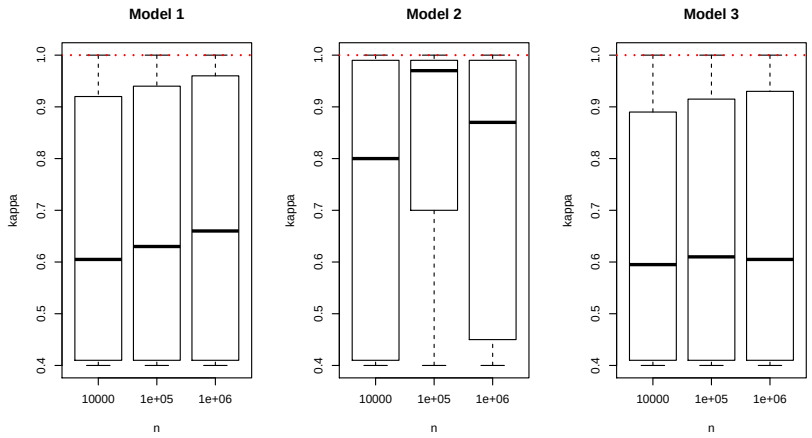

Figure 9: Boxplots of the cross-validated choice of κ for the models with constant calendar effect. The dotted red line shows the true $\kappa^* = 1$.

CCL method is identical to the benchmark approach with the true $\kappa^* = 1$ since there is no calendar effect. Hence, the CCL approach is the benchmark to beat in this scenario.

We first give a visual impression of the density forecasts produced by our approach. Figure 8 compares our density forecasts with the true densities f and the benchmark forecasts of the CCL approach. Specifically, it shows image plots of the true density (first column) and averaged density forecasts (second and third column) for the sample size $n = 10^4$, which are computed as follows: We first estimate/forecast the density f on the unit square for each of the 500 pseudo samples and then compute the average of the 500 density forecasts. The second column corresponds to the averaged density forecasts produced by the benchmark CCL method and the third column to those of our approach. The dotted line divides the observation region (lower triangle) and the forecast region (upper triangle). Inspecting the plots of Figure 8, our forecasts can be seen to approximate the true densities reasonably well for $n = 10^4$, that is, for the smallest sample size under consideration.

We next have a closer look at the performance measure MISE($\hat{f}$). Table 2 reports the MISE values (averaged over 500 simulation runs) that are produced by our method, the CCL approach and the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasters. As expected, the benchmark CCL approach produces slightly better results than our approach when there is no calendar effect. Moreover, the MISE values of our approach are overall comparable to those of the $I(0)$ and $I(1)$ forecasters. Hence, our approach exhibits a similar performance as the $I(0)$ and $I(1)$ forecasters in the models without a calendar effect. The $I(2)$ method, in contrast, produces less accurate results.

Figure 9 gives some details on our cross-validation method to choose κ. It shows

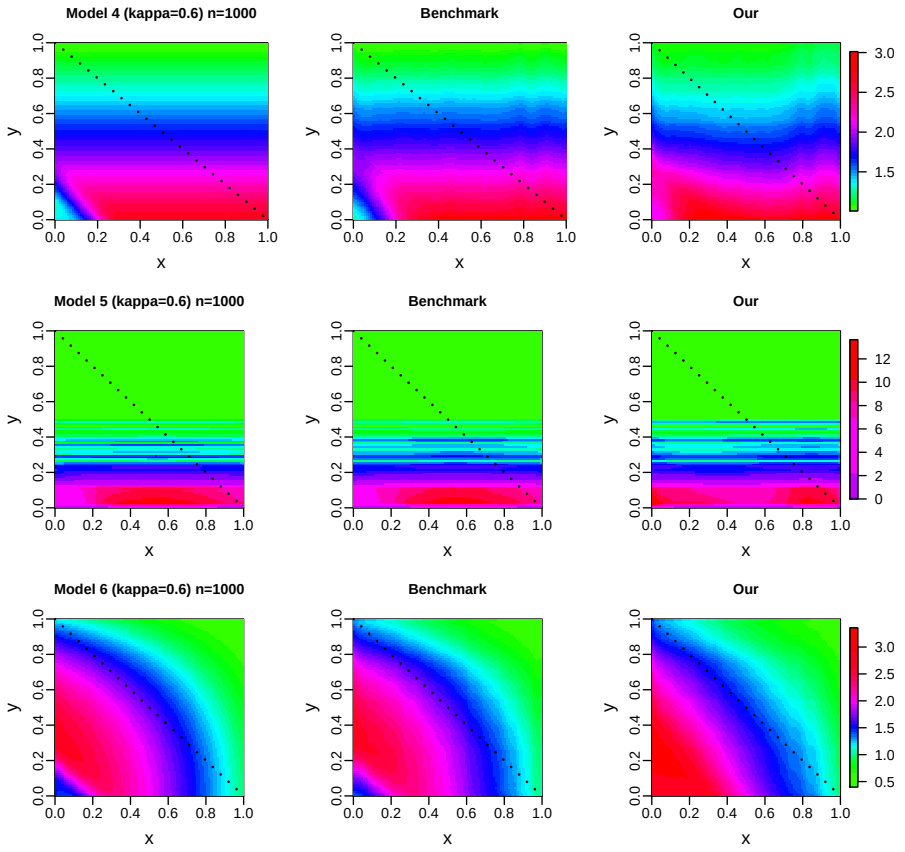

Figure 10: Models 4 to 6 with $\kappa^* = 0.6$ (non-constant calendar effect). True two-dimensional densities (first column), benchmark forecasts (second column) and our forecasts (third column).

boxplots of the cross-validated values of κ for each model and sample size, where we have considered a range of κ values between 0.4 and 1 when running the cross-validation method. As can be seen from the boxplots, the cross-validation method produces values of κ that are highly variable. Moreover, the cross-validated choice of κ moves rather below the true value $\kappa^* = 1$. This however does not seem to have a strong effect on the quality of the final forecast, as shown by the resulting MISE values of our approach in Table 2, which are quite close to those of the CCL approach.

To summarize, our method appears to be almost as good as the CCL method in Models 1–3, which is the benchmark to beat when there is no calendar effect. The simulation results for sample sizes smaller than 10^4 (not included in this paper) have shown that in these cases it is better to stick to the standard CCL method. Small sample sizes seem to be insufficient to estimate a calendar effect properly, due to the sparsity of data points.

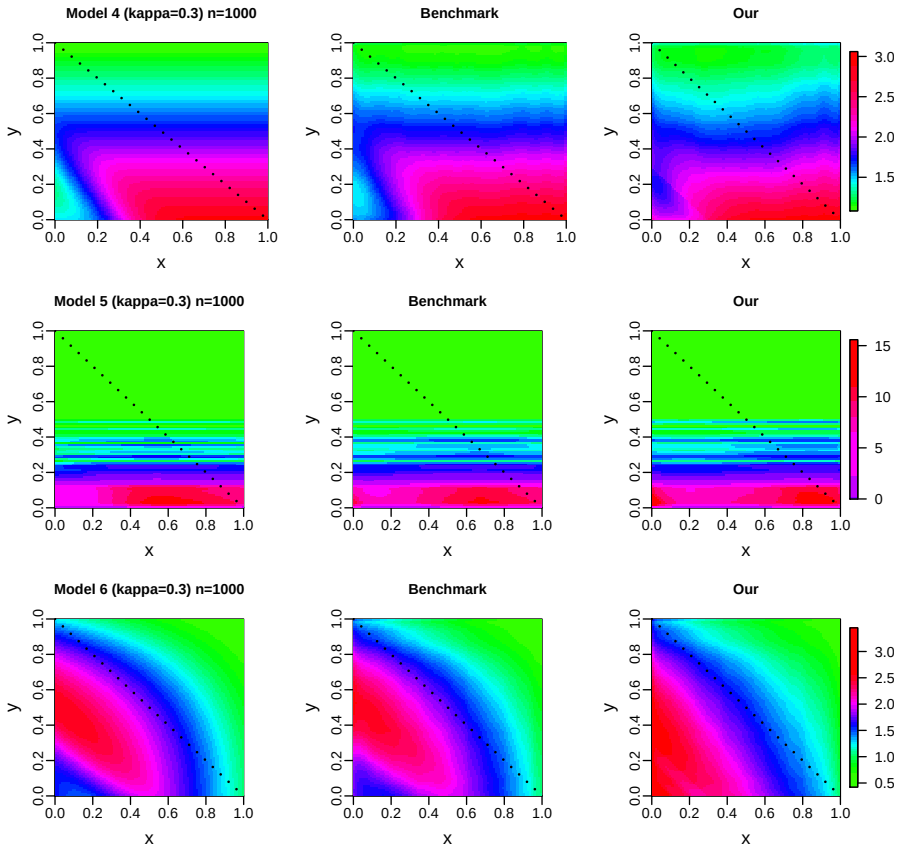

Figure 11: Models 4 to 6 (non-constant calendar effect) with $\kappa^* = 0.3$. True two-dimensional densities (first column), benchmark forecasts (second column) and our forecasts (third column).

7.3 Scenario 2: non-constant calendar effect

We now analyze some models with a non-constant calendar effect, which are referred to as Models 4–6. Models 4–6 have the same density components f_1 and f_2 as Models 1–3, respectively. The third component is $f_3(z) = c\{0.5F_\beta(z/(1-\kappa^*)) + 0.5\}$, where F_β is the cumulative distribution function of a Beta(4,4) random variable and c is a constant which is chosen such that the identification constraints in (IC1)–(IC3) are satisfied. We report simulation results for each model and sample size, considering $\kappa^* = 0.6$ and $\kappa^* = 0.3$.

We first present image plots to visualize the forecasting problems and to check whether our proposal is able to approximate the true densities reasonably well. Figures 10 and 11 show the true densities (first column), benchmark forecasts (second column), and our feasible forecasts (third column) for the sample size $n = 10^4$ averaged over 500 simulation runs. Note that the benchmark forecasts are not identical to the CCL forecasts in Models 4–6 with a calendar effect. As already mentioned in

Model	κ^*	n	Benchmark	Our	CCL	I(0)	I(1)	I(2)
4	0.6	1e+04	1.4101	2.9478	4.9210	31.0265	23.3819	8.7707
		1e+05	0.1884	0.5853	3.2857	24.9174	13.2668	1.3072
		1e+06	0.0285	0.0853	3.0034	23.5965	11.4839	0.2158
5		1e+04	4.3086	4.6054	5.1073	8.8085	4.8323	4.5381
		1e+05	0.4719	0.5846	1.2976	2.7381	0.5820	0.4929
		1e+06	0.2073	0.2377	1.0434	2.2727	0.2889	0.2121
6		1e+04	1.2030	2.3488	3.6743	14.6905	10.0212	10.8827
		1e+05	0.1486	0.4276	2.4192	11.4448	5.1631	1.0812
		1e+06	0.0209	0.0633	2.2354	10.9322	4.5059	0.1972
4	0.3	1e+04	1.9885	3.9444	15.0259	59.9799	20.0552	12.9594
		1e+05	0.2300	0.5996	12.6075	52.5815	12.7389	1.3383
		1e+06	0.0293	0.0815	12.2357	51.8809	11.8387	0.1991
5		1e+04	4.6628	4.8736	5.1887	10.9491	5.0686	4.7610
		1e+05	0.4823	0.5456	1.0943	4.2831	0.5895	0.4950
		1e+06	0.1432	0.1762	0.7622	3.7593	0.2256	0.1455
6		1e+04	1.5305	2.6785	9.2230	28.3376	9.6685	12.8880
		1e+05	0.1723	0.4047	7.5302	24.2104	5.3424	1.0762
		1e+06	0.0208	0.0518	7.2858	23.8338	4.9018	0.1399

Table 3: MISE values of the estimated density $\hat{f}$ (averaged over 500 simulation runs and multiplied by 10^6) for Models 4 to 6 (non-constant calendar effect) with $\kappa^* = 0.6$ and $\kappa^* = 0.3$.

Section 7.1, the benchmark forecasts are obtained by running our approach with the true κ^* parameter. The image plots of Figures 10 and 11 suggest that our density forecasts give a reasonable approximation to the true densities on the unit square in Models 4–6, even though the forecasts are somewhat less precise than those of the benchmark method.

The visual impression given by Figures 10 and 11 is confirmed by the MISE values reported in Table 3: The best forecast results are produced by the benchmark procedure. Our approach turns out to be second best, clearly outperforming the CCL method and the continuous versions of the $I(0)$, $I(1)$ and $I(2)$ forecasters. The MISE values produced by our approach are notably smaller than those of the CCL method and the $I(0)$ forecaster for all models and sample sizes under consideration. Similarly, they are markedly smaller than those of the $I(1)$ and $I(2)$ forecasters in all simulation scenarios concerning Models 4 and 6. Only in the simulation scenarios concerning Model 5, the $I(1)$ and $I(2)$ forecasters produce MISE values that are comparable in size to those of our approach.

Figure 12 shows boxplots of the cross-validated κ parameter that is used to

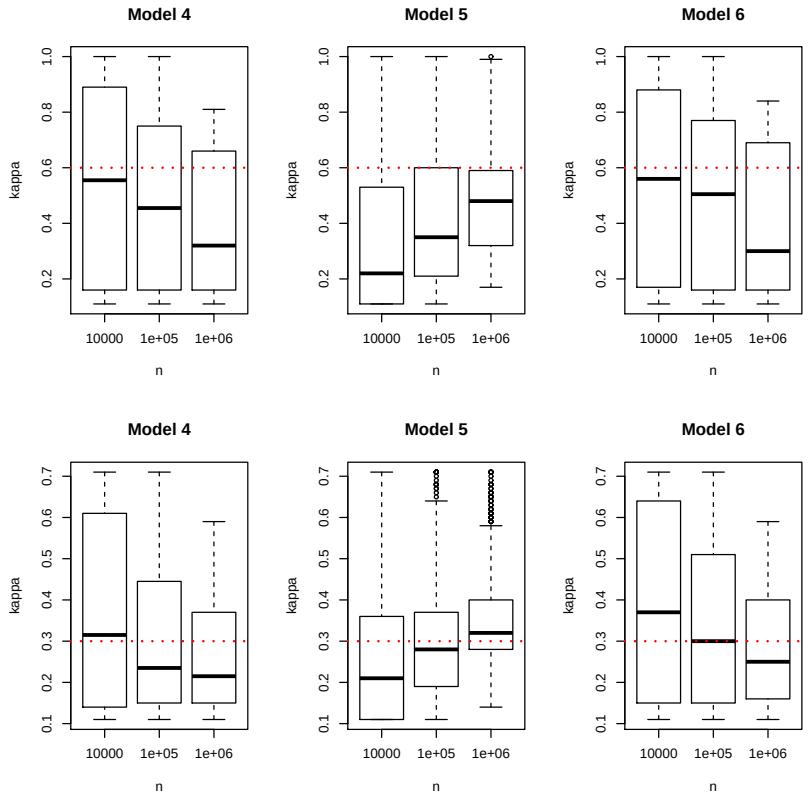

Figure 12: Boxplots of the cross-validated κ choice for Models 4 to 6 (non-constant calendar effect). The dotted red line shows the true κ^*.

derive our forecasts. When running the cross-validation method, we have considered a range of κ values between 0.1 and 1 for $\kappa^* = 0.6$, and between 0.1 and 0.7 for $\kappa^* = 0.3$. We notice that the distribution of the cross-validated κ values is highly variable, the central values in the boxplots hardly representing the true values κ^*. This has a moderate effect on the quality of the final forecasts, as reflected by the MISE values in Table 3: Our κ estimates lead to forecasts which are somewhat worse than the benchmark. Nevertheless, our forecasts are much better than those of the standard CCL approach and the $I(0)$, $I(1)$ and $I(2)$ methods.

8 Conclusion

In this paper, we have developed an extended version of the continuous chain ladder model introduced by Martínez-Miranda et al. (2013) and Mammen et al. (2015). The statistical problem underlying the model is to estimate and forecast a structured nonparametric density which decomposes into several multiplicative components. We have developed backfitting type methods to estimate the structured density and have derived asymptotic theory for our estimators. Moreover, we have proposed a

novel forecasting approach which is based on the following idea: By imposing appropriate identification constraints on the model, the component functions of the structured density can be normalized in a way which allows us to use the simplest possible forecasting strategy: constant extrapolation. Our estimation and forecasting methods are quite general in nature and are useful in a wide range of application contexts. In particular, they can be used to approach various empirical problems in economics. We have illustrated the broad applicability of our methods by two empirical examples. The first is concerned with claims reserving in non-life insurance, which is the original application of the chain ladder methodology. The second is an economic application on fertility forecasting.

9 Acknowledgements

The authors acknowledge the support from the Spanish Ministry of Economy and Competitiveness, through grant number MTM2016-76969P, which includes support from the European Regional Development Fund (ERDF). The authors also acknowledge the Human Fertility Database for freely providing part of the fertility data used in this paper, available at www.humanfertility.org (data downloaded on Sep. 2019), and thank the Centro de Servicios de Informática y Redes de Comunicaciones, Universidad de Granada, for providing the computing resources.

A Technical appendix

In this appendix, we prove the main theoretical results of the paper. Throughout the appendix, we use the symbol C to denote a generic real constant which may take a different value on each occurrence.

A.1 Proof of Theorem 1

To start with, we derive some theoretical properties of the operator $\mathcal{G}(\boldsymbol{\theta}, \boldsymbol{g})$ and its estimator $\hat{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{g})$. The derivative of $\mathcal{G}(\boldsymbol{\theta}, \boldsymbol{g})$ in the direction $(\boldsymbol{d}, \boldsymbol{\delta})$ is given by

$$\begin{aligned}\mathcal{G}'_{(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(x,y,z) = \big(&\mathcal{G}'_{1,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}), \mathcal{G}'_{2,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}), \mathcal{G}'_{3,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}),\\ &\mathcal{G}'_{4,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(x), \mathcal{G}'_{5,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(y), \mathcal{G}'_{6,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(z)\big)^\top,\end{aligned}$$

where

$$\begin{aligned}
\mathcal{G}'_{1,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}) &= -\int_{\mathcal{S}_1} f_1(x)\delta_1(x)dx \\
\mathcal{G}'_{2,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}) &= -\int_{\mathcal{S}_2} f_2(y)\delta_2(y)dy \\
\mathcal{G}'_{3,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta}) &= -\int_{\mathcal{S}} f(x,y)\kappa(x,y,x+y;\boldsymbol{g},\boldsymbol{\delta})dxdy \\
\mathcal{G}'_{4,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(x) &= \int_{\mathcal{J}_2(x)} f(x,y)\big\{d_1 - \kappa(x,y,x+y;\boldsymbol{g},\boldsymbol{\delta})\big\}dy \\
\mathcal{G}'_{5,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(y) &= \int_{\mathcal{J}_1(y)} f(x,y)\big\{d_2 - \kappa(x,y,x+y;\boldsymbol{g},\boldsymbol{\delta})\big\}dx \\
\mathcal{G}'_{6,(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})(z) &= 1_{[0,1-\kappa^*)}(z)\int_{\mathcal{J}_3(z)} f(x,z-x)\big\{d_3 - \kappa(x,z-x,z;\boldsymbol{g},\boldsymbol{\delta})\big\}dx \\
&\quad + 1_{[1-\kappa^*,1]}(z)\frac{1}{\kappa^*}\int_{1-\kappa^*}^{1}\int_{\mathcal{J}_3(v)} f(x,v-x)\big\{d_3 - \kappa(x,v-x,v;\boldsymbol{g},\boldsymbol{\delta})\big\}dxdv
\end{aligned}$$

with

$$\begin{aligned}
\kappa(x,y,z;\boldsymbol{g},\boldsymbol{\delta}) &= \delta_1(x)\{1+g_2(y)\}\{1+g_3(z)\} + \delta_2(y)\{1+g_1(x)\}\{1+g_3(z)\} \\
&\quad + \delta_3(z)\{1+g_1(x)\}\{1+g_2(y)\}.
\end{aligned}$$

Analogously to $\mathcal{G}'_{(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})$, we define $\hat{\mathcal{G}}'_{(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})$ to be the derivative of $\hat{\mathcal{G}}$ at $(\boldsymbol{\theta},\boldsymbol{g})$ in the direction of $(\boldsymbol{d},\boldsymbol{\delta})$. Endowing the space $\mathbb{R}^3 \times \mathcal{L}$ with the norm

$$\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty = \max\Big\{|d_1|,|d_2|,|d_3|,\|\delta_1\|_\infty,\|\delta_2\|_\infty,\|\delta_3\|_\infty\Big\},$$

where $\|\delta_j\|_\infty := \operatorname{ess\,sup}_{w\in\mathcal{S}_j}|\delta_j(w)|$ with ess sup denoting the essential supremum, we can derive the following result.

Lemma 1. *Let the conditions of Theorem 1 be fulfilled and assume in particular that $\hat{f}(x,y) - f(x,y) = O_p(\varepsilon_n)$ uniformly for $(x,y) \in \mathcal{S}$. Then*

(i) $\sup_{\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty=1}\big\|\hat{\mathcal{G}}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) - \mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})\big\|_\infty = O_p(\varepsilon_n)$.

(ii) The operator $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is invertible and has bounded inverse.

(iii) The operator $\hat{\mathcal{G}}'$ is Lipschitz continuous with probability tending to 1, that is, there exist constants $r, C > 0$ such that with probability tending to 1,

$$\sup_{\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty=1}\big\|\hat{\mathcal{G}}'_{(\boldsymbol{\theta}_1,\boldsymbol{g}_1)}(\boldsymbol{d},\boldsymbol{\delta}) - \hat{\mathcal{G}}'_{(\boldsymbol{\theta}_2,\boldsymbol{g}_2)}(\boldsymbol{d},\boldsymbol{\delta})\big\|_\infty \le C\big\|(\boldsymbol{\theta}_1,\boldsymbol{g}_1) - (\boldsymbol{\theta}_2,\boldsymbol{g}_2)\big\|_\infty$$

for all $(\boldsymbol{\theta}_1,\boldsymbol{g}_1)$, $(\boldsymbol{\theta}_2,\boldsymbol{g}_2) \in B_r(\mathbf{0},\mathbf{0})$, where $B_r(\boldsymbol{\theta},\boldsymbol{g})$ is an open ball with radius $r > 0$ and center $(\boldsymbol{\theta},\boldsymbol{g}) \in \mathcal{R}^3 \times \mathcal{L}$.

The main part of the proof of Theorem 1 consists in verifying Lemma 1. Given the result of Lemma 1, the proof of Theorem 1 proceeds analogously as the proof of Theorem 3 in Lee et al. (2015). For the sake of completeness, we provide the details in what follows. The proof of Lemma 1 is postponed until the arguments for Theorem 1 are complete.

With the help of statement (ii) of Lemma 1, we obtain that

$$\begin{aligned}\left\|\begin{pmatrix}\bar{\boldsymbol{\phi}}-\boldsymbol{\phi}\\(\bar{\boldsymbol{f}}-\boldsymbol{f})/\boldsymbol{f}\end{pmatrix}\right\|_\infty &= \left\|\mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}\begin{pmatrix}\mathbf{0}\\-\boldsymbol{f}_w\circ\hat{\boldsymbol{\mu}}\end{pmatrix}\right\|_\infty \le C\left\|\begin{pmatrix}\mathbf{0}\\-\boldsymbol{f}_w\circ\hat{\boldsymbol{\mu}}\end{pmatrix}\right\|_\infty\\ &\le C\max\Big\{\sup_{x\in\mathcal{S}_1}\int_{\mathcal{J}_2(x)}\big\{\hat{f}(x,y)-f(x,y)\big\}dy,\\ &\qquad \sup_{y\in\mathcal{S}_2}\int_{\mathcal{J}_1(y)}\big\{\hat{f}(x,y)-f(x,y)\big\}dx,\\ &\qquad \sup_{z\in\mathcal{S}_3}\int_{\mathcal{J}_3(z)}\big\{\hat{f}(x,z-x)-f(x,z-x)\big\}dx\Big\}, \qquad \text{(A.1)}\end{aligned}$$

which immediately implies that

$$\left\|\begin{pmatrix}\bar{\boldsymbol{\phi}}-\boldsymbol{\phi}\\(\bar{\boldsymbol{f}}-\boldsymbol{f})/\boldsymbol{f}\end{pmatrix}\right\|_\infty = O_p(\varepsilon_n). \qquad \text{(A.2)}$$

This together with statements (i) and (iii) of Lemma 1 yields that

$$\sup_{\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty=1}\left\|\hat{\mathcal{G}}'_{(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\{\bar{\boldsymbol{f}}-\boldsymbol{f}\}/\boldsymbol{f})}(\boldsymbol{d},\boldsymbol{\delta})-\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})\right\|_\infty = O_p(\varepsilon_n). \qquad \text{(A.3)}$$

From this and Lemma 1(ii), it further follows that $\hat{\mathcal{G}}'_{(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\{\bar{\boldsymbol{f}}-\boldsymbol{f}\}/\boldsymbol{f})}$ is invertible and

$$\sup_{\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty=1}\left\|\hat{\mathcal{G}}'^{-1}_{(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\{\bar{\boldsymbol{f}}-\boldsymbol{f}\}/\boldsymbol{f})}(\boldsymbol{d},\boldsymbol{\delta})\right\|_\infty \le C \qquad \text{(A.4)}$$

with probability tending to 1. Finally, it holds that

$$\begin{aligned}\hat{\mathcal{G}}\Big(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\frac{\bar{\boldsymbol{f}}-\boldsymbol{f}}{\boldsymbol{f}}\Big) &= \hat{\mathcal{G}}(\mathbf{0},\mathbf{0})+\hat{\mathcal{G}}'_{(\mathbf{0},\mathbf{0})}\Big(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\frac{\bar{\boldsymbol{f}}-\boldsymbol{f}}{\boldsymbol{f}}\Big)+O_p(\varepsilon_n^2)\\ &= \hat{\mathcal{G}}(\mathbf{0},\mathbf{0})+\mathcal{G}'_{(\mathbf{0},\mathbf{0})}\Big(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\frac{\bar{\boldsymbol{f}}-\boldsymbol{f}}{\boldsymbol{f}}\Big)+O_p(\varepsilon_n^2)\\ &= O_p(\varepsilon_n^2+n^{-1/2}), \qquad \text{(A.5)}\end{aligned}$$

where the first equality follows with the help of Lemma 1(iii), the second is a consequence of Lemma 1(i) and (A.2), and the third exploits the fact that

$$\hat{\mathcal{G}}(\mathbf{0},\mathbf{0})=\begin{pmatrix}0\\0\\\hat{\vartheta}-\vartheta\\\boldsymbol{f}_w\circ\hat{\boldsymbol{\mu}}\end{pmatrix} \qquad \text{and} \qquad \mathcal{G}'_{(\mathbf{0},\mathbf{0})}\Big(\bar{\boldsymbol{\phi}}-\boldsymbol{\phi},\frac{\bar{\boldsymbol{f}}-\boldsymbol{f}}{\boldsymbol{f}}\Big)=\begin{pmatrix}0\\0\\0\\-\boldsymbol{f}_w\circ\hat{\boldsymbol{\mu}}\end{pmatrix}.$$

With (A.2)–(A.5), we can apply the following version of the Newton-Kantorovich theorem (cp. Theorem 15.6 in Deimling (1985)).

Theorem. *Let B_1, B_2 be Banach spaces and $T : B_r(x_0) \subseteq B_1 \to B_2$ a continuously differentiable mapping, where $B_r(x_0)$ is the open ball with radius $r > 0$ and center $x_0 \in B_1$. Suppose that the derivative $T'_{(x_0)}$ of T at x_0 is bijective and has bounded inverse, that is, $\|T'^{-1}_{(x_0)}\| \leq \beta < \infty$. Moreover, assume that $\|T'^{-1}_{(x_0)} T(x_0)\| \leq \alpha$ and $\|T'_{(x)} - T'_{(y)}\| \leq k\|x - y\|$ for all $x, y \in B_r(x_0)$, where $q := 2k\alpha\beta < 1$ and $2\alpha < r$. Then the mapping T has a unique root z in the closed ball $\bar{B}_{2\alpha}(x_0)$ and the Newton iterates*

$$x_{m+1} = x_m - T'^{-1}_{(x_m)} T(x_m)$$

satisfy

$$\|x_m - z\| \leq \alpha 2^{-(m-1)} q^{2^m - 1}. \tag{A.6}$$

We now show that the conditions of this theorem are satisfied for $T = \hat{\mathcal{G}}$ and $x_0 = (\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, \{\bar{\boldsymbol{f}} - \boldsymbol{f}\}/\boldsymbol{f})$ with probability tending to 1: Note that by (A.4),

$$\sup_{\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty = 1} \left\| \hat{\mathcal{G}}'^{-1}_{(\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, \{\bar{\boldsymbol{f}} - \boldsymbol{f}\}/\boldsymbol{f})} (\boldsymbol{d}, \boldsymbol{\delta}) \right\|_\infty \leq \beta$$

for some sufficiently large β with probability tending to 1 and define

$$\alpha = \alpha_n := \beta \left\| \hat{\mathcal{G}} \Big(\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, \frac{\bar{\boldsymbol{f}} - \boldsymbol{f}}{\boldsymbol{f}} \Big) \right\|_\infty .$$

Since $\alpha_n = O_p(\varepsilon_n^2 + n^{-1/2})$ by (A.5) and Lemma 1(iii) holds for all $(\boldsymbol{\theta}_1, \boldsymbol{g}_1), (\boldsymbol{\theta}_2, \boldsymbol{g}_2) \in B_r(\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, \{\bar{\boldsymbol{f}} - \boldsymbol{f}\}/\boldsymbol{f})$ with some $r > 2\alpha_n$ for sufficiently large n, the conditions of the Newton-Kantorovich theorem are fulfilled with probability tending to 1. We thus obtain that with probability tending to 1, there exists a unique solution of the equation $\hat{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{g}) = (\mathbf{0}, \mathbf{0})$ in $\bar{B}_{2\alpha_n}(\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}, \{\bar{\boldsymbol{f}} - \boldsymbol{f}\}/\boldsymbol{f})$, which by definition is equal to $(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}, \{\hat{\boldsymbol{f}} - \boldsymbol{f}\}/\boldsymbol{f})$. By (A.6), it further holds that

$$\left\| \begin{pmatrix} \bar{\boldsymbol{\phi}} - \boldsymbol{\phi} \\ (\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f} \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\phi}} - \boldsymbol{\phi} \\ (\hat{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f} \end{pmatrix} \right\|_\infty = \left\| \begin{pmatrix} \bar{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}} \\ (\bar{\boldsymbol{f}} - \hat{\boldsymbol{f}})/\boldsymbol{f} \end{pmatrix} \right\|_\infty \leq C\alpha_n$$

with probability tending to 1, that is,

$$\left\| \begin{pmatrix} \bar{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}} \\ (\bar{\boldsymbol{f}} - \hat{\boldsymbol{f}})/\boldsymbol{f} \end{pmatrix} \right\|_\infty = O_p(\varepsilon_n^2 + n^{-1/2}).$$

This completes the proof of Theorem 1. It remains to verify Lemma 1.

Proof of Lemma 1. We now prove the three statements of Lemma 1.

Proof of (i). Inspecting the formulas for $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})$ and $\hat{\mathcal{G}}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})$, (i) can be seen to follow from the assumption that $\sup_{(x,y)\in\mathcal{S}}|\hat{f}(x,y)-f(x,y)| = O_p(\varepsilon_n)$.

Proof of (ii). We first prove that the mapping $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is injective. Suppose that $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$ for some $(\boldsymbol{d},\boldsymbol{\delta})$. We show that $(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$ must hold, implying that $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is injective. As a first step, we verify that $d_1 = d_2 = d_3 = 0$: Inspecting the third equation of the system $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$, we see that

$$\int_{\mathcal{S}} f(x,y)\delta_+(x,y)dxdy = 0.$$

Integrating the fourth equation of $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$, we thus obtain that

$$0 = \int_{\mathcal{S}} f(x,y)\big[d_1 - \delta_+(x,y)\big]dxdy = d_1\int_{\mathcal{S}} f(x,y)dxdy,$$

which implies that $d_1 = 0$. Proceeding analogously with the fifth and sixth equation, we get that $d_2 = d_3 = 0$ as well. We next prove that $\delta_1 = \delta_2 = \delta_3 = 0$: As $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$ by assumption, it holds that

$$\begin{aligned} 0 &= \int_{\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{S}_3} \Big(\mathbf{0}^\top, \frac{\delta_1(x)}{\ell_2\ell_3}, \frac{\delta_2(y)}{\ell_1\ell_3}, \frac{\delta_3(z)}{\ell_1\ell_2}\Big)\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})(x,y,z)dxdydz \\ &= -\int_{\mathcal{S}} f(x,y)\big[\delta_1(x)+\delta_2(y)+\delta_3(x+y)\big]^2 dxdy, \end{aligned}$$

where we use the notation $\ell_j = \int_0^1 1(w\in\mathcal{S}_j)dw$ for $j = 1,2,3$. This implies that

$$\delta_1(x)+\delta_2(y)+\delta_3(x+y) = 0 \quad \text{a.e. on } \mathcal{S}.$$

Exploiting that δ_3 is constant on the interval $[1-\kappa^*, 1]$ by definition, we can infer that δ_j is a constant function on $\mathcal{S}_j$ for $j = 1,2$. Since $\int_{\mathcal{S}_j} f_j(w)\delta_j(w)dw = 0$ for $j = 1,2$ by the first two equations of $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta}) = \mathbf{0}$, we further get that $\delta_1 = \delta_2 = 0$, which in turn implies that $\delta_3 = 0$.

We next show that $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is surjective. Define $\langle(\boldsymbol{c},\boldsymbol{\eta}),(\tilde{\boldsymbol{c}},\tilde{\boldsymbol{\eta}})\rangle = \sum_{j=1}^3 c_j\tilde{c}_j + \sum_{j=1}^3\int_{\mathcal{S}_j}\eta_j(w)\tilde{\eta}_j(w)dw$ for $(\boldsymbol{c},\boldsymbol{\eta}),(\tilde{\boldsymbol{c}},\tilde{\boldsymbol{\eta}})\in\mathbb{R}^3\times\mathcal{L}$ and suppose that for some $(\boldsymbol{c},\boldsymbol{\eta})\in\mathbb{R}^3\times\mathcal{L}$, it holds that

$$\Big\langle \begin{pmatrix}\boldsymbol{c}\\ \boldsymbol{\eta}\end{pmatrix}, \mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d},\boldsymbol{\delta})\Big\rangle = 0 \tag{A.7}$$

for all $(\boldsymbol{d},\boldsymbol{\delta})\in\mathbb{R}^3\times\mathcal{L}$. We show that $(\boldsymbol{c},\boldsymbol{\eta}) = (\mathbf{0},\mathbf{0})$ must hold in this case, implying that $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is surjective. Choosing $d_j = 1$ for some $j\in\{1,2,3\}$ and setting all other

components of $(\boldsymbol{d}, \boldsymbol{\delta})$ to zero in (A.7), we get that

$$0 = \int_{\mathcal{S}} \eta_1(x) f(x,y) dxdy \tag{A.8}$$

$$0 = \int_{\mathcal{S}} \eta_2(y) f(x,y) dxdy \tag{A.9}$$

$$0 = \int_{\mathcal{S}} \eta_3(x+y) f(x,y) dxdy. \tag{A.10}$$

Picking $\delta_j \equiv 1$ for some $j \in \{1,2,3\}$, setting all other components of $(\boldsymbol{d}, \boldsymbol{\delta})$ to zero and using the shorthand $\eta_+(x,y) = \eta_1(x) + \eta_2(y) + \eta_3(x+y)$, we further arrive at

$$0 = c_1 \int_{\mathcal{S}_1} f_1(x) dx + c_3 \int_{\mathcal{S}} f(x,y) dxdy + \int_{\mathcal{S}} f(x,y) \eta_+(x,y) dxdy \tag{A.11}$$

$$0 = c_2 \int_{\mathcal{S}_2} f_2(y) dy + c_3 \int_{\mathcal{S}} f(x,y) dxdy + \int_{\mathcal{S}} f(x,y) \eta_+(x,y) dxdy \tag{A.12}$$

$$0 = c_3 \int_{\mathcal{S}} f(x,y) dxdy + \int_{\mathcal{S}} f(x,y) \eta_+(x,y) dxdy. \tag{A.13}$$

We now combine (A.13) with (A.8)–(A.10) to obtain that $c_3 \int_{\mathcal{S}} f(x,y) dxdy = 0$, which in turn implies that $c_3 = 0$. Combining (A.11) and (A.12) with (A.8)–(A.10) in an analogous fashion, we additionally get that $c_1 = c_2 = 0$. We next set $d_1 = d_2 = d_3 = 0$ and $\delta_j = \eta_j$ for all $j = 1,2,3$ in (A.7). Taking into account that $c_1 = c_2 = c_3 = 0$, we obtain that

$$0 = \int_{\mathcal{S}} f(x,y) \eta_+^2(x,y) dxdy,$$

which implies that

$$\eta_+(x,y) = \eta_1(x) + \eta_2(y) + \eta_3(x+y) = 0 \quad \text{a.e. on } \mathcal{S}.$$

As in the first part of the proof, we now exploit that η_3 is constant on the interval $[1-\kappa^*, 1]$ to get that η_j is a constant function on $\mathcal{S}_j$ for $j = 1,2$. By (A.8) and (A.9), we can infer that $\eta_1 = \eta_2 = 0$, which in turn implies that $\eta_3 = 0$.

To verify that the inverse $\mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}$ is bounded, it is sufficient to prove that the bijective linear operator $\mathcal{G}'_{(\mathbf{0},\mathbf{0})}$ is bounded according to the bounded inverse theorem. Under our conditions, it obviously holds that $\|\mathcal{G}'_{(\mathbf{0},\mathbf{0})}(\boldsymbol{d}, \boldsymbol{\delta})\|_\infty \leq C \|(\boldsymbol{d}, \boldsymbol{\delta})\|_\infty$, which completes the proof.

Proof of (iii). The function κ defined at the beginning of the proof satisfies the inequality

$$\sup_{(x,y,z) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3} \big| \kappa(x,y,z; \boldsymbol{g}_2, \boldsymbol{\delta}) - \kappa(x,y,z; \boldsymbol{g}_1, \boldsymbol{\delta}) \big| \\ \leq 3 \|\boldsymbol{\delta}\|_\infty (2 + \|\boldsymbol{g}_1\|_\infty + \|\boldsymbol{g}_2\|_\infty) \|\boldsymbol{g}_2 - \boldsymbol{g}_1\|_\infty.$$

From this and the definition of $\hat{\mathcal{G}}'_{(\boldsymbol{\theta},\boldsymbol{g})}(\boldsymbol{d},\boldsymbol{\delta})$, it follows that for any given $r>0$,

$$\big\|\hat{\mathcal{G}}'_{(\boldsymbol{\theta}_1,\boldsymbol{g}_1)}(\boldsymbol{d},\boldsymbol{\delta})-\hat{\mathcal{G}}'_{(\boldsymbol{\theta}_2,\boldsymbol{g}_2)}(\boldsymbol{d},\boldsymbol{\delta})\big\|_\infty \le 6(1+r)\max_{1\le j\le 3}\sup_{v\in\mathcal{S}_j} f_{w,j}(v)\|\boldsymbol{g}_2-\boldsymbol{g}_1\|_\infty$$

for all $(\boldsymbol{\theta}_1,\boldsymbol{g}_1)$, $(\boldsymbol{\theta}_2,\boldsymbol{g}_2)\in B_r(\mathbf{0},\mathbf{0})$ and all $(\boldsymbol{d},\boldsymbol{\delta})$ with $\|(\boldsymbol{d},\boldsymbol{\delta})\|_\infty=1$. □

A.2 Proof of Theorem 2

The local linear estimator $\hat{f}$ with $h_1\sim h_2\sim n^{-1/5}$ has the property that

$$\sup_{(x,y)\in\mathcal{S}}\big|\hat{f}(x,y)-f(x,y)\big|=O_p(\varepsilon_n)$$

with $\varepsilon_n=n^{-3/10}\sqrt{\log n}$. By Theorem 1, it thus holds that

$$\sup_{w\in\mathcal{S}_j}|\hat{f}_j(w)-\bar{f}_j(w)|=O_p(n^{-3/5}\log n+n^{-1/2})=O_p(n^{-1/2})$$

for $j=1,2,3$. To complete the proof, it remains to show that

$$\sup_{w\in\mathcal{S}_j}|\bar{f}_j(w)-f_j(w)|=O_p(n^{-2/5}\sqrt{\log n}). \tag{A.14}$$

By (A.1), we know that

$$\left\|\begin{pmatrix}\bar{\boldsymbol{\phi}}-\boldsymbol{\phi}\\(\bar{\boldsymbol{f}}-\boldsymbol{f})/\boldsymbol{f}\end{pmatrix}\right\|_\infty \le C\max\Big\{\sup_{x\in\mathcal{S}_1}\int_{\mathcal{J}_2(x)}\big\{\hat{f}(x,y)-f(x,y)\big\}dy,\ \sup_{y\in\mathcal{S}_2}\int_{\mathcal{J}_1(y)}\big\{\hat{f}(x,y)-f(x,y)\big\}dx,\ \sup_{z\in\mathcal{S}_3}\int_{\mathcal{J}_3(z)}\big\{\hat{f}(x,z-x)-f(x,z-x)\big\}dx\Big\}.$$

Standard theory for kernel estimators shows that the right-hand side is of the order $O_p(n^{-2/5}\sqrt{\log n})$, which implies (A.14).

A.3 Proof of Theorem 3

By Theorem 1, we know that

$$\sup_{w\in\mathcal{S}_j}|\hat{f}_j(w)-\bar{f}_j(w)|=O_p(\varepsilon_n^2+n^{-1/2})=o_p(n^{-2/5})$$

with $\varepsilon_n=n^{-3/10}\sqrt{\log n}$ for $j=1,2,3$. Hence, for any fixed $w\in\mathcal{S}_j$,

$$\frac{\hat{f}_j(w)-f_j(w)}{f_j(w)}=\frac{\bar{f}_j(w)-f_j(w)}{f_j(w)}+o_p(n^{-2/5}),$$

which implies that the asymptotic distribution of $(\hat{f}_j(w) - f_j(w))/f_j(w)$ is identical to that of $(\bar{f}_j(w) - f_j(w))/f_j(w)$. To complete the proof, it thus suffices to derive the limit distribution of $(\bar{f}_j(w) - f_j(w))/f_j(w)$ for $j = 1, 2, 3$.

To start with, we decompose the term $(\bar{f}_j(w) - f_j(w))/f_j(w)$ into a variance and a bias part: Let $\hat{f}^A(x, y)$ be the first entry of the vector $\hat{\boldsymbol{\eta}}^A(x, y)$ which is defined analogous to $\hat{\boldsymbol{\eta}}$ in (3.10) with $\boldsymbol{b}$ replaced by $\boldsymbol{b} - \mathbb{E}\boldsymbol{b}$. Likewise, let $\hat{f}^B(x, y)$ be the first component of $\hat{\boldsymbol{\eta}}^B(x, y)$ which is defined as $\hat{\boldsymbol{\eta}}(x, y)$ with $\boldsymbol{b}(x, y)$ replaced by $\mathbb{E}\boldsymbol{b}(x, y) - (f(x, y),\, h_1\partial f(x, y)/\partial x,\, h_2\partial f(x, y)/\partial y)^\top$. With these definitions, we can decompose the local linear density estimator $\hat{f}$ according to

$$\hat{f}(x, y) - f(x, y) = \hat{f}^A(x, y) + \hat{f}^B(x, y),$$

where $\hat{f}^A$ and $\hat{f}^B$ play the role of the variance and the bias part of $\hat{f}$, respectively. For $k = A, B$, we define the quantities $\bar{\boldsymbol{\phi}}^k - \boldsymbol{\phi}$ and $\bar{\boldsymbol{f}}^k/\boldsymbol{f}$ by the operator equation

$$\begin{pmatrix} \bar{\boldsymbol{\phi}}^k - \boldsymbol{\phi} \\ \bar{\boldsymbol{f}}^k/\boldsymbol{f} \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})} \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ \hat{\boldsymbol{\mu}}^k \end{pmatrix}, \tag{A.15}$$

where

$$\begin{aligned}
\hat{\mu}_1^k(x) &= f_{w,1}^{-1}(x) \int_{\mathcal{J}_2(x)} \hat{f}^k(x, y)dy \\
\hat{\mu}_2^k(y) &= f_{w,2}^{-1}(y) \int_{\mathcal{J}_1(y)} \hat{f}^k(x, y)dx \\
\hat{\mu}_3^k(z) &= 1_{[0,1-\kappa^*)}(z)\, f_{w,3}^{-1}(z) \int_{\mathcal{J}_3(z)} \hat{f}^k(x, z - x)dx \\
&\quad + 1_{[1-\kappa^*,1]}(z)\, c_{w,3}^{-1} \frac{1}{\kappa^*} \int_{1-\kappa^*}^1 \int_{\mathcal{J}_3(v)} \hat{f}^k(x, v - x)dxdv.
\end{aligned}$$

The operator equation (A.15) parallels (5.6) which defines the quantities $\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}$ and $(\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f}$. From (A.15), it follows that $\bar{\boldsymbol{\phi}}^k - \boldsymbol{\phi}$ and $\bar{\boldsymbol{f}}^k/\boldsymbol{f}$ satisfy a system of backfitting equations. Specifically, for $k = A, B$, the quantities

$$\begin{aligned}
\bar{\boldsymbol{\delta}}^k &= (\bar{\delta}_1^k, \bar{\delta}_2^k, \bar{\delta}_3^k)^\top = \Big(\frac{\bar{f}_1^k}{f_1}, \frac{\bar{f}_2^k}{f_2}, \frac{\bar{f}_3^k}{f_3}\Big)^\top \\
\bar{\boldsymbol{d}}^k &= (\bar{d}_1^k, \bar{d}_2^k, \bar{d}_3^k)^\top = (\bar{\phi}_1^k - \phi_1, \bar{\phi}_2^k - \phi_2, \bar{\phi}_3^k - \phi_3)^\top
\end{aligned}$$

solve the backfitting equations

$$\bar{\delta}_1^k(x) = \bar{d}_1^k + \hat{\mu}_1^k(x) - \int_{\mathcal{J}_2(x)} \big\{\bar{\delta}_2^k(y) + \bar{\delta}_3^k(x + y)\big\} \frac{f(x, y)}{f_{w,1}(x)} dy \tag{A.16}$$

$$\bar{\delta}_2^k(y) = \bar{d}_2^k + \hat{\mu}_2^k(y) - \int_{\mathcal{J}_1(y)} \big\{\bar{\delta}_1^k(x) + \bar{\delta}_3^k(x + y)\big\} \frac{f(x, y)}{f_{w,2}(y)} dx \tag{A.17}$$

$$\bar{\delta}_3^k(z) = \bar{d}_3^k + \hat{\mu}_3^k(z) - 1_{[0,1-\kappa^*)}(z) \int_{\mathcal{J}_3(z)} \big\{\bar{\delta}_1^k(x) + \bar{\delta}_2^k(z-x)\big\} \frac{f(x,z-x)}{f_{w,3}(z)} dx$$
$$- 1_{[1-\kappa^*,1]}(z) \frac{1}{\kappa^*} \int_{1-\kappa^*}^1 \int_{\mathcal{J}_3(v)} \big\{\bar{\delta}_1^k(x) + \bar{\delta}_2^k(v-x)\big\} \frac{f(x,v-x)}{c_{w,3}} dx dv \quad \text{(A.18)}$$

subject to

$$0 = \int_{\mathcal{S}_1} f_1(x)\bar{\delta}_1^k(x)dx \quad \text{(A.19)}$$
$$0 = \int_{\mathcal{S}_2} f_2(y)\bar{\delta}_2^k(y)dy \quad \text{(A.20)}$$
$$0 = \int_{\mathcal{S}} f(x,y)\big[\bar{\delta}_1^k(x) + \bar{\delta}_2^k(y) + \bar{\delta}_3^k(x+y)\big]dxdy. \quad \text{(A.21)}$$

Analogous backfitting equations hold for the quantities $\bar{\boldsymbol{\phi}} - \boldsymbol{\phi}$ and $(\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f}$ defined via (5.6). Inspecting (A.16)–(A.18) and taking into account that $\hat{\mu}_j^A(w) + \hat{\mu}_j^B(w) = \hat{\mu}_j(w)$ for $j = 1, 2, 3$, it is easily seen that $\bar{\boldsymbol{\phi}} = \bar{\boldsymbol{\phi}}^A - \boldsymbol{\phi} + \bar{\boldsymbol{\phi}}^B$ and

$$\frac{\bar{\boldsymbol{f}} - \boldsymbol{f}}{\boldsymbol{f}} = \frac{\bar{\boldsymbol{f}}^A}{\boldsymbol{f}} + \frac{\bar{\boldsymbol{f}}^B}{\boldsymbol{f}},$$

where $\bar{\boldsymbol{f}}^A/\boldsymbol{f}$ and $\bar{\boldsymbol{f}}^B/\boldsymbol{f}$ are the variance and bias part of $(\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f}$, respectively. For our subsequent analysis, we write the backfitting equations (A.16)–(A.18) as

$$\bar{\boldsymbol{\delta}}^k = \bar{\boldsymbol{d}}^k + \hat{\boldsymbol{\mu}}^k - \boldsymbol{T}\bar{\boldsymbol{\delta}}^k, \quad \text{(A.22)}$$

or more explicitly as

$$\begin{pmatrix} \bar{\delta}_1^k(x) \\ \bar{\delta}_2^k(y) \\ \bar{\delta}_3^k(z) \end{pmatrix} = \begin{pmatrix} \bar{d}_1^k \\ \bar{d}_2^k \\ \bar{d}_3^k \end{pmatrix} + \begin{pmatrix} \hat{\mu}_1^k(x) \\ \hat{\mu}_2^k(y) \\ \hat{\mu}_3^k(z) \end{pmatrix} - \begin{pmatrix} (T_1\bar{\boldsymbol{\delta}}^k)(x) \\ (T_2\bar{\boldsymbol{\delta}}^k)(y) \\ (T_3\bar{\boldsymbol{\delta}}^k)(z) \end{pmatrix}$$

with an appropriately defined linear operator $\boldsymbol{T} = (T_1, T_2, T_3)$. In the sequel, we implicitly take for granted that the constraints (A.19)–(A.21) are satisfied whenever we talk about solutions of the backfitting equations (A.22). We now proceed in two steps: We first analyze the variance part $\bar{\boldsymbol{f}}^A/\boldsymbol{f}$ and then investigate the bias term $\bar{\boldsymbol{f}}^B/\boldsymbol{f}$.

To examine $\bar{\boldsymbol{f}}^A/\boldsymbol{f}$, we make us of the following lemma whose proof is provided below.

Lemma 2. *Under the conditions of Theorem 3, it holds that* $\boldsymbol{T}\hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ *uniformly on* $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$.

With the help of (A.22) and Lemma 2, we obtain that

$$\frac{\bar{\boldsymbol{f}}^A}{\boldsymbol{f}} - \hat{\boldsymbol{\mu}}^A = \big[\bar{\boldsymbol{\phi}}^A - \boldsymbol{\phi}\big] - \boldsymbol{T}\Big(\frac{\bar{\boldsymbol{f}}^A}{\boldsymbol{f}} - \hat{\boldsymbol{\mu}}^A\Big) + o_p(n^{-2/5})$$

uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$, that is,

$$\begin{pmatrix} \bar{\boldsymbol{\phi}}^A - \boldsymbol{\phi} \\ (\bar{\boldsymbol{f}}^A/\boldsymbol{f}) - \hat{\boldsymbol{\mu}}^A \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}\Big(o_p(n^{-2/5})\Big) = o_p(n^{-2/5}) \tag{A.23}$$

uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$, where the last equality follows from the fact that the operator $\mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})}$ is bounded by Lemma 1. According to (A.23), it in particular holds that

$$\frac{\bar{\boldsymbol{f}}^A}{\boldsymbol{f}} - \hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5}) \tag{A.24}$$

uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$, which implies that the asymptotic behaviour of $\bar{\boldsymbol{f}}^A/\boldsymbol{f}$ is equivalent (up to first order) to that of $\hat{\boldsymbol{\mu}}^A$. Standard calculations show that $n^{2/5}\hat{\mu}_j^A(w)$ is asymptotically normal with mean zero and variance $\sigma_j^2(w)$ for any w in the interior of $\mathcal{S}_j$ and $j = 1, 2, 3$. Hence, we obtain that $n^{2/5}\bar{f}_j^A(w)/f_j(w)$ is asymptotically normal with mean zero and variance $\sigma_j^2(w)$ as well.

We next turn to the analysis of $\bar{\boldsymbol{f}}^B/\boldsymbol{f}$. To do so, we let $\tilde{\mu}_j^B$ for $j = 1, 2, 3$ be defined as in Section 5. The following lemma whose proof is given below specifies how the terms $\hat{\boldsymbol{\mu}}^B = (\hat{\mu}_1^B, \hat{\mu}_2^B, \hat{\mu}_3^B)$ and $\tilde{\boldsymbol{\mu}}^B = (\tilde{\mu}_1^B, \tilde{\mu}_2^B, \tilde{\mu}_3^B)$ are related to each other.

Lemma 3. *For $j = 1, 2, 3$, it holds that*

$$\hat{\mu}_j^B = n^{-2/5}\tilde{\mu}_j^B + r_{j,n},$$

where $r_{j,n} = O(n^{-2/5})$ uniformly on $\mathcal{S}_j$ and $r_{j,n} = o(n^{-2/5})$ uniformly on $\mathcal{S}_j'$ with $\mathcal{S}_j' = \{w \in \mathcal{S}_j : w + th \in \mathcal{S}_j \text{ for all } t \in [-1, 1]\}$ and $h = \max\{h_1, h_2\}$.

With the help of Lemma 3, we obtain that $\boldsymbol{T}(\boldsymbol{r}_n) = o(n^{-2/5})$ uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$, where we use the notation $\boldsymbol{r}_n = (r_{1,n}, r_{2,n}, r_{3,n})$. Together with (A.22), this implies that

$$\frac{\bar{\boldsymbol{f}}^B}{\boldsymbol{f}} - \boldsymbol{r}_n = \big[\bar{\boldsymbol{\phi}}^B - \boldsymbol{\phi}\big] + n^{-2/5}\tilde{\boldsymbol{\mu}}^B - \boldsymbol{T}\Big(\frac{\bar{\boldsymbol{f}}^B}{\boldsymbol{f}} - \boldsymbol{r}_n\Big) + o(n^{-2/5}),$$

that is,

$$\begin{pmatrix} \bar{\boldsymbol{\phi}}^B - \boldsymbol{\phi} \\ (\bar{\boldsymbol{f}}^B/\boldsymbol{f}) - \boldsymbol{r}_n \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})} \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ n^{-2/5}\tilde{\boldsymbol{\mu}}^B \end{pmatrix} + o(n^{-2/5}) \tag{A.25}$$

uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$. Let $(\boldsymbol{d}, \boldsymbol{\beta})$ be the solution to the backfitting equation

$$\boldsymbol{\beta} = \boldsymbol{d} + n^{-2/5}\tilde{\boldsymbol{\mu}}^B - \boldsymbol{T}\boldsymbol{\beta}, \tag{A.26}$$

that is,

$$\begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{\beta} \end{pmatrix} = \mathcal{G}'^{-1}_{(\mathbf{0},\mathbf{0})} \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{f}_w \circ n^{-2/5}\tilde{\boldsymbol{\mu}}^B \end{pmatrix}. \tag{A.27}$$

Combining (A.25) and (A.27), we arrive at

$$\begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{\beta} \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{\phi}}^B - \boldsymbol{\phi} \\ (\bar{\boldsymbol{f}}^B/\boldsymbol{f}) - \boldsymbol{r}_n \end{pmatrix} = o(n^{-2/5}) \tag{A.28}$$

uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$. Together with Lemma 3, this implies that

$$\boldsymbol{\beta}(x,y,z) - (\bar{\boldsymbol{f}}^B/\boldsymbol{f})(x,y,z) = o(n^{-2/5})$$

for any fixed point $(x,y,z) \in \mathcal{S}_1' \times \mathcal{S}_2' \times \mathcal{S}_3'$. Hence, the asymptotic mean of $\bar{\boldsymbol{f}}^B/\boldsymbol{f}$, that is, the asymptotic bias of $(\bar{\boldsymbol{f}} - \boldsymbol{f})/\boldsymbol{f}$ is given by $\boldsymbol{\beta}$, which is the solution to the backfitting equation (A.26). This completes the proof of Theorem 3.

Proof of Lemma 2. Let $h = \max\{h_1, h_2\}$ and define $\mathcal{S}_1' = \{x \in \mathcal{S}_1 : x + th \in \mathcal{S}_1 \text{ for all } t \in [-1,1]\}$ along with $\mathcal{J}_2'(x) = \{y \in \mathcal{J}_2(x) : y + th \in \mathcal{J}_2(x) \text{ for all } t \in [-1,1]\}$. Similarly, let $\mathcal{S}_2' = \{y \in \mathcal{S}_2 : y + th \in \mathcal{S}_2 \text{ for all } t \in [-1,1]\}$ and $\mathcal{J}_1'(y) = \{x \in \mathcal{J}_1(y) : x + th \in \mathcal{J}_1(y) \text{ for all } t \in [-1,1]\}$. In addition, define

$$\mathcal{S}' = \{(x,y) \in \mathcal{S} : x \in \mathcal{S}_1' \text{ and } y \in \mathcal{J}_2'(x)\}$$

and set $\mathcal{S}_3' = \{z \in \mathcal{S}_3 : (x, z-x) \in \mathcal{S}' \text{ for some } x\}$ together with $\mathcal{J}_3'(z) = \{x \in \mathcal{J}_3(z) : (x, z-x) \in \mathcal{S}'\}$. Finally, set

$$\tilde{f}^A(x,y) = \frac{1}{n}\sum_{i=1}^n \Big\{ K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i - \mathbb{E}\big[K_{h_1}(X_i - x)K_{h_2}(Y_i - y)W_i\big] \Big\}$$

with $K_h(v) = h^{-1}K(v/h)$ and let $\tilde{\mu}_j^A$ be defined as $\hat{\mu}_j^A$ for $j = 1,2,3$ with $\hat{f}^A$ replaced by $\tilde{f}^A$.

Inspecting the definitions of $\hat{f}^A$ and $\tilde{f}^A$, it is easy to see that

$$\hat{f}^A(x,y) = \tilde{f}^A(x,y) \quad \text{for } (x,y) \in \mathcal{S}'. \tag{A.29}$$

Moreover, by standard arguments for kernel smoothers, it holds that

$$\sup_{(x,y)\in\mathcal{S}} |\hat{f}^A(x,y)| = O_p(n^{-3/10}\sqrt{\log n}) \tag{A.30}$$

$$\sup_{(x,y)\in\mathcal{S}} |\tilde{f}^A(x,y)| = O_p(n^{-3/10}\sqrt{\log n}). \tag{A.31}$$

With the help of (A.29)–(A.31), we obtain that

$$\hat{\mu}_j^A(w) = \tilde{\mu}_j^A(w) + o_p(n^{-3/10}\sqrt{\log n}) \quad \text{uniformly for } w \in \mathcal{S}_j \tag{A.32}$$

$$\hat{\mu}_j^A(w) = \tilde{\mu}_j^A(w) + o_p(n^{-1/2}\sqrt{\log n}) \quad \text{uniformly for } w \in \mathcal{S}_j' \tag{A.33}$$

for $j = 1, 2, 3$. Using (A.32)–(A.33) and noticing that the regions $\mathcal{S}_j \setminus \mathcal{S}_j'$ have Lebesgue measure of order $O(h)$, we can further infer that

$$\boldsymbol{T}(\hat{\boldsymbol{\mu}}^A - \tilde{\boldsymbol{\mu}}^A) = o_p(n^{-2/5}) \quad \text{uniformly on } \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3. \tag{A.34}$$

Standard calculations for kernel estimators yield that $\boldsymbol{T}\tilde{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$. This together with (A.34) allows us to conclude that $\boldsymbol{T}\hat{\boldsymbol{\mu}}^A = o_p(n^{-2/5})$ uniformly on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3$ as well. □

Proof of Lemma 3. It holds that $\hat{f}^B(x, y) = n^{-2/5}\tilde{f}^B(x, y) + o(n^{-2/5})$ uniformly for $(x, y) \in \mathcal{S}'$ and $\hat{f}^B(x, y) = O(n^{-2/5})$ uniformly for $(x, y) \in \mathcal{S}$. From this, we obtain that $\hat{\mu}_j^B(w) = n^{-2/5}\tilde{\mu}_j^B(w) + o(n^{-2/5})$ uniformly over $w \in \mathcal{S}_j'$ and $\hat{\mu}_j^B(w) = n^{-2/5}\tilde{\mu}_j^B(w) + O(n^{-2/5})$ uniformly over $w \in \mathcal{S}_j$ for $j = 1, 2, 3$. □

A.4 Derivation of equation (3.10)

The first order conditions of the minimization problem (3.9) are

$$\lim_{b_1,b_2\to 0} \int_{\mathcal{S}} \big[\tilde{f}_{b_1,b_2}(v, w) - \boldsymbol{a}(v, w; x, y)^\top \boldsymbol{\eta}(x, y)\big] \times \boldsymbol{a}(v, w; x, y) K\Big(\frac{v - x}{h_1}\Big) K\Big(\frac{w - y}{h_2}\Big) dvdw = \mathbf{0},$$

which gives that

$$\lim_{b_1,b_2\to 0} \int_{\mathcal{S}} \tilde{f}_{b_1,b_2}(v, w)\boldsymbol{a}(v, w; x, y) h_1^{-1} h_2^{-1} K\Big(\frac{v - x}{h_1}\Big) K\Big(\frac{w - y}{h_2}\Big) dvdw = \boldsymbol{A}(x, y)\boldsymbol{\eta}(x, y) \tag{A.35}$$

with $\boldsymbol{A}(x, y)$ defined in (3.11). Plugging the definition of the kernel density estimator $\tilde{f}_{b_1,b_2}(v, w) = (nb_1b_2)^{-1} \sum_{i=1}^n K(\frac{X_i - v}{b_1}) K(\frac{Y_i - w}{b_2}) W_i$ into (A.35), we further obtain that

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{Q}(X_i, Y_i, x, y) = \boldsymbol{A}(x, y)\boldsymbol{\eta}(x, y) \tag{A.36}$$

with

$$\boldsymbol{Q}(X_i, Y_i, x, y) = \lim_{b_1,b_2\to 0} \int_{\mathcal{S}} b_1^{-1} b_2^{-1} K\Big(\frac{X_i - v}{b_1}\Big) K\Big(\frac{Y_i - w}{b_2}\Big) \times \boldsymbol{a}(v, w; x, y) h_1^{-1} h_2^{-1} K\Big(\frac{v - x}{h_1}\Big) K\Big(\frac{w - y}{h_2}\Big) W_i dvdw.$$

Elementary arguments yield that

$$\boldsymbol{Q}(X_i, Y_i, x, y) = \lim_{b_1,b_2\to 0} \int_{\mathcal{S}} b_1^{-1} b_2^{-1} K\Big(\frac{X_i - v}{b_1}\Big) K\Big(\frac{Y_i - w}{b_2}\Big) dvdw \times \boldsymbol{a}(X_i, Y_i; x, y) h_1^{-1} h_2^{-1} K\Big(\frac{X_i - x}{h_1}\Big) K\Big(\frac{Y_i - y}{h_2}\Big) W_i.$$

Moreover, since (i) the n observations (X_i, Y_i) are interior points of $\mathcal{S}$ with probability 1 under our technical assumptions and (ii) $\int_{\mathcal{S}} b_1^{-1} b_2^{-1} K(\frac{X_i - v}{b_1}) K(\frac{Y_i - w}{b_2}) dv dw = 1$ for b_1, b_2 small enough and any interior point (X_i, Y_i) of $\mathcal{S}$, it follows that

$$\boldsymbol{Q}(X_i, Y_i, x, y) = \boldsymbol{a}(X_i, Y_i; x, y) h_1^{-1} h_2^{-1} K\Big(\frac{X_i - x}{h_1}\Big) K\Big(\frac{Y_i - y}{h_2}\Big) W_i$$

almost surely. Plugging this into (A.36), we arrive at

$$\boldsymbol{A}(x, y)\boldsymbol{\eta}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{a}(X_i, Y_i; x, y) h_1^{-1} h_2^{-1} K\Big(\frac{X_i - x}{h_1}\Big) K\Big(\frac{Y_i - y}{h_2}\Big) W_i = \boldsymbol{b}(x, y)$$

almost surely with $\boldsymbol{b}(x, y)$ defined in (3.12), which yields (3.10).

References

Aaronson, D., Lange, F. and Mazumder, B. (2014). Fertility transitions along the extensive and intensive margins. *American Economic Review* **104**, 3701-3724.

An, R. and Xiang, X. (2016). Age-period-cohort analyses of obesity prevalence in US adults. *Public Health* **141**, 163–169.

Anderson, R.W.G and Searson D.J. (2015). Use of age-period-cohort models to estimate effects of vehicle age, year of crash and year of vehicle manufacture on driver injury and fatality rates in single vehicle crashes in New South Wales, 2003–2010. *Accidents and Prevention* **75**, 202–210.

Antonczyk, D., Fitzenberger, B., Mammen, E. and Yu, K. (2018). A nonparametric approach to identify age, time, and cohort effect. *Preprint.*

Baudin, T., de la Croix, D. and Gobbi, P. (2015). Fertility and childlessness in the United States. *American Economic Review* **105**, 1852-1882.

Baum, C. L. and Ruhm, C. J. (2009). Age, socioeconomic status and obesity growth. *Journal of Health Economics* **28**, 635–648.

Bischofberger, S.M. (2020). In-Sample Hazard Forecasting Based on Survival Models with Operational Time. *Risks* **8**, 1–17.

Bischofberger, S.M., Hiabu, M. and Isakson, A. (2019a). Continuous chain-ladder with paid data. *Scandinavian Actuarial Journal*, https://doi.org/10.1080/03461238.2019.1694973.

Bischofberger, S.M., Hiabu, M., Mammen, E. and Nielsen, J.P. (2019b). A comparison of in-sample forecasting methods. *Computational Statistics and Data Analysis* **137**, 133–154.

Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in Medicine* **26**, 3018–3045.

Carstensen, B., Plummer, M., Laara, E. and Hills, M. (2018). Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.32. `https://CRAN.R-project.org/package=Epi`.

Cooley, T. and Henriksen, E. (2018). The demographic deficit. *Journal of Monetary Economics* **93**, 45-62.

Cooley, T., Henriksen, E. and Nusbaum, C. (2019). Demographic obstacles to European growth. *Preprint*.

Connor, G., Hagmann, M. and Linton, O. (2012). Efficient estimation of a semiparametric characteristic-based factor model of security returns. *Econometrica* **18**, 730–754.

Deimling, K. (1985). Nonlinear Functional Analysis. *Springer*.

Fannon, Z., Monden, C. and Nielsen, B. (2018). Age-period-cohort modelling and covariates, with an application to obesity in England 2001-2014. *Mimeo, University of Oxford*.

Fengler, M., Mammen, E. and Vogt, M. (2015). Specification and structural break tests for additive models with applications to realized variance data. *Journal of Econometrics* **188**, 196–218.

Harnau, J. (2018a). Misspecification tests for log-normal and over-dispersed Poisson chain-ladder models. *Risk* **6**, article 25.

Harnau, J. (2018b). Log-normal or over-dispersed Poisson. *Risk* **6**, article 70.

Harnau, J. and Nielsen, B. (2018). Over-dispersed age-period-cohort models. *Journal of the American Statistical Association* **113**(524), 1722-1733.

Heckman, J. and Robb, R. (1985) Using longitudinal data to estimate age, period and cohort effects in earnings equations. In *Cohort Analysis in Social Research*, Mason, W.M. and Fienberg, S.E. (eds.) New York: Springer, 137–150.

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

Holford, T.R.(1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39**, 311–324.

Holford, T.R. (1985). An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases* **38**, 831–836.

Holford, T.R. (2005). Age-period-cohort analysis. In *Encyclopedia of Biostatistics*, J. Wiley & Sons. Reissued in Wiley StatsRef: Statistics Reference Online, 2014.

Holford, T.R. (2006). Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine* **25**, 977–993.

Hvidtfeldt, U., Gerster, M., Knudsen, L. and Keiding, N. (2010). Are low Danish fertility

rates explained by changes in timing of births? *Scandinavian Journal of Public Health* **38**, 426–433.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics and Data Analysis* **51**, 4942–4956.

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.

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.

Kuang, D., Nielsen, B. and Nielsen, J. P. (2009). Chain Ladder as Maximum Likelihood Revisited. *Annals of Actuarial Science* **4**(1), 105–121.

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.

Kupper, L. L., Janis, J. M., Karmous, A. and Greenberg, B. G. (1985). Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases* **38**, 811–830.

Lee, R. D. (1993). Modeling and forecasting the time series of US fertility: age distribution, range, and ultimate level. *International Journal of Forecasting* **9** 187–202.

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B.U. (2015). Asymptotics for in-sample density forecasting. *The Annals of Statistics* **43**(2), 620–651.

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B.U. (2017). Operational time and in-sample density forecasting. *The Annals of Statistics* **45**(3), 1312–1341.

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B.U. (2018). In-sample forecasting: a brief review and new algorithms. *ALEA-Latin American Journal of Probability and Mathematical Statistics* **15**(2), 875–895.

Linton, O., Mammen, E., Nielsen, J. and Tanggaard, C. (2001). Estimating yield curves by kernel smoothing methods. *Journal of Econometrics* **105**, 185–223.

Linton, O. and Mammen, E. (2008). Nonparametric transformation to white noise. *Journal of Econometrics* **142**, 241–264.

Luo, L. (2013). Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem. *Demography* **50**, 1945–1967.

Mammen, E., Linton, O. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics* **27**, 1443–1490.

Mammen, E., Støve, B. and Tjøstheim, D. (2009). Nonparametric additive models for panels of time series. *Econometric Theory* **25**, 442–481.

Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. (2015). In-sample forecasting applied to reserving and mesothelioma mortality. *Insurance: Mathematics and Economics* **61**, 76–86.

Martínez-Miranda, M. D., Nielsen, J. P Sperlich, S. and Verrall, R. (2013). Continuous Chain Ladder: Reformulating and generalising a classical insurance problem. *Expert Systems with Applications* **40**, 5588–5603.

Martínez-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 the Royal Statistical Society A* **178**, 29–55.

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

Momota, A. (2016). Intensive and extensive margins of fertility, capital accumulation, and economic welfare. *Journal of Public Economics* **133**, 90-110.

Nielsen, B. (2015). apc: An R package for age-period-cohort analysis. *R Journal* **7**, 52–64.

Nielsen, B. (2018). apc: Age-Period-Cohort Analysis. R package version 1.4. `https://CRAN.R-project.org/package=apc`.

Nielsen, J. P. (1999). Multivariate boundary kernels from local linear estimation. *Scandinavian Actuarial Journal* **1**, 93–95.

Nielsen, B. and Nielsen, J.P. (2014). Identification and forecasting in mortality models. *The Scientific World Journal*, article ID 347043, 24 pages.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics* **25**, 186–211.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137**, 134–161.

Shang, H. L. (2019). Visualizing rate of change: an application to age-specific fertility rates. *Journal of the Royal Statistical Society: Series A* **182**, 249–262.

Yang, Y. and Land, K.C. (2006). Age-period-cohort analysis of repeated cross-section surveys. *Sociological Methodology* **36**, 297–326.

Yang, Y. and Land, K.C. (2013). Age-Period-Cohort Analysis: New Models, Methods and Empirical Applications. Boca Raton, FL: CRC Press.

Yang, Y., Fu, W.J. and Land, K.C. (2004). A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology* **34**, 75–110.

Yu, K., Park, B.U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *The Annals of Statistics* **36**, 228–260.

Wilke, R. (2018). Forecasting macroeconomic labour market flows: what can we learn from micro-level analysis? *Oxford Bulletin of Economics and Statistics* **80**, 822–842.