ALEA, Lat. Am. J. Probab. Math. Stat. **15**, 875–895 (2018)
DOI: 10.30757/ALEA.v15-33

Invited paper

In-sample forecasting: A brief review and new algorithms

Y. K. Lee, E. Mammen, J. P. Nielsen and B. U. Park

Department of Statistics, Kangwon National University,
Chuncheon 200-701, Republic of Korea
E-mail address: youngklee@kangwon.ac.kr

Institute for Applied Mathematics, Heidelberg University,
Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
E-mail address: mammen@math.uni-heidelberg.de

Cass Business School, City, University of London,
106 Bunhill Row, London EC1Y8TZ, United Kingdom
E-mail address: Jens.Nielsen.1@city.ac.uk

Department of Statistics, Seoul National University,
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
E-mail address: bupark@stats.snu.ac.kr, bupark2000@gmail.com

Abstract. Statistical methods often distinguish between in-sample and out-of-sample approaches. In particular this is the case when time is involved. Then often time series methods are proposed that extrapolate past patterns into the future via complicated recursion formulas. Standard statistical inference is on the other hand concerned with estimating parameters within the given sample. This review paper is about a statistical methodology, where all parameters are estimated in-sample while producing a forecast out-of-sample without recursion or extrapolation. A new super-simulation algorithm ensures a faster implementation of the simplest and perhaps most important version of in-sample forecasting.

Received by the editors January 10th, 2018; accepted July 18th, 2018.
2010 *Mathematics Subject Classification.* 62G07, 62G20.
Key words and phrases. Structured nonparametric models; age-cohort models; density estimation; kernel smoothing; backfitting; Chain Ladder; UK mesothelioma mortality.
Research of Young K. Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2018R1A2B6001068) and by 2016 Research Grant from Kangwon National University (520160394). Research of Jens P. Nielsen was supported by the Institute and Faculty of Actuaries, London, UK. Research of B. U. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001753).

1. Introduction

Mammen et al. (2015) and Lee et al. (2015) defined the term "in-sample forecasting" to mean forecasting a structured function in regions where the function is not observed but where it is determined by its values in the observed region. There have been many modeling approaches that connect the underlying distributions in the observed and un-observed areas via some common structure. One of the best known models of this class are perhaps age-cohort models often applied in epidemiology, biometrical and industrial forecasting. Here functions of interest depend on age effects and cohort effects that can be estimated using past observations. Outcomes for the future values of the function can be achieved by plugging in the fitted effects. Thus the age-cohort model is an in-sample forecaster because future age-cohort driven mean values are determined by age effects and cohort effects that can be estimated using available data.

There are several specifications of age-cohort models. In longevity studies, mortality rates have been modeled as products of age-effects and age-specific period trends, see e.g. Lee and Carter (1992) and Lee and Miller (2001), and see Renshaw and Haberman (2006) for an extension that also includes age-specific cohort effects. In medicine the cohort-effect can be onset of disease and the age-effect can be duration. In insurance the former can be the time of registering an insurance claim and the latter can be the duration until the claim is fully paid, see Kuang et al. (2009) among many others. It turns out that continuous age-cohort models can be formulated via something as simple as a combination of two independent stochastic variables. Let us for example assume that we have one variable X representing the start of something. It could be onset of some infection, underwriting of an insurance contract, reporting of an insurance claim, birth of a new member of a cohort or an employee losing his job in the labour market. Let then Y be a stochastic variable independent of X representing the development or delay to some event from this starting point. It could be incubation period of some disease, development of an insurance claim, age of a cohort member or time spent looking for a new job. Then, $X+Y$ is the calendar time of the relevant event. This event is observed if and only if it has already happened until a calendar time, say t_0. The forecasting exercise is about predicting the density of future events in calendar times after t_0.

In the continuous age-cohort model the forecasting density is specified in terms of the densities of X and Y. The most typical example of a structured density age-cohort model has a simple multiplicative form and has been studied by Martínez-Miranda et al. (2013) and Mammen et al. (2015). The first group of authors called it "continuous chain-ladder", because of its relation to the chain-ladder method that is omnipresent in applied non-life insurance. The chain-ladder method is an actuarial loss reserving technique that is based on the estimation of age-to-age factors using past data to predict future loss development patterns. In a continuous chain-ladder model it is assumed that X and Y have smooth densities f_1 and f_2 and are independent, leading to a multiplicative density model. When f_1 and f_2 are estimated by histograms, our in-sample forecasting approach could be formulated via a parametric model. This version of in-sample density forecasting is omnipresent in academic studies as well as in business forecasting, see Martínez-Miranda et al. (2013) for more details and references in insurance and in statistics. Extensions of such parametric histogram type of models can often be understood as structured density models modeled via histograms. However, in-sample forecasting is more

general in scope than methods based on the simple multiplicative model. Any generalised structured density or regression function that can be estimated from the past and that covers outcomes of the function in the future can be used. A generalised structured function is defined as a known function of lower-dimensional unknown underlying functions, see Mammen and Nielsen (2003) for a formal definition of generalised structured models.

Under the assumption that the model is true, our forecasts make use of the estimated lower-dimensional functions. The forecast is achieved by plugging the fits of these functions into the structured equation that is valid for the considered future date. The forecasting technique does not make use of an approximative extrapolation method that is based on Taylor expansions, for example, to get approximations for near future outcomes. And it does not use methods from time series analysis to model the further development of some random parameters. This is why the methodology is called "in-sample forecasting": a structured nonparametric estimator forecasting the future without using approximate extrapolations or time series forecasts. For letting the method work the structural assumptions are essential. The validity of these assumptions for the past can be checked by goodness-of-fit tests. For the above mentioned multiplicative density model, for example, tests can be constructed that question the multiplicative form. This can be done by omnibus tests or by tests that compare the fit with estimates in extended models. We will mention some model extensions for this model below. A more rigorous work on testing for in-sample forecasting models is still missing.

More formal description of in-sample forecasting can be found in Mammen et al. (2015) and Lee et al. (2015). We call the problem of estimating a nonparametric function f "in-sample forecasting" if it is to estimate the values of the function $f(\mathbf{z})$ for $\mathbf{z} \in \tilde{\mathcal{I}}$ only with noisy observations of $f(\mathbf{z})$ for $\mathbf{z}$ in a set $\mathcal{I}$ that is disjoint to $\tilde{\mathcal{I}}$. This makes sense under structural assumptions that identify the values of $f(\mathbf{z})$ for $\mathbf{z} \in \tilde{\mathcal{I}}$ by the values of $f(\mathbf{z})$ for $\mathbf{z} \in \mathcal{I}$. This is the case with the continuous chain-ladder model, $f(x,y) = f_1(x)f_2(y)$ with $\mathcal{I}$ being equal to the triangle $\{(x,y) : x,y \in [0,t_0], x+y \le t_0\}$ and $\tilde{\mathcal{I}} = [0,t_0]^2 \backslash \mathcal{I}$. The name "chain-ladder" probably came from the shape of the run-off triangle $\mathcal{I}$ in the discrete case. The component functions f_1 and f_2 can be estimated by observing truncated observations $(X_i, Y_i) \in \mathcal{I}$. This identifies $f(x,y)$ for $(x,y) \in \tilde{\mathcal{I}}$. We come back to this model in Section 4 with a general support set $\mathcal{I}$. There are some extensions of this model. In Lee et al. (2015) a seasonal effect is added to the model. The seasonal effect can be estimated from the past because of its recurring character. Here the density has the form $f(x,y) = f_1(x)f_2(x)f_3(\phi(x+y))$, where ϕ represents the unknown recurrent seasonal effect. On the other hand, Lee et al. (2017) assume $f(x,y) = f_1(x)f_2(y\phi(x))$, where ϕ represents an unknown effect called "operational time". This accounts for data where the speed of aging measured by Y develops in time. Other examples of structured models for in-sample forecasting include additive regression models $Y = m_1(X_1) + \cdots + m_d(X_d) + \varepsilon$, where observations for the function $m(\mathbf{x}) = m_1(x_1) + \cdots + m_d(x_d)$ are available for $\mathbf{x}$ in the support of $\mathbf{X} = (X_1, \ldots, X_d)^\top$ but the function m is identified in the larger set $S_1 \times \cdots \times S_d$ with S_j equal to the support of X_j. For further examples related to additive models, see also Mammen and Nielsen (2003) and Mammen et al. (2014). It is of course not necessary that all entering functions are nonparametric. Lee (2016) pointed out that parametrising one component could stabilise estimation and forecast at

the cost of introducing a model bias in case the semiparametric model assumption is violated.

Several algorithms have been used for the calculation of in-sample forecasters. In a number of papers the calculation has been done by solving empirical integral equations. For additive models an alternative is to use backfitting algorithms. Recently, for the continuous chain-ladder model it has been proposed to consider the problem as survival density estimation, see Hiabu et al. (2016a). They use a reversing time argument to reduce the original two-dimensional projection problem to that of estimating two one-dimensional survival densities. The present paper introduces a new algorithm for the multiplicative density in-sample forecaster, which also reduces the complexity of the problem. The new so-called super-simulation-algorithm of this paper works with any density estimator based on independent and identically distributed data. The algorithm enables a wider range of density estimation options for applied statisticians, actuaries and econometricians who use the methodology. The super-simulation-algorithm first ignores that data are not available on the entire support of the population model. Data are only available in the past, not in the future. In the first step, each component function in a structured model is estimated as if full information was available. In the second step, the super-simulation-algorithm simulates data in the no-data-region using the estimated components. In the third step, the component functions are re-estimated using both the original and simulated data, and this iteration continues until convergence. A theorem is supplied proving that this computationally tractable super-algorithm does work as intended.

There are similarities and differences between the approach based on operational time and the one that adds the period effect to the age-cohort model. Both approaches allow for some calendar time dependency. However, the operational time model is clearly an in-sample forecaster, while the period effect might be something different. For a review of age-period-cohort models in the discrete universe, see O'Brien (2014) and the many reference therein. There is an identifiability issue in age-period-cohort models, see Antonczyk et al. (2017) among many others, and one is often left with second order differences in the discrete case and second order derivatives in our continuous case when working with canonical and well-defined parametrisation, see Riebler et al. (2012), Smith and Wakefield (2016) and Beutner et al. (2017) for some further understanding on this. There is therefore a practical reason to consider operational time in-sample forecasting as an alternative to age-period-cohort models: the estimation, the identification and the forecasting with operational time are all simpler than with age-period-cohort models.

This paper is structured as follows. Section 2 is a brief overview of some of the historical reasons leading to the development of in-sample forecasting. We also discuss some of the mathematical statistics literature on complicated censoring and truncation patterns and on redistributing mass to truncated or censored areas. Section 3 discusses a data set example where the model with operational time is compared with the simple multiplicative model and it points out the double truncated nature of this data set as well as the lack of exposure data. Section 4 presents the new super-simulation-algorithm for in-sample forecasting. This is worked out for a general type of support sets including the continuous chain-ladder model. We argue that this simulation algorithm is approximately equivalent to an iterative deterministic algorithm. Section 5 contains the theoretical properties of the new

approach. Section 6 is devoted to concluding remarks, and all technical proofs are deferred to the Appendix.

2. Redistribution of un-observed mass

This review paper introduces a new interpretation of in-sample forecasting algorithms as a method of redistributing un-observed mass, which we detail in Section 4. Redistributing mass to the right is not a new idea. Efron (1967) and Dinse (1985) pointed out that an alternative interpretation of the estimator of Kaplan and Meier (1958) was to consider it as an iterative procedure redistributing mass to the right at each step. The Kaplan-Meier estimator is able to adjust for right censoring when estimating a cumulative distribution function. When adjusting for right censoring, the Efron's algorithm starts with estimating a distribution function, ignoring the censoring. It distributes the mass at the first censored time to the right of the censored point, moves to the next censored time to distribute again to the right the accumulated mass at the censored time, and continues the redistribution procedure until the last censored time. Efron proved that the final estimator of this procudure is self-consistent meaning that this estimator does not change anymore from further iterations. This type of iterative procedures have later been generalised to more complicated truncation and censoring patterns. It has been shown that redistribution procedures (now not only to the right) are closely related to the EM-algorithm and imputation methodologies, see for example Turnbull (1976). A brief review of some of the original ideas in the invention of in-sample forecasting may illustrate better why this paper analyzes the redistribution-of-mass algorithm in detail.

The early development of the in-sample forecasting idea was an example of what one today would call robotification, automatisation, machine learning or something else indicating that expensive manual procedures are being overtaken by computer intensive methodology. A particular case considered was the estimation of outstanding liabilities in non-life insurance, which is considered the most labour intensive actuarial methodology. Various statistical problems in estimating reserves in non-life insurance have been dealt in the actuarial science literature. Some notable examples include Kuang et al. (2009), Verrall et al. (2010), Martínez-Miranda et al. (2011, 2012, 2013). The first of these works established the maximum-likelihood version of the forecasting problem that turned out to be the estimation of canonical parameters in a nice smooth exponential family. Verrall et al. (2010) considered a model that takes into account the delay from when an insurance claim is incurred to when it is reported, as well as the delay from when a claim is reported to when it is fully paid. Martínez-Miranda et al. (2011) discussed the distributional properties of the method proposed in Verrall et al. (2010), and Martínez-Miranda et al. (2012) presented an extension to the model formulated by Verrall et al. (2010) and developed a new method of estimating outstanding claims. Finally, Martínez-Miranda et al. (2013) introduced how prior knowledge could be incorporated into the framework of Martínez-Miranda et al. (2012). Later, Hiabu et al. (2016b,c) showed how prior knowledge could be used most efficiently via a redistribution-to-the-right approach. The prior knowledge in non-life insurance most often comes as historical payments of already settled claims and the predicted severities of reported but un-settled claims that are based on expert opinion, see Hiabu et al.

(2016b,c). Redistributing this information to the right turned out to improve forecasts considerably. At the same time, Martínez-Miranda et al. (2013) introduced the multiplicative model $f(x, y) = f_1(x)f_2(y)$ as a continuous version of the discrete model where the classical chain-ladder method is based, which is more aligned with modern statistical analyses. This naturally led to the theoretical works, Mammen et al. (2015) and Lee et al. (2015, 2017), which developed sound theoretical backgrounds for in-sample forecasting mentioned in the introduction.

From the above brief historical outline of the development of in-sample forecasting, it is clear why redistribution-of-mass algorithms are interesting for the future. While redistribution-of-mass is equivalent to the maximum likelihood principle when the latter is relevant and the maximum likelihood is a purely statistical concept requiring observations analysed via statistical distributions, redistribution-of-mass lends to an analysis beyond mathematical statistics that incorporates prior knowledge. Malani (1995) briefly indicated how disease markers could be added to the redistribution-to-the right algorithm. Many applications of this type of methodology have been introduced later, see e.g. Chen and Zhao (2013a,b) that also add the element of estimating various health costs when complicated missing data problems are present. The latter problem seems related to the insurance cost problem mentioned above. Future research might lead to a framework incorporating both insurance costs and health care costs in the same kind of in-sample forecasting model. Redistribution-of-mass algorithms might be a key element in such developments. In the next section a data set is considered that is relevant from both the points of view of health care costs and insurance costs.

3. An illustrative forecasting example

To illustrate an in-sample forecasting exercise we present an illustrative application where we analyse an asbestos data set that is double-truncated and has year of birth as a potential covariate. The data set we consider here is on UK mesothelioma mortality. In this application a death case is only observed if it happened after the study began and before the study ended. This is a case of double truncation. Year of birth is driving the timing of the two truncations and could potentially play the role as a covariate. One study is considered where year of birth is indeed a covariate defining operational time and another study is presented where year of birth is not used as a covariate. Double truncation is related to double censoring. The latter is perhaps easier to analyze. For studies of double censoring, see Gehan (1965), Turnbull (1974), Chang and Yang (1987), Gu and Zhang (1993), Efron and Petrosian (1999) and the elegant self-consistency algorithm of Mykland and Ren (1996). Redistribution of mass to the censored areas is one elegant approach to tackling censoring and double censoring and it is closely related to all the above algorithms solving the double censoring challenges. Double truncation might be more tricky, see Moreira and de Uña Álvarez (2012), Moreira et al. (2016) and Moreira and Van Keilegom (2013) for studies of double truncation including smoothing. Our case allows for double truncation while at the same time lacking exposure data. Our problem is therefore a really complicated missing data problem. It is treated below in the estimation part of our in-sample forecasting problem. We consider forecasting of asbestos related deaths, following the earlier works of Martínez-Miranda et al. (2015); Martínez-Miranda et al. (2016) based on discrete

data. Compared to these earlier studies we consider continuous rather than a discrete modelling approach while implementing the generalised structured densities $f(x, y) = f_1(x)f_2(x\phi(y))$, where ϕ represents an unknown operational effect, see Lee et al. (2017).

The UK mesothelioma mortality data set consists of the counts of deaths caused by exposure to asbestos, given by year (1968–2012) and age (25–94) at the time of death. The total number of deaths during the period and in the range of age is 49,447. Basically, for this data set one may take the variable X to be the cohort and Y the age at death. Thus $X = (\text{year of death}) - Y$. To put the support of the data as a subset of the unit rectangle, we made the following transformation.

$$Y = \frac{(\text{AGE}) - 25 + U_2}{70}, \quad X = \frac{(\text{YEAR}) - 1968 + U_1 + 70 - 70\,Y}{70 + 45},$$

with two independent U_1 and U_2 that are uniformly distributed on $[0, 1]$. In the above transformation, the lowest possible year of birth, $1968 - 94 = 1874$ for those who died in 1968 at the age 94, is transformed to the cohort value $X = 0$ and the highest, $2012 - 25 = 1987$ for those who died in 2012 at the age 25, to $X = 1$. The support set $\mathcal{I}$ of the transformed (X, Y) is a parallelogram surrounded by the four lines represented by the equations $y = 0$, $y = 1$, $x = (70 - 70\,y)/115$ and $x = (115 - 70\,y)/115$.

The results of the application of our method to the mortality data are shown in Figures 3.1. For the result in Figure 3.1 we used the 10-fold cross-validated bandwidths described in Section 6 of Lee et al. (2017). The (operational) age component f_2 looks like we would expect and reflects an exponentially increasing mortality with age. Here, we note that f_2, sitting on the values of $\{y\phi(x) : (x, y) \in \mathcal{I}\}$, is not fully supported on the unit interval $[0, 1]$ because of the operational time ϕ and the shape of $\mathcal{I}$. The estimated density f_2 drops to zero near the end point 1.0 since the values of $Y_i\hat{\phi}(X_i)$ ranges from 0.2 to 0.97. Thus, it is due to the low density near the end point rather than due to a boundary bias. The asbestos exposure part, f_1, also looks as we would expect from earlier studies using UK import of asbestos as a surrogate for exposure, see Peto et al. (1995) for example. See also Hodgson et al. (2005), Rake et al. (2009) and Tan et al. (2010, 2011) for other recent inputs to the modeling of asbestos related death in the UK. The operational time component is increasing for the most part indicating that time was going slow at the beginning, where exposure first should take place before a long duration period towards dying of asbestos. Later asbestos exposure already took place and time to dying of asbestos is shorter. The non-increasing parts around close to the two boundaries do not represent many actual deaths. That operational time at first is slowing down might be because this is early days and people have to become old enough to be able to die. The slowing down in the later cohort might be due to heterogeneity or the advance in medical technologies. If we think of it as all the individuals having an unobserved frailty parameter, then the right boundary of operational time represents the few with very low frailty leading to a slow down in operational time at the right boundary.

One may use our estimated model to forecast the density on an unobserved area. In general, let S be a subset of $[0, 1]^2$, outside of the observed area $\mathcal{I}$, where one wants to forecast the density. With the estimated density model $\hat{f}(x, y) = \hat{f}_1(x)\hat{f}_2(y\hat{\phi}(x))$, the relative mass of the probability on S with respect to that on $\mathcal{I}$

is estimated by

$$A(S) = \int_S \hat{f}_1(x)\hat{f}_2(y\hat{\phi}(x))\,dx\,dy. \tag{3.1}$$

The number of future observations that fall in the area S is then forecasted by $N(S) = n \cdot A(S)$, where n is the sample size, i.e., the total number of observations in $\mathcal{I}$.

To apply the forecasting method to the mortality data set and evaluate its accuracy, we re-estimated the model $f(x, y) = f_1(x)f_2(y\phi(x))$, now using the data observed until the year 2010. In this case, we note that the cohort on the scale $[0, 1]$ is given by

$$X = \frac{(\text{YEAR}) - 1968 + U_1 + 70 - 70\,Y}{70 + 43}.$$

We forecasted the number of deaths for the years 2011 and 2012 according to the formula at (3.1). In this application of the formula,

$$S = \{(x, y) \in [0, 1]^2 : (112 + \alpha - 70\,y)/113 < x \le (113 + \alpha - 70\,y)/113\},$$

where $\alpha = 1$ corresponds to the year 2011 and $\alpha = 2$ to the year 2012. The actual numbers of deaths in the years 2011 and 2012 were 2,311 and 2,535, respectively. Our approach produced fairly accurate forecasting results, 2,316 and 2,465, respectively.

The forecasting based on the simple product model $f(x, y) = f_1(x)f_2(y)$, without considering the operational time, gave results that are far off the targets. The predicted counts of death were 1,721 for the year 2011 and 1,693 for the year 2012, which shows the great benefit of using the model with the operational time.

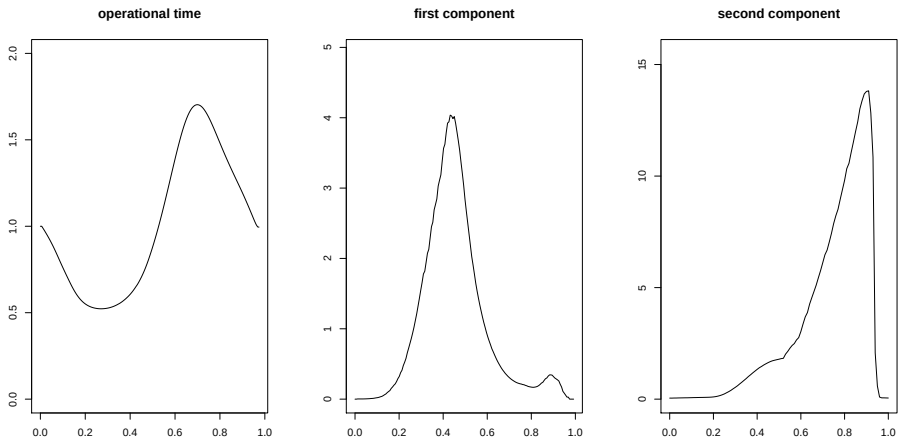

FIGURE 3.1. *Estimates of the time transformation ϕ(left), the first component function f_1 (middle) and the second component function f_2 (right) obtained by applying the model of Example 2.3 to the mortality data.*

4. A super-simulation-algorithm

This section introduces a simpler, faster and more flexible algorithm to calculate the basic estimator of the multiplicative density $f(x,y) = f_1(x)f_2(y)$. The suggested estimator of Martínez-Miranda et al. (2013) and Mammen et al. (2015) requires a two-dimensional local linear density estimator as its starting point before projecting it down on the multiplicative density model. The suggested estimator of this section is more flexible and it only uses one-dimensional density estimators. The underlying idea of the algorithm is to redistribute mass to truncated areas and it is inspired by the interpretation of the Kaplan-Meier estimator discussed in Efron (1967) and Dinse (1985) as were briefly described in Section 2. In the algorithm, at any given step in the iteration the two independent densities are estimated as if all data would be available, both the past data and the future data. Before each iteration step simulated data are added to the data representing future values. The reason this simple algorithm converges is that future data are lifted from first nothing at all to contain something and then after a few more steps to finally contain the best estimate available. Basically, the main idea of the algorithm can be applied to various problems when we want to estimate a model with truncated data. In this paper, we explore the idea for the multiplicative density model.

Let the density $f(x,y) = f_1(x)f_2(y)$ be supported on the unit rectangle $[0,1]^2$, where f_j are univariate densities supported on $[0,1]$. We wish to estimate this simple model based on truncated observations $(X_i, Y_i) \in \mathcal{I}$. We assume that the projections of $\mathcal{I}$ onto x- and y-axis equal $[0,1]$. Define the sections of $\mathcal{I}$ and $\tilde{\mathcal{I}}$ as follows.

$$I_1(y) = \{x \in [0,1] : (x,y) \in \mathcal{I}\}, \quad I_2(x) = \{y \in [0,1] : (x,y) \in \mathcal{I}\},$$
$$\tilde{I}_1(y) = \{x \in [0,1] : (x,y) \in \tilde{\mathcal{I}}\}, \quad \tilde{I}_2(x) = \{y \in [0,1] : (x,y) \in \tilde{\mathcal{I}}\}.$$

We note that $I_1(y) \cup \tilde{I}_1(y) = [0,1]$ for all $y \in [0,1]$, and $I_2(x) \cup \tilde{I}_2(x) = [0,1]$ for all $x \in [0,1]$. The main advantage of the methods we propose is that they are based only on one-dimensional estimation. Let $\hat{g}_1$ and $\hat{g}_2$ be any one-dimensional estimators of f_1 and f_2 based on the marginal observations $\{X_i : 1 \le i \le n\}$ and $\{Y_i : 1 \le i \le n\}$, respectively.

Algorithm S: Let $\mathcal{D}^{[0]} = \{(X_i, Y_i) : 1 \le i \le n\}$. Set $\hat{f}_1^{[0]} = \hat{g}_1$ and $\hat{f}_2^{[0]} = \hat{g}_2$. For $k = 0, 1, 2, \ldots$, do the following steps until convergence.

(1) Generate n_k pseudo observations in the region $\tilde{\mathcal{I}}$ according to the density $\hat{f}_1^{[k]}$ and $\hat{f}_2^{[k]}$, where

$$n_k = n \cdot \frac{\int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x)\hat{f}_2^{[k]}(y)\,dx\,dy}{\int_{\mathcal{I}} \hat{f}_1^{[k]}(x)\hat{f}_2^{[k]}(y)\,dx\,dy}. \tag{4.1}$$

(2) Add the pseudo observations in (1) to $\mathcal{D}^{[0]}$ and let $\mathcal{D}^{[k+1]}$ denote the combined data.

(3) Construct $\tilde{f}_1^{[k+1]}$ and $\tilde{f}_2^{[k+1]}$ using the marginal observations $\{X_i\}$ and $\{Y_i\}$, respectively, with $(X_i, Y_i) \in \mathcal{D}^{[k+1]}$.

(4) Repeat steps (1)-(3) L-times and denote the average values of $\tilde{f}_1^{[k+1]}$ and $\tilde{f}_2^{[k+1]}$ by $\hat{f}_1^{[k+1]}$ and $\hat{f}_2^{[k+1]}$.

We used this algorithm with $L = 1$ where it already gave reasonable results. For large values of L, the main idea of Algorithm S leads to the following mathematical formulation, which gives rise to an alternative algorithm. Put

$$A = P((X, Y) \in \mathcal{I}) = \int_{\mathcal{I}} f_1(x) f_2(y)\, dx\, dy. \tag{4.2}$$

First, we note that the pseudo observations, say (X_i^s, Y_i^s), added to $\mathcal{D}^{[0]}$, have a joint density $\tilde{f}(\cdot \,|\, \tilde{\mathcal{I}})$ defined as

$$\tilde{f}(x, y \,|\, \tilde{\mathcal{I}}) = \left[\int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy \right]^{-1} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y), \quad (x, y) \in \tilde{\mathcal{I}}. \tag{4.3}$$

Their marginal densities are given by

$$\begin{aligned}
\tilde{f}_1(x \,|\, \tilde{\mathcal{I}}) &= \left[\int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy \right]^{-1} \int_{\tilde{I}_2(x)} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dy, \quad x \in [0, 1], \\
\tilde{f}_2(y \,|\, \tilde{\mathcal{I}}) &= \left[\int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy \right]^{-1} \int_{\tilde{I}_1(y)} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx, \quad y \in [0, 1].
\end{aligned}$$

On the other hand, the marginal densities f_j restricted to the region $\mathcal{I}$, say $f_j(\cdot \,|\, \mathcal{I})$, are estimated from the original observations (X_i, Y_i) by $\tilde{f}_j(\cdot \,|\, \mathcal{I})$. This means that the marginal density estimators of f_j may be updated according to

$$\begin{aligned}
\hat{f}_1^{[k+1]}(x) &= \hat{A}^{[k]} \cdot \tilde{f}_1(x \,|\, \mathcal{I}) + (1 - \hat{A}^{[k]}) \cdot \tilde{f}_1(x \,|\, \tilde{\mathcal{I}}), \\
\hat{f}_2^{[k+1]}(y) &= \hat{A}^{[k]} \cdot \tilde{f}_2(y \,|\, \mathcal{I}) + (1 - \hat{A}^{[k]}) \cdot \tilde{f}_2(y \,|\, \tilde{\mathcal{I}}),
\end{aligned} \tag{4.4}$$

where $\hat{A}^{[k]}$ is an estimator of A defined at (4.2) in the kth update. Taking

$$\hat{A}^{[k]} = \int_{\mathcal{I}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy = \frac{n}{n + n_k}$$

gives

$$\begin{aligned}
\hat{f}_1^{[k+1]}(x) &= \left(\int_{\mathcal{I}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy \right) \cdot \tilde{f}_1(x \,|\, \mathcal{I}) + \int_{\tilde{I}_2(x)} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dy, \\
\hat{f}_2^{[k+1]}(y) &= \left(\int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy \right) \cdot \tilde{f}_2(y \,|\, \mathcal{I}) + \int_{\tilde{I}_1(y)} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx.
\end{aligned} \tag{4.5}$$

It is worthwhile to note that, if $\tilde{f}_j(\cdot \,|\, \mathcal{I})$ are densities, i.e., $\tilde{f}_j(\cdot \,|\, \mathcal{I}) \geq 0$ and $\int_0^1 \tilde{f}_j(u \,|\, \mathcal{I})\, du = 1$, then are all the updates $\hat{f}_j^{[k+1]}$ for $k \geq 0$ as well. This is clear since

$$\int_{\mathcal{I}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy + \int_{\tilde{\mathcal{I}}} \hat{f}_1^{[k]}(x) \hat{f}_2^{[k]}(y)\, dx\, dy = 1.$$

The algorithm (4.5) is basically equivalent to Algorithm S. The only difference is that Algorithm S actually generates pseudo data from $\tilde{f}(\cdot \,|\, \tilde{\mathcal{I}})$ at (4.3), while (4.5) uses $\tilde{f}(\cdot \,|\, \tilde{\mathcal{I}})$ directly, to update the estimators of the marginal densities $f_j = A f_j(\cdot \,|\, \mathcal{I}) + (1 - A) f_j(\cdot \,|\, \tilde{\mathcal{I}})$. Contrary to Algorithm S, the algorithm (4.5) does not require simulating pseudo data in $\tilde{\mathcal{I}}$. In our theoretical development to be presented in the next section, we focus on the latter.

Define

$$f_{w,1}(x) = \int_{I_2(x)} f_1(x) f_2(v)\, dv, \quad f_{w,2}(y) = \int_{I_1(y)} f_1(u) f_2(y)\, du.$$

With these definitions we note that $A = \int_0^1 f_{w,1}(x)\,dx = \int_0^1 f_{w,2}(y)\,dy$. For notational convenience we set $g_j = f_j(\cdot\,|\,\mathcal{I})$ and $\hat{g}_j = \tilde{f}_j(\cdot\,|\,\mathcal{I})$. Thus,

$$g_1(x) = A^{-1} \cdot f_{w,1}(x), \quad g_2(y) = A^{-1} \cdot f_{w,2}(y), \tag{4.6}$$

The estimating equation for $\hat{f}_j$ as estimators of f_j that corresponds to the iteration scheme (4.5) is given by

$$\begin{aligned} \hat{f}_1(x) &= \left(\int_{\mathcal{I}} \hat{f}_1(x)\hat{f}_2(y)\,dx\,dy \right) \cdot \hat{g}_1(x) + \int_{\tilde{I}_2(x)} \hat{f}_1(x)\hat{f}_2(y)\,dy, \\ \hat{f}_2(y) &= \left(\int_{\mathcal{I}} \hat{f}_1(x)\hat{f}_2(y)\,dx\,dy \right) \cdot \hat{g}_2(y) + \int_{\tilde{I}_1(y)} \hat{f}_1(x)\hat{f}_2(y)\,dx. \end{aligned} \tag{4.7}$$

The population version of the estimating equation (4.7) is then

$$\begin{aligned} f_1(x) &= \left(\int_{\mathcal{I}} f_1(x)f_2(y)\,dx\,dy \right) \cdot g_1(x) + \int_{\tilde{I}_2(x)} f_1(x)f_2(y)\,dy, \\ f_2(y) &= \left(\int_{\mathcal{I}} f_1(x)f_2(y)\,dx\,dy \right) \cdot g_2(y) + \int_{\tilde{I}_1(y)} f_1(x)f_2(y)\,dx, \end{aligned} \tag{4.8}$$

which is clearly satisfied by the true component functions f_j. For a triangular set $\mathcal{I} = \{(x,y) : 0 \le x, y \le 1, x + y \le 1\}$ a version of the estimator defined by (4.7) based on a two-dimensional density estimator has been discussed in Mammen et al. (2015). In this paper the backfitting algorithm (4.5) has been used for the calculation of the estimators.

We close this section by reporting a brief simulation result for the new estimators defined through the equations (4.7). We took $\{(x,y) \in [0,1]^2 : 0 \le x + y \le 1\}$ for the support set $\mathcal{I}$. For the marginal density functions, we set $f_1(x) = (3/2) - x$ and $f_2(y) = (5/4) - (3/4)y^2$. We generated 100 pseudo samples (X_i, Y_i) of sizes $n = 400$ and $1,000$ from the joint density $p(x,y) = A^{-1} f_1(x) f_2(y) I((x,y) \in \mathcal{I})$. We used the local linear estimators for $\hat{g}_j$ in (4.7) as defined below. For a baseline kernel function K and a bandwidth $h > 0$, let $K_h(v) = K(v/h)/h$ and define

$$\mathbf{A}_j(u) = \int_0^1 \begin{pmatrix} 1 & (v-u)/h_j \\ (v-u)/h_j & (v-u)^2/h_j^2 \end{pmatrix} K_{h_j}(v-u)\,dv. \tag{4.9}$$

Also, define

$$\hat{\mathbf{b}}_j(u) = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ (W_i^{(j)} - u)/h_j \end{pmatrix} K_{h_j}(W_i^{(j)} - u), \tag{4.10}$$

where $W_i^{(j)} = X_i$ for $j = 1$ and Y_i for $j = 2$. Our estimators $\hat{g}_j$ of g_j are the first entries of the vectors $\mathbf{A}_j^{-1}\hat{\mathbf{b}}_j$. We chose the Epanechnikov kernel $K(u) = (3/4)(1-u^2)I_{[-1,1]}(u)$.

	new algorithm		old algorithm	
	f_1	f_2	f_1	f_2
$n = 400$	0.01279	0.01195	0.01902	0.00579
$n = 1000$	0.00946	0.00746	0.01870	0.00523

TABLE 4.1. Mean integrated squared error (MISE) of the component function estimators $\hat{f}_j$ based on 100 MC samples of sizes $n = 400$ and $n = 1,000$.

Table 1 shows the mean integrated squared errors of the new estimators $\hat{f}_j$ based on the algorithm (4.5) and those based on the two-dimensional local linear density estimator as in Mammen et al. (2015). The numbers in the table are the mean integrated squared errors for the optimal bandwidth choices $h = h_1 = h_2$ that gave the minimal $\mathrm{MISE}(\hat{f}_1) + \mathrm{MISE}(\hat{f}_2)$ in a range of preselected bandwidths. The optimal bandwidths for the new algorithm were different from those for the old algorithm. Also, these common optimal bandwidths may be better for one component, but worse for the other. Indeed, the MISE results indicate that the new algorithm is better for estimating f_1, but not for f_2. If we compare the sum when $n = 400$, the value of $\mathrm{MISE}(\hat{f}_1) + \mathrm{MISE}(\hat{f}_2)$ equals 0.02474 for the new algorithm and 0.02481 for the old, so there is not much difference between them. However, when $n = 1000$, it equals 0.01692 for the new and 0.02393 for the old.

Figure 4.2 depicts the distributions of the computing times in seconds for the two methods. These results strongly suggest that the new algorithm outperforms the old one in terms of computing time. There appears a bimodal structure in the distribution of computing times for both the new and old methods. This is clearer for the new algorithm but there is some evidence for the old as well. We note that the area around the first mode for the old algorithm is dominating that around the second mode. We found that the mass around the second mode was roughly 7%, while it was around 16% in the case of the new algorithm. We also found that both algorithms converged in 4-7 iterations. This means that the elapsed time for a single iteration was roughly 0.02 sec for the new algorithm, and roughly 20 sec for the old. Thus, the gap between the first and second modes in the case of the new algorithm is just a matter of one or two additional iterations, while the gap in the case of the old algorithm seems mostly due to a small fluctuation in computing the two-dimensional local linear density estimates.

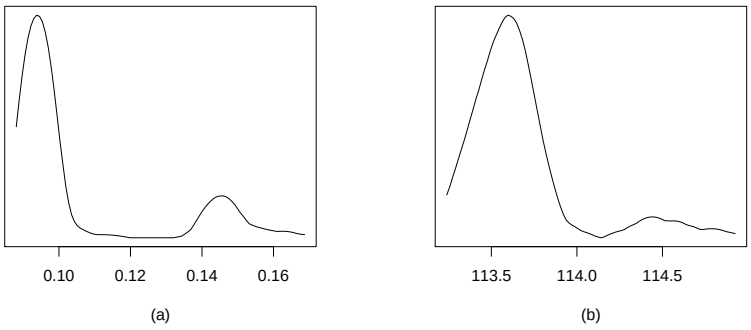

FIGURE 4.2. *The distributions of the times (in seconds) for computing the component estimates per sample, based on 100 MC samples of size* $n = 1,000$. *Panel (a) is for the new algorithm and (b) for the old one.*

5. Asymptotic theory

We study the statistical properties of the estimators $\hat{f}_j$ that satisfy the system of equations at (4.7). The theory developed in this section generalises the results of Mammen et al. (2015) to a general class of sets $\mathcal{I}$. In the latter paper only triangular sets $\mathcal{I} = \{(x,y)^\top : 0 \le x, y \le 1, x+y \le 1\}$ have been discussed. We also consider now estimators of f_1 and f_2 based on the local linear estimators of g_1 and g_2, defined after the equations (4.9) and (4.10).

Basically we make the following assumptions.

(A1) The marginal densities f_j are continuous and bounded away from zero and infinity on $[0,1]$;

(A2) The projections of $\mathcal{I}$ onto x- and y-axis equal $[0,1]$. Also, there exist sequences $x_0 = 0 < x_1 < ... < x_k = 1$ and $y_0 = 1 > y_1 > ... > y_k = 0$ with $(x, y_j) \in \mathcal{I}$ for $x_j \le x \le x_{j+1}$, or with $(x_j, y) \in \mathcal{I}$ for $y_j \le y \le y_{j+1}$, for all $0 \le j \le k-1$.

The condition (A2) means that the support set $\mathcal{I}$ contains a ladder that traverses the entire interval $[0,1]$ along the x- or y-axis.

We write $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2)^\top$ and $\mathbf{f} = (f_1, f_2)^\top$. Let $\mathcal{F}$ denote the class of tuples of univariate functions $\boldsymbol{\eta} \equiv (\eta_1, \eta_2)^\top$ such that η_j are nonnegative, continuous on $[0,1]$ and $\int_0^1 \eta_j = 1$. Also, let $\mathcal{F}_0$ be the class of tuples of univariate functions such that η_j are continuous on $[0,1]$ and $\int_0^1 \eta_j = 0$. Define $\mathbf{F}(\boldsymbol{\eta}) = (F_1(\boldsymbol{\eta}), F_2(\boldsymbol{\eta}))^\top$ as a map from $\mathcal{F}$ to $\mathcal{F}_0$ by

$$F_1(\boldsymbol{\eta})(x) = \eta_1(x) - \left(\int_{\mathcal{I}} \eta_1(u)\eta_2(v)\, du\, dv \right) \cdot g_1(x) - \int_{\tilde{I}_2(x)} \eta_1(x)\eta_2(v)\, dv,$$
$$F_2(\boldsymbol{\eta})(y) = \eta_2(y) - \left(\int_{\mathcal{I}} \eta_1(u)\eta_2(v)\, du\, dv \right) \cdot g_2(y) - \int_{\tilde{I}_1(y)} \eta_1(u)\eta_2(y)\, du.$$

Likewise, define $\hat{\mathbf{F}}$ with $\hat{g}_j$ replacing g_j, respectively. We note that

$$\mathbf{F}(\mathbf{f}) = \mathbf{0}, \quad \hat{\mathbf{F}}(\hat{\mathbf{f}}) = \mathbf{0}. \tag{5.1}$$

Now, define

$$\mathbf{G}(\boldsymbol{\eta}) = \mathbf{F}(\mathbf{f} \cdot (\mathbf{1} + \boldsymbol{\eta})), \quad \hat{\mathbf{G}}(\boldsymbol{\eta}) = \hat{\mathbf{F}}(\mathbf{f} \cdot (\mathbf{1} + \boldsymbol{\eta})),$$

where, for vectors $\mathbf{a} = (a_1, a_2)^\top$ and $\mathbf{b} = (b_1, b_2)^\top$, we write $\mathbf{a} \cdot \mathbf{b}$ for $(a_1 b_1, a_2 b_2)^\top$, $\mathbf{a}/\mathbf{b}$ for $(a_1/b_1, a_2/b_2)^\top$ and $\mathbf{a} \pm \mathbf{b}$ for $(a_1 \pm b_1, a_2 \pm b_2)^\top$. Both $\mathbf{G}$ and $\hat{\mathbf{G}}$ map $\mathcal{S}$ to $\mathcal{F}_0$, where

$$\mathcal{S} = \{\boldsymbol{\eta} \in C[0,1] \times C[0,1] : \int_0^1 \eta_j\, f_j = 0\}.$$

Then, the two equations at (5.1), respectively, are equivalent to

$$\mathbf{G}(\mathbf{0}) = \mathbf{0}, \quad \hat{\mathbf{G}}((\hat{\mathbf{f}} - \mathbf{f})/\mathbf{f}) = \mathbf{0}. \tag{5.2}$$

Both the maps $\mathbf{G}$ and $\hat{\mathbf{G}}$ are nonlinear. To analyze $\hat{\boldsymbol{\delta}} := (\hat{\mathbf{f}} - \mathbf{f})/\mathbf{f}$ as the solution of the second equation at (5.2), we consider the linear approximation of $\hat{\mathbf{G}}$ based on its Fréchet derivative. Let $\hat{\mathbf{G}}'(\mathbf{0}) : \mathcal{S} \to \mathcal{F}_0$ denote the Fréchet derivative of $\hat{\mathbf{G}}$

at $\mathbf{0}$. It is given by

$$\begin{aligned}\hat{\mathbf{G}}'(\mathbf{0})_1(\boldsymbol{\delta})(x) &= \delta_1(x)f_1(x) - \hat{g}_1(x)\cdot\int_{\mathcal{I}}[\delta_1(u)+\delta_2(v)]\,f_1(u)f_2(v)\,du\,dv \\ &\quad - \int_{\tilde{I}_2(x)}[\delta_1(x)+\delta_2(v)]\,f_1(x)f_2(v)\,dv, \\ \hat{\mathbf{G}}'(\mathbf{0})_2(\boldsymbol{\delta})(y) &= \delta_2(y)f_2(y) - \hat{g}_2(y)\cdot\int_{\mathcal{I}}[\delta_1(u)+\delta_2(v)]\,f_1(u)f_2(v)\,du\,dv \\ &\quad - \int_{\tilde{I}_1(y)}[\delta_1(u)+\delta_2(y)]\,f_1(u)f_2(y)\,du.\end{aligned} \tag{5.3}$$

Similarly, we get $\mathbf{G}'(\mathbf{0})(\boldsymbol{\delta})$ by simply replacing $\hat{g}_j$ by g_j in the expression for $\hat{\mathbf{G}}'(\mathbf{0})$. We note that with $\hat{\boldsymbol{\delta}} = (\hat{\mathbf{f}} - \mathbf{f})/\mathbf{f}$

$$\begin{aligned}\mathbf{0} = \hat{\mathbf{G}}(\hat{\boldsymbol{\delta}}) &\simeq \hat{\mathbf{G}}(\mathbf{0}) + \hat{\mathbf{G}}'(\mathbf{0})(\hat{\boldsymbol{\delta}}) \\ &\simeq \hat{\mathbf{G}}(\mathbf{0}) + \mathbf{G}'(\mathbf{0})(\hat{\boldsymbol{\delta}}) \\ &= \hat{\mathbf{G}}(\mathbf{0}) - \mathbf{G}(\mathbf{0}) + \mathbf{G}'(\mathbf{0})(\hat{\boldsymbol{\delta}}).\end{aligned} \tag{5.4}$$

Recall the definitions of A at (4.2) and g_j at (4.6). We get

$$-\left[\hat{\mathbf{G}}(\mathbf{0}) - \mathbf{G}(\mathbf{0})\right] = A\cdot(\hat{\mathbf{g}} - \mathbf{g}).$$

The approximation (5.4) motivates an approximation of $\hat{\mathbf{f}}$, which is easier to analyze. Define $\bar{\mathbf{f}} = (\bar{f}_1, \bar{f}_2)^\top$ by

$$\mathbf{G}'(\mathbf{0})(\bar{\boldsymbol{\delta}}) = A\cdot(\hat{\mathbf{g}} - \mathbf{g}), \tag{5.5}$$

where $\bar{\boldsymbol{\delta}} = (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}$. Our first theorem demonstrates that $\mathbf{G}'(\mathbf{0})$ is invertible, so that $\bar{\boldsymbol{\delta}}$ and thus $\bar{\mathbf{f}}$ are well-defined. The theorem requires an additional assumption.

(A3) For $j = 1$ and 2, $\mathrm{mes}(I_j(u)) > 0$ except for a finite number of points $u \in [0,1]$.

Before stating the theorem, we introduce some terminologies that are used throughout this section. Note that g_1 and thus $f_{w,1}$ equal zero only at points u such that $\mathrm{mes}(I_2(u)) = 0$ due to assumption (A1), where $\mathrm{mes}(I)$ for a set I denotes its Lebesgue measure. Similarly, g_2 and thus $f_{w,2}$ equal zero only at points u such that $\mathrm{mes}(I_1(u)) = 0$. Define

$$I_1^o = \{x \in [0,1] : \mathrm{mes}(I_2(x)) > 0\}, \quad I_2^o = \{y \in [0,1] : \mathrm{mes}(I_1(u)) > 0\}.$$

In the case of the run-off triangular support $\mathcal{I} = \{(x,y) \in [0,1]^2 : 0 \le x+y \le 1\}$, which is typical for insurance claim data, $I_j^o = [0,1)$ for $j = 1$ and 2. In the case of the asbestos data example, $\mathcal{I}$ is a parallelogram such that $\mathcal{I} = \{(x,y) \in [0,1]^2 : -ax + 1 \le y \le -ax + a\}$ for some $a > 1$. In this case $I_1^o = (0,1)$ but $I_2^o = [0,1]$.

Theorem 5.1. *Assume the conditions (A1)–(A3). Then, the linear operator* $\mathbf{G}'(\mathbf{0}) : \mathcal{S} \to \mathcal{F}_0$ *is invertible.*

Our next theorem demonstrates that $\hat{\boldsymbol{\delta}} = (\hat{\mathbf{f}} - \mathbf{f})/\mathbf{f}$ is well approximated by $\bar{\boldsymbol{\delta}} = (\bar{\mathbf{f}} - \mathbf{f})/\mathbf{f}$ defined to be the solution of the equation (5.5).

Theorem 5.2. *Assume that the conditions of Theorem 5.1 hold. Furthermore, assume that f_j are continuously differentiable. Suppose that* $\sup_{u\in[0,1]} \left|\hat{g}_j(u) - g_j(u)\right| = O_p(\varepsilon_n)$, $j = 1, 2$, *for some sequence of real numbers* $\varepsilon_n \to 0$. *Then,*

$$\sup_{x\in[0,1]} \left|\hat{f}_j(x) - \bar{f}_j(x)\right| = O_p(\varepsilon_n^2), \quad j = 1, 2.$$

Next, we discuss the asymptotic distribution of $\hat{f}_j$. For $\hat{g}_j$, we consider the local linear estimators that are the first entries of the vectors $\mathbf{A}_j^{-1}\hat{\mathbf{b}}_j$, respectively, where $\mathbf{A}_j$ and $\hat{\mathbf{b}}_j$ are defined at (4.9) and (4.10), respectively. From the standard theory of local linear kernel smoothing, it holds that $\sup_{u\in[0,1]} \left|\hat{g}_j(u) - g_j(u)\right| = O_p(n^{-2/5}\sqrt{\log n})$ with $h_j \sim n^{-1/5}$.

To state the theorem, define

$$\tilde{g}_j^B(u) = \frac{1}{2}\left(\int u^2 K\right) c_j^2\, g_j''(u),$$

where $c_j \asymp n^{1/5}h_j$. Also, define $\boldsymbol{\beta} \in \mathcal{S}$ to be the solution of $\mathbf{G}'(\mathbf{0})(\boldsymbol{\beta}) = A \cdot \tilde{\mathbf{g}}^B$. Let $\sigma_1^2(x) = c_1^{-1} \int K^2/g_1(x)$ and $\sigma_2^2(y) = c_2^{-1} \int K^2/g_2(y)$.

Theorem 5.3. *Assume that the conditions of Theorem 5.1 hold. Furthermore, assume that f_j are twice continuously differentiable, that K is supported on $[-1, 1]$, symmetric and Lipschitz continuous, and that $n^{1/5}h_j \to c_j$ for some $0 < c_j < \infty$. Let x and y be fixed points in $I_1^o \cap (0, 1)$ and $I_2^o \cap (0, 1)$. Then, $n^{2/5}(\hat{f}_1(x) - f_1(x))/f_1(x)$ and $n^{2/5}(\hat{f}_2(y) - f_2(y))/f_2(y)$, respectively, converges to $N(\beta_1(x), \sigma_1^2(x))$ and $N(\beta_2(y), \sigma_2^2(y))$. Furthermore, $n^{2/5}(\hat{f}_1(x) - f_1(x))/f_1(x)$ and $n^{2/5}(\hat{f}_2(y) - f_2(y))/f_2(y)$ are asymptotically independent.*

6. Concluding remarks

In-sample forecasting, as reviewed in this paper, is a recent generalisation of a long list of practitioner methods - often based on discrete histogram type of methodology - to a modern structured nonparametric smoothing approach. The term in-sample forecasting is new and adds one more method in one single concept to our toolbox of forecasting procedures. The two other major methods of forecasting are time series forecasting and simple deterministic extrapolation. We believe that the stability of in-sample forecasting, that do not extrapolate any parameters used for forecasting, will serve as a useful alternative to the often less stable time series methodology. Clearly, however, in-sample forecasting cannot model everything, and sometimes there really is a time series to be estimated and forecasted. One can even imagine a latent time series to be present within an in-sample forecasting study, for example a calendar effect. This time series is best dealt with the traditional time series methodology. One can therefore imagine a future blend of in-sample forecasting and time series methodology and using the best from both worlds of forecasting to provide an output that is stable as well as flexible. One can also imagine much more flexible in-sample forecasting methods in the future involving many more structured nonparametric models than those considered in this paper. This in turn asks for the development of a general statistical testing rules to pick among the available model options.

Appendix

A.1. *Proof of Theorem 5.1.* Let $\boldsymbol{\delta} \in \mathcal{S}$. Then, $\int_0^1 \delta_j\, f_j = 0$. From the versions of the two equations at (5.3) for $\mathbf{G}'(\mathbf{0})$, we get that

$$\begin{aligned}
\frac{\mathbf{G}'(\mathbf{0})_1(\boldsymbol{\delta})(x)}{f_{w,1}(x)} &= \delta_1(x) + \int_{I_2(x)} \delta_2(v) \frac{f_1(x) f_2(v)}{f_{w,1}(x)}\, dv \\
&\quad - \int_0^1 \delta_1(u) g_1(u)\, du - \int_0^1 \delta_2(v) g_2(v)\, dv, \quad x \in I_1^0, \\
\frac{\mathbf{G}'(\mathbf{0})_2(\boldsymbol{\delta})(y)}{f_{w,2}(y)} &= \delta_2(y) + \int_{I_1(y)} \delta_1(u) \frac{f_1(u) f_2(y)}{f_{w,2}(y)}\, du \\
&\quad - \int_0^1 \delta_1(u) g_1(u)\, du - \int_0^1 \delta_2(v) g_2(v)\, dv, \quad y \in I_2^o.
\end{aligned} \tag{A.1}$$

Due to assumption (A3), for each of the two functions $\mathbf{G}'(\mathbf{0})_j(\boldsymbol{\delta})/f_{w,j}$ there exists a unique function that is continuous on the whole interval $[0,1]$ and coincides with $\mathbf{G}'(\mathbf{0})_j(\boldsymbol{\delta})/f_{w,j}$ on I_j^o. We continue to denote the extended continuous functions by $\mathbf{G}'(\mathbf{0})_j(\boldsymbol{\delta})/f_{w,j}$. Let

$$\mathcal{S}' = \{\boldsymbol{\eta} \in C[0,1] \times C[0,1] : \int_0^1 \eta_j\, g_j = 0\}.$$

Then, writing $\mathbf{f}_w = (f_{w,1}, f_{w,2})^\top$, it holds that $\mathbf{G}'(\mathbf{0})_1(\boldsymbol{\delta})/\mathbf{f}_w \in \mathcal{S}'$ for all $\boldsymbol{\delta} \in \mathcal{S}$.

We now consider a linear operator $\mathbf{L} : \mathcal{S} \to \mathcal{S}'$ such that $\mathbf{L}(\boldsymbol{\eta}) = (L_1(\eta_1), L_2(\eta_2))^\top$ and

$$L_j(\eta_j) = \eta_j - \int_0^1 \eta_j\, g_j, \quad j = 1, 2.$$

Its inverse map $\mathbf{L}^{-1} : \mathcal{S}' \to \mathcal{S}$ is given by $\mathbf{L}^{-1}(\boldsymbol{\eta}) = (L_1^{-1}(\eta_1), L_2^{-1}(\eta_2))^\top$ with

$$L_j^{-1}(\eta_j) = \eta_j - \int_0^1 \eta_j\, f_j, \quad j = 1, 2.$$

We define $\mathbf{T} : \mathcal{S}' \to \mathcal{S}'$ by

$$\mathbf{T}(\boldsymbol{\eta}) = \frac{\mathbf{G}'(\mathbf{0})(\mathbf{L}^{-1}(\boldsymbol{\eta}))}{\mathbf{f}_w}. \tag{A.2}$$

Then, it follows from (A.1) that

$$\begin{aligned}
T_1(\boldsymbol{\eta})(x) &= \eta_1(x) + \int_{I_2(x)} \eta_2(v) \frac{f_1(x) f_2(v)}{f_{w,1}(x)}\, dv, \\
T_2(\boldsymbol{\eta})(y) &= \eta_2(y) + \int_{I_1(y)} \eta_1(u) \frac{f_1(u) f_2(y)}{f_{w,2}(y)}\, du.
\end{aligned} \tag{A.3}$$

From the definition of the map $\mathbf{T}$ at (A.2), the invertibility of $\mathbf{G}'(\mathbf{0})$ is equivalent to the invertibility of $\mathbf{T}$. We prove $\mathbf{T}$ is invertible. We endow $\mathcal{S}'$ with an inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle \boldsymbol{\eta}, \boldsymbol{\delta} \rangle = \int_{[0,1]^2} \boldsymbol{\eta}(x,y)^\top \begin{pmatrix} f_{w,1}(x) & 0 \\ 0 & f_{w,2}(y) \end{pmatrix} \boldsymbol{\delta}(x,y)\, dx\, dy.$$

Suppose that $\mathbf{T}(\boldsymbol{\eta}) = \mathbf{0}$ for some $\boldsymbol{\eta} \in \mathcal{S}'$. Then, it holds that

$$0 = \langle \boldsymbol{\eta}, \mathbf{T}(\boldsymbol{\eta}) \rangle = \int_{\mathcal{I}} [\eta_1(x) + \eta_2(y)]^2\, f_1(x) f_2(y)\, dx\, dy. \tag{A.4}$$

This implies that

$$\eta_1(x) + \eta_2(y) = 0 \quad \text{for all } (x, y) \in \mathcal{I}. \tag{A.5}$$

Because of assumption (A2), either η_1 or η_2 is piecewise constant, so that either η_1 or η_2 is a zero function since η_j are continuous and satisfy $\int_0^1 \eta_j\, g_j = 0$. This implies that both are zero functions. This proves that $\mathbf{T}$ is one-to-one.

Now, we prove $\mathbf{T}$ is onto. Similarly as in deriving (A.4), we get that, for any $\boldsymbol{\eta}, \boldsymbol{\delta} \in \mathcal{S}'$,

$$\begin{aligned} \langle \boldsymbol{\eta}, \mathbf{T}(\boldsymbol{\delta}) \rangle &= \int_{\mathcal{I}} [\eta_1(x) + \eta_2(y)]\, [\delta_1(x) + \delta_2(y)]\, f_1(x) f_2(y)\, dx\, dy \\ &= \langle \mathbf{T}(\boldsymbol{\eta}), \boldsymbol{\delta} \rangle. \end{aligned} \tag{A.6}$$

This implies that $\mathbf{T}$ is self-adjoint, so that $\text{Image}(\mathbf{T})^{\perp} = \text{Null}(\mathbf{T}) = \{\mathbf{0}\}$. It suffices to show that $\text{Image}(\mathbf{T})$ is closed. Suppose that $\{\boldsymbol{\delta}_n\} \subset \mathcal{S}'$ and $\mathbf{T}(\boldsymbol{\delta}_n) \to \boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \mathcal{S}'$. We prove that $\boldsymbol{\eta} \in \text{Image}(\mathbf{T})$. Note that $\mathbf{T} : \mathcal{S}' \to \text{Image}(\mathbf{T})$ is invertible. Its inverse denoted by $\mathbf{T}^{-1} : \text{Image}(\mathbf{T}) \to \mathcal{S}'$ is also linear and continuous due to Banach Inverse Theorem. Thus, $\boldsymbol{\delta}_n = \mathbf{T}^{-1}(\mathbf{T}(\boldsymbol{\delta}_n))$ is Cauchy in $\mathcal{S}'$ so that there exists $\boldsymbol{\delta} \in \mathcal{S}'$ such that $\boldsymbol{\delta}_n \to \boldsymbol{\delta}$. Now,

$$\mathbf{T}(\boldsymbol{\delta}) = \mathbf{T}\Big(\lim_{n \to n} \boldsymbol{\delta} \Big) = \lim_{n \to n} \mathbf{T}(\boldsymbol{\delta}_n) = \boldsymbol{\eta}.$$

This completes the proof of the invertibility of $\mathbf{T}$, and thus of $\mathbf{G}'(\mathbf{0})$.

A.2. *Proof of Theorem 5.2.* Clearly from the expression of $\hat{\mathbf{G}}'(\mathbf{0})$ in (5.3) it follows that there exists a constant $0 < C_1 < \infty$ such that

$$\big\| \hat{\mathbf{G}}'(\mathbf{0})(\boldsymbol{\delta}) - \mathbf{G}'(\mathbf{0})(\boldsymbol{\delta}) \big\|_{\infty} \le C_1 \cdot \|\hat{\mathbf{g}} - \mathbf{g}\|_{\infty} \cdot \|\boldsymbol{\delta}\|_{\infty},$$

where $\|\boldsymbol{\eta}\|_{\infty} = \sup_{x \in [0,1]} |\eta_1(x)| + \sup_{y \in [0,1]} |\eta_2(y)|$. Thus,

$$\sup_{\|\boldsymbol{\delta}\|_{\infty} = 1} \big\| \hat{\mathbf{G}}'(\mathbf{0})(\boldsymbol{\delta}) - \mathbf{G}'(\mathbf{0})(\boldsymbol{\delta}) \big\|_{\infty} = O_p(\varepsilon_n). \tag{A.7}$$

We may also prove that there exist constants $0 < r, C_2 < \infty$ such that, with probability tending to one,

$$\sup_{\|\boldsymbol{\delta}\|_{\infty} = 1} \big\| \hat{\mathbf{G}}'(\boldsymbol{\eta}_1)(\boldsymbol{\delta}) - \hat{\mathbf{G}}'(\boldsymbol{\eta}_2)(\boldsymbol{\delta}) \big\|_{\infty} \le C_2 \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\|_{\infty} \quad \text{for all } \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in B_r(\mathbf{0}), \tag{A.8}$$

where $B_r(\mathbf{0})$ denotes the ball in $\mathcal{S}$ with radius r centered at $\mathbf{0}$. Along the lines of the proof of Theorem 3 in Lee et al. (2015) with (A.7) and (A.8), we may prove the theorem.

A.3. *Proof of Theorem 5.3.* Due to Theorem 5.2, the asymptotic distributions of $\hat{f}_j$ are determined by $\bar{f}_j$. For a given $\epsilon > 0$, put

$$\begin{aligned} I_1^{\epsilon} &= \{x \in [0,1] : \text{mes}(I_2(x)) \ge \epsilon, \inf_{y \in I_2(x)} \text{mes}(I_1(y)) \ge \epsilon\}, \\ I_2^{\epsilon} &= \{y \in [0,1] : \text{mes}(I_1(y)) \ge \epsilon, \inf_{x \in I_1(y)} \text{mes}(I_2(x)) \ge \epsilon\}. \end{aligned}$$

We note that for the triangular support $I_1^{\epsilon} = I_2^{\epsilon} = [\epsilon, 1-\epsilon]$, and for the parallelogram support $I_1^{\epsilon} = [c \cdot \epsilon, 1 - c \cdot \epsilon]$ and $I_2^{\epsilon} = [C \cdot \epsilon, 1 - C \cdot \epsilon]$ for some $0 < c, C < \infty$.

To analyze $\bar{f}_j$, we recall that $\bar{\boldsymbol{\delta}}$ is defined by $\mathbf{G}'(\mathbf{0})(\bar{\boldsymbol{\delta}}) = A \cdot (\hat{\mathbf{g}} - \mathbf{g})$, see (5.5). We decompose $\hat{g}_j - g_j$ as $\hat{g}_j - g_j = \hat{g}_j^A + \hat{g}_j^B$, where

$$\hat{g}_j^A = \hat{g}_j - E(\hat{g}_j), \quad \hat{g}_j^B = E(\hat{g}_j) - g_j.$$

Note that $\hat{g}_j^B$ are non-stochastic. For $s = A$ and B, define $\bar{\boldsymbol{\delta}}^s$ to be the solution of $\mathbf{G}'(\mathbf{0})(\bar{\boldsymbol{\delta}}^s) = A \cdot \hat{\mathbf{g}}^s$. Then, it holds that

$$\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{\delta}}^A + \bar{\boldsymbol{\delta}}^B. \tag{A.9}$$

We may prove that, for any $\epsilon > 0$,

$$\int_{I_2(x)} \hat{g}_2^A(v) \left(\int_{I_1(v)} f_1(u)\, du \right)^{-1} dv = o_p(n^{-2/5}),$$
$$\int_{I_1(y)} \hat{g}_1^A(u) \left(\int_{I_2(u)} f_2(v)\, dv \right)^{-1} du = o_p(n^{-2/5})$$

uniformly for $(x, y) \in I_1^\epsilon \times I_2^\epsilon$. The latter may be proved as in the proof of Lemma 2 in Lee et al. (2015). From this and the expression of $\mathbf{G}'(\mathbf{0})$, a version of $\hat{\mathbf{G}}'(\mathbf{0})$ at (5.3) with $\hat{g}_j$ being replaced by g_j, it holds that, for any $\epsilon > 0$,

$$\bar{\boldsymbol{\delta}}^A = \mathbf{G}'(\mathbf{0})^{-1}(A \cdot \hat{\mathbf{g}}^A) = \hat{\mathbf{g}}^A / \mathbf{g} + o_p(n^{-2/5}) \tag{A.10}$$

uniformly for $(x, y) \in I_1^\epsilon \times I_2^\epsilon$.

Now, let x and y be fixed points in I_1^o and I_2^o, respectively. Then, there exists $\epsilon_0 > 0$ such that $x \in I_1^{\epsilon_0}$ and $y \in I_2^{\epsilon_0}$. Because of (A.10) this implies that

$$\bar{\delta}_1^A(x) = \hat{g}_1^A(x)/g_1(x) + o_p(n^{-2/5}), \quad \bar{\delta}_2^A(y) = \hat{g}_2^A(y)/g_2(y) + o_p(n^{-2/5}). \tag{A.11}$$

The first-order asymptotic properties of $\bar{\delta}_1^A(x)$ and $\bar{\delta}_2^A(y)$ are readily obtained from those of $\hat{g}_1^A(x)$ and $\hat{g}_2(y)$, respectively.

Next, we consider $\bar{\boldsymbol{\delta}}^B$ in the decomposition (A.9). We first note that $\hat{g}_j^B = n^{-2/5}\tilde{g}_j^B + r_j$ for $j = 1, 2$, where r_j are generic terms such that

$$\sup_{u \in [0,1]} |r_j(u)| = O(n^{-2/5}), \qquad \sup_{u \in [h, 1-h]} |r_j(u)| = o(n^{-2/5}).$$

Writing $\mathbf{r} = (r_1, r_2)^\top$, we get that, for any $\epsilon > 0$, $\bar{\boldsymbol{\delta}}^B = n^{-2/5} \cdot \boldsymbol{\beta} + \mathbf{r}$ uniformly for $(x, y) \in I_1^\epsilon \times I_2^\epsilon$. Now, let x and y be fixed points in $I_1^o \cap (0, 1)$ and $I_2^o \cap (0, 1)$, respectively. Then, it holds that

$$\bar{\delta}_1^B(x) = n^{-2/5} \cdot \beta_1(x) + o(n^{-2/5}), \quad \bar{\delta}_2^B(y) = n^{-2/5} \cdot \beta_2(y) + o(n^{-2/5}). \tag{A.12}$$

The expansions (A.11) and (A.12) give the theorem.

References

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

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

M. N. Chang and G. L. Yang. Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15** (4), 1536–1547 (1987). MR913572.

S. Chen and H. Zhao. Estimating incremental cost-effectiveness ratios and their confidence intervals with different terminating events for survival time and costs. *Biostatistics* **14** (3), 422–432 (2013a). DOI: 10.1093/biostatistics/kxt002.

S. Chen and H. Zhao. Generalized redistribute-to-the-right algorithm: application to the analysis of censored cost data. *J. Stat. Theory Pract.* **7** (2), 304–323 (2013b). MR3196602.

G. E. Dinse. An alternative to efron's redistribution-of-mass construction of the kaplanmeier estimator. *The American Statistician* **39** (4), 299–300 (1985). DOI: 10.1080/00031305.1985.10479453.

B. Efron. The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health*, pages 831–853. University of California Press, Berkeley, Calif. (1967).

B. Efron and V. Petrosian. Nonparametric methods for doubly truncated data. *J. Amer. Statist. Assoc.* **94** (447), 824–834 (1999). MR1723343.

E. A. Gehan. A generalized two-sample wilcoxon test for doubly censored data. *Biometrika* **52** (3/4), 650–653 (1965). DOI: 10.2307/2333721.

M. G. Gu and C.-H. Zhang. Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* **21** (2), 611–624 (1993). MR1232508.

M. Hiabu, E. Mammen, M. D. Martínez-Miranda and J. P. Nielsen. In-sample forecasting with local linear survival densities. *Biometrika* **103** (4), 843–859 (2016a). DOI: 10.1093/biomet/asw038.

M. Hiabu, C. Margraf, M. D. Martínez-Miranda and J. P. Nielsen. Cash flow generalisations of non-life insurance expert systems estimating outstanding liabilities. *Expert Systems with Applications* **45**, 400 – 409 (2016b). DOI: 10.1016/j.eswa.2015.09.021.

M. Hiabu, C. Margraf, M. D. Martínez-Miranda and J. P. Nielsen. The link between classical reserving and granular reserving through double chain ladder and its extensions. *British Actuarial Journal* **21** (1), 97–116 (2016c). DOI: 10.1017/S1357321715000288.

J. T. Hodgson, D. M. McElvenny, A. J. Darnton, M. J. Price and J. Peto. The expected burden of mesothelioma mortality in Great Britain from 2002 to 2050. *British Journal Of Cancer* **92**, 587 (2005). DOI: 10.1038/sj.bjc.6602307.

D. Kuang, B. Nielsen and J. P. Nielsen. Chain-ladder as maximum likelihood revisited. *Annals of Actuarial Science* **4** (1), 105–121 (2009). DOI: 10.1017/S1748499500000610.

R. Lee and T. Miller. Evaluating the performance of the lee-carter method for forecasting mortality. *Demography* **38** (4), 537–549 (2001). DOI: 10.1353/dem.2001.0036.

R. D. Lee and L. R. Carter. Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association* **87** (419), 659–671 (1992). DOI: 10.2307/2290201.

Y. K. Lee. Estimation of a semiparametric multiplicative density model. *J. Korean Statist. Soc.* **45** (4), 647–653 (2016). MR3566169.

Y. K. Lee, E. Mammen, J. P. Nielsen and B. U. Park. Asymptotics for in-sample density forecasting. *Ann. Statist.* **43** (2), 620–651 (2015). MR3319138.

Y. K. Lee, E. Mammen, J. P. Nielsen and B. U. Park. Operational time and in-sample density forecasting. *Ann. Statist.* **45** (3), 1312–1341 (2017). MR3662456.

H. M. Malani. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* **82** (3), 515–526 (1995). DOI: 10.2307/2337530.

E. Mammen, M. D. Martínez-Miranda and J. P. Nielsen. In-sample forecasting applied to reserving and mesothelioma mortality. *Insurance Math. Econom.* **61**, 76–86 (2015). MR3324046.

E. Mammen and J. P. Nielsen. Generalised structured models. *Biometrika* **90** (3), 551–566 (2003). MR2006834.

E. Mammen, B. U. Park and M. Schienle. Additive models: extensions and related models. In *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*, pages 176–211. Oxford Univ. Press, Oxford (2014). MR3306926.

M. D. Martínez-Miranda, B. Nielsen and J. P. Nielsen. Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. *J. Roy. Statist. Soc. Ser. A* **178** (1), 29–55 (2015). MR3291760.

M. D. Martínez-Miranda, B. Nielsen and J. P. Nielsen. Simple benchmark for mesothelioma projection for great britain. *Occupational and Environmental Medicine* (2016). DOI: 10.1136/oemed-2015-103303.

M. D. Martínez-Miranda, B. Nielsen, J. P. Nielsen and R. J. Verrall. Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers. *ASTIN Bulletin* **41** (1), 107–129 (2011).

M. D. Martínez-Miranda, J. P. Nielsen, S. Sperlich and R. J. Verrall. Continuous chain ladder: Reformulating and generalizing a classical insurance problem. *Expert Systems with Applications* **40** (14), 5588 – 5603 (2013). DOI: 10.1016/j.eswa.2013.04.006.

M. D. Martínez-Miranda, J. P. Nielsen and R. Verrall. Double chain ladder. *Astin Bull.* **42** (1), 59–76 (2012). MR2931855.

M. D. Martínez-Miranda, J. P. Nielsen and R. J. Verrall. Double chain ladder and bornhuetter-ferguson. *North American Actuarial Journal* **17** (2), 101–113 (2013). DOI: 10.1080/10920277.2013.793158.

C. Moreira and J. de Uña Álvarez. Kernel density estimation with doubly truncated data. *Electron. J. Stat.* **6**, 501–521 (2012). MR2988417.

C. Moreira, J. de Uña Álvarez and L. Meira-Machado. Nonparametric regression with doubly truncated data. *Comput. Statist. Data Anal.* **93**, 294–307 (2016). MR3406213.

C. Moreira and I. Van Keilegom. Bandwidth selection for kernel density estimation with doubly truncated data. *Comput. Statist. Data Anal.* **61**, 107–123 (2013). MR3063004.

P. A. Mykland and J.-J. Ren. Algorithms for computing self-consistent and maximum likelihood estimators with doubly censored data. *Ann. Statist.* **24** (4), 1740–1764 (1996). MR1416658.

R. O'Brien. *Age-Period-Cohort Models: Approaches and Analyses with Aggregate Data*. Chapman & Hall/CRC Press, London, 1st edition (2014).

J. Peto, F. E. Matthews, J.T. Hodgson and J.R. Jones. Continuing increase in mesothelioma mortality in britain. *The Lancet* **345** (8949), 535–539 (1995). DOI: 10.1016/S0140-6736(95)90462-X.

C. Rake, C. Gilham, A. Hatch, J.and Darnton, J. Hodgson and J. Peto. Occupational, domestic and environmental mesothelioma risks in the british population: a case-control study. *British Journal of Cancer* **100**, 1175–83 (2009). DOI: 10.1038/sj.bjc.6604879.

A.E. Renshaw and S. Haberman. A cohort-based extension to the leecarter model for mortality reduction factors. *Insurance: Mathematics and Economics* **38** (3), 556 – 570 (2006). DOI: 10.1016/j.insmatheco.2005.12.001.

A. Riebler, L. Held and H. Rue. Estimation and extrapolation of time trends in registry data-borrowing strength from related populations. *Ann. Appl. Stat.* **6** (1), 304–333 (2012). DOI: 10.1214/11-AOAS498.

T. R. Smith and J. Wakefield. A review and comparison of age-period-cohort models for cancer incidence. *Statist. Sci.* **31** (4), 591–610 (2016). MR3598741.

E. Tan, N. Warren, A. J. Darnton and J. T. Hodgson. Projection of mesothelioma mortality in Britain using Bayesian methods. *British Journal Of Cancer* **103**, 430 (2010). DOI: 10.1038/sj.bjc.6605781.

E. Tan, N. Warren, A. J. Darnton and J. T. Hodgson. Modelling mesothelioma mortality in great britain using the two-stage clonal expansion model. *Occupational and Environmental Medicine* **68** (2011).

B. W. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69**, 169–173 (1974). MR0381120.

B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** (3), 290–295 (1976). MR0652727.

R. J. Verrall, J. P. Nielsen and A. H. Jessen. Prediction of rbns and ibnr claims using claim amounts and claim counts. *ASTIN Bulletin* **40** (2), 871–887 (2010). DOI: 10.2143/AST.40.2.2061139.