Mammen, E., Martinez-Miranda, M. D. & Nielsen, J. P. (2015). In-Sample Forecasting Applied to Reserving and Mesothelioma Mortality. Insurance: Mathematics and Economics, 61, pp. 76-86. doi: 10.1016/j.insmatheco.2014.12.001

In-Sample Forecasting Applied to Reserving and Mesothelioma Mortality

Enno Mammen
Universität Mannheim, Abteilung Volkswirtschaftslehre
L7, 3-5, 68131-Mannheim, Germany

María Dolores Martínez Miranda
University of Granada, Spain
Cass Business School, City University London
106 Bunhill Row, UK - London EC1Y 8TZ, U.K.

Jens Perch Nielsen
Cass Business School, City University London
106 Bunhill Row, UK - London EC1Y 8TZ, U.K.

November 8, 2014 *

*Second author acknowledges the support of the European Commission under the Marie Curie Intra-European Fellowship FP7-PEOPLE-2011-IEF Project number 302600 and the Spanish "Ministerio de Ciencia e Innovación" by the grant MTM2008-03010. Authors thank the Centro de Servicios de Informática y Redes de Comunicaciones (CSIRC), Universidad de Granada, for providing the computing time.

Abstract

This paper shows that recent published mortality projections with unobserved exposure can be understood as structured density estimation. The structured density is only observed on a sub-sample corresponding to historical calendar time. The mortality forecast is obtained by extrapolating the structured density to future calendar times using that the components of the density are identified within sample. The new method is illustrated on the important practical problem of forecasting mesothelioma for the UK population. Full asymptotic theory is provided. The theory is given in such generality that it also introduces mathematical statistical theory for the recent continuous chain ladder model. This allows a modern approach to classical reserving techniques used every day in any non-life insurance company around the globe. Applications to mortality data and non-life insurance data are provided along with relevant small sample simulation studies.

Keywords: Non-parametric; Kernel density estimation; Reserve risk; Multiplicative; Chain Ladder.

1 Introduction

Let us assume that we have a structured density defined as a density that is a known function of one-dimensional densities, see Mammen and Nielsen (2003) for the equivalent definition of structured regression. Assume furthermore that observations are available from this structured density on a restricted support only. Finally assume that the character of this restricted support is such that in-sample information is available for all the one-dimensional functions defining the original structured density. In this situation, an extrapolation or forecast is immediately available for that part of the support without observations. It turns out that one of the most important problems in non-life insurance, estimation of outstanding liabilities in reserving, has exactly this form. The structured density is most often a multiplicative density in this case. The support with observations represents insurance claims until the current calendar time, and the support without observations represents future insurance claims. This forecast method has traditionally been called the chain ladder technique in actuarial science and the multiplicative density has been estimated as a structured histogram or equivalently from maximum likelihood assuming a multiplicative Poisson structure, see Wüthrich and Merz (2008) for and overview and Kuang et al. (2009), Verrall et al. (2010), Martínez-Miranda et al. (2011, 2012, 2013a,b,c), for recent reformulations of classical chain ladder in mathematical statistical terms published in the actuarial literature. Other recent contributions in

reserving considering statistical models based on individual claims include Antonio and Plat (2014) and Pigeon et al. (2013,2014). The longevity problem is another important application of structured density forecasting and as in non-life insurance a histogram type of approach is widely used and analysed, see Haberman and Renshaw (2012) and Hatzopoulos and Haberman (2013). In this paper we propose to use our alternative approach based on structured non-parametric models and we will illustrate its power by applying it to mesothelioma mortality forecasts. We compare our empirical findings with Martínez-Miranda et al. (2014) who used a classical approach based on a multiplicative Poisson structure.

While we stick to the multiplicative density structure in this paper, it is evident that important generalizations are possible. One could add a variety of one-dimensional densities to the overall structure leading to non-multiplicative structures. It would also be interesting to generalize the approach of this paper to other sources of mortality than age and cohort. One example would be to add calendar time effects generalizing the histogram approach to calendar effect estimation developed in Kuang et al.(2008a,b). Another would be to add time independent or time dependent covariates. It would be also interesting to consider the work of Zhang et al. (2013) to develop distribution free prediction sets, see Lei et al. (2013). Finally, the approach of projecting a multivariate density smoother down on the structure of interest is not restricted to local linear density smoothers and could be generalised to other multivariate density smoothers including Panaretosa and Konis (2012), Xiao et al. (2013) and Lu et al. (2013).

The paper is organized as follows. In Section 2 the structured density model is formulated in the special multiplicative case. A projection approach based on local linear density estimation is defined. The asymptotic properties of the suggested method is provided in Section 6 (with more details and proofs deferred to the Appendix). Applications to non-life insurance and mesothelioma mortality forecasting are explained in Section 3. While these two applications rely on the multiplicative density structure, observations are available on very different underlying supports. However, for both applications the entering one-dimensional densities are identified by the observed data. The analyses of two datasets are described in Section 4. Section 5 includes a brief simulation study with simulation settings defined to be close to real life situations. All the calculations in the paper have been performed with R, R Development Core Team (2011).

2 Multiplicative density forecasting

2.1 Model formulation

Let us consider n i.i.d. observations $\{(X_i, Y_i), i = 1, \ldots, n\}$ from a two-dimensional variable (X, Y) having a density f with support on a subset $\mathcal{I}_f$ of the rectangle $\mathcal{S}_f = \{(x, y) : 0 \leq x \leq T_1, 0 \leq y \leq T_2\}$, with $T_1, T_2 > 0$. The aim is to forecast the density of (X, Y) in $\mathcal{S}_f$ from the given observations that are only available in the set $\mathcal{I}_f$. To this goal let us assume that f is multiplicative, this is, it is of the form:

$$f(x, y) = c_f f_1(x) f_2(y), \tag{1}$$

where f_1 and f_2 are probability densities on $[0, T_1]$ and $[0, T_2]$, respectively. The constant c_f is chosen such that

$$\int_{\mathcal{I}_f} c_f f_1(x) f_2(y) dx \; dy = 1. \tag{2}$$

This formulation transforms the original forecasting problem to an estimation problem of the densities f_1 and f_2. The approach of this paper is developed for a general support $\mathcal{I}_f$ including the two different support structures that came up in our two applications (mortality studies and insurance reserving). See also Nielsen and Linton (1998) for related projection methods in structured nonparametric regression.

Note that if the support where the densities are observed is a rectangle, then the estimation problem would be trivial and both components could be estimated separately. The non-rectangular supports considered in this paper implies that the estimation problem is more complicated. However, we are only considering non-rectangular supports, where the multiplicative components are still estimable in-sample. While the term in-sample forecasting is defined in this paper, the in-sample forecasting trick is an old one and has been used in non-life insurance in actuarial science as long as anyone remembers. In actuarial science the non-rectangular support is a triangle and the multiplicative structure is estimated via a parametric approach related to maximum-likelihood estimation, see Kuang et al. (2009). It has recently been pointed out that this classical actuarial forecasting methodology can be understood as first estimating a multiplicatively structured histogram and then extrapolating into the future, see Martínez-Miranda et al. (2013a). More complicated structures violating the independence assumption between X and Y could also be considered. This is, however, beyond the scope of this paper. Among many examples one could imagine that a calendar time effect enters the model in some multiplicative way, see Kuang et al. (2011)for a classical structured histogram approach to forecasting including such a calendar effect.

2.2 The projection approach

Consider the density f with support $\mathcal{I}_f$ and consider one point $(x, y) \in \mathcal{I}_f$. The local linear estimator introduced in Nielsen (1999) and Müller and Stadtmüller (1999) is derived by solving the following minimization problem:

$$\widehat{\Theta} = \min_{\Theta} \lim_{b \to 0} \int_{\mathcal{I}_f} \left\{ \widetilde{f}_b(u, v) - \theta_1 - \theta_{2,1}(u - x) - \theta_{2,2}(v - y) \right\}^2 K_{h_1}(u-x) K_{h_2}(v-y) du\, dv, \tag{3}$$

where $\Theta = (\theta_1, \theta_{2,1}, \theta_{2,2})$, $\widehat{\Theta} = (\widehat{\theta}_1, \widehat{\theta}_{2,1}, \widehat{\theta}_{2,2})$ and $\widetilde{f}_b(u, v) = n^{-1} \sum_{i=1}^n K_b(X_i - u) K_b(Y_i - v)$. Here $K_b(u) = b^{-1} K(u/b)$, for a one-dimensional symmetric kernel function K and bandwidth parameters $b > 0$, $h_1 > 0$, $h_2 > 0$. The local linear density of $f(x, y)$ is given by $\widehat{f}_{h;\mathcal{I}_f}(x, y) = \widehat{\theta}_1$, which is defined for any given vector of bandwidth parameters $h = (h_1, h_2) \in I\!R^2_+$.

Note that $\widehat{f}_{h;\mathcal{I}_f}(x, y)$ is an estimator of the density f of (X, Y) restricted to the support $\mathcal{I}_f$. Forecasting into the future amounts to extrapolating our estimated density to the full support $\mathcal{S}_f$. This forecast or extrapolation is only possible under some assumptions on the functional form of $f(x, y)$. In this paper, we consider one of the simplest structured density options (1) and project the unrestricted local linear estimator down on the relevant multiplicative space. Specifically c_f, f_1 and f_2 are estimated by minimizing the following expression:

$$\min_{c_f, f_1, f_2} \int_{\mathcal{I}_f} \left(\widehat{f}_{h;\mathcal{I}_f}(x, y) - c_f f_1(x) f_2(y) \right)^2 w(x, y) dx\, dy, \tag{4}$$

under the constrain that $\int_0^{T_1} f_1(x) dx = 1$ and $\int_0^{T_2} f_2(y) dy = 1$. Here w is some weighting function such that $w(x, y) > 0$.

In practice, the above minimization can be done by using the following iterative algorithm:

1. Let consider the estimator $\widehat{f}_{h;\mathcal{I}_f}$ derived above and $\widehat{f}^{(0)}_{1,h}$ being an initial estimator of f_1.

2. Calculate an estimator the f_2 as

$$\widehat{f}^{(1)}_{2,h}(y) = \frac{\int_{\mathcal{I}_y} \widehat{f}_{h;\mathcal{I}_f}(x, y) w(x, y) dx}{\int_{\mathcal{I}_y} \widehat{f}^{(0)}_1(x) w(x, y) dx},$$

   where $\mathcal{I}_y = \{x : (x, y) \in \mathcal{I}_f\}$.

3. From $\widehat{f}_{2,h}^{(1)}$ given above, let update the estimator of f_1 by

$$\widehat{f}_{1,h}^{(1)}(x) = \frac{\int_{\mathcal{I}_x} \widehat{f}_{h;\mathcal{I}_f}(x,y)w(x,y)dy}{\int_{\mathcal{I}_x} \widehat{f}_2^{(1)}(y)w(x,y)dy},$$

where $\mathcal{I}_x = \{y : (x,y) \in \mathcal{I}_f\}$.

4. Repeat steps 2 and 3 until convergence.

The constraints in the algorithm are neglected and it suffices to adjust the estimators to hold them at the end. Let denote the final adjusted (to hold the constraints) estimates by $\widehat{f}_1$ and $\widehat{f}_2$, then the constant c_f in (4) is estimated such that $c_f \int_{\mathcal{I}_f} \widehat{f}_1(x)\widehat{f}_2(y)dx\ dy = 1$. A convenient choice for that the initial estimator $\widehat{f}_{1,h}^{(0)}$ in step 1 can be a constant function (actually we have made this choice for the empirical illustrations in the paper).

Note that the above algorithm is similar to the backfitting algorithm used for fitting additive models and it has also been proposed for the continuous chain-ladder model in Martínez-Miranda et al. (2013a). Note that in the algorithm it is not needed that the two-dimensional density $f(x,y)$ is estimated. The algorithm makes only use of estimates of the marginal quantities $\int_{\mathcal{I}_y} f(x,y)w(x,y)dx$ and $\int_{\mathcal{I}_x} f(x,y)w(x,y)dy$. In our asymptotic theory we will discuss alternative estimation approaches where these quantities are directly estimated without using estimates of $f(x,y)$.

An important remaining practical problem is data-adaptive choice of the bandwidths (h_1, h_2). One way would be to use a data-adaptive bandwidth selector for the estimation of the unstructured two-dimensional density f. In our simulations and data-examples we used a least squares cross-validation approach. The cross-validated bandwidth tuple $(\widehat{h}_1, \widehat{h}_2)$ is defined as the minimizer of

$$\text{LSCV}(h_1,h_2) = \int_{\mathcal{I}_f} \widehat{f}_{h;\mathcal{I}_f}(x,y)^2 dx\ dy - 2\sum_{i=1}^{n} \int_{\mathcal{I}_f} \widehat{f}_{h;\mathcal{I}_f}^{[-i]}(x,y)d\widetilde{F}_n(x,y), \qquad (5)$$

where $\widehat{f}_{h;\mathcal{I}_f}^{[-i]}$ is the leave-one-out version of the estimator $\widehat{f}_{h;\mathcal{I}_f}$, and $\widetilde{F}_n$ is the empirical distribution function. This approach has the disadvantage that the resulting bandwidths are optimal for the nonparametric estimation problem of a two-dimensional density but not for the one-dimensional components of our structured model, see our asymptotic theory below. The development of a cross-validation bandwidth selector that is designed for a structured model is still missing and deserves further research, which is beyond the scope of this paper.

3 Two examples: reserving and mesothelioma mortality forecasting

In this section the two main forecasting applications are explained. The first is about the so called reserve estimating outstanding liabilities. This number is perhaps the most important number in the accounts of a non-life insurance company and some of the worst solvency problems non-life insurance company historically have faced are results of insufficient reserves for outstanding liabilities. The second application is on forecasting of mesothelioma mortality (Peto et al. (1995), Martínez-Miranda et al. (2013b)). While this application is also of immense importance for insurance companies as well, it also has implications for political decisions on economics.

Our practical data are provided in discrete form. The data is not discrete by nature, but have been aggregated by data providers. Standard methods in this area are designed around aggregated methods and there is a tendency to consider aggregated data as the original data (Kuang et al. (2009), Hatzopoulos and Haberman (2013), Clayton and Schifflers (1987)). Statisticians, insisting on working on the original continuous data only, will have the disadvantage of not being able to work on many of the most important problems in the two fields. It is therefore encouraging that our purely continuous approach adapts to aggregated data in a simple, efficient and robust way. Carroll et al. (2013) concluded in a slightly different context that cross-validation methods seem more robust to discretization than methods based on asymptotic expansions such as plug-in methods and they give some theoretical background for this conclusion. We follow their advice and stick to cross-validation when choosing the level of smoothing.

3.1 Claims reserving in non-life insurance

Here the data arrange is a triangle support defined as $\mathcal{I}_f = \{(x, y) : 0 \leq x, y \leq T, x + y \leq T\}$, where x is the underwriting time, y is the claims development time and $[0, T]$ (with $T > 0$) is the time observation window (see Martínez-Miranda et al. (2013a)). Traditionally the actuaries work with the data in an aggregated way, defining the traditional run-off triangles. These triangles can be represented as sets such as $\aleph_m = \{N_{ij} : (i, j) \in \mathcal{I}_m\}$, where $\mathcal{I}_m = \{(i, j) : i = 1, \ldots, m, j = 1, \ldots, m;\ i + j - 1 \leq m\}$. The available data are then indexed in the set $\mathcal{I}_m$, which is just a discretization in periods such as quarters or years of the continuous triangle $\mathcal{I}_f$ defined above. The values N_{ij} correspond to aggregated values such as the total number of claims of insurance incurred in period i, which are reported in period $i + j - 1$ i.e. with $j - 1$ periods delay from year i, or the total quantity paid for such

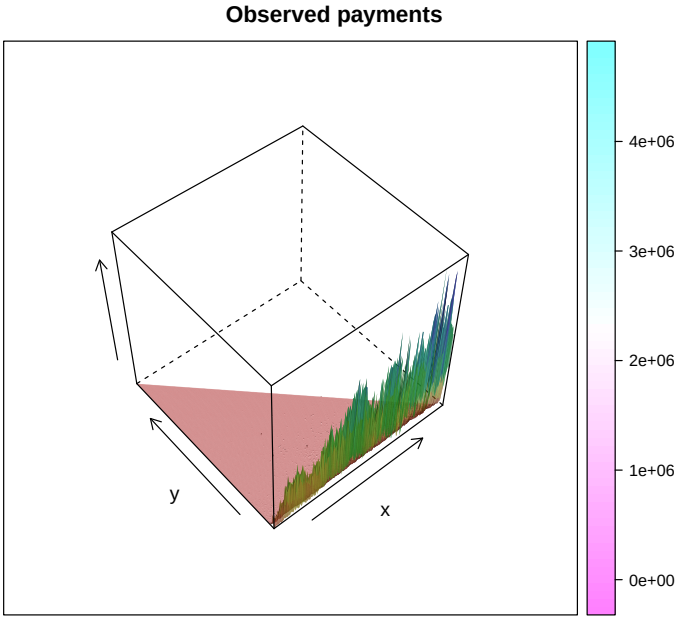

Figure 1: Triangle support in Reserving: paid run-off triangle from a major insurer.

claims. Here $n = \sum_{i=1}^{m} \sum_{j=1}^{m} N_{ij}$ is the sample size, which represent the number of observed claims or the total paid quantity. Figure 1 shows an example of this kind of data and the special triangle support is visualized. In this case, the available information are the total paid quantities for claims incurred between 1990 and 2011. The payments are arranged into a triangle by the occurrence or underwriting month and the development month. During these 264 months (22 years) it was paid a total of 1,362,222,980 pounds (sample size). The classical aim is to forecast in the lower triangle given by $\mathcal{J}_m = \{(i,j) : i = 1,\ldots,m, j = 1,\ldots,m;\ i+j-1 > m\}$. Note that $\mathcal{J}_m$ represents the future liabilities for the company arising for claims which happened in the m observed origin periods ($i = 1,\ldots,m$). In this paper we describe how this type of data can be analysed and predicted by a multiplicatively structured density $f(x,y) = f_1(x)f_2(y)$, where f_1 and f_2 are respectively the underwriting time and the development time density components. The support of the observations is the triangle $\mathcal{I}_f$ and the forecasting goal is to extrapolate the multiplicative density to the triangle $[0,T]^2 \backslash \mathcal{I}_f$.

3.2 Mesothelioma mortality forecasting

The age-period is one of the most common arranges in mortality data. In this setting, the data consists of the number of people with age y who died at a period p. Then, a typical data set can be written as $\{(p_i, a_j, N_{ij}) | i = 1,\ldots,P, j = 1,\ldots,A\}$, with N_{ij} being the number of deaths at period p_i for the age a_j, and $n = \sum_{i=1}^{P} \sum_{j=1}^{A} N_{ij}$ is the

sample size. We use a data set of this type provided by UK Health Service Executive that consists of annual aggregated counts of deaths caused by exposure to asbestos in Great Britain. The data are given by age levels and year (periods) of death between 1967-2007. The data array has dimensions $A = 65$ (age levels) and $P = 41$ periods. The observed total number of deaths is 31902 (sample size). Previous analysis of these data revealed that the main source of asbestos death is the cohort and the age and it has been argued that a (parametric) age-cohort model should be suitable for forecasting purposes. We now reformulate the problem as a structured density estimation problem with restricted support. To translate the (parametric) age-cohort model to a density formulation we decompose the two-dimensional density into a cohort and an age component. The model would be exactly the same as in the above reserving problem if the data would come in an age-cohort arrange. However the data have been recorded by periods instead of cohorts and therefore the support of the densities is now a parallelogram. In the asbestos data the cohort is defined as the transformation $x = p - y$, and can be indexed by $k = A - i + j$, where $i = 1, \dots, P = 41$ is the index period and $a = 1, \dots, A = 65$ is the age index. Let denote by $f(x, y)$ the two-dimensional density, with x and y being the cohort and the age, respectively. The age-cohort model lead again to a structured (multiplicative) density, $f(x, y) = f_1(x)f_2(y)$, where f_1 and f_2 are the cohort and age density components. But, since the data come in an aggregated way, the support of the density $f(x, y)$ is indexed in the discrete set $\mathcal{I}_{P,A} = \{(k, j) : k = 1, \dots, K, j = 1, \dots, A;\ 1 \leq A - k + j \leq P\}$. Figure 2 shows a histogram of the observed data and represent the parallelogram support. The forecast purposes in this particular case, we focus on a horizon of $\tau = 40$ future periods from the latest observed year 2007. Also, since it is expected that no more cohorts will be affected by asbestos, there is no reason to extrapolate the cohorts. Then the forecast set can be written as $\mathcal{J}_{P+\tau,A} = \{(k, j) : k = 1, \dots, K, j = 1, \dots, A;\ P + 1 \leq A - k + j \leq P + \tau\}$. The same data set has also been analysed by Martínez-Miranda et al. (2014) using maximum likelihood on a Poisson multiplicative structure.

4 Structured density forecasting for outstanding liabilities and mesothelioma mortality

This section provides our empirical results on structured density forecasting. The empirical findings in our two examples are not too different from the results already obtained from the structured histogram approach and the multiplicative Poisson approach. This holds for outstanding liabilities as well as for mesothelioma mortality. This is comforting and convince us that our intuitive and visual structured density

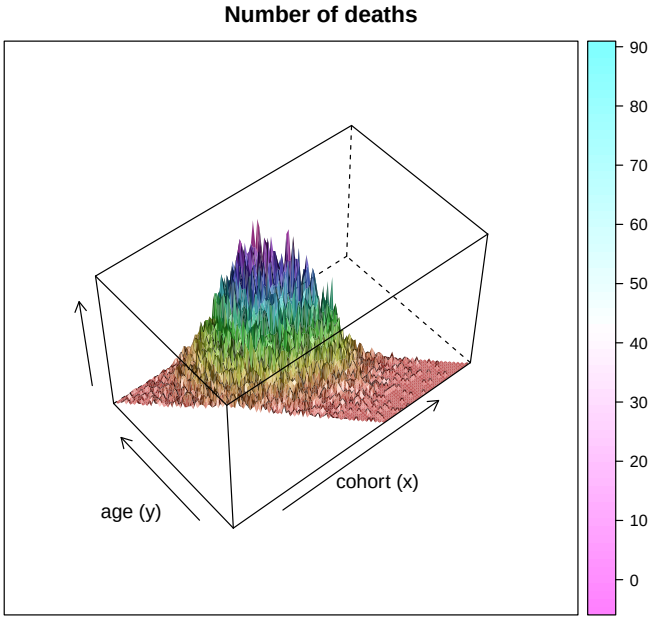

Figure 2: Parallelogram support in mesothelioma mortality forecasting.

forecasting methodology provide the type of results we want. The structured density approach also allows us to derive mathematical statistical asymptotic theory on the performance of the underlying one-dimensional functions, see Section 6. This type of mathematical statistical asymptotic theory has been missing in the literature so far and has restricted the implementation of statistical methods of inference. Furthermore, the parametric approach does not allow a theoretical investigation of the important trade off of variance versus bias, that structured non-parametric theory provide.

4.1 Structured forecasting outlying liabilities in non-life insurance: a data study

In Figure 1 the raw data of insurance payments (given in pounds) are plotted on the observed triangle. Note that the claims development is quite fast for this data and most of the mass is around small values of y. This is also clear from studying the two underlying multiplicative densities given in Figure 3. These densities have been estimated using the methods described in Section 2.2, using the Epanechnikov kernel. The mass of the outstanding liabilities is obtained by integrating the estimator in the unobserved part of the triangle. The results are shown in Table 1. The structured kernel density approach predicts a total reserve of $20,045,534$, which is close but a bit higher to the $17,265,400$ obtained using classical reserving methods (see Martínez-Miranda et al. (2013a) for more details). It as apparent from Table 1,

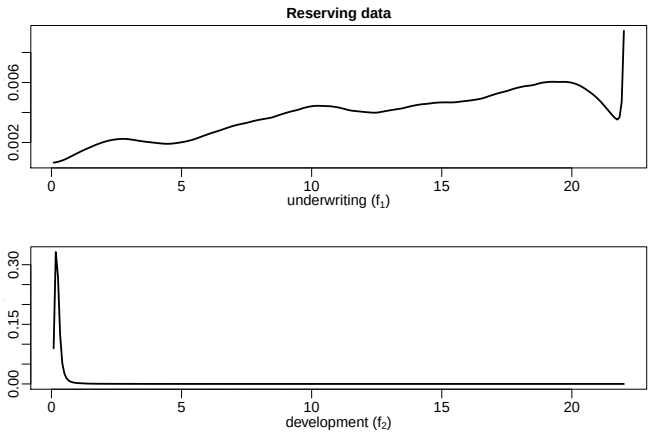

Figure 3: Estimated time effects.

that most of the difference of the reserved estimates between classical chain ladder on the new continuous in-sample forecaster origins from differences in estimating the last years reserve. It is well known in the reserving literature that this last year is both the most fragile and the most important year to estimate well. Stabilizing strategies have been developed to improve the stability of classical chain ladder via expert knowledge using so called Bornhuetter-Fergursson techniques, see for example Martínez-Miranda et al. (2013b). While smoothing seems to add to the stability improving on classical chain ladder, then Bornhuetter-Fergursson methodologies also seem relevant for future consideration in the continuous chain ladder context considered in this paper.

4.2 Structured forecasting of mesothelioma mortality: a data study

In Figure 2 the parallelogram of observed mesothelioma mortality is plotted. While the support is not a triangle, the one-dimensional densities underlying the multiplicative model are still in-sample. Figure 4 shows the two estimated underlying densities using the kernel approach. We have used again the cross-validated bandwidth and the Epanechnikov kernel. Figure 5 plots the final two-dimensional forecasting density. Notice that this estimated multivariate density graph provides a smooth fit in-sample, where data are available, at the same time as predictions are provided in the areas where data are not available. As it is shown in Figure 6, the predicted future mesothelioma mortality peak is 2194 cases in the year 2019, which is very close to the prediction by Martínez-Miranda et al. (2014) of 2220 deaths at the same year (see Table 2 in the paper in the case of using the full dataset).

Table 1: Outstanding liabilities forecasts. Comparison between the kernel structured density with cross-validated bandwidth (LL–LSCV) and the classical reserving approach (Pois−ML).

Year	LL-LSCV	Pois−ML
1	18,256,059	15,414,076
2	868,681	893,305
3	416,166	432,549
4	223,499	236,865
5	128,257	133,692
6	72,906	76,662
7	39,334	37,491
8	20,903	22,054
9	12,318	11,934
10	3627	3284
11	1715	1993
12	957	809
13	544	385
14	365	302
15	202	0
Tot.	20,045,534	17,265,400

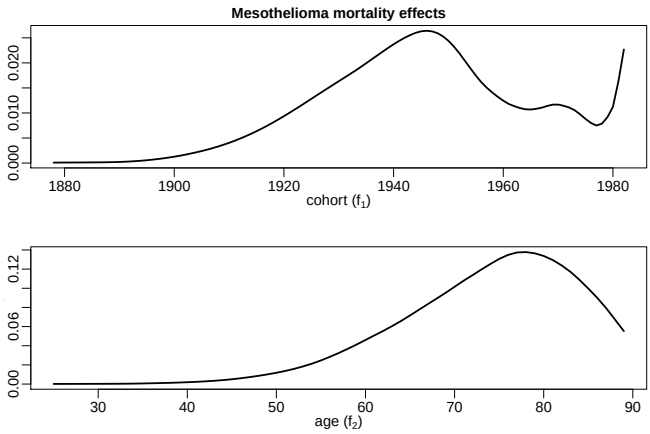

Figure 4: Estimated cohort and age effects in mesothelioma mortality.

This figure plots the shape of the future predicted mortality obtained by using our density approach (solid line). Notice that our forecast is quite close to the forecast by Martínez-Miranda et al. (2014) (dashed line).

5 Simulation study

In this section we describe a brief simulation study where we have compared the structured kernel density forecast with classical approaches. In the simulations the nonparametric approach clearly outperformed the latter ones. We have simulated three models assuming the multiplicative structure (1). The density components for each model are shown in Figure 7. The chosen models were motivated by our data examples from reserving and mortality prediction. The first two models mimic reserving problems, where the density f_1 corresponds to the underwriting time effect and f_2 is the development time density. The third model represents the underlying structure in the mesothelioma mortality data, where f_1 and f_2 are the densities corresponding to cohort and age effects, respectively. Compare Model 1 with the shapes in Figure 3 and Model 3 with the shapes in Figure 4. As in the data examples, we have simulated the data in an aggregated base. For Models 1 and 2 we have simulated monthly reserving data during 10 years. The data exhibits a triangle support such as shown in Figure 1. For the mortality model we have simulated data with a parallelogram support as in Figure 2 and we have used 105 cohorts and 65 ages as was also the case in our data set. We have run the simulations for sample sizes $n = 10^3, 10^4$ and 10^5.

For each model and each sample size, we have generated 250 samples. In the simulations we have compared the structured kernel density approach with Pois-

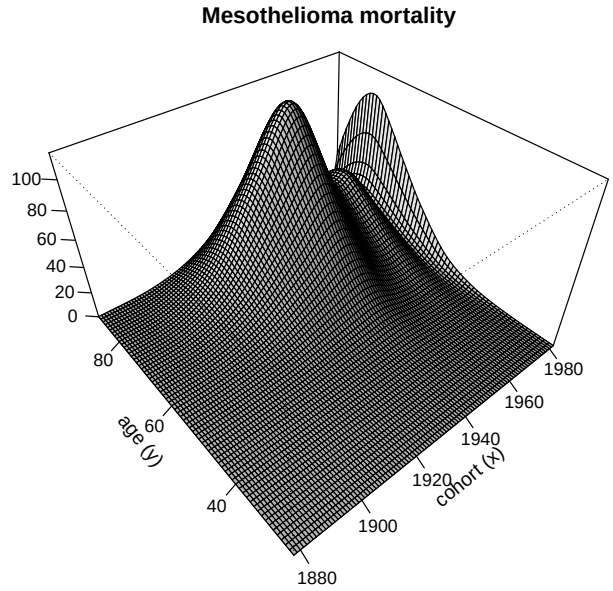

Figure 5: Mesothelioma mortality forecasts by age and cohort.

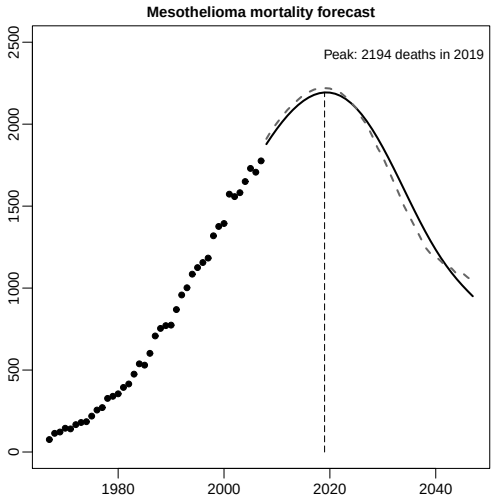

Figure 6: Mesothelioma mortality forecasts in the future years. The solid line represents the kernel density forecast. The dashed line is the forecast using the Poisson maximum likelihood approach by Martínez-Miranda et al. (2014).

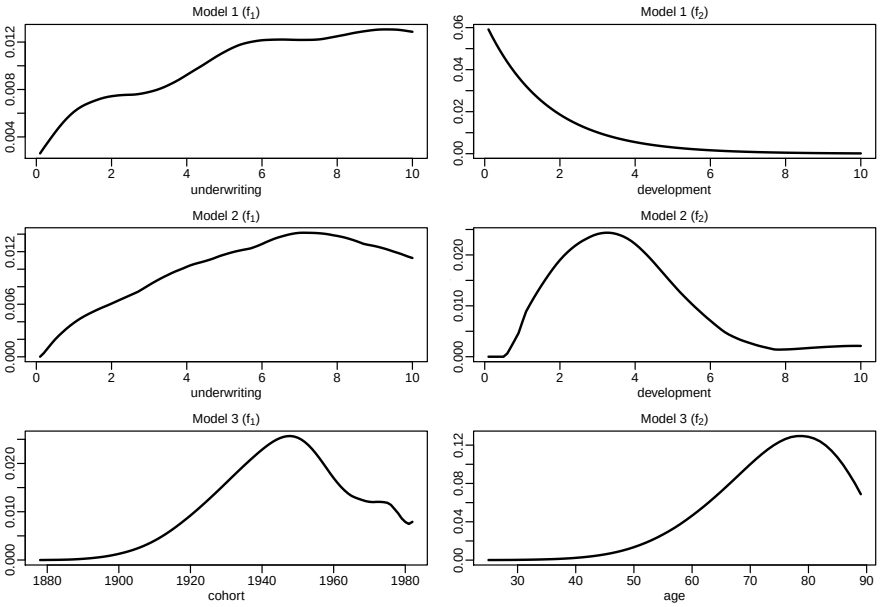

Figure 7: Simulated models.

son maximum likelihood, that is the popular chain ladder technique on reserving data aggregated into a yearly run-off triangle; and the recent age-cohort model by Martínez-Miranda et al. (2014) for the mortality data. The performance of the methods has been evaluated by using the following ISE criterion:

$$ISE(\widehat{f}) = \int_{\mathcal{S}_f \backslash \mathcal{I}_f} \left\{ \widehat{f}(x,y) - f(x,y) \right\}^2 \, dx \, dy, \tag{6}$$

where $\widehat{f}(x,y) = \widehat{f_1}(x)\widehat{f_2}(y)$ is the estimator of the actual multiplicative density, $f(x,y) = f_1(x)f_2(y)$. Notice that we are measuring the quality of the forecasts in the set $\mathcal{S}_f \backslash \mathcal{I}_f$.

The structured kernel density estimator has been calculated using the Epanechnikov kernel for two different bandwidth choices. On the one hand we consider the optimal bandwidths for each simulated sample in the ISE sense (6). This is an infeasible choice that we consider here as a benchmark. Furthermore, we used the least squares cross-validation criterion defined in (5). In both cases, the minimization has been done using a grid of 400 vectors of bandwidths $h = (h_1, h_2)$.

Figures 8-10 show box plots of the empirical distribution of the square root of the performance measure (6) (square root multiplied by 10^4), obtained from the 250 simulated samples, for each model and for each sample size. Table 2 shows a numerical summary of the distribution. The overall conclusion is clear, structured kernel density forecasting beats the Poisson maximum likelihood approaches (Pois–ML) in all the scenarios. The improvement is quite remarkable using the ISE-optimal

Table 2: Summary of the ISE errors for each simulated model and sample size.

		Model 1			Model 2			Model 3		
n		LL−ISE	LL-LSCV	Pois-ML	LL-ISE	LL-LSCV	Pois-ML	LL-ISE	LL-LSCV	Pois-ML
10^3	Median	0.074	0.123	0.382	0.221	0.640	0.850	0.793	1.302	3.668
	Mean	0.079	0.145	0.391	0.256	1.542	0.872	0.911	1.862	5.003
	SD	0.035	0.078	0.054	0.124	8.958	0.439	0.519	2.134	3.742
10^4	Median	0.030	0.054	0.357	0.126	0.421	0.398	0.446	0.668	1.686
	Mean	0.032	0.058	0.359	0.150	0.490	0.430	0.479	1.025	2.040
	SD	0.010	0.021	0.015	0.073	0.470	0.093	0.209	1.389	1.302
10^5	Median	0.013	0.017	0.355	0.100	0.288	0.357	0.232	0.301	0.675
	Mean	0.015	0.019	0.355	0.112	0.289	0.361	0.252	0.352	0.806
	SD	0.006	0.008	0.005	0.050	0.053	0.015	0.087	0.203	0.550

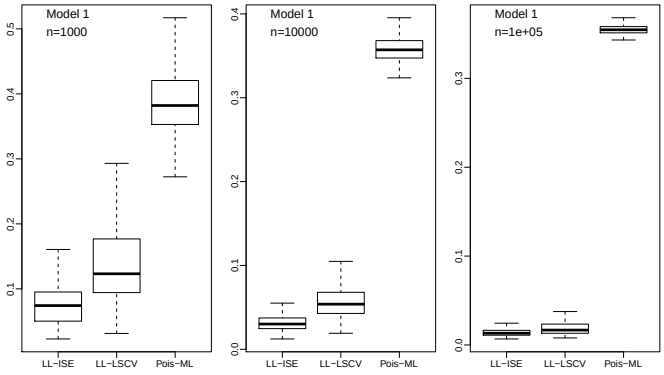

Figure 8: Model 1: box plots of the ISE errors at each simulated sample. Comparison between the kernel structured density (LL-ISE, LL-LSCV) and the Poisson maximum likelihood approach (Pois-ML).

bandwidth choice (LL–ISE) but, even using the simple cross-validated estimators (LL–LSCV), the quality of the forecasts is quite impressive. The differences in the performance between the ISE-optimal bandwidth choice (LL–ISE) and the cross-validated estimators (LL–LSCV) suggests that there is need for investigating more efficient bandwidth selectors that are better suited for the structured estimation problem. As remarked above the cross-validated bandwidth is designed for the estimation of the unstructured density and not for our structured model.

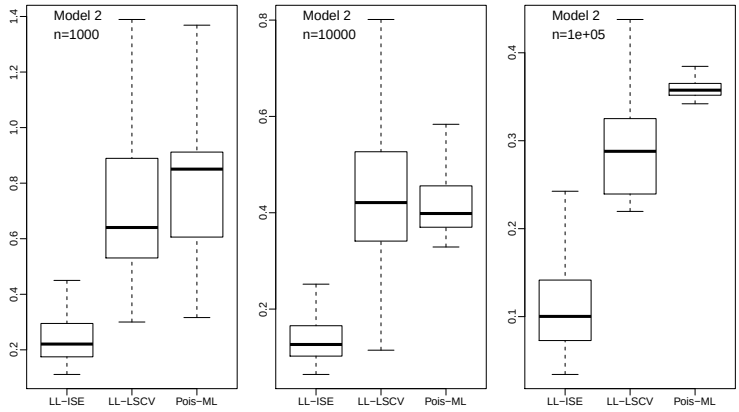

Figure 9: Model 2: box plots of the ISE errors at each simulated sample. Comparison between the kernel structured density (LL-ISE, LL-LSCV) and the Poisson maximum likelihood approach (Pois-ML).

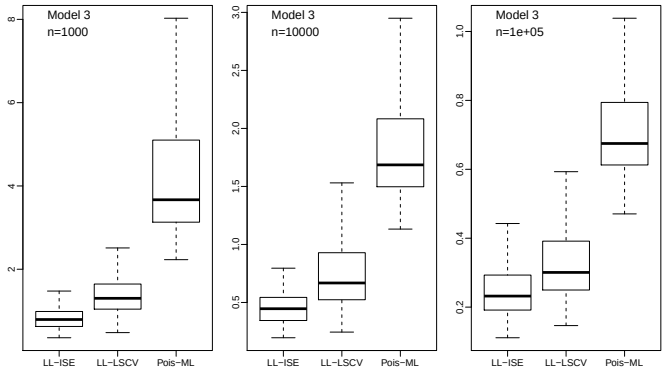

Figure 10: Model 3: box plots of the ISE errors at each simulated sample. Comparison between the kernel structured density (LL-ISE, LL-LSCV) and the Poisson maximum likelihood approach (Pois-ML).

6 Asymptotic theory

We restrict our theory to the triangular support $\mathcal{I}_f = \{(x,y) : 0 \leq x, y \leq 1, x+y \leq 1\}$, and, for simplicity, we consider the time into the interval $[0,1]$. Note that more general supports as the one we find in the mesothelioma mortality data set follow the same type of derivations.

We observe n i.i.d. observations (X_i, Y_i) with density f on $\mathcal{I}_f = \{(x,y) : 0 \leq x, y \leq 1, x+y \leq 1\}$. We assume that f is of the form:

$$f(x,y) = c_f f_1(x) f_2(y),$$

where f_1 and f_2 are probability densities on $[0,1]$. The constant c_f is chosen such that

$$\int_{\mathcal{I}_f} c_f f_1(x) f_2(y) dx\ dy = 1. \tag{7}$$

We choose $g_1(x) = \int_0^{1-x} f(x,y) w(x,y)\ dy$, $g_2(y) = \int_0^{1-y} f(x,y) w(x,y)\ dx$ and some weight function $w(x,y) > 0$ and we assume that some estimators $\hat{g}_1$ and $\hat{g}_2$ are given for the functions g_1 and g_2. Note that if the weight function w fulfills $w(x,y) \equiv 1$ we get that g_1 is the marginal density of X and g_2 is the marginal density of Y. There are several options for the estimation of g_1 and g_2. A first option is a weighted kernel density estimator

$$\begin{aligned}
\hat{g}_1^{LC}(x) &= \hat{g}_{1,h_1}^{LC}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(X_i, x) w(X_i, Y_i), \\
\hat{g}_2^{LC}(x) &= \hat{g}_{2,h_2}^{LC}(y) = \frac{1}{n}\sum_{i=1}^{n} K_{h_2}(Y_i, y) w(X_i, Y_i),
\end{aligned}$$

where $K_h(u,v)$ is a kernel function that is equal to $K_h(u-v)$ for v in an appropriate interior of $[0,1]$ and equal to a boundary kernel otherwise. Here, $K_h(u) = h^{-1}K(h^{-1}u)$ is a kernel function with bandwidth h and probability density function K. One could also choose a local linear version where $\hat{g}_1^{LL}(x) = \hat{g}_{1,h_1}^{LL}(x) = \theta_{1,1}$ and $\hat{g}_2^{LL}(y) = \hat{g}_{2,h_2}^{LL}(y) = \theta_{2,1}$ with

$$\begin{aligned}
\begin{pmatrix} \theta_{1,1} \\ \theta_{1,2} \end{pmatrix} &= \arg\min_{\theta_{1,1},\theta_{1,2}} \Big\{ \lim_{w\to 0} \int \big[\hat{g}_{1,w}^{LC}(u) - \theta_{1,1} - \theta_{1,2}(u-x)\big]^2 \\
&\qquad \times K_{h_1}(u-x)\ du \Big\}, \\
\begin{pmatrix} \theta_{2,1} \\ \theta_{2,2} \end{pmatrix} &= \arg\min_{\theta_{2,1},\theta_{2,2}} \Big\{ \lim_{w\to 0} \int \big[\hat{g}_{2,w}^{LC}(u) - \theta_{2,1} - \theta_{2,2}(u-x)\big]^2 \\
&\qquad \times K_{h_2}(u-x)\ du \Big\}.
\end{aligned}$$

Another option would be that we choose $\hat{g}_1(x) = \int_0^{1-x} \hat{f}(x,y)w(x,y)\,dy$ and $\hat{g}_2(y) = \int_0^{1-y} \tilde{f}(x,y)w(x,y)\,dx$ where $\hat{f}$ and $\tilde{f}$ are estimators of f that may differ or that may be equal. Because of its relation to marginal integration we call this the marginal integration estimator. In our data example and in our simulations we used $\hat{f}(x,y) = \tilde{f}(x,y) = \hat{f}_{h;\mathcal{I}_f}(x,y)$ where $\hat{f}_{h;\mathcal{I}_f}$ is the local linear estimator defined in Subsection 2.2. With this estimator we get the following estimator of g_1 and g_2

$$\hat{g}_1^{MI}(x) = \int_0^{1-x} \hat{f}_{h;\mathcal{I}_f}(x,y)w(x,y)\,dy, \tag{8}$$

$$\hat{g}_2^{MI}(y) = \int_0^{1-y} \hat{f}_{h;\mathcal{I}_f}(x,y)w(x,y)\,dx. \tag{9}$$

In our first theorem we do not assume that any of these three pilot estimators is chosen. We only use the main assumption is that the estimators $\hat{g}_1$ and $\hat{g}_2$ allow the following expansions

$$\frac{\hat{g}_1(x) - g_1(x)}{g_1(x)} = O_P(\varepsilon_n), \tag{10}$$

$$\frac{\hat{g}_2(y) - g_2(y)}{g_2(y)} = O_P(\varepsilon_n), \tag{11}$$

uniformly for $0 \le x, y \le 1$, where ε_n is some sequence converging to 0.

If f allows two derivatives kernel density estimators of g_1 and g_2 have bias terms of order h^2 and a variance of order $(nh)^{-1}$ with $h = h_1$ or $h = h_2$, respectively. This holds in the interior $[0,1]$. At the boundary we get a variance of order $(nh^2)^{-1}$ and bias terms of order h or h^2, depending on the chosen kernel estimator. At the boundary the bias and variance terms are balanced by bandwidth choices of order $n^{-1/6}$ or $n^{-1/4}$. This results in a rate of convergence of order $n^{-1/3}$ or $n^{-1/4}$, respectively. The uniform expansion then holds with $\varepsilon_n = n^{-1/3}(\log n)^{1/2}$. Local linear kernel density estimators with multiplicative bias correction have bias terms of order h^4, under higher order smoothness assumptions. With a bandwidth choice of order $n^{-1/10}$ this gives estimators with point wise rate of convergence $n^{-2/5}$. Here we get that the expansion holds with $\varepsilon_n = n^{-2/5}(\log n)^{1/2}$.

Let define

$$\mathcal{F}(c, r_1, r_2)(x,y) = \begin{pmatrix} c\; r_1(x)\frac{1}{g_1(x)} \int_0^{1-x} r_2(v)w(x,v)dv - 1 \\ c\; r_2(y)\frac{1}{g_2(y)} \int_0^{1-y} r_1(u)w(u,y)du - 1 \end{pmatrix}. \tag{12}$$

Our estimators $\hat{c}_f$, $\hat{f}_1$ and $\hat{f}_2$ of c_f, f_1 and f_2 are given as solution of the equation

$$\hat{\mathcal{F}}(\hat{c}_f, \hat{f}_1, \hat{f}_2) = 0$$

under the constraint $\int_0^1 \hat{f}_1(u)du = 1$ and $\int_0^1 \hat{f}_2(v)dv = 1$. Here $\hat{\mathcal{F}}$ denotes the sample analogue to $\mathcal{F}$ in (12). In Subsection 2.2 we have described an iterative backfitting

algorithm to solve the equation $\hat{\mathcal{F}}(\hat{c}_f, \hat{f}_1, \hat{f}_2)$. There we used the choice $\hat{g}_1 = \hat{g}_1^{MI}$ and $\hat{g}_2 = \hat{g}_2^{MI}$, see (8)–(9). Later, we will make use of the fact that the functional (12) is approximately equal to

$$\mathcal{F}(c, r_1, r_2)(x, y) = \begin{pmatrix} c\ r_1(x)\frac{1}{g_1(x)} \int_0^{1-x} r_2(v) w(x, v) dv - 1 \\ c\ r_2(y)\frac{1}{g_2(y)} \int_0^{1-y} r_1(u) w(u, y) du - 1 \end{pmatrix}.$$

For $(\delta_0, \delta_1(\cdot), \delta_2(\cdot)) \in \mathcal{M} = \{(\mu_0, \mu_1, \mu_2) : \int_0^1 f_1(u)\mu_1(u)du = 0 \text{ and } \int_0^1 f_2(v)\mu_2(v)dv = 0\}$ we put

$$\mathcal{G}_0(\delta_0, \delta_1, \delta_2)(x, y) = \mathcal{F}(c_f(1 + \delta_0), f_1(1 + \delta_1), f_2(1 + \delta_2))(x, y).$$

The derivative of the functional at $\delta_0 = 0, \delta_1 = 0, \delta_2 = 0$ in direction $(\mu_0, \mu_1(\cdot), \mu_2(\cdot)) \in \mathcal{M}$ is equal to

$$\begin{aligned} \mathcal{G}_0'(\mu_0, \mu_1, \mu_2)(x, y) &= \begin{pmatrix} \mu_0 + \mu_1(x) \\ \mu_0 + \mu_2(y) \end{pmatrix} + \mathcal{H}_0(\mu_0, \mu_1, \mu_2)(x, y), \\ \mathcal{H}_0(\mu_0, \mu_1, \mu_2)(x, y) &= \begin{pmatrix} g_1(x)^{-1} \int_0^{1-x} \mu_2(v) f(x, v) w(x, v) dv \\ g_2(y)^{-1} \int_0^{1-y} \mu_1(u) f(u, y) w(u, y) du \end{pmatrix}. \end{aligned}$$

We denote the inverse of this functional by ${\mathcal{G}_0'}^{-1}$. We will show below that the inverse exists under our assumptions. We now define $\tilde{c}, \tilde{f}_1(x)$ and $\tilde{f}_2(y)$ by

$$\begin{pmatrix} \frac{\tilde{c}_f - c_f}{c_f} \\ \frac{\tilde{f}_1(x) - f_1(x)}{f_1(x)} \\ \frac{\tilde{f}_2(y) - f_2(y)}{f_2(y)} \end{pmatrix} = {\mathcal{G}_0'}^{-1} \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{pmatrix} (x, y),$$

where

$$\begin{aligned} \tilde{\mu}_1(x) &= -\frac{\hat{g}_1(x) - g_1(x)}{g_1(x)}, \\ \tilde{\mu}_2(y) &= -\frac{\hat{g}_2(y) - g_2(y)}{g_2(y)}. \end{aligned}$$

For the marginal integration estimator we use the following definition for the choices $x = 1$ and $y = 1$: $\tilde{\mu}_1(1) = -\frac{\hat{f}(1,0) - f(1,0)}{f(1,0)}$, and $\tilde{\mu}_2(1) = -\frac{\hat{f}(0,1) - f(0,1)}{f(0,1)}$.

Our first result shows that $\tilde{c}_f, \tilde{f}_1(x)$ and $\tilde{f}_2(y)$ provide a first order approximation to $\hat{c}_f, \hat{f}_1(x)$ and $\hat{f}_2(y)$.

Theorem 1 *Let us assume that assumptions (7)–(11) apply and that $w(x, y)$ and $f(x, y)$ are bounded from below and from above for $0 \leq x, y \leq 1$, $x + y \leq 1$. Then*

with probability tending to one, there exists a solution $\hat{c}_f$, $\hat{f}_1$ *and* $\hat{f}_2$ *of the equation* $\hat{\mathcal{F}}(\hat{c}_f, \hat{f}_1, \hat{f}_2) = 0$ *with*

$$\begin{aligned} |\hat{c}_f - \tilde{c}_f| &= O_P\left(\varepsilon_n^2\right), \\ \sup_{0\le x\le 1} |\hat{f}_1(x) - \tilde{f}_1(x)| &= O_P\left(\varepsilon_n^2\right), \\ \sup_{0\le y\le 1} |\hat{f}_2(y) - \tilde{f}_2(y)| &= O_P\left(\varepsilon_n^2\right). \end{aligned}$$

Note that the reminder term in the theorem suffices for the above discussed examples for appropriate choices of bandwidths.

For a further analysis of the asymptotics of $\hat{f}_1(x)$ and $\hat{f}_2(y)$ we assume that $\hat{g}_1(x)$ and $\hat{g}_2(y)$ allows the following decomposition

$$\begin{aligned} \hat{g}_1(x) &= g_1(x) + \hat{g}_1^A(x) + \hat{g}_1^B(x) + o_P(\varepsilon_n^*), \\ \hat{g}_2(y) &= g_1(x) + \hat{g}_2^A(y) + \hat{g}_2^B(y) + o_P(\varepsilon_n^*) \end{aligned}$$

uniformly for $0 \le x, y \le 1, x + y \le 1$, where $\hat{g}_1^B$ and $\hat{g}_2^B$ are some deterministic bias terms, and $\hat{g}_1^A$ and $\hat{g}_2^A$ are the asymptotic stochastic part of $\hat{g}_1$ or $\hat{g}_2$, respectively. Furthermore, ε_n^* is the rate of convergence of $\hat{f}_1$ and $\hat{f}_2$. With arguments as in the proof of Theorem 1 one can show that

$$\begin{aligned} |\hat{c} - c - \tilde{c}^B - \tilde{c}^A| &= o_P\left(\varepsilon_n^*\right), \\ \sup_{0\le x\le 1} |\hat{f}_1(x) - f_1(x) - \tilde{f}_1^A(x) - \tilde{f}_1^B(x)| &= o_P\left(\varepsilon_n^*\right), \\ \sup_{0\le y\le 1} |\hat{f}_2(y) - f_2(y) - \tilde{f}_2^A(y) - \tilde{f}_2^B(y)| &= o_P\left(\varepsilon_n^*\right), \end{aligned}$$

where for $j = A, B$

$$\begin{pmatrix} \frac{\tilde{c}^j}{c} \\ \frac{\tilde{f}_1^j(x)}{f_1(x)} \\ \frac{\tilde{f}_2^j(y)}{f_2(y)} \end{pmatrix} = \mathcal{G}_0'^{-1} \begin{pmatrix} \tilde{\mu}_1^j \\ \tilde{\mu}_2^j \end{pmatrix}(x, y),$$

with

$$\tilde{\mu}_1^j(x) = -\frac{\hat{g}_1^j(x)}{g_1^j(x)}, \qquad \tilde{\mu}_2^j(y) = -\frac{\hat{g}_2^j(y)}{g_2^j(y)}.$$

A close inspection of the operator $\mathcal{G}_0'^{-1}$, see also the proof of Theorem 1 shows that $(\tilde{c}^j, \tilde{f}_1^j(x), \tilde{f}_2^j(y))^\top$ is of the same order as $(\tilde{\mu}_1^j, \tilde{\mu}_2^j)^\top$. Thus we get that the estimators $\hat{f}_1$ and $\hat{f}_2$ have bias terms of the same order as $\hat{g}_1$ and $\hat{g}_2$ with explicit formulas given by the above equations. For the stochastic terms of $\hat{f}_1$ and $\hat{f}_2$ the rate depends on x or y, respectively. For fixed x or y with $0 \le x < 1$, $0 \le y < 1$ one gets for kernel

smoothers under regularity conditions that $\tilde{\mu}_1^A(x)$ and $\tilde{\mu}_2^A(y)$ has one-dimensional rate $(nh)^{-1/2}$. Moreover, one can show that for $0 \leq x < 1$, $0 \leq y < 1$ it holds that

$$\mathcal{G}_0'^{-1} \begin{pmatrix} \tilde{\mu}_1^A \\ \tilde{\mu}_2^A \end{pmatrix}(x,y) - \begin{pmatrix} \tilde{\mu}_1^A \\ \tilde{\mu}_2^A \end{pmatrix}(x,y) = o_P((nh)^{-1/2}),$$

compare also similar results for smooth backfitting in Mammen, Linton and Nielsen (1999). This allows a direct calculation of the pointwise asymptotic limit of $\hat{f}_1$ and $\hat{f}_2$ for x or y with $0 \leq x < 1$, $0 \leq y < 1$.

We summarize these findings in the following corollary, where the assumptions needed for a rigorous argumentation are also specified.

Corollary 2 *Suppose that the kernel K has support $[-1,1]$ abnd that it is symmetric and Lipschitz continuous. We suppose that $\int_0^1 K_h(u,x)\ du = 1$ for $x \in [0,1]$, that $|K_h(u,x)| \leq ch^{-1}$ for some $c > 0$, that $K_h(u,x) = 0$ for $|u-x| > h$, and that $K_h(u,x) = h^{-1}K(h^{-1}(u-x))$ for $c^*h \leq x \leq 1 - c^*h$ for some $c^* > 0$. Furthermore, we assume that the densities f_1 and f_2 are twice continuously differentiable and bounded away from 0 on $[0,1]$. The bandwidths fulfill $n^{1/5}h_j \to c_j$ for some constants $c_j > 0$. Choose $\hat{g}_j = \hat{g}_j^r$ for $j = 1,2$ with $r = LC$, $r = LL$ or $r = MI$.*

Then for $0 < x, y < 1$, $n^{2/5}f_1(x)^{-1}(\hat{f}_1(x) - f_1(x))$ and $n^{2/5}f_2(y)^{-1}(\hat{f}_2(y) - f_2(y))$ are asymptotically independent and have an asymptotic normal limit with mean $\beta_1(x)$ or $\beta_2(y)$, respectively, and with variance $\sigma_1^2(x)$ or $\sigma_2^2(y)$, respectively, where

$$\begin{aligned}
\beta_1(x) &= \frac{\tilde{f}_1^B(x)}{f_1(x)}, \\
\beta_2(y) &= \frac{\tilde{f}_2^B(y)}{f_2(y)}, \\
\sigma_1^2(x) &= \frac{\int_0^{1-x} w^2(x,v) f(x,v)\ dv}{\left[\int_0^{1-x} w(x,v) f(x,v)\ dv\right]^2}, \\
\sigma_2^2(y) &= \frac{\int_0^{1-y} w^2(u,y) f(u,y)\ du}{\left[\int_0^{1-y} w(u,y) f(u,y)\ du\right]^2}.
\end{aligned}$$

Note that $\sigma_1^2(x)$ or $\sigma_2^2(y)$ are the asymptotic variances of $\tilde{\mu}_1^A$ and $\tilde{\mu}_2^A$, respectively. Furthermore, because of

$$\left[\int_0^{1-x} w(x,v) f(x,v)\ dv\right]^2 \leq \left[\int_0^{1-x} w^2(x,v) f(x,v)\ dv\right] \left[\int_0^{1-x} f(x,v)\ dv\right],$$

we get that $\sigma_1^2(x)$ is minimal for the choice $w(x,y) \equiv 1$. The same holds for $\sigma_2^2(y)$. Thus, the weighting $w(x,y) \equiv 1$ is optimal.

Let us shortly mention an alternative estimator of f_1 and f_2. The idea of this estimator is based on the observation that $\log f(x,y) = \log c + \log f_1(x) + \log f_2(y)$. This motivates the estimator

$$\begin{aligned}(\hat{c}, \hat{f}_1, \hat{f}_2) &= \arg\min_{c,f_1,f_2} \int_{\mathcal{I}_f} \Big[\log \hat{f}(x,y) - \log c - \log f_1(x) - \log f_2(y)\Big]^2 \\ &\quad \times w(x,y)dx\ dy.\end{aligned}$$

This estimator can be calculated by the iterations

$$\begin{aligned}\log \hat{c}^{[l+1,a]} + \log \hat{f}_1^{[l+1]}(x) &= \frac{\int_0^{1-x} \Big[\log \hat{f}(x,y) - \log \hat{f}_2^{[l]}(y)\Big] w(x,y)dy}{\int_0^{1-x} w(x,y)dy}, \\ \log \hat{c}^{[l+1,b]} + \log \hat{f}_2^{[l+1]}(x) &= \frac{\int_0^{1-y} \Big[\log \hat{f}(x,y) - \log \hat{f}_1^{[l+1]}(x)\Big] w(x,y)dx}{\int_0^{1-y} w(x,y)dx}.\end{aligned}$$

A Proof of Theorem 1

In the proof we will make use of the following theorem.

Theorem 3 *(Newton-Kantorovich theorem). Suppose that there exist constants α, β, k, r and a value ξ_0 such that a functional $\mathcal{T}$ has a derivative $\mathcal{T}'(\xi)$ for $\|\xi - \xi_0\| \leq r$, $\mathcal{T}'$ is invertible,*

$$\begin{aligned}\|\mathcal{T}'(\xi_0)^{-1}\mathcal{T}(\xi_0)\| &\leq \alpha, \\ \|\mathcal{T}'(\xi_0)^{-1}\| &\leq \beta, \\ \|\mathcal{T}'(\xi) - \mathcal{T}'(\xi')\| &\leq k\|\xi - \xi'\|,\end{aligned}$$

for all ξ, ξ' with $\|\xi - \xi_0\| \leq r$, $\|\xi' - \xi_0\| \leq r$, $2k\alpha\beta < 1$ and $2\alpha < r$. Then $\mathcal{T}(\xi) = 0$ has a unique solution ξ^* in $\{\xi : \|\xi - \xi_0\| \leq 2\alpha\}$. Furthermore ξ^* can be approximated by Newtons iterative method

$$\xi_{l+1} = \xi_l - \mathcal{T}'(\xi_l)^{-1}\mathcal{T}(\xi_l).$$

This algorithm converges with geometric rate

$$\|\xi_{l+1} - \xi^*\| \leq \alpha 2^{-(l-1)} q^{2^l - 1}$$

with $q = 2\alpha\beta k < 1$.

For a discussion of this theorem see for example chapter 7 of Deimling (1985). We will apply this theorem with $\xi_0 = (\xi_0^0, \xi_0^1, \xi_0^2)^\top = (\frac{\tilde{c}_f - c_f}{c_f}, \frac{\tilde{f}_1 - f_1}{f_1}, \frac{\tilde{f}_2 - f_2}{f_2})^\top$ and $\mathcal{T} = \hat{\mathcal{G}}_{\xi_0}$ where for $\xi = (\xi^0, \xi^1, \xi^2)^\top$ with $\xi^0 \in \mathbb{R}$, and functions $\xi^1, \xi^2 : [0,1] \to \mathbb{R}$ we put

$$\hat{\mathcal{G}}_\xi(\delta_0, \delta_1, \delta_2)(x,y) = \hat{\mathcal{F}}(c_f(1+\xi^0+\delta_0), f_1(1+\xi^1+\delta_1), f_2(1+\xi^2+\delta_2))(x,y)$$

for $(\delta_0, \delta_1, \delta_2) \in \mathcal{M}$. We will use the Newton-Kantorovich theorem with the following norm $\|\xi\|_\infty = |\xi^0| + \sup_{0\le x\le 1} |\xi^1(x)| + \sup_{0\le y\le 1} |\xi^2(y)|$. The derivative of the functional $\hat{\mathcal{G}}_\xi$ at $\delta_0 = 0, \delta_1 = 0, \delta_2 = 0$ in direction $(\mu_0, \mu_1(\cdot), \mu_2(\cdot)) \in \mathcal{M}$ is equal to

$$\begin{aligned}
\hat{\mathcal{G}}'_\xi(\mu_0,\mu_1,\mu_2)(x,y) &= \begin{pmatrix} g_1(x)\hat{g}_1(x)^{-1}(\mu_0+\mu_1(x)) \\ g_2(y)\hat{g}_2(y)^{-1}(\mu_0+\mu_2(y)) \end{pmatrix} + \hat{\mathcal{H}}_\xi(\mu_0,\mu_1,\mu_2)(x,y), \\
\hat{\mathcal{H}}_\xi(\mu_0,\mu_1,\mu_2)(x,y) &= \hat{\mathcal{H}}_0(\mu_0,\mu_1,\mu_2)(x,y) \\
&\quad + \begin{pmatrix} \hat{g}_1(x)^{-1}\int_0^{1-x} d(x,v)f(x,v)w(x,v)dv \\ \hat{g}_2(y)^{-1}\int_0^{1-y} d(u,y)f(u,y)w(u,y)du \end{pmatrix}, \\
\hat{\mathcal{H}}_0(\mu_0,\mu_1,\mu_2)(x,y) &= \begin{pmatrix} \hat{g}_1(x)^{-1}\int_0^{1-x} \mu_2(v)f(x,v)w(x,v)dv \\ \hat{g}_2(y)^{-1}\int_0^{1-y} \mu_1(u)f(u,y)w(u,y)du \end{pmatrix}, \\
d(x,y) &= \mu_0[\xi^1(x) + \xi^2(y) + \xi^1(x)\xi^2(y)] \\
&\quad + \mu_1(x)[\xi^0 + \xi^2(y) + \xi^0\xi^2(y)] \\
&\quad + \mu_2(y)[\xi^0 + \xi^1(x) + \xi^0\xi^1(x)].
\end{aligned}$$

Furthermore, the operators $\mathcal{G}_\xi$, $\mathcal{G}'_\xi$ and $\mathcal{H}_0$ are defined as $\hat{\mathcal{G}}_\xi$, $\hat{\mathcal{G}}'_\xi$ and $\hat{\mathcal{H}}_0$ but with $\hat{g}_1$ and $\hat{g}_2$ replaced by g_1 and g_2.

For an application of the Newton-Kantorovich theorem, we first argue that $\hat{\mathcal{G}}'_\xi$ has a (uniformly in n) bounded inverse, with probability tending to one, for $\|\xi\|_\infty$ small enough. Note that $\hat{\mathcal{G}}_0$ converges to $\mathcal{G}_0$ and $\hat{\mathcal{G}}'_0$ converges to $\mathcal{G}'_0$ in operator norm, because of (10)–(11). At this stage we use the operator norm with respect to the norm $\|(\mu_0,\mu_1,\mu_2)\|_w^2 = \mu_0^2 + \|\mu_1\|_w^2 + \|\mu_2\|_w^2$ where $\|\cdot\|_w$ is the norm of the space $\mathrm{L}_2(P^w)$ for the probability measure P^w on $\{(x,y) : 0 \le x, y \le 1, x+y \le 1\}$ with density $\rho w(x,y)f(x,y)$. Here, ρ is a norming constant.

We first argue that $\mathcal{G}'_0$ has a bounded inverse. For this purpose we proceed similarly as in the proof of Lemma 1 in Mammen, Linton and Nielsen (1999). Suppose that $\mathcal{G}'_0(\mu_0,\mu_1,\mu_2)^\top = 0$. We will argue that this implies that $(\mu_0,\mu_1,\mu_2)^\top = 0$. We have that $\Pi_1\mu^+(x) = E^{P^w}[\mu^+(X,Y)|X=x] = 0$ and $\Pi_2\mu^+(y) = E^{P^w}[\mu^+(X,Y)|Y=y] = 0$ where $\mu^+(x,y) = \mu_0 + \mu_1(x) + \mu_2(y)$. We now consider Π_1, restricted to the set $R_y = \{r : r(x,y) \equiv r^{**}(y)$ for some function $r^{**} : [0,1] \to \mathbb{R}, \int_0^1 r^{**}(y)f_2(y)\,dy = 0, \int_0^1 r^{**}(y)^2 g_2(y)\,dy < \infty\}$. We will show below that for the kernel $\pi_1(x,y) = g_1^{-1}(x)g_2^{-1}(y)f(x,y)w(x,y)$ of $\Pi_1\big|_{R_y}$ it holds that

$$\int \pi_1^2(x,y)g_1(x)g_2(y)\,dx\,dy < \infty. \tag{13}$$

This implies that for some constant $C_H > 0$

$$\|\Pi_1(r)\|_w \leq C_H \|r^{**}\|_w \tag{14}$$

for $r \in R_y$ with $r(x,y) = r^{**}(y)$ and that $\Pi_1\big|_{R_y}$ is a compact operator. Furthermore, arguing as in Appendix A.4 of Bickel, Klaassen, Ritov and Wellner (1993), see also the proof of Lemma 1 in Mammen, Linton and Nielsen (1999), one gets that for some constant $c > 0$

$$\|m_1 + m_2\|_w \geq c \max\{\|m_1\|_w, \|m_2\|_w\}$$

for $m_1 \in R_x = \{r : r(x,y) \equiv r^*(x)$ for some function $r^* : [0,1] \to \mathbb{R}, \int_0^1 r^*(x) f_1(x)\, dx = 0, \int_0^1 r^*(x)^2 g_1(x)\, dx < \infty\}$ and $m_2 \in R_y$. Applied to $m_1 = \mu_1$ and $m_2 = \Pi_1 \mu_2$ this gives $\mu_1 = 0$ and $\Pi_1 \mu_2 = 0$, and thus $\mu_0 = 0$. By a symmetric argument we get that $\mu_2 = 0$. Thus $\mathcal{G}_0'$ is invertible. We also get that (14) holds with $C_H < 1$, see Theorem 2 (B) in Appendix A.4 of Bickel, Klaassen, Ritov and Wellner (1993). Thus all eigenvalues of the operator $\mathcal{H}_0$ are absolutely bounded by $C_H < 1$. We get that the operator norm of $\mathcal{G}_0^{-1} = (I + \mathcal{H}_0)^{-1}$ with respect to the norm $\|\cdot\|_w$ is bounded by $(1 - C_H)^{-1}$.

We now argue that the operator norm of $\mathcal{G}_0^{-1} = (I + \mathcal{H}_0)^{-1}$ is bounded with respect to the norm $\|\cdot\|_\infty$. We already know that $\mathcal{G}_0$ is bijective because the operator norm $\mathcal{G}_0^{-1} = (I + \mathcal{H}_0)^{-1}$ with respect to $\|\cdot\|_w$ is bounded. Thus, according to the bounded inverse theorem it suffices to show that $\mathcal{G}_0$ is bounded with respect to the norm $\|\cdot\|_\infty$. This can be seen by a simple calculation.

From the last fact we get that we have for our choice of ξ_0 that $\|\xi_0\|_\infty = O_P(\varepsilon_n)$. Using similar arguments as above and some approximation arguments one can show that the operator norm of $\hat{\mathcal{G}}_{\xi_0}^{-1}$ with respect to the norm $\|\cdot\|_\infty$ is bounded by a fixed constant, with probability tending to one. Thus we have verified the second condition of the Newton-Kantorovich theorem. Also by approximation arguments one can verify Lipschitz continuity of $\hat{\mathcal{G}}_{\xi_0}'$ (with probability tending to one), as it is required in the third assumption of the Newton-Kantorovich theorem.

We will show that

$$\|\hat{\mathcal{G}}_{\xi_0}(\frac{\tilde{c} - c}{c}, \frac{\tilde{f}_1 - f_1}{f_1}, \frac{\tilde{f}_2 - f_2}{f_2})\|_\infty = O_P(\varepsilon_n^2). \tag{15}$$

This implies the first assumption of the Newton-Kantorovich theorem. Thus we can apply the Newton-Kantorovich theorem. The statement of our theorem then follows from this theorem.

It remains to show (13) and (15). For the proof of (13) note that for some constant $C > 0$ it holds that

$$\int \pi_1^2(x,y) g_1(x) g_2(y)\ dx\ dy \leq C \int_0^1 \int_0^{1-y} (1-x)^{-1}(1-y)^{-1}\ dx\ dy$$
$$= -C \int_0^1 \ln(y)(1-y)^{-1}\ dy = -C\ \mathrm{Li}_2(1-y)\Big|_0^1 = C\ \mathrm{Li}_2(1) < \infty,$$

where Li_2 is the dilogarithm.

We now come to the proof of (15). We will show that

$$\|\hat{\mathcal{G}}_0(\frac{\tilde{c}-c}{c}, \frac{\tilde{f}_1 - f_1}{f_1}, \frac{\tilde{f}_2 - f_2}{f_2})\|_\infty = O_P(\varepsilon_n^2).$$

The proof of (15) follows by a slight extension of the arguments. Note that

$$\hat{\mathcal{G}}_0(\frac{\tilde{c}-c}{c}, \frac{\tilde{f}_1 - f_1}{f_1}, \frac{\tilde{f}_2 - f_2}{f_2}) = \hat{\mathcal{F}}(\tilde{c}, \tilde{f}_1, \tilde{f}_2)$$
$$= (\hat{\mathcal{F}} - \mathcal{F})(\tilde{c}, \tilde{f}_1, \tilde{f}_2) + [\mathcal{F}(\tilde{c}, \tilde{f}_1, \tilde{f}_2) - \mathcal{F}(c, f_1, f_2)].$$

By using Taylor expansions and crude bounds one can show that, up to terms of order $O_P(\varepsilon_n^2)$, the first term is equal to $-(\tilde{\mu}_1, \tilde{\mu}_2)^\top$ and the second term is equal to $(\tilde{\mu}_1, \tilde{\mu}_2)^\top$. This gives the desired result.

References

Antonio, K. and Plat, H. (2014). *Micro-level stochastic loss reserving in general insurance. Scand. Act. J.*, **7**, 649–669.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* New York:Springer-Verlag.

Carroll, R. J., Delaigle, A. and Hall, P. (2013). Unexpected properties of bandwidth choice when smoothing dicrete data for constructing a functional data classifier. *Ann. Stat.* **41**, 2739–2767.

Clayton, D. and Schifflers, E. (1987). Models for temporal variation in cancer rates. II: Age-period-cohort models. *Stat. Med.*, **6**, 469–481.

Deimling, N. (1985). *Nonlinear Functional Analysis.* New York:Springer-Verlag.

Haberman, S. and Renshaw, A. (2012). Parametric mortality improvement rate modelling and projecting. *Insur. Math. Econ.*, **50**(3), 309–333.

Hatzopoulos, P. and Haberman, S. (2013). Common mortality modelling and coherent forecasts - an empirical analysis of worldwide mortality data. *Insur. Math. Econ.*, **52**(2), 320–337.

Kuang, D., Nielsen B. and Nielsen J. P. (2011). Forecasting in an extended chain-ladder-type model. *J. Risk. Insur.*, **78**(2), 345–359.

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

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.

Lei, J., Robins, J. and Wasserman, L. (2013). Distribution-free prediction sets. *J. Am. Stat. Assoc.*, **108**, 278–287.

Lu, L., Jiang, H. and Wong, W. H. (2013). Multivariate density estimation by Bayesian Sequential Partitioning. *J. Am. Stat. Assoc.*, DOI:10.1080/01621459.2013.813389.

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

Mammen, E. and Nielsen, J. P. (2003). Generalised Structured Models. *Biometrika*, **90**, 551–566.

Martínez-Miranda, M. D., Nielsen, B. and Nielsen, J. P. (2014). Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. *J.R. Statist. Soc.*. DOI: 10.1111/rssa.12051.

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

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

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

Martínez-Miranda, M. D., Nielsen, J. P. and Verrall, R. (2013b). Double Chain Ladder and Bornhuetter-Fergursson. *N. Am. Actuar. J.*, 17, 101–113.

Martínez-Miranda, M. D., Nielsen, J. P. Verrall, R. and Wüthrich, M. (2013c). Double Chain Ladder, Claims Development Inflation and Zero Claims. *Scand. Actuar. J.*. DOI:10.1080/03461238.2013.823459.

Müller, H.G. and Stadtmüller, U. (1999). Multivariate boundary kernels and a continuous least squares principle. *J.R. Statist. Soc.* B, **61**, 439–458.

Nielsen, J. P. (1999). Multivariate Boundary Kernels from Local Linear Estimation *Scand. Act. J.* **1**, 93–95.

Nielsen, J. P. and Linton, O. (1998). An optimization interpretation of integration and backfitting estimators for separable non-parametric models. *J.R. Statist. Soc.* B, **60**, 217–222.

Panaretosa, V. M. and Konis, K. (2012). Nonparametric Construction of Multivariate Kernels. *J. Am. Stat. Assoc.*, **107**, 1085–1095.

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

Pigeon, M., Antonio, K., Denuit, M. (2014). Individual loss reserving using paid-incurred data. *Insur. Math. Econ.*, **58**, 121–131.

Pigeon, M., Antonio, K., Denuit, M. (2013). Individual loss reserving with the multivariate skew normal framework. *Ast. Bull.*, **43**, 399–428.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL: http://www.R-project.org.

Verrall, R., Nielsen, J. P and Jessen, A. (2010). Prediction of RBNS and IBNR Claims Using Claim Amounts and Claim Counts. *Astin Bull.*, **40**(2), 871–887.

Wüthrich, M. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*. Wiley Finance.

Xiao, L., Li, Y. and Ruppert, D. (2013). Fast Bivariate P-splines: the sandwich smoother. *J.R. Statist. Soc.* B, **75**, 577–599.

Zhang, X., Park, B. U. and Wang, J. (2013). Time-Varying Additive Models for Longitudinal Data. *J. Am. Stat. Assoc.*, **108**, 983–998.