



City Research Online

City St George's, University of London

Citation: Hiabu, M. (2016). In-sample forecasting: structured models and reserving. (Unpublished Doctoral thesis, City, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://city-ref-test.eprints-hosting.org/id/eprint/17082/>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

CITY, UNIVERSITY OF LONDON

DOCTORAL THESIS

In-sample Forecasting: Structured Models and Reserving

Supervisors:

Prof. Jens P. NIELSEN

Dr. María D. MARTÍNEZ MIRANDA

Prof. Enno MAMMEN

Author:

Munir HIABU

Examiners:

Prof. Richard J. VERRALL

Prof. Oliver B. LINTON

*A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy*

in the

Faculty of Actuarial Science and Insurance
Cass Business School



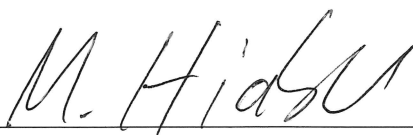
December 2016

Declaration of Authorship

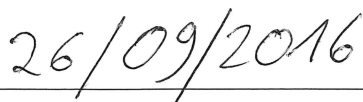
I, Munir HIABU, declare that this thesis titled, 'In-sample Forecasting: Structured Models and Reserving' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only simple copies made for study purpose, subject to the normal conditions of acknowledgement.

Signed:



Date:



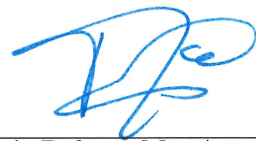
Co-authors declaration

Munir Hiabu was the driving force behind the two papers “In-sample forecasting with local linear survival densities” and “Smooth backfitting of multiplicative structured hazards”. However, the practical usefulness, the computational accuracy and the theoretical insight were closely supervised and monitored by the three supervisors. Jens Perch Nielsen supervising the practical usefulness, María Dolores Martínez Miranda supervising the computational aspects and Enno Mammen supervising the theoretical insight. In terms of working hours, most of the work was done by Munir Hiabu. But the help and supervision from his three supervisors were of course important for the final product.

Signatures:



(Jens Perch Nielsen)



(María Dolores Martínez Miranda)



(Enno Mammen)

CITY UNIVERSITY LONDON

Abstract

Faculty of Actuarial Science and Insurance

Cass Business School

Doctor of Philosophy

In-sample Forecasting: Structured Models and Reserving

by Munir HIABU

In most developed countries, the insurance sector accounts for around eight percent of the GDP. In Europe alone the insurers liabilities are estimated at around €900 billion. Every insurance company regularly estimates its liabilities and reports them, in conjunction with statements about capital and assets, to the regulators. The liabilities determine the insurers solvency and also its pricing and investment strategy. The new EU directive, Solvency II, which came into effect in the beginning of 2016, states that those liabilities should be estimated with ‘realistic assumption’ using ‘relevant actuarial and statistical methods’. However, modern statistics has not found its way in the reserving departments of today’s insurance companies. This thesis attempts to contribute to the connection between the world of mathematical statistics and the reserving practice in general insurance. As part of this thesis, it is in particular shown that today’s reserving practice can be understood as a non-parametric estimation approach in a structured model setting. The forecast of future claims is done without the use of exposure information, i.e., without knowledge about the number of underwritten policies. New statistical estimation techniques and properties are derived which are build from this motivating application.

Acknowledgements

Many people have contributed directly or indirectly to the work presented in this thesis.

I would especially like to thank:

- ... Jens P. Nielsen for having been a truly unique (that's good!) supervisor introducing me to the world of insurance and research in general. I am especially thankful for being taught how important it is to have a clear idea about the specific applicability and usefulness when developing statistical models.
- ... María D. (Lola) Martínez Miranda for her constant support in computational issues, and always having motivating words.
- ... Enno Mammen for being an incredible teacher with many fruitful discussions, and always having an analogue in the regression world at hand.
- ... Julian Zell for helping me in many IT and coding problems I have encountered.
- ... Andrew Hunt and Mikael Homanen for proofreading and improving lots of drafts of this thesis.
- ... the PhD department and the Faculty of Actuarial Science and Insurance of Cass Business School for providing an supportive environment.
- ... the Research Training Group (RTG) 1953 "Statistical Modeling of Complex Systems and Processes" in Mannheim/Heidelberg and the statistics group at the University of Heidelberg.

Finally, I would like to thank all the people who morally and socially supported me throughout my PhD:

- ... my family: my parents Settelbenat and Abdu, and my sisters Anisa and Wintana.
- ... all the fellow PhD students at Cass, especially Andrés Villegas, André Silva, Andrew Hunt, Anran Chen, Carolin Margraf, Christopher Benoit, Emanuel Kastl, Katerina Papoutsi, Kevin Curran, Mikael Homanen, Orkun Saka, Riccardo Borghi and Robert Schumacher.

... and the people in Mannheim/Heidelberg, especially Alexander Kreiß, Artem Makarov, Claudia Strauch, Dominic Edelmann, Jan Johannes, Johannes Dueck, Karl Gregory, Martin Wahl, Mehmet Madensoy, Nopporn Thamrongrat, Stefan Richter and Xavier Loizeau.

Contents

Declaration of Authorship	ii
Co-author declaration	iii
Abstract	iv
Acknowledgements	v
Contents	vii
1 Introduction	1
2 In-sample forecasting with local linear survival densities	7
2.1 Introduction	9
2.2 In-sample forecasting and related work	10
2.3 Model	12
2.4 Local linear density estimator in reversed time	14
2.5 Bandwidth selection in reversed time	17
2.5.1 Cross-validation and do-validation	17
2.5.2 Weighting for application in claims reserving	18
2.6 Asymptotic properties of weighted combinations of indirect cross-validation	20
2.7 Illustration	24
2.8 Simulation study	27
2.9 Concluding Remarks	29
Appendix 2.A Asymptotic properties and proofs	30
Appendix 2.B Discretization	36
3 Smooth backfitting of multiplicative structured hazards	41
3.1 Introduction	43
3.2 Aalen’s multiplicative intensity model	45
3.2.1 Left truncation and right censoring time as covariates	45
3.3 Estimation	47
3.3.1 Unstructured estimation of the hazard	47
3.3.2 Structured estimator by solution weighted minimization	48
3.4 Properties of the estimator	51
3.5 Application: Outstanding loss liabilities	57

Appendix 3.A	Bandwidth selection	62
Appendix 3.B	Proofs	63
3.B.1	Proof of Proposition 3.1	63
3.B.2	Proof of Proposition 3.2	66
3.B.3	Proof of Theorem 3.3	66
Appendix 3.C	Discrete data	70
4	On the relationship between classical chain ladder and granular reserving	73
4.1	Introduction	75
4.2	The continuous model	78
4.2.1	Model formulation	78
4.2.2	Estimation in the continuous framework	81
4.3	Discretization of the continuous model	86
4.3.1	The model	86
4.3.2	The histogram estimator and chain ladders development factors	90
4.3.3	Local polynomial estimator	92
4.4	Simulation study	93
4.5	Concluding remarks	94
Appendix 4.A	Computational complexity	99
Appendix 4.B	A martingale CLT	100
Appendix 4.C	Proofs	101
4.C.1	Proof of Proposition 4.1	101
4.C.2	Proof of Proposition 4.5	102
4.C.3	Proof of Proposition 4.7	102
4.C.4	Proof of Proposition 4.4	103
5	Continuous chain-ladder with paid data	109
5.1	Introduction	111
5.2	Model formulation	113
5.3	Reserving and In-sample forecasting	114
5.4	Local polynomial estimation	116
5.5	Concluding remarks	119
Appendix 5.A	Proofs	120
5.A.1	Proof of Proposition 5.1	120
5.A.2	Estimation of the weighted survival function	120
5.A.3	Proof Proposition 5.2	122
5.A.4	Proof of Proposition 5.3	123

1

Introduction

The idea behind insurance is simple. If the risk of rare events is spread from the individual to a larger community, everyone feels, and in fact is, safer. This basic idea makes insurance companies useful and important. Insurance companies charge premiums to their customers in exchange for covering their risk. They are multi-billion dollar entities that invest their clients' premiums into the financial market and the real economy. Overall, the insurance sectors in most developed economies earn premiums amounting to approximately eight percent of their GNPs (ESRB, [2015](#); FIO, [2015](#)).

The size of insurance sectors, while reflecting their importance, makes economic systems vulnerable to them. The failure of just one large insurance company does not only harm its policyholders, but could also disrupt the financial market and the real economy. Therefore, the insurance market remains highly regulated. Since early 2016 in the EU, regulation has been set by the Solvency II directive. Its main content is the new 'Solvency II balance sheet', which lists the insurers assets, liabilities, and capital. The largest item on the general insurers balance sheet is often liabilities, which determine the solvency and investment strategy of the company. Liabilities are composed of the future costs for reported claims that have not been settled yet, and also for incurred claims that have not yet been reported. In Europe, liabilities amount to approximately €900 billion (IE, [2013](#)). With regards to the reserve, i.e., the best estimate of the liabilities, Article 77 'Calculation of technical provisions' states:

“The calculation of the best estimate shall be based upon up-to-date and credible information and realistic assumptions and be performed using adequate, applicable and relevant actuarial and statistical methods.”

While everyone should agree on the content of said article, this statement is actually quite vague. Without detailed guidelines, it is perhaps not too surprising that modern statistics has not found its way in the reserving departments of today’s insurance companies.

The reasons for these are manifold and the discussion is not a subject of this thesis. However, a main reason for this absence might be the lack of sufficient exchange and interaction between actuaries and statisticians and also between practicing and academic actuaries. The first is most evident from the fact that statisticians rarely publish in actuarial journals and vice versa.

This thesis aims to contribute to the connection between the world of mathematical statistics with the reserving practice in general insurance. Chapter 2 & 3 present new contributions in mathematical statistics, but explain how those are useful for the actuarial field of reserving. Chapter 4 is written for an actuarial audience; it elaborates how the mathematical objects of Chapter 2 & 3 can be explained as traditional objects known to actuaries.

The connection between the actuarial and statistical world is done by introducing the notion of in-sample forecasting, which will be repeatedly explained throughout all chapters. Consider the estimation problem of a two-dimensional function, either a density or a hazard, supported on a rectangle. If these functions are known, then full information about the distribution, and, therefore, uncertainty is available. The difficulty is that observations are only available on a subspace of that rectangle. The reason is that points on the rectangle represent dates, some of which correspond to the future which is not known at time of data collection. Without further parametric assumptions, this problem can only be solved if the univariate components of the density or hazard are separable. This separability assumption is known as structured model in non-parametric statistics. Under certain assumptions on the subspace, this is already enough to estimate the original functions with support on the full rectangle and in particular to get information about the ‘future part’ of the rectangle.

These considerations are interesting from both statistical and actuarial perspectives. From a statistical point of view, in-sample forecasting allows structured models to get another justification and application besides their traditional motivation of visualization in higher dimensions and solution of the curse of dimensionality. Note that in-sample forecasting has potential to be applied to other fields as well, as is for instance already done for asbestos mortality forecasting, see Mammen, Martínez-Miranda, and Nielsen (2015).

More importantly, from an actuarial perspective, it is shown (Chapter 4) that what actuaries are doing today when setting reserves can be understood in a non-parametric structured model setting: A one dimensional component of the two dimensional hazard is shown to be the ‘actuarial’ development factors. These factors are often the central object in the reserving departments of general insurance companies. Hence, actuaries have a deep understanding of this function with respect to different situations and businesses. Via the identification to a hazard function one now gets a better understanding of the statistical estimation: In Chapter 4 & 5, it is for the first time discussed which assumptions on the data generating process have to be made for the classical chain ladder estimation technique to be consistent. Under these assumptions, we develop improved estimators of the development factors, based on the theory of Chapter 2. It also turns out that the assumptions are often quite restrictive. In Chapter 3, we develop a new statistical estimation technique for relaxed assumptions.

This thesis is composed of four self-contained chapters stemming from four research papers. The first two papers were developed in collaboration with my supervisors as co-authors who are mentioned in the beginning of those chapters. Being self-contained, each chapter has its own introduction, notation, conclusions and references.

A brief description of the contributions of each chapter follows.

Chapter 2: In-sample forecasting with local linear survival densities

In this chapter, we introduce in-sample forecasting via a counting process approach and describe how to estimate the resulting survival densities with local linear smoothers motivated by a least squares criterium. For that, we also provide:

- a class of data-driven bandwidth selectors with full asymptotic theory,

- a weighting in the bandwidth selection when the task is in-sample forecasting of reserves in general insurance,
- an application and simulation study in the field of reserving in general insurance.

Chapter 3: Smooth backfitting of multiplicative structured hazards

This chapter generalizes the setting of Chapter 2. The assumption of independent components is relaxed. This makes the one dimensional hazard of Chapter 2 multivariate. We introduce smooth backfitting, known from regression (Mammen, Linton, and Nielsen, 1999), to hazards in a survival analysis setting. Smooth backfitting efficiently estimates the one-dimensional components of a multiplicative separable hazard. Given a local linear pilot estimator of the d -dimensional hazard, the backfitting algorithm is motivated from a least squares criterion and converges to the closest multiplicative function. We show that the one dimensional components are estimated with a one-dimensional convergence rate, and hence do not suffer from the curse of dimensionality. The setting is very similar to Linton, Nielsen, and Van de Geer (2003), but has two significant improvements. First, our approach works without the use of higher order kernels. With them, one can theoretically derive nearly $n^{-1/2}$ -consistency (with growing order), but they often fail to show good performance in practice. Second, the support of the multivariate hazard does not need to be rectangular. In the provided in-sample forecasting application of reserving in general insurance, the support is indeed triangular.

Chapter 4: On the relationship between classical chain ladder and granular reserving

This chapter explains how the contributions of Chapters 2 & 3 are related to today's reserving practice in general insurance. It is shown that the one dimensional hazard has a one to one correspondence to the 'development factors' originating from the most widely used reserving method, chain-ladder. This is done by modeling the claims data used in the chain-ladder technique as arising from individual *iid* observations. As a side result, we also show that the level of aggregation has an effect on the underlying assumptions and often is a bias-variance trade-off.

Chapter 5: Continuous chain-ladder with paid data

The theory of Chapter 4 can only be used to forecast claim numbers. This chapter extends that model to be suitable for claim amounts by introducing a methodology to estimate a cost weighted density. The message is that practitioners can essentially do the same whether the data are claim counts or claim amounts. This corresponds to the fact that in practice, the chain-ladder method is used in both cases. However, when claim amounts are considered, this comes with the cost of an additional assumption, which is that the influences of development delay and underwriting date on the claim severity are independent to each other.

References

- ESRB (2015). *ESRB report on systemic risks in the EU insurance sector*. Tech. rep. European Systemic Risk Board.
- FIO (2015). *Annual Report on the Insurance Industry*. Tech. rep. Federal Insurance Office – US Department of the Treasury.
- IE (2013). *Funding the future – Insurers’ role as institutional investors*. Tech. rep. Insurance Europe and Oliver Wyman.
- Linton, O. B., J. P. Nielsen, and S. Van de Geer (2003). “Estimating Multiplicative and Additive Hazard Functions by Kernel Methods”. In: *Ann. Stat.* 31, pp. 464–492.
- Mammen, E., O. B. Linton, and J. P. Nielsen (1999). “The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions”. In: *Ann. Stat* 27, pp. 1443–1490.
- Mammen, E., M. D. Martínez-Miranda, and J. P. Nielsen (2015). “In-sample forecasting applied to reserving and mesothelioma”. In: *Insurance Math. Econom.* 61, pp. 76–86.

2

In-sample forecasting with local linear survival densities

This chapter has been accepted for publication in *Biometrika*. As of today, there is no publication date.

It is joint work with my supervisors Jens P. Nielsen, E. Mammen and Maria D. Martínez Miranda.

Previous versions of this chapter or parts of it were presented at the following conferences:

- June 2014. 2nd International Society for NonParametric Statistics (ISNPS) Conference, Cadiz, Spain.
- January 2015. 2nd Perspectives on Actuarial Risks in Talks of Young researchers (PARTY), Liverpool, UK.
- June 2015 19th International Congress on Insurance: Mathematics and Economics (IME), Liverpool, UK.
- August 2015 50th Actuarial Research Conference (ARC), Toronto, Canada.
- March 2016 12th German Probability and Statistics Days 2016, Bochum, Germany.

In-sample forecasting with local linear survival densities

M. Hiabu^a, E. Mammen^b, María D. Martínez Miranda^a, Jens P. Nielsen^a

^a*Cass Business School, City, University of London, United Kingdom*

^b*Institute for Applied Mathematics, Heidelberg University, Germany*

Abstract

In this paper, in-sample forecasting is defined as forecasting a structured density to sets where it is unobserved. The structured density consists of one-dimensional in-sample components that identify the density on such sets. We focus on the multiplicative density structure, which has recently been seen as the underlying structure of non-life insurance forecasts. In non-life insurance the in-sample area is defined as one triangle and the forecasting area as the triangle that added to the first triangle produces a square. Recent approaches estimate two one-dimensional components by projecting an unstructured two-dimensional density estimator onto the space of multiplicatively separable functions. We show that time-reversal reduces the problem to two one-dimensional problems, where the one-dimensional data are left-truncated and a one-dimensional survival density estimator is needed. This paper then uses the local linear density smoother, with weighted cross-validated and do-validated bandwidth selectors. Full asymptotic theory is provided, with and without time reversal. Finite sample studies and an application to non-life insurance are included.

Keywords: Aalen's multiplicative model; Cross-validation; Do-validation; Density estimation; Local linear kernel estimation; Survival data.

2.1 Introduction

This paper develops a dimension-reduction procedure in order to forecast an age-cohort structure. Our motivating example is taken from non-life insurance where the estimation of outstanding liabilities involves an age-cohort model. In non-life insurance such a structure is called chain ladder: cohorts are based on the year of underwriting the insurance policy and age is the development of claims. Age-cohort and chain ladder models have often been formulated as discrete models aggregating observations in months, quarters or years. Martínez-Miranda et al. (2013) identified the chain ladder method as a structured histogram in the vocabulary of non-parametric smoothing, and suggested replacing the structured histogram smoothers by continuous kernel smoothers, which are more efficient.

We assume that our data are sampled from two independent distributions, one for cohort and one for age, but are truncated if cohort plus age is greater than the calendar time of data collection. Future observations remain unobserved, and the forecasting exercise is to predict them. Visualized, the historical data belong to a triangle and the forecasting exercise is to predict the densities on the triangle that added to the first completes a square. We call this forecasting structure in-sample forecasting, because information on the two relevant densities of the multiplicative structure is indeed in the sample. The independence assumption for the unfiltered data will be discussed in the next section. Our model is thus that we have independent and identically distributed truncated observations sampled from the two-dimensional random variable, (X, Y) , with values on the triangle $\mathcal{I} = \{(x, y) : x + y \leq T, x, y \geq 0\}$, $T \in \mathbb{R}_+$. These observations are truncated from the complete set with support on the square $[0, T]^2$. We wish to make in-sample forecasts of the density with support on the second triangle, $\mathcal{J} = [0, T]^2 \setminus \mathcal{I}$, which completes the square. Furthermore, for unfiltered (X, Y) , the joint density, f , has support on the whole square, $[0, T]^2$ and is multiplicative, i.e., $f(x, y) = f_1(x)f_2(y)$. Given this multiplicative structure, the truncated observations provide in-sample information about the density in the forecasting triangle. Estimating only the survival functions or cumulative hazards is not enough when integrating the forecasts considered in this paper, since $\mathcal{J}$ is non-rectangular.

We estimate the two multiplicative components without first having to estimate the two-dimensional density. This is possible due to the reinterpretation of the forecasting aim as two distinct one-dimensional right-truncated density estimation problems, which can be solved in a counting process framework. It is well-known that intractable right-truncation can be replaced by more tractable left-truncation by reversing the time scale; see for example Ware and DeMets (1976) and Lagakos, Barraj, and De Gruttola (1988). The time-reversal approach requires estimates of the survival densities, for which we use the local linear survival kernel density estimator of Nielsen, Tanggaard, and Jones (2009) with cross-validated or do-validated bandwidths, see Mammen et al. (2011), Gámiz et al. (2013) and Gámiz et al. (2016). We introduce full asymptotic theory of the corresponding bandwidth selectors with and without weighting, and with and without time reversal. Reducing the forecasting to a one-dimensional problem enables us to introduce a new measure of forecasting accuracy that is equivalent to an importance-weighted loss function. The bandwidths chosen by this new measure focus on the areas of the one-dimensional functions that are most important for the forecast. When estimating outstanding liabilities, least information is available for the most recent years but they are the most important ones to estimate accurately. The new approach leads to larger bandwidths than classical goodness-of-fit loss measures. This better reflects the nature of the underlying problem, and improves forecasting accuracy.

2.2 In-sample forecasting and related work

While we use counting process theory in this paper to reduce the number of dimensions, the problem can also be formulated via independent stochastic variables X and Y and their density on a triangular support; see Martínez-Miranda et al. (2013), Mammen, Martínez-Miranda, and Nielsen (2015) and Lee et al. (2015), where in the two latter papers the triangular support is one special case. The independence assumption of X and Y have direct analogues to survival analysis. The density f_1 of X measures exposure, i.e., the number of individuals at risk, while the density f_2 of Y corresponds to duration. While classical counting process theory in survival analysis operates with observed exposure, in-sample forecasting estimates f_1 and does not need observed exposure. This has the advantage of operating on less data. Simple model assumptions are often preferable

when forecasting, therefore in-sample forecasting might be preferable even in situations where more data, including exposure, is available.

For example when reserves for outstanding liabilities are to be estimated in insurance companies, there is usually no follow-up data of individuals in the portfolio available and reported claims, categorized in different businesses and other baseline characteristics, are the only records. The reason that insurers do not use classical biostatistical exposure data, i.e., they do not follow every underwritten policy, might be because of the bad quality and complexity of such exposure data with many potential causes of failure which heavily affect the actual cost of a claim. When claim numbers are considered, then X is the underwriting date of the policy, and Y is the time between underwriting date and the report of a claim, the reporting delay. Truncation occurs when $X + Y$ is smaller than the date of data collection. The mass of the unobserved, future triangle, $\mathcal{J}$, then corresponds to the proportion of claims underwritten in the past which are not reported yet. The assumption of a multiplicative density means that the reporting delay does not depend on the underwriting date. Thus, calendar time effects like court rulings, emergence of latent claims, or changes in operational time cannot be accommodated in the model before further generalisations of the model are introduced. Nevertheless we restrict our discussion to the multiplicative model for several reasons. It has its justification as baseline for generalisations in many directions. It also approximates the data structure well enough in many applications. We will come back to this point when discussing our data example. The relevance of the multiplicative model also lies in the fact that it helps to understand discrete related versions that are used every day in all non-life insurance companies, see England and Verrall (2002) for an overview of those discrete models.

The underlying model before filtering is the same in forward and backward time, namely that the underlying sampled random variables, X and Y , are independent with joint multiplicative density $f(x, y) = f_1(x)f_2(y)$. This multiplicative structure based on partially observed independent random variables is well known in biostatistical theory and the fulfillment of multiplicity can be checked via independence tests of Tsai (1990), Mandel and Betensky (2007) and Addona, Atherton, and Wolfson (2012). Brookmeyer and Gail (1987) aimed at understanding the estimation of outstanding numbers of onset AIDS cases from a given population. They considered prevalent cohorts, where time of origin is not known, and discussed the resulting biases from just using the prevalent time available

instead of infection time of each observed individual. Wang (1989) works with prevalent cohort data, but where time of origin is known, and points out that this sampling boils down to a random truncation model. Both these two well known biostatistical papers work in usual forward moving time but nevertheless could have taken advantage of the filtered non-parametric density approach of this paper, see §2.6, had it existed.

In the in-sample forecasting application two sampling details are different, leading us to reverse the time and using the non-parametric density approach in reversed time. One sampling detail is that less is known than in the paper of Wang (1989), because exposure, i.e., the number of people at risk, is unobserved. Another is that more is known than in the paper of Wang (1989), because all failures are observed, without exception. In reversed time, the future numbers of failures, the past number of failures in regular time, is exactly the exposure needed for estimation. Therefore, the extra bit of information that all failures are observed up to a point can alleviate the challenge of unobserved exposure, and the technique doing this is to reverse the direction of time.

2.3 Model

Consider the probability space $\{\mathcal{S}, \mathcal{B}(\mathcal{S}), P\}$, where $\mathcal{S}$ is the square $\{(x, y) : 0 \leq x, y \leq T\}$. We are interested in estimating the density, $f = dP/d\lambda$, where λ is the two-dimensional Lebesgue measure. We will assume that f is multiplicative, i.e., $f(x, y) = f_1(x)f_2(y)$, and that observations are only sampled on a subset of the full support of this density, f . The truncated density is assumed to be supported on the triangle, $\mathcal{I}$. In this case, we consider observations of the independent and identically distributed pairs, $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, with $X_i \leq T - Y_i$, or equivalently $Y_i \leq T - X_i$, where T is the calendar time at which the data are collected. Both observation schemes can be understood as random right-truncation targeting only X or Y , respectively, and so both can be formulated in the following counting process framework. We define two counting processes, one indicating the occurrence of X , and the other indicating the occurrence of Y . By reversing the times of the counting processes, the right-truncation becomes left-truncation (Lagakos, Barraj, and De Gruttola, 1988).

We define the two time reversed counting processes as

$$N_1^i(t) = I(T - X_i \leq t), \quad N_2^i(t) = I(T - Y_i \leq t) \quad (i = 1, \dots, n),$$

with respect to the filtrations

$$\begin{aligned} \mathcal{F}_{1,t}^i &= \sigma \left(\left\{ T - X_i \leq s : s \leq t \right\} \cup \left\{ Y_i \leq s : s \leq t \right\} \cup \mathcal{N} \right), \\ \mathcal{F}_{2,t}^i &= \sigma \left(\left\{ T - Y_i \leq s : s \leq t \right\} \cup \left\{ X_i \leq s : s \leq t \right\} \cup \mathcal{N} \right), \end{aligned}$$

satisfying the usual conditions (Andersen et al., 1993, p. 60), and where $\mathcal{N} = \{A : A \subseteq B, B \in \mathcal{B}(\mathcal{S}), \text{pr}(B) = 0\}$. Adding the null set, $\mathcal{N}$, to the filtration guarantees its completeness. This is a technically useful construction, but it has been argued that it is not necessary; see Jacod (1979) and Jacod and Shiryaev (1987). We keep the assumption because we use results that rely on it.

Both counting processes live on a reversed timescale, so all the usual estimators derived from these counting processes will be estimators based on $T - X$ and $T - Y$, rather than on X and Y . To minimize any potential confusion, we will mark all functions corresponding to $T - X$ or $T - Y$ with an superscript, R . The desired estimators will then be linear transformations of the time-reversed versions.

The advantage of this time reversal can be seen by identifying the random intensity of N_l^i, λ_l^i , which is well-defined since X and Y have bounded densities. Thus it holds, almost surely, that $\lambda_l^i(t) = \lim_{h \downarrow 0} h^{-1} E [N_l^i \{(t+h)-\} - N_l^i(t-)| \mathcal{F}_{t-}]$ ($l = 1, 2$), see Aalen (1978). Straightforward computations lead to intensities satisfying Aalen's multiplicative intensity model (Aalen, 1978):

$$\lambda_l^i(t) = \alpha_l(t) Z_l^i(t),$$

where the hazard ratios α_1 , α_2 and the predictable processes, Z_1^i and Z_2^i , are

$$\begin{aligned}\alpha_1(t) &= \lim_{h \downarrow 0} h^{-1} \text{pr} \{(T - X) \in [t, t + h) \mid (T - X) \geq t\} = \frac{f_1(T - t)}{F_1(T - t)} = \frac{f_1^R(t)}{S_1^R(t)}, \\ Z_1^i(t) &= I\{Y_i < t \leq (T - X_i)\}, \\ \alpha_2(t) &= \lim_{h \downarrow 0} h^{-1} \text{pr} \{(T - Y) \in [t, t + h) \mid (T - Y) \geq t\} = \frac{f_2(T - t)}{F_2(T - t)} = \frac{f_2^R(t)}{S_2^R(t)}, \\ Z_2^i(t) &= I\{X_i < t \leq (T - Y_i)\},\end{aligned}$$

and $F_l = \int_0^\cdot f_l(x)dx$ ($l = 1, 2$) are the cumulative distribution functions. As the hazard function, α_1 , does not depend on f_2 , and the hazard function, α_2 , does not depend on f_1 , we can estimate f_1 and f_2 as one-dimensional densities.

2.4 Local linear density estimator in reversed time

Due to the symmetry between $T - X$ and $T - Y$, all of the following results hold for both f_1 and f_2 . For clarity, therefore, we suppress the subscript l , which indicates the coordinate. Furthermore, we will denote the exposure or risk process by $Z(t) = \sum_{i=1}^n Z^i(t)$.

Following Nielsen, Tanggaard, and Jones (2009), our proposed estimator of the density function, f^R , will involve a pilot estimator of the survival function, $S^R(t)$. Here, for simplicity, we choose the Kaplan–Meier product-limit estimator,

$$\hat{S}^R(t) = \prod_{s \leq t} \{1 - \Delta \hat{A}(s)\},$$

where $\hat{A}(t) = \sum_{i=1}^n \int_0^t \{Z(s)\}^{-1} dN^i(s)$ is the Aalen estimator of the integrated hazard function, $A(t) = \int_0^t \alpha(s)ds$. We define the local linear estimator $\hat{f}_{h,K}^R(t)$ of $f^R(t)$ as the minimizer $\hat{\theta}_0$ in the equation

$$\begin{aligned}\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} &= \arg \min_{\theta_0, \theta_1 \in \mathbb{R}} \sum_{i=1}^n \left[\int K_h(t - s) \{\theta_0 + \theta_1(t - s)\}^2 Z^i(s) W(s) ds \right. \\ &\quad \left. - 2 \int K_h(t - s) \{\theta_0 + \theta_1(t - s)\} \hat{S}^R(s) Z^i(s) W(s) dN^i(s) \right] \quad (2.1)\end{aligned}$$

Here and below, an integral $\int$ with no limits denotes integration over the whole support, i.e., $\int_0^T$. In addition, for kernel K and bandwidth h , $K_h(t) = h^{-1}K(t/h)$. The definition

of the local linear estimator as the minimizer of (5.3) can be motivated by the fact that the sum on the right hand side of (5.3) equals the limit of

$$\sum_{i=1}^n \int \left[\left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \hat{S}^R(u) dN^i(u) - \theta_0 - \theta_1(t-s) \right\}^2 - \xi(\varepsilon) \right] K_h(t-s) Z^i(s) W(s) ds,$$

for ε converging to zero. Here, $\xi(\varepsilon) = \{\varepsilon^{-1} \int_s^{s+\varepsilon} \hat{S}^R(u) dN^i(u)\}^{-2}$ is a vertical shift subtracted to make the expression well-defined. Because $\xi(\varepsilon)$ does not depend on (θ_0, θ_1) , $\hat{\theta}_0$ is defined by a local weighted least squares criterion. The function, W , is an arbitrary predictable weight function on which the pointwise first order asymptotics will not depend. There exist two popular weightings: the first being the natural unit weighting, $W(s) = 1$, while the second is the Ramlau–Hansen weighting, $W(s) = \{n/Z(s)\} I\{Z(s) > 0\}$. The latter becomes the classical kernel density estimator in the simple unfiltered case. However, in the framework of filtered observations the natural unit weighting, $W(s) = 1$, tends to be more robust (Nielsen, Tanggaard, and Jones, 2009), so we use it. For this, the solution of (5.3) (Nielsen, Tanggaard, and Jones, 2009; Gámiz et al., 2013) is

$$\hat{f}_{h,K}^R(t) = n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) \hat{S}^R(s) dN^i(s), \quad (2.2)$$

where

$$\begin{aligned} \bar{K}_{t,h}(t-s) &= \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2} K_h(t-s), \\ a_j(t) &= n^{-1} \int K_h(t-s)(t-s)^j Z(s) ds \quad (j = 0, 1, 2). \end{aligned}$$

If K is a second-order kernel, then $n^{-1} \int \bar{K}_{t,h}(t-s) Z(s) ds = 1$, $n^{-1} \int \bar{K}_{t,h}(t-s)(t-s) Z(s) ds = 0$, $n^{-1} \int \bar{K}_{t,h}(t-s)(t-s)^2 Z(s) ds > 0$, so that $\bar{K}_{t,h}$ can be interpreted as a second-order kernel with respect to the measure, μ , where $d\mu(s) = n^{-1} Z(s) ds$. This is essential in understanding the pointwise asymptotics of the local linear estimator $\hat{f}_{h,K}(t) = \hat{f}_{h,K}^R(T-t)$ which, as we will see, coincides with the kernel estimator $\sum_{i=1}^n \int K_{t,h}(t-s) \hat{S}^R(s) \{Z_1(s)\}^{-1} dN^i(s)$.

We introduce the following notation. For every kernel, K , let

$$\mu_j(K) = \int s^j K(s) ds, \quad R(K) = \int K^2(s) ds, \quad \bar{K}^*(u) = \frac{\mu_2(K) - \mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2} K(u).$$

For an interval I , $C_d(I)$, denotes the space of d -times continuously differentiable function on I . We make the following assumptions.

- S1. The bandwidth $h = h(n)$ satisfies $h \rightarrow 0$ and $n^{1/4}h \rightarrow \infty$ for $n \rightarrow \infty$.
- S2. The density f is strictly positive and it holds that $f \in C_2([0, T])$.
- S3. The kernel K is symmetric, has bounded support and has finite second moment.

Assumptions (S2) and (S3) are standard in smoothing theory. In contrast to the unfiltered case, (S1) assumes more than just the bandwidth h converging to zero. This is required, otherwise the estimation error of the survival function would determine the first-order asymptotic properties of the bias, since $n^{-1/2}/h^2 \rightarrow 0$ would not hold.

The key in obtaining the pointwise limit distribution of $\hat{f}_{h,K}(t) - f(t)$ is to split the estimation error into a sum of a stable part and a martingale part,

$$B^R(t) = f_{h,K}^{R,*}(t) - f^R(t), \quad V^R(t) = \hat{f}_{h,K}^R(t) - f_{h,K}^{R,*}(t),$$

where $f_{h,K}^{R,*}(t) = n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) Z^i(s) \hat{S}^R(s) \alpha(s) ds$. The estimation error can then be described as

$$\hat{f}(t) - f(t) = B^R(T-t) + V^R(T-t) = B(t) + V(t).$$

Proposition 2.1. Under (S1)–(S3), for $t \in (0, T)$,

$$(nh)^{1/2} \left\{ \hat{f}_l(t) - f_l(t) - B_l(t) \right\} \rightarrow N \left\{ 0, \sigma_l^2(t) \right\} \quad (l = 1, 2), \quad n \rightarrow \infty,$$

in distribution, where $B_l(t) = \frac{1}{2} \mu_2(\bar{K}^*) f_l''(t) h^2 + o(h^2)$, $\sigma_l^2(t) = \lim_{n \rightarrow \infty} nh \langle V_l \rangle_t = R(\bar{K}^*) f_l(t) F_l(t) \gamma_l(t)^{-1}$, $\gamma_l(t) = \text{pr}(Z_l^1(t) = 1)$.

Proposition 4.3 is proved in the Supplemental Material.

2.5 Bandwidth selection in reversed time

2.5.1 Cross-validation and do-validation

For a kernel estimator, the bandwidth is a positive scalar parameter controlling the smoothing degree. Data-driven cross-validation in density estimation goes back to Rudemo (1982) and Bowman (1984). Nowadays, a slightly modified version (Hall, 1983) is used intended to minimize the integrated squared error. By adding a general weighting, w , and the exposure, Z , which acknowledges the filtered observations, the aim is to find the minimizer of the integrated squared error $\Delta_K(h) = \int \left\{ \hat{f}_{h,K}^R(t) - f^R(t) \right\}^2 Z(t)w(t) dt$, which has the same minimizer as $\int \left\{ \hat{f}_{h,K}^R(t) \right\}^2 Z(t)w(t) dt - 2 \int \hat{f}_{h,K}^R(t) f^R(t) Z(t)w(t) dt$. Only the second integral of this term needs to be estimated. For the survival density estimator defined in §2.4, Nielsen, Tanggaard, and Jones (2009) propose choosing the bandwidth estimator, $\hat{h}_{CV}^K$, as the minimizer of

$$\hat{Q}_{K,w}(h) = \int \left\{ \hat{f}_{h,K}^R(t) \right\}^2 Z(t)w(t) dt - 2 \sum_{i=1}^n \int \hat{f}_{h,K}^{R,[i]}(t) \hat{S}^R(t)w(t) dN^i(t), \quad (2.3)$$

where $\hat{f}_{h,K}^{R,[i]}(t) = n^{-1} \sum_{j \neq i} \int \bar{K}_{t,h}(t-s) \hat{S}^R(s) dN^j(s)$. This can be seen as a generalization of classical cross-validation.

Over the last 20 years, many new methods have been developed to improve cross-validation; see Heidenreich, Schindler, and Sperlich (2013). One of the strongest bandwidth selectors of this review is so-called one-sided cross-validation (Hart and Yi, 1998; Martínez-Miranda, Nielsen, and Sperlich, 2009), which uses the fact that, under mild regularity conditions, the ratio of asymptotically optimal bandwidths of two estimators with different kernels, K and L , is a feasible factor, $\rho(K, L) = \{R(K)\mu_2^2(L)/\mu_2^2(K)R(L)\}^{1/5}$, which depends only on the two kernels; see also (2.6) and (2.7) below. The authors replace the kernel K used for the kernel estimator in (5.4), by its right-sided version $L = K_R = 2K(\cdot)I(\cdot \geq 0)$ when minimizing (2.3) and multiply the resulting cross-validation bandwidth by the feasible factor, $\rho(K, K_R)$, to derive a bandwidth for a kernel estimator with kernel, K . Such a construction makes sense if cross-validation for a one-sided kernel estimator works better than cross-validating with the original kernel, K . One can generalize this idea by defining indirect cross-validation as a method

where a kernel, L , can be arbitrarily chosen. We denote such bandwidth estimator by $h_{\text{ICV}}^L = \rho(K, L)h_{\text{CV}}^L$.

Savchuk, Hart, and Sheater (2010) propose an indirect cross-validation procedure where one choses a linear combination of two Gaussian kernels as kernel, L . Mammen et al. (2011) introduce the do-validation method, which performs indirect cross-validation twice by using two one-sided kernels, $L_1 = K_L = 2K(\cdot)I(\cdot \leq 0)$ and $L_2 = K_R$, as indirect kernels in (2.3). The do-validation bandwidth is the average of the two resulting bandwidths, $h_{\text{DO}} = 0.5(h_{\text{ICV}}^{K_L} + h_{\text{ICV}}^{K_R})$. Cross-validation for kernels K_L and K_R works better than for K because the asymmetry of the kernels K_L and K_R leads to larger optimal bandwidths. An empirical study in favour of do-validation in our survival setting has been performed in Gámiz et al. (2013). Asymptotic theory for weighted and unweighted cross-validation and do-validation, with and without time reversal, is developed in this paper in our general survival density framework. Below we discuss how the weighting, w , in (2.3) can be chosen when the aim is to estimate outstanding loss liabilities.

2.5.2 Weighting for application in claims reserving

In Gámiz et al. (2013), standard cross-validation is defined as the minimizer of (2.3) with $w(t) = 1$. Hence, standard cross-validation can be formulated as an in-sample technique, which aims to estimate the optimal bandwidth for the estimator calculated from the given sample. However, the situation in the forecasting problem motivating this paper is different, since our interest focuses on the unobserved region.

In this section, we illustrate how to choose a reasonable weighting scheme to estimate the outstanding liabilities for a non-life insurance company. The most relevant data for this relate to the most recent time-periods, for which only a small number of data are available. This is a well-known challenge for actuaries, who generally tackle it by using expert opinion and manual adjustments to the data. Bornhuetter and Ferguson (1972), Mack (2008) and Alai, Merz, and Wüthrich (2010) give a flavour of the Bornhutter–Ferguson method used by actuaries. Our smoothing methodology, based on continuous data, could be used as an alternative to these less rigorous approaches, and so replace expert opinion and manual adjustments by using information from relevant neighbourhoods according to an optimal smoothing criteria.

Unfortunately, the trivial weighting, $w = 1$, implies that the recent years only have small influence on the size of the bandwidth, due to the lack of sufficient data. In contrast, we want the weighting, $w(t)$, to depend on the estimated size of the liabilities at t , in order to give greatest weight to the most recent period. Assume that T is an integer indicating for instance months or years, then for a period, $p = 1, \dots, T$, the reserve, $R(p)$, is given as $R(p) = n \int_{p-1}^p f_1(s) S_2(T-s) ds / \int_{\mathcal{I}} f(x, y) dx dy$, which is proportional to $\int_{p-1}^p f_1(s) F_2^R(s) ds$. Hence if this is the quantity of interest, for short periods, we propose the following weighted integrated squared error to be the optimality criteria for estimating f_1 ,

$$\begin{aligned} \Delta_{1,K}(h) &= n^{-1} \int \left\{ f_1(s) F_2^R(s) - \hat{f}_{1,h,K}(s) \hat{F}_2^R(s) \right\}^2 ds \\ &= n^{-1} \int \left\{ f_1^R(s) S_2(s) - \hat{f}_{1,h,K}^R(s) \hat{S}_2(s) \right\}^2 ds. \end{aligned}$$

The estimator $\hat{S}_2$ converges to S_2 uniformly with rate $n^{-1/2}$ Andersen et al. (1993, p. 261). Thus, we can substitute $S_2(s)$ by its estimator $\hat{S}_2(s) = 1 - \hat{S}_2^R(T-s)$, and define

$$\tilde{\Delta}_{1,K}(h) = n^{-1} \int \left\{ f_1^R(s) - \hat{f}_{1,h,K}^R(s) \right\}^2 \left\{ \hat{S}_2(s) \right\}^2 ds.$$

But, since f_1 and $\hat{S}_2$ do not depend on h , minimizing $\tilde{\Delta}_{1,K}$ in h is equivalent to minimizing

$$\begin{aligned} Q_K(h) &= \tilde{\Delta}_{1,K}(h) - \int \left\{ f_1^R(t) \hat{S}_2(t) \right\}^2 dt \\ &= \int \left\{ \hat{f}_{1,h,K}^R(t) \right\} \left\{ \hat{S}_2(t) \right\}^2 dt - 2 \int f_1^R(t) \hat{f}_{1,h,K}^R(t) \left\{ \hat{S}_2(t) \right\}^2 dt. \end{aligned}$$

Therefore, we choose the weight $w_1(t) = \hat{S}_2(t)^2 / Z_1(t)$ in (2.3), and the cross-validation estimator of $Q_K(h)$ becomes

$$\begin{aligned} \hat{Q}_{K,w_1}(h) &= \int \left\{ \hat{f}_{1,h,K}^R(t) \right\}^2 \left\{ \hat{S}_2(t) \right\}^2 dt \\ &\quad - 2 \sum_{i=1}^n \int \hat{f}_{1,h,K}^{R,[i]}(t) \hat{S}_1^R(t) \left\{ \hat{S}_2(t) \right\}^2 \left\{ Z_1(t) \right\}^{-1} dN^i(t). \end{aligned}$$

By symmetry, the weighting for f_2 can be derived in a similar fashion, with $w_2(t) = \hat{S}_1(t)^2 / Z_2(t)$.

2.6 Asymptotic properties of weighted combinations of indirect cross-validation

In this section we formulate the asymptotic theory of the bandwidth selectors in the original time direction. This gives statisticians using cross-validation or do-validation with the local linear density estimator of Nielsen, Tanggaard, and Jones (2009); as in Gámiz et al. (2013), the asymptotic theory needed to support their approach. We then provide the theory for the reversed time direction.

We first briefly describe the general model in the original time direction (Nielsen, Tanggaard, and Jones, 2009; Gámiz et al., 2013). When observing n individuals, let N_i be a $\{0, 1\}$ -valued counting process, which observes the failures of the i th individual in the time interval, $[0, T]$. We assume that N_i is adapted to a filtration, $\mathcal{F}_t$, which satisfies the usual conditions, see §2.3. We also observe the $\{0, 1\}$ -valued predictable process, Z_i , which equals unity when the i th individual is at risk. It is assumed that Aalen's multiplicative intensity model, $\lambda_i(t) = \alpha(t)Z_i(t)$, is satisfied. This formulation contains the case of a longitudinal study with left-truncation and right-censoring. In this case, we observe triplets (Y_i, X_i, δ_i) ($i = 1, \dots, n$) where Y_i is the time at which an individual enters the study, X_i is the time he/she leaves the study and δ_i is binary and equals 1 if death is the reason for leaving the study. Hence, $Y_i \leq X_i$, and the counting process formulation would be $N_i(t) = I(X_i \leq t)\delta_i$ and $Z_i(t) = I(Y_i \leq t < X_i)$.

The local linear survival density estimator in the original time direction is then defined as $\hat{f}(t) = n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) \hat{S}(s) dN_i(s)$, where $\hat{S}(s)$ is the Kaplan–Meier estimator of the survival function. The integrated squared error, $\Delta_K(h)$, and the cross-validation criterion, $\hat{Q}_{K,w}(h)$, then become

$$\Delta_K(h) = n^{-1} \sum_{i=1}^n \int \left\{ \hat{f}(t) - f(t) \right\}^2 w(t) Z_i(t) dt, \quad (2.4)$$

$$\hat{Q}_{K,w}(h) = \sum_{i=1}^n \int \left\{ \hat{f}(t) \right\}^2 Z_i(t) w(t) dt - 2 \sum_{i=1}^n \int \hat{f}^{[i]}(t) \hat{S}(t) w(t) dN_i(t), \quad (2.5)$$

where $\hat{f}^{[i]}(t) = n^{-1} \sum_{j \neq i} \int \bar{K}_{t,h}(t-s) \hat{S}(s) dN_j(s)$.

We will derive the asymptotic properties of weighted combinations of indirect cross-validation bandwidths and in particular of the do-validation approach. In Lemma 2.3 of

Appendix 2.A, we prove that the integrated squared error in (2.4) is uniformly asymptotically equivalent to $M_K(h) = (nh)^{-1}R(\bar{K}^*) \int f(t)S(t)w(t)dt + h^4\mu_2^2(\bar{K}^*) \int \{f''(t)/2\}^2\gamma(t)w(t)dt$, which leads to the optimal deterministic bandwidth selector

$$h_{\text{MISE}} = C_0 n^{-1/5}, \quad C_0 = \left\{ \frac{R(\bar{K}^*)}{\mu_2^2(\bar{K}^*)} \frac{\int f(t)S(t)w(t)dt}{\int f''(t)^2\gamma(t)w(t)dt} \right\}^{1/5}, \quad (2.6)$$

where $\gamma(t) = n^{-1}E\{Z(t)\}$. To simplify the discussion, we assume that h_{ISE} , is defined as the minimizer of (2.5) over the interval $I_n^* = [a_1^* n^{-1/5}, a_2^* n^{-1/5}]$, where the constants $a_2^* > a_1^* > 0$ are chosen such that $a_1^* < C_0 < a_2^*$.

We will study the asymptotic properties of the weighted combinations of indirect cross-validation selectors introduced in Section 2.5.1,

$$\hat{h}_{\text{ICV}} = \sum_{j=1}^J m_j \rho_j h_{\text{CV}}^{L_j}, \quad \rho_j = \rho(L_j) = \left\{ \frac{R(K)\mu_2^2(L_j)}{\mu_2^2(K)R(L_j)} \right\}^{1/5}, \quad (2.7)$$

where L_j are arbitrary kernels and m_j are weights with $\sum_{j=1}^J m_j = 1$. For K symmetric, $J = 2, L_1 = K_L, L_2 = K_R$, and $m_1 = m_2 = 0.5$ we get the do-validation bandwidth estimator. We make the following assumptions.

T1. Let $Z = \sum_{i=1}^n Z_i$. The expected relative exposure function, $\gamma(t) = n^{-1}E\{Z(t)\}$ is strictly positive, satisfies $\gamma \in C_2([0, T])$, and $\sup_{s \in [0, T]} |Z(s)/n - \gamma(s)| = o_P\{(\log n)^{-1}\}$, $\sup_{s, t \in [0, T], |t-s| \leq C_K h} |\{Z(t) - Z(s)\}/n - \{\gamma(t) - \gamma(s)\}| = o_P\{(nh)^{-1/2}\}$, where the constant C_K is defined in (T2).

T2. The kernels, K and L_j ($j = 1, \dots, J$), are compactly supported, i.e., the support lies within $[-C_K, C_K]$ for some constant, $C_K > 0$. The kernels are continuous on $\mathbb{R} \setminus \{0\}$ and have one-sided derivatives that are Hölder continuous on $\mathbb{R}^- = \{x : x < 0\}$ and $\mathbb{R}^+ = \{x : x > 0\}$. Thus, there exist constants c and δ such that $|g(x) - g(y)| \leq c|x - y|^\delta$ for $x, y < 0$ or $x, y > 0$ with g equal to K' or L_j' ($j = 1, \dots, J$). The left and right-sided derivatives differ at most on a finite set. The kernel K is symmetric.

T3. It holds that $f \in C_2([0, T])$. The second derivative of f is Hölder continuous with exponent $\delta > 0$ and f is strictly positive.

T4. There exists a function $\tilde{w} \in C_1([0, T])$, with $\sup_{t \in [0, T]} |\tilde{w}(t) - w(t)| = o_P(1)$.

TABLE 2.1: The factor Ψ^K in (2.8) as comparison of asymptotic variances among bandwidth selection methods.

Method	Epanechnikov	Quartic	Sextic
Do-validation	2.19	1.89	2.36
Cross-validation	7.42	5.87	6.99
Plug-in	0.72	0.83	1.18

Assumption (T2) is a weak standard condition on kernels. Assumption (T3) differs from standard smoothness conditions only by the mild additional assumption that the second derivative of the density function fulfils a Hölder condition. Assumption (T1) is also rather weak. In the special framework considered in §1–4, and also in the framework of longitudinal data described previously, (T1) is easily verified by setting $\gamma(t) = \text{pr}(Z_i(t) = 1)$.

Theorem 2.2. *Under (T1)–(T4), the bandwidth selector $\hat{h}_{\text{ICV}}$ of the local linear survival density estimator in the original time direction satisfies*

$$n^{3/10} (\hat{h}_{\text{ICV}} - h_{\text{MISE}}) \rightarrow N(0, \sigma_1^2), \quad n^{3/10} (\hat{h}_{\text{ICV}} - h_{\text{ISE}}) \rightarrow N(0, \sigma_2^2), \quad n \rightarrow \infty,$$

where

$$\begin{aligned} \sigma_1^2 &= S_1 \int \left\{ \sum_{j=1}^J m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) \right\}^2 du, \\ \sigma_2^2 &= S_1 \int \left\{ \sum_{j=1}^J m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) - H_K(u) \right\}^2 du + S_2, \\ S_1 &= \frac{2}{25} \frac{\int S^2(t) f^2(t) \tilde{w}^2(t) dt}{R^{7/5}(K) \mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(t) \tilde{w}(t) dt \right\}^{3/5} \left\{ \int f(t) S(t) \tilde{w}(t) dt \right\}^{7/5}}, \\ S_2 &= \frac{4}{25} \frac{\int f''(t)^2 S(t) f(t) \tilde{w}^2(t) \gamma(t) dt - \int \left\{ \int_t^T f''(u) f(u) \tilde{w}(u) \gamma(u) du \right\}^2 \alpha(t) \gamma^{-1}(t) dt}{R^{2/5}(K) \mu_2^{6/5}(K) \left\{ \int f(t) S(t) \tilde{w}(t) dt \right\}^{2/5} \left\{ \int f''(t)^2 \gamma(t) \tilde{w}(t) dt \right\}^{8/5}}, \end{aligned}$$

and $G_K(u) = I(u \neq 0) \left\{ \bar{K}^{**}(u) - \bar{K}^{**}(-u) \right\}$, and

$H_K(u) = I(u \neq 0) \int \bar{K}^*(v) \left\{ \bar{K}^{**}(u+v) - \bar{K}^{**}(-u+v) \right\} dv$, with

$$\bar{K}^{**}(u) = -\frac{\mu_2(K) - \mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2} \{K(u) + uK'(u)\} + \frac{\mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2} K(u).$$

Theorem 2.2 is proved in the Appendix 2.A. The theorem states that the relative differ-

ence between the bandwidths h_{CV} , h_{MISE} and h_{ISE} is of order $n^{-1/10}$. This can be explained intuitively by the fact that a bounded interval contains $O(n^{1/5})$ non-overlapping subintervals of length h , and the kernel estimators are thus asymptotically independent if their argument differs by a magnitude of order $O(n^{-1/5})$. The rate $n^{-1/10} = (n^{-1/5})^{1/2}$ can then be explained by a central limit theorem.

The result generalizes the asymptotic properties of do-validation established by Mammen et al. (2011) in the unfiltered case. If the observations, $X_1, \dots, X_n$, are unfiltered, i.e., $Z_i(t) = I(t \leq X_i)$, then the Kaplan–Meier estimator becomes $\hat{S}(t) = n^{-1} \sum_i Z_i(t)$, which implies that $\gamma(t) = S(t)$. Then, by choosing the weighting $w(t) = \hat{S}(t)^{-1}$, the integrated squared error (2.4) and the cross-validation criterion (2.5) are identical to the unfiltered case and, thus, Theorem 2.2 is Theorem 1 in Mammen et al. (2011).

For a fixed kernel K and different choices of weighted indirect kernels (m_j, L_j) , the variances, σ_2^2 , only differ in the feasible factor

$$\Psi_{ICV}^K(m_1, \dots, m_J, L_1, \dots, L_J) = \int \left\{ \sum_{j=1}^J m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) - H_K(u) \right\}^2 du. \quad (2.8)$$

The asymptotic variance of a plug-in estimation error, $(h_{MISE} - h_{ISE})$, is obtained by replacing the factor Ψ_{ICV}^K in σ_2^2 by $\Psi_{MISE}^K = \int \{H_K(u)\}^2 du$. Plug-in estimators are those derived by estimating the infeasible quantities of h_{MISE} and achieve the same asymptotic limit as h_{MISE} under appropriate conditions. The values of Ψ^K can be used to compare the asymptotic performance of different methods. Table 2.1 shows these values for do-validation, cross-validation and the plug-in method using the Epanechnikov, quartic and sextic kernels. Once the asymptotic properties in the original time direction are derived, it is straightforward to derive a similar result in the reversed time direction.

Corollary 2.3. *Under assumption (T1)–(T3), the bandwidth selector, $\hat{h}_{ICV}$, of the local linear survival density estimator in the reversed time direction satisfies*

$$n^{3/10} (\hat{h}_{ICV} - h_{MISE}) \rightarrow N(0, \sigma_1^2), \quad n^{3/10} (\hat{h}_{ICV} - h_{ISE}) \rightarrow N(0, \sigma_2^2), \quad n \rightarrow \infty,$$

where

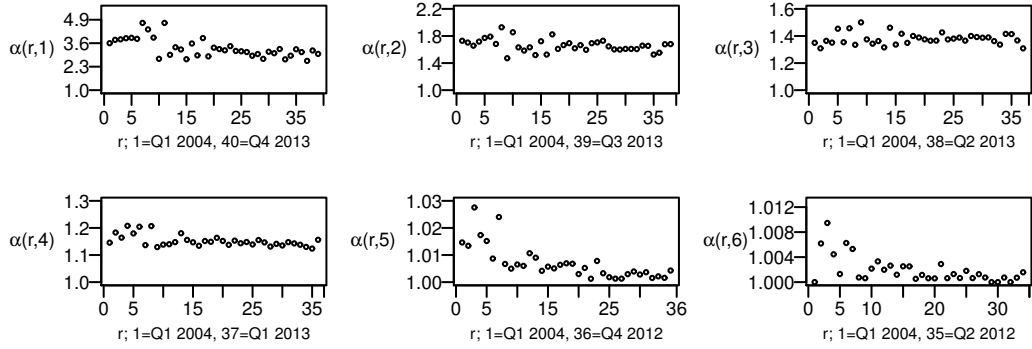
$$\begin{aligned}\sigma_1^2 &= S_1 \int \left\{ \sum_{j=1}^J m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) \right\}^2 du, \\ \sigma_2^2 &= S_1 \int \left\{ \sum_{j=1}^J m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) - H_K(u) \right\}^2 du + S_2, \\ S_1 &= \frac{2}{25} \frac{R^{-7/5}(K) \int F^4(t) \alpha^2(T-t) \tilde{w}^2(T-t) dt}{\mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(T-t) \tilde{w}(T-t) dt \right\}^{3/5} \left\{ \int f(t) F(t) \tilde{w}(T-t) dt \right\}^{7/5}}, \\ S_2 &= \frac{4}{25} \left[\frac{\int f''(t)^2 F(t) f(t) \tilde{w}^2(T-t) \gamma(T-t) dt}{R^{2/5}(K) \mu_2(K)^{6/5} \left\{ \int f''(t)^2 \gamma(T-t) \tilde{w}(T-t) dt \right\}^{8/5} \left\{ \int f(t) F(t) \tilde{w}(T-t) dt \right\}^{2/5}} \right. \\ &\quad \left. - \frac{\int \left\{ \int_t^T f''(u) f(u) \tilde{w}(T-u) \gamma(T-u) du \right\}^2 \alpha(t) \gamma^{-1}(t) dt}{R^{2/5}(K) \mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(T-t) \tilde{w}(T-t) dt \right\}^{8/5} \left\{ \int f(t) F(t) \tilde{w}(T-t) dt \right\}^{2/5}} \right].\end{aligned}$$

2.7 Illustration

We now analyse a data set of reported and outstanding claims from a motor business in Cyprus. All the calculations in this and the next section have been performed with R (R Development Core Team, 2014). The data of this section consist of $n = 58180$ claims reported between 2004 and 2013. The data are $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where X_i denotes the underwriting date of claim i , and Y_i the reporting delay in days. The data exist on a triangle, with $X_i + Y_i \leq 31$ December 2013. Our aim is to forecast the number of future claims from contracts underwritten in the past which have not yet been reported. It is implicitly assumed that the triangle is fully run off, such that the maximum reporting delay is ten years. This is reasonable, see Figure 2.2, since f_2 has a strong decay already after one year. According to the theory, we use a multiplicative structured density, $f(x, y) = f_1(x)f_2(y)$, where the components f_1 and f_2 are the underwriting date density and the development time density, respectively.

For justification of this assumption, we performed several tests which all indicated that the assumption might be violated. We then did a more pragmatic step which is motivated from actuarial practice. We transformed the data into a triangle with dimension 3654×3654 , $\mathcal{N}_{x,y} = \sum_{i=1}^n I(X_i = x, Y_i = y)$, $(x, y) \in \{1, \dots, 3654\}^2$, and then aggregated the data into a quarterly triangle, $(\mathcal{N}_{r,s}^Q)$, with dimension 40×40 , which is the form usually

FIGURE 2.1: Development factors of the first six quarter for individual underwriting quarter.



available in a reserving department; see the Supplemental Material. For $s = 1, \dots, 6$, we calculated the quantities $\alpha(r, s) = \sum_{l=1}^{s+1} \mathcal{N}_{r,l}^Q / \sum_{l=1}^s \mathcal{N}_{r,l}^Q$, known as development factors in actuarial sciences (Kuang, Nielsen, and Nielsen, 2009). The values of $\alpha(r, s)$ are displayed in Figure 3.2. If the multiplicativity assumption is satisfied, then $\alpha(r, s)$ is approximately equal to $\{\sum_{l=1}^{s+1} f_1(x_r) f_2(y_l)\} / \{\sum_{l=1}^s f_1(x_r) f_2(y_l)\}$ which does not depend on r . Here, x_r lies in the r th quarter and y_l in the l th quarter. Hence, the points in each plot should lie around horizontal lines.

Only considering the first four plots, one could argue that discrepancy from constancy is only caused by white noise from the stochastic nature of the observations. However, there seems to be a negative drift in the 5th and 6th plots. Non constancy is caused in particular by the first 7 underwriting quarters which correspond to the first 7 points in each plot. Re-evaluating the first four plots, one can also spot the drift there; despite the noise. The relative drift size in the different plots seems of similar magnitude when the values are subtracted by 1. This indicates that the data do indeed not satisfy the independence assumption. A pragmatic solution would be to throw away the data of the first 7 underwriting quarters, as it is often done by actuaries when using the chain-ladder method. We preferred to keep the whole data set because there are not many data observed after the fourth quarter. A better strategy might be to look for extensions of our model where the reporting delay density f_2 depends on calendar time. This is topic of ongoing research. Additional seasonal effects are considered in Lee et al. (2015). Other calendar time effects will often involve the need of extrapolation of a time series; see also Kuang, Nielsen, and Nielsen (2008) for the discrete-time case. Accounting for the spotted drift in the data example leads only to a slight change of the total number

of forecasted claim numbers but to larger differences in the forecasted delay times.

We have calculated the local linear density estimators of the two underlying multiplicative densities, f_1 and f_2 , using the Epanechnikov kernel and weighted cross-validated and do-validated bandwidth selectors. For the density f_1 , cross-validation chose a bandwidth of 408 days and do-validation a bandwidth of 1,860 days, while, for f_2 , the minimizer of the cross- and do-validation criteria were 15 days and 72 days, respectively. Figure 2.2 shows the estimated densities.

The left plot indicates that there is no trend in the amount of underwritten policies. In the right plot, consistent with the policy duration of one year and our experience of other motor insurance, we find that most of the claims are reported within 1.4 years. There is a sharp increase and decrease at the beginning and at the end of the first year, respectively, and a near-uniform development in between. It seems plausible that boundary and bias correction techniques would be useful in future analyses. One could for example consider multiplicative bias correction (Nielsen, Tanggaard, and Jones, 2009) or asymmetric kernels (Hirukawa and Sakudo, 2014).

In this application, we encounter the usual problem with standard cross-validation which sometimes picks bandwidths which are much too small. Do-validation seems to have estimated a reasonable bandwidth.

The number of outstanding claims for the future quarters, obtained by integrating the multiplicative estimator over diagonals in the unobserved part, are shown in Table 3.1. As a benchmark, we have calculated the total reserve using the standard chain ladder method by aggregating the data on a quarterly basis. The chain ladder method is the most widely used reserving method in practice, and can be interpreted as a Poisson maximum likelihood estimator with multiplicative mean structure (Kuang, Nielsen, and Nielsen, 2009). It predicts a smaller number than the continuous approaches. Under a Poisson approximation with an approximated standard deviation of 48 we get significant differences between the predicted future claims.

FIGURE 2.2: Estimated underwriting and development densities in the real data application: Cross-validation (dashed), do-validation (solid).

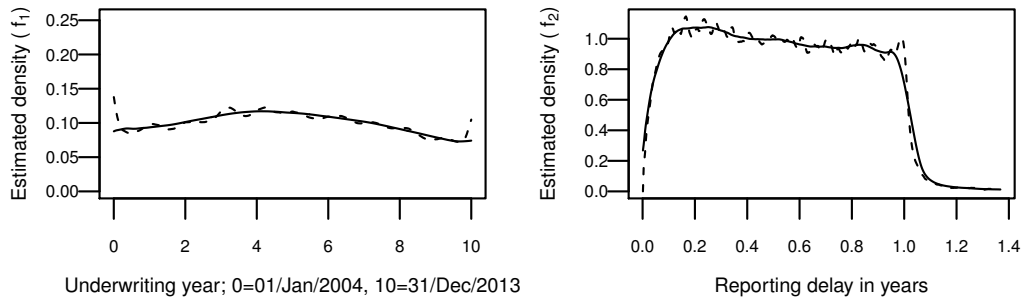


TABLE 2.2: Number of claims forecasts in the real data application. In quarters; 1 = 2014 Q1, 39 = 2022 Q3.

Future quarter:	1	2	3	4	5	6	7	8	9	10	11 – 39	Total
Cross-validation	1027	733	465	201	15	5	3	2	1	1	1	2452
Do-validation	970	684	422	166	14	5	3	2	1	1	1	2270
Chain ladder	948	651	387	148	12	5	3	2	1	1	1	2160

2.8 Simulation study

We now describe a simulation study to show that the local linear estimator is a good strategy for reserve forecasting. We simulated the two do-validated densities from the application section, shown in Figure 2.2, assuming the multiplicative structure $f(x, y) = f_1(x)f_2(y)$. These models have been chosen to illustrate realistic situations in claims reserving. Furthermore, for computational reasons, we simulated data by aggregating the occurrence of claims in bin sizes of three days; see Appendix 2.B. We consider four sample sizes corresponding to 0.5, 1.0, 1.5 and 2.0 times the sample size, $n = 58180$, from the application.

For each sample size, we generated 500 samples and have solved the forecasting problem using the methods described in this paper. Since the data are generated in discrete time, the methods were applied using the discrete expressions in Appendix 2.B. The performance of the methods for each simulated data set was evaluated using the discrete approximation of the integrated squared error.

The local linear estimators were calculated using the Epanechnikov kernel with four different bandwidth choices. Firstly the infeasible integrated squared error optimal bandwidth which changes in each simulated sample and secondly the mean of those integrated squared error optimal bandwidths of the 500 simulated samples for every

TABLE 2.3: Summary of the integrated squared errors multiplied by 10^5 , along the 500 simulated samples. Four different bandwidths: optimal bandwidth (ISE), averaged optimal out of the 500 samples (MISE), cross-validation (CV), do-validation (DO).

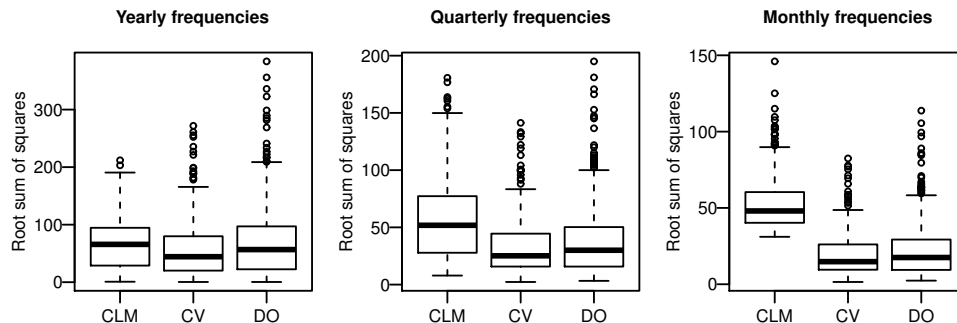
		f_1				f_2			
n		ISE	MISE	CV	DO	ISE	MISE	CV	DO
29090	Median	0.84	2.45	5.87	6.49	1.40	1.44	1.57	1.50
	Mean	1.50	3.31	18.40	17.65	1.49	1.53	1.66	1.58
	SD	1.72	2.39	33.07	39.64	0.59	0.60	0.68	0.60
58180	Median	0.56	2.29	4.65	4.47	0.84	0.86	0.91	0.87
	Mean	1.12	2.81	11.24	7.21	0.87	0.89	0.95	0.89
	SD	1.30	1.58	17.27	9.09	0.29	0.29	0.34	0.29
87270	Median	0.52	2.42	4.04	3.74	0.62	0.63	0.67	0.65
	Mean	0.99	2.71	7.49	5.29	0.64	0.65	0.69	0.66
	SD	1.14	1.24	11.55	5.68	0.20	0.20	0.22	0.20
116360	Median	0.43	2.35	3.42	3.74	0.49	0.51	0.53	0.53
	Mean	0.89	2.64	5.97	6.15	0.51	0.53	0.55	0.54
	SD	1.06	1.08	8.54	9.00	0.15	0.15	0.17	0.15

run. These two infeasible choices are compared to the two data-driven bandwidths, weighted cross-validation and weighted do-validation.

Table 2.3 shows that weighted cross-validation and do-validation perform reasonably well. The results support the asymptotic theory ranking cross-validation as more volatile than do-validation. For the development density, f_2 , note that, for larger sample sizes, there is nearly no difference between the optimal infeasible methods and the two validated bandwidth selectors. In any event, the feasible approaches seem to be doing very well at picking appropriate bandwidths.

We also simulated the development of the claims according to Table 3.1. Let R_p be the true reserve for the future period p and $\hat{R}_p$ its estimator. Then, the error was calculated as $\{\sum (R_p - \hat{R}_p)^2\}^{1/2}$. Figure 2.3 shows box plots of the errors in the future count development, obtained from the 500 simulated samples. For comparison, we calculated density estimates based on the chain ladder method, with data aggregated in years, quarters, and months, respectively. Chain ladder modelling is competitive for yearly numbers, but breaks down for more detailed quarterly, monthly, or daily numbers. It is not included in Table 2.3.

FIGURE 2.3: Prediction errors of simulated monthly (right panel), quarterly (middle panel) and yearly (left panel) data along the 500 simulated samples. Sample size is $n = 58180$. Three different methods: Chain ladder method (CLM), local linear density estimator with cross-validation (CV) and do-validation (DO) bandwidth.



2.9 Concluding Remarks

This paper produces a simpler alternative to the in-sample forecasting approach of Mammen, Martínez-Miranda, and Nielsen (2015) and Lee et al. (2015). This is done by reversing the time, and it works because all failures are observed until some calendar time. Obviously the simple multiplicative structure of the model could be questioned, see England and Verrall (2002) for some actuarial discussion on the short-comings of the multiplicative chain ladder model. One possible generalisation of our model would be to let the development density depend on calendar time. Another generalisation would be to include covariates, as has been done e.g. by Wells (1994) for counting process intensities. An example would be to incorporate claim severities. This could be done by extending the counting process set-up of this paper to the marked point processes approach (Norberg, 1993). This could also help to generalise the recent double chain ladder technique of Verrall, Nielsen, and Jessen (2010), Martínez-Miranda et al. (2011) and Martínez-Miranda, Nielsen, and Verrall (2012) to continuous time. In this paper we developed detailed asymptotic theory for the estimation of the density $f(x, y)$. Discussions of plug-in estimators of integrals of the density over triangles and/or diagonals need further theory.

Acknowledgements

We gratefully acknowledge support by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953, by a subsidy granted to the National Research University Higher School of Economics, Moscow, Russian Federation by the Government of the Russian Federation for the implementation of the Global Competitiveness Program, by an IEF grant of the European Community's Seventh Framework Programme, and by a grant of the Spanish "Ministry of Economy and Competitiveness" (European Regional Development Fund).

2.A Asymptotic properties and proofs

For readability, we will write most quantities without indices, i.e., we will not write the obvious dependence on bandwidth and kernel. We start by stating a central limit theorem for martingales, which was proved in Ramlau-Hansen (1983) and will be used in the proofs of Proposition 1 and Theorem 1.

Theorem 2.2 (Ramlau-Hansen (1983)). *Consider a predictable process $W_n(t)$ and assume that for some $\sigma^2 \geq 0$*

$$\int W_n^2(t)Z(t)\alpha(t)dt = \sigma^2 + o_p(1), \quad \int W_n^2(t)I\{W_n^2(t) > \varepsilon\}Z(t)\alpha(t)dt = o_p(1), \quad \varepsilon > 0.$$

Then, it holds that $\int W_n(u)dM(u) \rightarrow N(0, \sigma^2)$, in distribution, as $n \rightarrow \infty$.

Proof of Proposition 1

First we define the relative exposure $\gamma(t) = \text{pr}\{Z^1(t) = 1\}$. Note that γ is strictly positive and $\gamma \in C_2([0, T])$, since f_1 and f_2 have these properties. Furthermore,

$$\sup_{s \in [0, T]} |Z(s)/n - \gamma(s)| = o_p(1). \quad (2.9)$$

$$\sup_{t \in [0, T]} |\hat{S}(t) - S(t)| = O_p(n^{-1/2}), \quad (2.10)$$

$$\sup_{t \in [h, T-h]} |a_j(t) - h^j \mu_j(K) \gamma(t)| = o_p(1) \quad (j = 1, 2, 3). \quad (2.11)$$

The proof consist of two parts. First, we have to show that $B(t) = \frac{1}{2}\mu_2(\bar{K}^*)f''(t)h^2 + o(h^2)$, and then that $(nh)^{1/2}V(t) \rightarrow N\{0, R(\bar{K}^*)f(t)F(t)\gamma(t)^{-1}\}$, for $n \rightarrow \infty$. We start with $B(t)$. The uniform convergence property in (2.10), together with (2.9) and (S1) yield $B(t) = n^{-1} \int \bar{K}_{t,h}(t-s) \{f(s) - f(t)\} Z^{(n)}(s) ds + O_p(n^{-1/2})$. Then, a Taylor expansion gives $B(t) = h^2 n^{-1} f''(t) \int \bar{K}_{t,h}(t-s)(t-s)^2 Z^{(n)}(s) ds + o_p(h^2)$. Finally, from (5.5), we derive $B(t) = h^2 f''(t) \int \bar{K}_{t,h}^*(t-s) \{(t-s)\}^2 ds + o_p(h^2)$, which concludes the first part of the proof. For $V(t)$, again, (2.10) and (5.5) yield $V(t) = n^{-1} \int \bar{K}_{t,h}^*(t-s) S^R(s) \gamma^{-1}(s) dM(s) + O_p(n^{-1/2})$, and, with Theorem 4.8, we conclude that $(nh)^{1/2}V(t) \rightarrow N\{0, \sigma^2(t)\}$, where $\sigma^2(t) = R(\bar{K}^*)f(t)F(t)\gamma(t)^{-1}$.

Proof of Theorem 1

The kernel, L , will denote a generic kernel with $L = K$ or $L = L_j$ satisfying assumption (T2). Recall that

$$\begin{aligned} B(t) &= f^*(t) - f(t) = n^{-1} \int \bar{L}_{t,h}(t-s) \{\hat{S}(s)\alpha(s) - f(t)\} Z(s) ds, \\ V(t) &= \hat{f}(t) - f^*(t) = n^{-1} \int_0^T \bar{L}_{t,h}(t-s) \hat{S}(s) dM(s), \end{aligned}$$

where $f^*(t) = n^{-1} \sum_{i=1}^n \int_0^T \bar{L}_{t,h}(t-s) \hat{S}(s) Z_i(s) \alpha(s) ds$, $dM(t) = dN(t) - \alpha(t)Z(t)dt$, $N(t) = \sum_{i=1}^n N_i(t)$. We first state a uniform asymptotic expansion for the integrated squared error. Hereby it is necessary that the quantities we are dealing with are predictable. Thus, we approximate V by $\tilde{V}$, with

$$\begin{aligned} \tilde{V}(t) &= n^{-1} \int \tilde{L}_{t,h}(t-s) S(s) dM(s), \\ \tilde{L}_{t,h}(u) &= \frac{\tilde{a}_{2,h}^L(t) - \tilde{a}_{1,h}^L(t)u}{\tilde{a}_{0,h}^L(t)\tilde{a}_{2,h}^L(t) - \{\tilde{a}_{1,h}^L(t)\}^2} L_h(u), \\ \tilde{a}_{l,h}^L(t) &= n^{-1} \int L_h(t-s)(t-s)^l [Z(t) + n\{\gamma(s) - \gamma(t)\}] ds. \end{aligned}$$

Using assumption (B1), we have uniformly for $0 \leq t \leq T$ and $h \in I_n^*$ that

$$\{\log(n)nh\}^{1/2}|V(t) - \tilde{V}(t)| = o_P(1).$$

Now, for the weighted integrated squared error $\Delta_L(h)$, we obtain the following asymptotic expansion.

Lemma 2.3. *Under Assumption (T1) – (T4), it holds that $\Delta_L(h) = M_L(h) + o_P(n^{-4/5})$, uniformly for $h \in I_n^*$, with*

$$M_L(h) = (nh)^{-1} R(\bar{L}^*) \int f(t) S(t) w(t) dt + h^4 \mu_2^2(\bar{L}^*) \int \left\{ \frac{f''(t)}{2} \right\}^2 \gamma(t) w(t) dt$$

Proof. We decompose the integrated squared error into

$$\Delta_L(h) = n^{-1} \int B^2(t) Z(t) w(t) dt + 2n^{-1} \int B(t) V(t) Z(t) w(t) dt + n^{-1} \int_0^T V^2(t) Z(t) w(t) dt.$$

Now, note that $\sup_{t \in [0, T]} |\tilde{V}(t)| = O_P\{n^{-2/5}(\log n)^{1/2}\}$, $\sup_{t \in [0, T]} |B(t)| = O_P(n^{-2/5})$, and together with (T1), (2.A), we conclude that uniformly for $h \in I_n^*$,

$$\begin{aligned} \Delta_L(h) &= \int \tilde{V}(t)^2 \gamma(t) \tilde{w}(t) dt + 2 \int \tilde{V}(t) B(t) \gamma(t) \tilde{w}(t) dt + \int B^2(t) \gamma(t) \tilde{w}(t) dt + o_P(n^{-4/5}) \\ &= S_{L,1}(h) + S_{L,2}(h) + T_{L,1}(h) + T_{L,2}(h) + o_P(n^{-4/5}), \end{aligned}$$

where

$$\begin{aligned} S_{L,1}(h) &= \int \int \bar{H}_{L,h}(u, v) dM(u) dM(v) - \int_0^T \bar{H}_{L,h}(u, u) \alpha(u) Z(u) du, \\ S_{L,2}(h) &= 2 \int \delta_{L,h}(u) dM(u), \\ T_{L,1}(h) &= \int \bar{H}_{L,h}(u, u) \alpha(u) Z(u) du, \\ T_{L,2}(h) &= \int B^2(u) \gamma(u) \tilde{w}(u) du, \\ \bar{H}_{L,h}(u, v) &= n^{-2} \int \tilde{L}_{t,h}(t-u) \tilde{L}_{t,h}(t-v) S(u) S(v) \gamma(t) \tilde{w}(t) dt, \\ \delta_{L,h}(u) &= n^{-1} \int \tilde{L}_{t,h}(t-u) S(u) B(t) \gamma(t) \tilde{w}(t) dt. \end{aligned}$$

First, we define

$$\begin{aligned} S_{L,1,t}(x) &= n^{4/5} \int_0^t \int_0^t \bar{H}_{L, xn^{-1/5}}(u, v) dM(u) dM(v) - \int \bar{H}_{L, xn^{-1/5}}(u, u) \alpha(u) Z(u) du, \\ S_{L,2,t}(x) &= 2n^{4/5} \int_0^t \delta_{L, xn^{-1/5}}(u) dM(u). \end{aligned}$$

Now $t \mapsto S_{L,1,t}(x)$ and $t \mapsto S_{L,2,t}(x)$ are martingales. Applying Theorem 4.8 to these processes, gives pointwise convergence to zero. Following on from this, we show that

the functions $x \mapsto S_{L,1,T}(x)$ and $x \mapsto S_{L,2,T}(x)$ are tight, so that uniformly for $h \in I_n^*$, $S_{L,1}(h) = o_P(n^{-4/5})$, $S_{L,2}(h) = o_P(n^{-4/5})$. Finally, with standard smoothing theory arguments we conclude that uniformly for $h \in I_n^*$,

$$\begin{aligned} T_{L,1}(h) &= n^{-4/5} R(\bar{L}^*) \int f(t) S(t) \tilde{w}(t) dt + o_P(n^{-4/5}), \\ T_{L,2}(h) &= n^{-4/5} \mu_2^2(\bar{L}^*) \int \left\{ \frac{f''(t)}{2} \right\}^2 \gamma(t) \tilde{w}(t) dt + o_P(n^{-4/5}), \end{aligned}$$

which concludes the proof. $\square$

For the asymptotic discussion of the cross-validation method, note that the minimizer of $\hat{Q}_L(h)$ equals the minimizer of $\hat{\Delta}_L(h)$ with

$$\begin{aligned} \hat{\Delta}_L(h) &= n^{-1} \hat{Q}_L(h) - n^{-1} \int f(t)^2 Z(t) w(t) dt + 2n^{-1} \int f(t) \hat{S}(t) w(t) dM(t) \\ &\quad + 2n^{-1} \int f(t) \hat{S}(t) \alpha(t) Z(t) w(t) dt. \end{aligned}$$

Furthermore, almost surely

$$\hat{Q}_L(h) = \int \hat{f}(t)^2 Z(t) w(t) dt - 2 \int \hat{f}^-(t) \hat{S}(t) w(t) dN(t),$$

with $\hat{f}^-(t) = n^{-1} \int \bar{L}_{t,h}(t-s) \hat{S}(s) I(s \neq t) dN(s)$. We define $D_L(h) = \Delta_L(h) - \hat{\Delta}_L(h)$.

The next lemma states consistency of cross-validation.

Lemma 2.4. *Under Assumption (T1) – (T4), we get $D_L(h) = o_P(n^{-4/5})$, uniformly for $h \in I_n^*$. In particular, we have that $\hat{h}_{CV} = h_{MISE} + o_P(n^{-1/5})$.*

Proof. Simple computations lead to

$$\begin{aligned} D_L(h) &= 2n^{-1} \left[\int \left\{ \hat{f}^-(s) - f(s) \right\} \hat{S}(s) w(s) dM(s) \right. \\ &\quad \left. + \int \left\{ \hat{f}(s) - f(s) \right\} \left\{ \hat{S}(s) \alpha(s) - f(s) \right\} w(s) Z(s) ds \right] \\ &= 2n^{-1} \int \tilde{V}^-(s) \hat{S}(s) \tilde{w}(s) dM(s) + 2n^{-1} \int B(s) \hat{S}(s) \tilde{w}(s) dM(s) \\ &\quad + 2n^{-1} \int \left\{ \hat{f}(s) - f(s) \right\} \left\{ \hat{S}(s) \alpha(s) - f(s) \right\} \tilde{w}(s) Z(s) ds + o_P(n^{-4/5}) \\ &= o_P(n^{-4/5}), \end{aligned}$$

uniformly for $h \in I_n^*$, where $\tilde{V}^-(t) = n^{-1} \int \bar{L}_{t,h}(t-s) \hat{S}(s) I(s \neq t) dM(s)$. $\square$

Next, to develop a linear expansion of $\hat{h}_{ISE}^K$ we state the following Lemma.

Lemma 2.5. *Under Assumptions (T1) – (T4), we get uniformly for $h \in I_n^*$*

$$\Delta_L''(h) = M_L''(h) + o_P(n^{-2/5}), \quad D_L''(h) = o_P(n^{-2/5}),$$

as well as

$$\begin{aligned} M_L''(h) &= 12h^2\mu_2^2(\bar{L}^*) \int \left\{ \frac{f''(t)}{2} \right\}^2 \gamma(t)\tilde{w}(t)dt \\ &\quad + 2n^{-1}h^{-3}R(\bar{L}^*) \int f(t)S(t)\tilde{w}(t)dt + o_p(n^{-2/5}), \\ D_L'(h) &= -n^{-2}h^{-2} \int \int G_L\left(\frac{u-v}{h}\right) S(v)S(u)\gamma^{-1}(u)\tilde{w}(u) dM(u) dM(v) \\ &\quad + 2n^{-1}h\mu_2(\bar{L}^*) \int f''(u)S(u)\tilde{w}(u) dM(u) \\ &\quad - 2n^{-1}h\mu_2(\bar{L}^*) \int \int_u^T f''(s)f(s)\tilde{w}(s)\gamma(s) ds \gamma^{-1}(u) dM(u) + o_p(n^{-7/10}), \\ \Delta_L'(h_{MISE}) &= -n^{-2}h_{MISE}^{-2} \int \int H_L\left(\frac{u-v}{h}\right) S(u)S(v)\tilde{w}(u)\gamma^{-1}(u) dM(u)dM(v) \\ &\quad + 2n^{-1}h_{MISE}\mu_2(\bar{L}^*) \int S(u)f''(u)\tilde{w}(u)dM(u) \\ &\quad - 2n^{-1}h_{MISE}\mu_2(\bar{L}^*) \int \int f''(s)f(s)\tilde{w}(s)\gamma(s) ds \gamma^{-1}(u)dM(u) + o_p(n^{-7/10}), \end{aligned}$$

where $G_L(u) = I(u \neq 0)\{\bar{L}^{**}(u) - \bar{L}^{**}(-u)\}$ and $H_L(u) = I(u \neq 0) \int \bar{L}^*(v)\{\bar{L}^{**}(u+v) - \bar{L}^{**}(-u+v)\}dv$, with

$$\begin{aligned} \bar{L}^*(u) &= \frac{\mu_2(L) - \mu_1(L)u}{\mu_2(L) - \{\mu_1(L)\}^2}L(u), \\ \bar{L}^{**}(u) &= -\frac{\mu_2(L) - \mu_1(L)u}{\mu_2(L) - \{\mu_1(L)\}^2}\{L(u) + uL'(u)\} + \frac{\mu_1(L)u}{\mu_2(L) - \{\mu_1(L)\}^2}L(u). \end{aligned}$$

Proof. This follows by straightforward computations, similar to those for Lemma 2.3 and Lemma 2.4. Note that following Mammen and Nielsen (2007) we can replace the kernels $\bar{L}_{t,h}(u)$ and $\partial_h \bar{L}_{t,h}(u)$ by the kernels $\gamma(t)^{-1}\bar{L}_h^*(u)$ and $\{\gamma(t)h\}^{-1}\bar{L}_h^{**}(u)$, respectively. Also note that while in all prior computations we could simply replace $B(t) = n^{-1} \int \bar{L}_{t,h}(t-s) \{\hat{S}(s)\alpha(s) - f(t)\} Z(s) ds$ by $n^{-1} \int \bar{L}_{t,h}(t-s) \{f(s) - f(t)\} Z(s) ds$, this is not the case in $\Delta_L'(h_{MISE})$. Here one gets an additional error term arising from the estimation error $\hat{S}(t) - S(t) = -S(t) \int_0^t \hat{S}(s-)\{S(s)Z(s)\}^{-1}dM(s)$. $\square$

Now, with the continuity of M'' , a simple Taylor expansion gives

$$\begin{aligned} h_{\text{ISE}} &= h_{\text{MISE}} - M_L''(h_{\text{MISE}})^{-1} \Delta_L'(h_{\text{MISE}}) + o_p(n^{-3/10}), \\ \hat{h}_{\text{CV}} &= h_{\text{MISE}} - M_L''(h_{\text{MISE}})^{-1} \hat{\Delta}_L'(h_{\text{MISE}}) + o_p(n^{-3/10}), \end{aligned}$$

and together with Lemma 2.5 we conclude

$$\begin{aligned} h_{\text{ISE}} - h_{\text{MISE}} &= C_{1,L}^{-1} n^{-8/5} h_{\text{MISE}}^{-2} \int \int H_L \left(\frac{u-v}{h_{\text{MISE}}} \right) S(u) S(v) \tilde{w}(u) \gamma^{-1}(u) \, dM(u) dM(v) \\ &\quad - 2C_{1,L}^{-1} \mu_2(\bar{L}^*) n^{-3/5} h_{\text{MISE}} \\ &\quad \times \int S(u) f''(u) \tilde{w}(u) - \left\{ \int_u^T f''(s) f(s) \tilde{w}(s) \gamma(s) \, ds \right\} \gamma^{-1}(u) \, dM(u) \\ &\quad + o_p(n^{-3/10}), \\ \hat{h}_{\text{CV}} - h_{\text{MISE}} &= C_{1,L}^{-1} n^{-8/5} h_{\text{MISE}}^{-2} \\ &\quad \times \int \int (H_L - G_L) \left(\frac{u-v}{h_{\text{MISE}}} \right) S(u) S(v) \tilde{w}(u) \gamma^{-1}(u) \, dM(u) dM(v) \\ &\quad + o_p(n^{-3/10}), \end{aligned}$$

where

$$\begin{aligned} C_{1,L} &= n^{2/5} M_L''(h_{\text{MISE}}) \\ &= 5R^{2/5}(\bar{L}^*) \mu_2^{6/5}(\bar{L}^*) \left\{ \int f(t) S(t) \tilde{w}(t) dt \right\}^{2/5} \left\{ \int \{f''(t)\}^2 \gamma(t) \tilde{w}(t) dt \right\}^{3/5}. \end{aligned}$$

That results directly in the conclusion

$$\hat{h}_{\text{ICV}} - h_{\text{MISE}} = U_1(T) + o_p(n^{-3/10}), \quad \hat{h}_{\text{ICV}} - h_{\text{ISE}} = U_2(T) + o_p(n^{-3/10}),$$

where

$$\begin{aligned}
U_1(t) &= n^{-8/5} h_{\text{MISE}}^{-2} \\
&\quad \times \int_0^t \int_0^t \sum_{j=1}^J m_j \rho_j^3 C_{1,L_j}^{-1} (H_{L_j} - G_{L_j}) \left\{ \frac{\rho_j(u-v)}{h_{\text{MISE}}} \right\} S(u) S(v) \tilde{w}(u) \gamma^{-1}(u) \, dM(u) dM(v) \\
U_2(t) &= n^{-8/5} h_{\text{MISE}}^{-2} \\
&\quad \times \int_0^t \int_0^t \sum_{i=1}^J m_i \rho_i^3 C_{1,L_i}^{-1} (H_{L_i} - G_{L_i}) \left\{ \frac{\rho_i(u-v)}{h_{\text{MISE}}} \right\} S(u) S(v) \tilde{w}(u) \gamma^{-1}(u) \, dM(u) dM(v) \\
&\quad - C_{1,K}^{-1} n^{-8/5} h_{\text{MISE}}^{-2} \int_0^t \int_0^t H_K \left(\frac{u-v}{h_{\text{MISE}}} \right) S(u) S(v) \tilde{w}(u) \gamma^{-1}(u) \, dM(u) dM(v) \\
&\quad + 2C_{1,K}^{-1} \mu_2(\bar{K}^*) n^{-3/5} h_{\text{MISE}} \\
&\quad \times \int_0^t \left[S(u) f''(u) \tilde{w}(u) - \left\{ \int_u^t f''(s) f(s) \tilde{w}(s) \gamma(s) \, ds \right\} \gamma^{-1}(u) \right] \, dM(u).
\end{aligned}$$

Now, U_1 and U_2 are martingales and their variances σ_1^2 and σ_2^2 can be computed with Theorem 4.8.

2.B Discretization

In this section, we will describe how we discretized the continuous approach, in order to be suitable for a simulation study. The discrete triangle is described as $\mathcal{I}^d = \{(r, s) : r = 1, \dots, T; s = 1, \dots, T; r + s \leq T + 1\}$, where $T \in \mathbb{N}$, in the chosen unit, denotes the last time point where data are aggregated. Then, given observations, $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, a discrete data set $(\mathcal{N}_{r,s})_{(r,s) \in \mathcal{I}^d}$, is obtained by defining

$$\mathcal{N}_{r,s} = \sum_{i=1}^n I\{X_i \in [r-1, r), \quad X_i + Y_i \in [r+s-2, r+s-1)\}.$$

We then define the occurrence and the exposure as

$$\begin{aligned}
O_r &= \sum_{i=1}^n \int_r^{r+1} dN_1^i(s) = \sum_{l=1}^{r+1} \mathcal{N}_{(T-r),l}, \\
E_r &= \sum_{i=1}^n \int_{r-0.5}^{r+0.5} Z_1^i(s) \, ds = \sum_{i=1}^n Z_1^i(r+0.5) = \sum_{\substack{k \leq (T-r) \\ l \leq r+1}} \mathcal{N}_{k,l}, \quad (r = 0, \dots, T-1).
\end{aligned}$$

Using this, the local linear estimator becomes

$$\hat{f}_{1,h,K}^R(t) = n^{-1} \sum_{r=0}^{T-1} \bar{K}_{t,h} \{-(r+0.5)\} \hat{S}_1^R(r+0.5) O_r. \quad (2.12)$$

The Kaplan–Meier estimator becomes

$$\hat{S}_1^R(r+0.5) = \prod_{l=1}^r \left(1 - \frac{O_l}{E_l}\right),$$

and is constant around these grid points. The local linear kernel becomes

$$\bar{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2} K_h(t-s),$$

with $a_j(t) = n^{-1} \sum_{r=0}^{T-1} E_r K_h(t-r)(t-r)^j$ ($j = 0, 1, 2$). Furthermore, the final estimator, $\hat{f}_1$, of f_1 , is then $\hat{f}_1(t) = \hat{f}_{1,h,K}^R(T-t)$. The integrated squared error can be written as

$$\Delta_{1,K}(h) = n^{-1} \sum_{r=0}^{T-1} \left\{ \hat{f}_{1,h,K}^R(r+0.5) - f_1(T-r-0.5) \right\}^2 E_r w_1(r+0.5),$$

for our preferred weighting function, $w_1(r+0.5) = \left\{1 - \hat{S}_2^R(T-r-0.5)\right\}^2 / E_r$. Finally, the weighted cross-validation score becomes

$$\hat{Q}_{K,w_1}(h) = \sum_{r=0}^{T-1} \left\{ \hat{f}_{1,h,K}^R(r+0.5) \right\}^2 E_r w_1(r+0.5) - 2 \sum_{r=0}^{T-1} \hat{f}_{1,h,K}^{R,[r]}(r+0.5) \hat{S}_1^R(r+0.5) O_r w_1(r+0.5),$$

where $\hat{f}_{1,h,K}^{R,[r]}$ is the estimator arising from (2.12) by setting $O_r = O_r - 1$.

References

- Aalen, O. O. (1978). “Non-parametric inference for a family of counting processes”. In: *Ann. Stat.* 6, pp. 701–726.
- Addona, V., J. Atherton, and D. B. Wolfson (2012). “Testing the assumption of independence of truncation time and failure time”. In: *Int. J. Biostat.* 8.
- Alai, D., M. Merz, and M. V. Wüthrich (2010). “Prediction uncertainty in the Bornhuetter–Ferguson claims reserving method: revisited”. In: *Ann. Actuar. Sci.* 5, pp. 7–17.

- Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Bornhuetter, R. L. and R. E. Ferguson (1972). “The actuary and IBNR”. In: *Casualty Actuarial Society Proceedings* LIX, pp. 181–195.
- Bowman, A. W. (1984). “An Alternative Method of Cross-Validation for the Smoothing of Density Estimates”. In: *Biometrika* 71, pp. 353–360.
- Brookmeyer, R. and M. G. Gail (1987). “Biases in Prevalent Cohorts”. In: *Biometrics* 43, pp. 739–749.
- England, P. D. and R. J. Verrall (2002). “Stochastic Claims Reserving In General Insurance”. In: *British Actuarial Journal* 8, pp. 443–544.
- Gámiz, M. L., L. Janys, M. D. Martínez-Miranda, and J. P. Nielsen (2013). “Bandwidth selection in marker dependent kernel hazard estimation”. In: *Comput. Stat. Data An.* 68, pp. 155–169.
- Gámiz, M. L., E. Mammen, M. D. Martínez-Miranda, and J. P. Nielsen (2016). “Double one-sided cross-validation of local linear hazards”. In: *J. Roy. Statist. Soc. Ser. B* 78, pp. 1–26.
- Hall, P. (1983). “Large sample optimality of least squares cross-validation in density estimation”. In: *Ann. Stat.* 11, pp. 1156–1174.
- Hart, J.D. and S. Yi (1998). “One-Sided Cross-Validation”. In: *J. Am. Stat. Assoc.* 93, pp. 620–631.
- Heidenreich, N. B., A. Schindler, and S. Sperlich (2013). “Bandwidth selection for kernel density estimation: a review of fully automatic selectors.” In: *AStA Adv. Statist. Anal.* 97, pp. 403–433.
- Hirukawa, M. and M. Sakudo (2014). “Nonnegative bias reduction methods for density estimation using asymmetric kernels”. In: *Comput. Stat. Data An.* 75, pp. 112–123.
- Jacod, J. (1979). *Calcul stochastique et problemes de martingales*. Berlin: Springer.
- Jacod, J. and A. N. Shiryaev (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer.
- Kuang, D., B. Nielsen, and J. P. Nielsen (2008). “Forecasting with the age-period-cohort model and the extended chain-ladder model”. In: *Biometrika* 95, pp. 987–991.
- (2009). “Chain-ladder as maximum likelihood revisited”. In: *Ann. Actuar. Sci* 4, pp. 105–121.
- Lagakos, S. W., L. M. Barraj, and V. De Gruttola (1988). “Nonparametric Analysis of Truncated Survival Data, with Application to AIDS”. In: *Biometrika* 75, pp. 515–523.

- Lee, Y. K., E. Mammen, J. P. Nielsen, and B. Park (2015). “Asymptotics for In-Sample Density Forecasting”. In: *Ann. Stat.* 43, pp. 620–651.
- Mack, T. (2008). “The prediction error of Bornhuetter–Ferguson”. In: *Astin Bull.* 38, pp. 87–103.
- Mammen, E., M. D. Martínez-Miranda, and J. P. Nielsen (2015). “In-sample forecasting applied to reserving and mesothelioma”. In: *Insurance Math. Econom.* 61, pp. 76–86.
- Mammen, E., M. D. Martínez-Miranda, J. P. Nielsen, and S. Sperlich (2011). “Do-validation for kernel density estimation”. In: *J. Am. Stat. Assoc.* 106, pp. 651–660.
- Mammen, E. and J. P. Nielsen (2007). “A general approach to the predictability issue in survival analysis with applications”. In: *Biometrika* 94, pp. 873–892.
- Mandel, M. and R. A. Betensky (2007). “Testing goodness of fit of a uniform truncation model”. In: *Biometrics* 63, pp. 405–412.
- Martínez-Miranda, M. D., B. Nielsen, J. P. Nielsen, and R. Verrall (2011). “Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers”. In: *Astin Bull.* 41, pp. 107–129.
- Martínez-Miranda, M. D., J. P. Nielsen, and S. Sperlich (2009). “One sided cross-validation for density estimation with an application to operational risk”. In: *Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation*. Ed. by G. N. von Gregorion. New Jersey: John Wiley and Sons, pp. 177–195.
- Martínez-Miranda, M. D., J. P. Nielsen, S. Sperlich, and R. Verrall (2013). “Continuous Chain Ladder: Reformulating and generalising a classical insurance problem”. In: *Expert. Syst. Appl.* 40, pp. 5588–5603.
- Martínez-Miranda, M. D., J. P. Nielsen, and R. Verrall (2012). “Double Chain Ladder”. In: *Astin Bull.* 42, pp. 59–76.
- Nielsen, J. P., C. Tanggaard, and M. C. Jones (2009). “Local linear density estimation for filtered survival data”. In: *Statistics* 43, pp. 176–186.
- Norberg, R. (1993). “Prediction of Outstanding Liabilities in Non-Life Insurance”. In: *Astin Bull.* 23, pp. 95–115.
- R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0, <http://www.R-project.org>. Vienna: R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org>.
- Ramlau-Hansen, H. (1983). “Smoothing counting process intensities by means of kernel functions”. In: *Ann. Stat.* 11, pp. 453–466.

- Rudemo, M. (1982). “Empirical Choice of Histograms and Kernel Density Estimators”. In: *Scand. J. Stat.* 9, pp. 65–78.
- Savchuk, O. Y., J. D. Hart, and S. J. Sheater (2010). “Indirect Cross Validation for Density Estimation”. In: *J. Am. Stat. Assoc.* 105, pp. 415–423.
- Tsai, W.-Y (1990). “Testing the assumption of independence of truncation time and failure time”. In: *Biometrika* 77, pp. 169–177.
- Verrall, R., J. P. Nielsen, and A. Jessen (2010). “Including Count Data in Claims Reserving”. In: *Astin Bull.* 40, pp. 871–887.
- Wang, M.-C (1989). “A Semiparametric Model for Randomly Truncated Data”. In: *J. Am. Stat. Assoc.* 84, pp. 742–748.
- Ware, J. H. and D. L. DeMets (1976). “Reanalysis of some baboon descent data”. In: *Biometrics* 32, pp. 459–463.
- Wells, M. T. (1994). “Nonparametric Kernel Estimation in Counting Processes with Explanatory Variables”. In: *Biometrika* 81, pp. 795–801.

3

Smooth backfitting of multiplicative structured hazards

This chapter is a working paper. It is joint work with my supervisors Jens P. Nielsen, E. Mammen and Maria D. Martínez Miranda.

Previous versions of this chapter or parts of it were presented at the following conferences:

- June 2015 19th International Congress on Insurance: Mathematics and Economics (IME), Liverpool, UK.
- August 2015 50th Actuarial Research Conference (ARC), Toronto, Canada.
- March 2016 12th German Probability and Statistics Days 2016, Bochum, Germany.
- June 2016. 3rd International Society for NonParametric Statistics (ISNPS) Conference, Avignon, France.

Smooth backfitting of multiplicative structured hazards

M. Hiabu^a, E. Mammen^b, María D. Martínez Miranda^a, Jens P. Nielsen^a

^a*Cass Business School, City, University of London, United Kingdom*

^b*Institute for Applied Mathematics, Heidelberg University, Germany*

Abstract

We propose a smooth backfitting approach to non-parametrically estimate a multiplicative separable hazard function. Motivated by Mammen, Linton, and Nielsen (1999), the approach is based on a least square criterium projecting a pilot estimator to the space of multiplicative separable functions. We achieve optimal one-dimensional convergence rates, which are independent of the dimension of the hazard function. Compared to existing, literature our approach only needs second order kernels and derivatives and also allows the hazard to have non-rectangular support. We provide an application for the estimation of the reserve in general insurance where the data has triangular support.

Keywords: Structured model; Multiplicative hazard; Hazard estimation; Local linear kernel estimation; Survival data.

3.1 Introduction

Consider a non-negative random variable T . One might think of T as a survival time, that is the occurrence time of death or failure of any kind. Let $Z = (Z_1, \dots, Z_d)$ be a d -dimensional covariate process which is observed until the survival time. We are interested in the conditional hazard

$$a(t|Z) = \lim_{h \downarrow 0} h^{-1} \Pr [T \in [t, t+h) | T \geq t, \{Z(s), s \leq t\}]. \quad (3.1)$$

We assume that

$$a(t|Z) = \alpha(t, Z(t)), \quad (3.2)$$

where α is some unknown smooth function depending on the time t and the value of the covariate at only the time point t . In many cases, T might be subject to some filtering. Filtered observations are present in a vast variety of topics including right censoring in experimental studies like clinical trials or left truncation in insurance loss data. A first version of the non-parametric model (3.2) was introduced in Beran (1981) where the author only considered time independent covariates and a filtering scheme of only right censoring. Dabrowska (1987) derives weak converges of the estimator presented there. The more general model, that is with time dependent covariates and also more general filtering patterns, are analysed in McKeague and Utikal (1990) and Nielsen and Linton (1995) as part of a counting process model. Here, one observes n independent and identically distributed copies of the process (N, Y, Z) , where Y is a predictable process and N a counting process with intensity

$$\lambda(t) = \alpha(t, Z(t))Y(t), \quad (3.3)$$

The multiplicative intensity assumption (3.3) of the counting process is known as Aalen's multiplicative intensity model. Andersen et al. (1993) give a comprehensive overview of how to embed various survival data, including model (3.1), into this counting process formulation.

Non-parametric approaches like (3.2) and (3.3) are often favoured since they have minimal assumptions on the underlying model and are thus more robust than a parametric

approach. However, the optimal convergence rate decreases rapidly with higher dimensions which is also known as curse of dimensionality. This weakness can be overcome by separable structure assumptions on the underlying hazard, see also Stone (1985). It also gives the advantage of better visualisation and interpretations of the components. In this paper we will assume that the conditional hazard is multiplicative, i.e.,

$$\alpha(t, z) = \alpha_0(t)\alpha_1(z_1) \cdots \alpha_d(z_d). \quad (3.4)$$

Model (3.4) is considered in Gámiz et al. (2013) and Linton, Nielsen, and Van de Geer (2003). Linton, Nielsen, and Van de Geer (2003) estimate the components of (3.4) based on marginal integration (Linton and Nielsen, 1995), and derive the optimal one-dimensional convergence rate of $n^{-2/5}$. Since marginal integration estimators are not efficient, an additional backfitting step (Hastie and Tibshirani, 1990; Linton, 1997; Linton, 2000) is applied afterwards to overcome that drawback.

In this paper, we will estimate the components of (3.4) by a projection approach based on least squares. It is motivated by the smooth backfitting approach of Mammen, Linton, and Nielsen (1999) in regression. Compared to Linton, Nielsen, and Van de Geer (2003), we achieve two major improvements. Firstly, we do not need higher derivatives of the hazard function and higher order kernels. In Linton, Nielsen, and Van de Geer (2003), it is assumed that $(2r+1)/3 > d+1$, where r is the order of the used kernel and also the required order of continuous differentiability of the hazard function α . Despite having asymptotic advantage, higher order kernels are known to often have poor performance for reasonable sample sizes (Marron and Wand, 1992; Marron, 1994). In our approach we only need second order kernels and only two-times differentiability of the hazard, independent of the dimension, d , of the covariates. Secondly, marginal integration has a weak point arising from its inner idea. It only works if the support of the hazard is rectangular. The approach of this paper works for quite general supports, see the assumptions in Section 3.4. A rectangular support will for instance not be given in those cases where T is subject to truncation with respect to Z . In Section 3.5 we will present an application where this is the case and the support is a triangle.

3.2 Aalen's multiplicative intensity model

We consider Aalen's multiplicative intensity model. It allows for very general observations schemes. It covers filtered observations arising from left truncation and right censoring but also more complicated changes of occurrence and exposure. In the next section we describe how to embed left truncation and right censoring into this framework. In contrast to Linton, Nielsen, and Van de Geer (2003) we will hereby allow the filtering to be correlated to the survival time and be represented in the covariate process. We briefly summarise the general model we are assuming.

We observe n iid copies of the stochastic processes $(N(t), Y(t), Z(t))$, $t \in [0, R_0]$, $R_0 > 0$. Here, N denotes a right-continuous counting process which is zero at time zero and has jumps of size one. The process Y is left-continuous and takes values in $\{0, 1\}$ where the value 1 indicates that the i 'th individual is under risk. Finally, Z is a d -dimensional left-continuous covariate process with values in a rectangle $\prod_{j=1}^d [0, R_j] \subset \mathbb{R}^d$, $j = 1, \dots, d$. The multivariate process $((N_1, Y_1, Z_1), \dots, (N_n, Y_n, Z_n))$, $i = 1, \dots, n$, is adapted to the filtration $\mathcal{F}_t$ which satisfies the *usual conditions*. Now we assume that N_i satisfies Aalen's multiplicative intensity model, that is

$$\lambda_i(t) = \lim_{h \downarrow 0} h^{-1} E[N_i((t+h)-) - N_i(t-)| \mathcal{F}_{t-}] = \alpha(t, Z_i(t))Y_i(t).$$

The deterministic function $\alpha(t, z)$ is called hazard function and is the failure rate of an individual at time t given the covariate $Z_i(t) = z$.

3.2.1 Left truncation and right censoring time as covariates

Let us assume that we have iid observations $(T_i, Z_i(t))$, $i = 1, \dots, n$. In Linton, Nielsen, and Van de Geer (2003) it is assumed that the support of $(T_1, Z_1(T_1))$ equals the whole rectangle $\mathcal{R} = \prod_{j=1}^d R_j$. This is necessary in the approach of Linton, Nielsen, and Van de Geer (2003), since the marginal integration estimator of Linton and Nielsen (1995) it is based on would otherwise be inconsistent.

The most prominent example for Aalen's multiplicative intensity model is filtered observation due to left truncation and right censoring. If the censoring and truncation variables carry information about the hazard function, i.e., they are not independent to

the survival time T , one would like to have them included in the covariates. But this implies that the support of $(T_1, Z_1(T_1))$ will not equal $\mathcal{R} = \prod_{j=0}^d R_j$. The approach of this paper allows the observations to have support on only a subset, say $\mathcal{X} \subseteq \mathcal{R} = \prod_{j=0}^d R_j$.

We now show how to embed covariates with truncation and censoring information into Aalen's multiplicative intensity model. Every covariate coordinate can carry individual truncation information as long as it corresponds to left truncation. To be more precise, we combine time and the covariates into one $d + 1$ -dimensional vector $X = (T, Z)$, and assume that the observation X is left truncated. That is, we observe X if and only if $(T, Z(T)) \in \mathcal{I}$, where the set $\mathcal{I}$ is compact and it holds that $(t_1, Z(t_1)) \in \mathcal{I}$ and $t_2 \geq t_1$, then $(t_2, Z(t_2)) \in \mathcal{I}$, *a.s.*. The set $\mathcal{I}$ is allowed to be random but is independent to T given the given the covariate process Z . Furthermore, T is subject to right censoring with censoring time C . We assume that also T and C are conditional independent given the covariate process Z . This includes the case that censoring time equals one covariate coordinate. Concluding, we observe n *iid* copies of $(\tilde{T}, Z^*, \mathcal{I}, \delta)$ where $\delta = \mathbb{1}(T^* < C)$, $\tilde{T} = \min(T^*, C)$, and where (T^*, Z^*) is the truncated version of X , i.e., $(T^*, Z^*(T^*)) \in \mathcal{I}$.

Then, we can define the counting process N_i as

$$N_i(t) = \mathbb{1} \left\{ \tilde{T}_i \leq t, \delta_i = 1 \right\},$$

with respect to the filtration $\mathcal{F}_{i,t} = \sigma \left(\left\{ \tilde{T}_i \leq s, Z_i^*(s), \mathcal{I}_i, \delta_i : s \leq t \right\} \cup \mathcal{N} \right)$, where $\mathcal{N} = \{A \mid A \subseteq B, \text{ with } B \in \mathcal{B}(\mathcal{S}), \Pr(B) = 0\}$. With straight forward computations one can conclude that under the setting above, including (3.2), it is straight forward to verify that Aalen's multiplicative intensity model is satisfied with

$$\alpha_z(t) = \alpha(t, z_1, \dots, z_d) = \lim_{h \downarrow 0} h^{-1} \Pr\{T \in [t, t+h) \mid T \geq t, Z_i(t) = z\},$$

$$Y_i(t) = \mathbb{1}\{(t, Z_i^*(t)) \in \mathcal{I}_i, t \leq \tilde{T}_i\}.$$

3.3 Estimation

3.3.1 Unstructured estimation of the hazard

We introduce the notation $X_i(t) = (t, Z_i(t))$. We also set $x = (t, z)$, with $x_0 = t$, $x_1 = z_1, \dots, x_d = z_d$.

To estimate the components of the structured hazard in (3.6) below, we will need an unstructured pilot estimator of the hazard α first. We propose the local linear kernel estimator, $\hat{\alpha}^{LL}(x)$, based on least squares (cf. Nielsen (1998)). Its value in x is defined as the minimiser $\hat{\theta}_0$ in the equation

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} = \arg \min_{\theta_0 \in \mathbb{R}, \theta_1 \in \mathbb{R}^{d+1}} \sum_{i=1}^n \int \left[\left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} dN_i(u) - \theta_0 - \theta_1^T (x - X_i(s)) \right\}^2 - \xi(\varepsilon) \right] \times K_b(x - X_i(s)) Y_i(s) ds. \quad (3.5)$$

In the following, we restrict ourselves to a multiplicative kernel $K(u_0, \dots, u_d) = \prod_{j=0}^d k(u_j)$ and a one-dimensional bandwidth b with $K_b(u) = \prod_{j=0}^d b^{-1} k(b/u_j)$. More general choices would have been possible with the cost of extra notation. The local linear estimator includes boundary corrections so that the bias is of same order at the boundary as in the interior of the support, namely $O(\max_{1 \leq i \leq d+1} b_i^2)$. The local constant estimator achieves only slower rates at the boundary region and local polynomial estimators of higher order, like in regression, have the usual drawback known from higher order kernels, that they only perform poorly as long as sample sizes are not very large.

The solution of the least square minimisation (3.5) can be rewritten as the ratio of smooth estimators of the number of occurrence and the exposure Gámiz et al. (2013).

$$\begin{aligned} \hat{O}^{LL}(x) &= \frac{1}{n} \sum_{i=1}^n \int \left\{ 1 - (x - X_i(s)) D(x)^{-1} c_1(x) \right\} K_b(x - X_i(s)) dN_i(s), \\ \hat{E}^{LL}(x) &= \frac{1}{n} \sum_{i=1}^n \int \left\{ 1 - (x - X_i(s)) D(x)^{-1} c_1(x) \right\} K_b(x - X_i(s)) Y_i(s) ds, \end{aligned}$$

where the components of the $(d + 1)$ –dimensional vector c_1 are

$$c_{1j}(x) = n^{-1} \sum_{i=1}^n \int K_b(x - X_i(s))(x_j - X_{ij}(s))Y_i(s)ds, \quad j = 0, \dots, d,$$

and the entries (d_{jk}) of the $(d + 1) \times (d + 1)$ –dimensional matrix $D(x)$ are given by

$$d_{jk}(x) = n^{-1} \sum_{i=1}^n \int K_b(x - X_i(s))(x_j - X_{ij}(s))(x_k - X_{ik}(s))Y_i(s)ds.$$

In this respect, the local linear estimator compares to the the local constant version that can be defined as

$$\begin{aligned} \hat{O}_{LC}(x) &= \kappa_n(x) \sum_{i=1}^n \int K_b(x - X_i(s))dN_i(s), \\ \hat{E}_{LC}(x) &= \kappa_n(x) \sum_{i=1}^n \int K_b(x - X_i(s))Y_i(s)ds, \\ \kappa_n(x) &= \left[\int K_b(x - u) du \right]^{-1} \end{aligned}$$

and

$$\hat{\alpha}_{LC}(x) = \frac{\hat{O}_{LC}(x)}{\hat{E}_{LC}(x)}.$$

Under standard smoothing conditions, if b is chosen of order $n^{-1/(4+d+1)}$, then the bias of $\hat{\alpha}^{LL}(x)$ and $\hat{\alpha}^{LC}(x)$ is of order $n^{-2/(4+d+1)}$ and the variance is of order $n^{-4/(4+d+1)}$, which is the optimal rate of convergence in the corresponding regression problem Stone (1982). For an asymptotic theory of these estimators see Linton, Nielsen, and Van de Geer (2003).

3.3.2 Structured estimator by solution weighted minimization

In the sequel we will assume a multiplicative structure of the hazard α , i.e.,

$$\alpha(x) = \alpha^* \prod_{j=0}^d \alpha_j(x_j), \tag{3.6}$$

where α_j , $j = 0, \dots, d$, are some functions and α^* is a constant. For identifiability of the components, we make the following further assumption:

$$\int \alpha_j(x_j) w_j(x_j) dx_j = 1, \quad j = 0, \dots, d, \quad (3.7)$$

where w_j is some weight function.

We also need the following notation:

$$F_t(z) = \Pr(Z_1(t) \leq z | Y_1(t) = 1), \quad y(t) = E[Y_1(t)].$$

By denoting $f_t(z)$ the density corresponding to $F_t(z)$ with respect to the Lebesgue measure, we also define

$$E(x) = f_t(z)y(t)$$

and $O(x) = E(x)\alpha(x)$.

We define the estimators $\hat{\alpha}^*$ and $\hat{\alpha} = (\hat{\alpha}_0, \dots, \hat{\alpha}_d)$ of the hazard components in (3.6) as solution of the following equation

$$\hat{\alpha}_k(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \hat{O}(x) dx_{-k}}{\int_{\mathcal{X}_{x_k}} \hat{\alpha}^* \prod_{j \neq k} \hat{\alpha}_j(x_j) \hat{E}(x) dx_{-k}} \quad k = 0, \dots, d, \quad (3.8)$$

under the constraint (3.7) with

$$w_k(x_k) = \int_{\mathcal{X}_{x_k}} \prod_{j \neq k} \hat{\alpha}_j(x_j) \hat{E}(x) dx_{-k}.$$

Here $\mathcal{X}_{x_k}$ denotes the set $\{(x_0, \dots, x_{k-1}, x_{k+1}, \dots, x_d) | (x_0, \dots, x_d) \in \mathcal{X}\}$. Furthermore, $\hat{E}$ and $\hat{O}$ are some full-dimensional estimators of E and O and

$$\hat{\alpha}^* = \frac{\int_{\mathcal{X}} \hat{O}(x) dx}{\int_{\mathcal{X}} \prod_{j=0}^d \hat{\alpha}_j(x_j) \hat{E}(x) dx}.$$

We will discuss below that the equation has a solution with probability tending to one. In the next section we will show asymptotic properties of the estimator. We will see that we do not require that the full-dimensional estimators $\hat{E}$ and $\hat{O}$ are consistent. We will

only need asymptotic consistency of marginal averages of the estimators, see below. This already highlights that our estimator efficiently circumvents the curse of dimensionality.

In practise, system (3.8) can be solved by the following iterative procedure:

$$\hat{\alpha}_k^{(r+1)}(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \hat{O}(x) dx_{-k}}{\int_{\mathcal{X}_{x_k}} \prod_{j=0}^{k-1} \hat{\alpha}_j^{(r+1)}(x_j) \prod_{j=k+1}^{d+1} \hat{\alpha}_j^{(r)}(x_j) \hat{E}(x) dx_{-k}}, \quad k = 0, \dots, d \quad (3.9)$$

After a finite number of cycles or after a termination criterion applies the last values of $\hat{\alpha}_k^{(r+1)}(x_k)$, $k = 0, \dots, d$, are multiplied by a factor such that the constraint (3.7) is fulfilled with the above choice of $w_k(x_k)$. This can always be achieved by multiplication with constants. This gives the backfitting approximations of $\hat{\alpha}_k(x_k)$ for $k = 0, \dots, d$.

The estimator $\hat{\alpha}$ can be motivated as a weighted least squares estimator with random data adaptive weights. To see this consider the estimator $\bar{\alpha}_j$ that minimizes

$$\min_{\bar{\alpha}_j} \int_{\mathcal{X}} \left\{ \tilde{\alpha}(x) - \bar{\alpha}^* \prod_{j=0}^d \bar{\alpha}_j(x_j) \right\}^2 w(x) dx, \quad (3.10)$$

where $w(x)$ is some weighting and where $\tilde{\alpha}(x) = \hat{O}(x)/\hat{E}(x)$ is an unconstrained full-dimensional estimator of α . This gives

$$\bar{\alpha}^* = \frac{\int_{\mathcal{X}} \tilde{\alpha}(x) \prod_{j=0}^d \bar{\alpha}_j(x_j) w(x) dx}{\int_{\mathcal{X}} \left\{ \prod_{j=0}^d \bar{\alpha}_j(x_j) \right\}^2 w(x) dx},$$

and $(\bar{\alpha}_0, \dots, \bar{\alpha}_d)$ can be described via the backfitting equation

$$\bar{\alpha}_k(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \tilde{\alpha}(x) \prod_{j \neq k} \bar{\alpha}_j(x_j) w(x) dx_{-k}}{\int_{\mathcal{X}_{x_k}} \bar{\alpha}^* \left\{ \prod_{j \neq k} \bar{\alpha}_j(x_j) \right\}^2 w(x) dx_{-k}}, \quad k = 0, \dots, d. \quad (3.11)$$

The asymptotic variance of kernel estimators of α is proportional to $\alpha(x)/E(x)$, see e.g. Linton and Nielsen (1995). This motivates the choice $w(x) = E(x)/\alpha(x)$. However, this choice is not possible because $E(x)$ and $\alpha(x)$ are unknown. One could use $w(x) = \check{E}(x)/\check{\alpha}(x)$ where $\check{E}(x)$ and $\check{\alpha}(x)$ are some pilot estimators of E and α . We follow another idea and we propose to weight the minimization (3.10) with its solution. We choose

$$w(x) = \frac{\hat{E}(x)}{\prod_i \hat{\alpha}_i(x)}, \quad (3.12)$$

and heuristically, by putting $\bar{\alpha}_j = \hat{\alpha}_j$ and by plugging (3.12) into (3.11) we get (3.8). The next section discusses existence and asymptotic properties of the solution $\hat{\alpha}_j$ of (3.8).

3.4 Properties of the estimator

The estimator $\hat{\alpha}_j$ is defined as solution of a nonlinear operator equation. We will approximate this equation by a linear equation that can be interpreted as equation that arises in nonparametric additive regression models. We will show that the solution of the linear equation approximates $\hat{\alpha}_j$. The linear equation and its solution is well understood from the theory of additive models. This will be our essential step to arrive at an asymptotic understanding of our estimator $\hat{\alpha}_j$.

For our main theorem we make the following assumptions. We hereby do not make assumptions on the full support $\mathcal{R}$ but only on a subset $\mathcal{X} \subseteq \mathcal{R}$.

A1 The function $E(x)$ is two times continuously differentiable and $\inf_{x \in \mathcal{X}} E(x) > 0$.

A2 The hazard α is two times continuously differentiable and $\inf_{x \in \mathcal{X}} \alpha(x) > 0$.

A3 The kernel K has compact support which is without loss of generality supposed to be $[-1, 1]$. Furthermore it is symmetric and continuous.

A4 It holds that $nb^5 \rightarrow c_b$ for a constant $0 < c_b < \infty$ as $n \rightarrow \infty$.

A5 It holds that

$$\int_{\mathcal{X}_{x_j, x_k}} \frac{1}{O_j(x_j)O_k(x_k)} dx_j dx_k < \infty$$

for $j, k = 0, \dots, d$, $j \neq k$, where $O_j(x_j) = \int p(x) dx_{-j}$ and $p(x) = \prod_{j=0}^d \alpha_j(x_j)E(x)$ and where $\mathcal{X}_{x_j, x_k}$ denotes the set $\{(x_l : l \in \{0, \dots, d\} \setminus \{j, k\}) \mid (x_0, \dots, x_d) \in \mathcal{X}\}$.

A6 It holds that the two-dimensional marginal densities $O_{j,k}(x_j, x_k) = \int p(x) dx_{-(j,k)}$ are bounded from above and bounded away from 0.

A7 The projections of $\mathcal{X}$ and $\mathcal{R} = [0, R_0] \times \prod_{i=1}^d [0, R_i]$ to their j 'th ($j = 0, \dots, d$) coordinate are equal, that is

$$\bigcup_{x_j \in [0, R_j]} \mathcal{X}_{x_j} = [0, R_0] \times \prod_{k \neq j} [0, R_k], \quad j = 0, \dots, d.$$

A8 For some $\delta > 0$ it holds that for $j, k = 0, \dots, d, j \neq k$

$$\begin{aligned} \int_{\mathcal{X}_{x_j, x_k}} \frac{1}{O_j^{1+\delta}(x_j)O_k(x_k)} dx_j dx_k &< \infty, \\ \sup_{x_k \in \mathcal{X}_k} \int_{\mathcal{X}_{j(x_k)}} \frac{1}{O_j^{1-\delta}(x_j)O_k(x_k)} dx_j &< \infty, \\ \sup_{x_k \in \mathcal{X}_k} \int_{\mathcal{X}_{j(x_k)}} \frac{1}{O_j^{1/2}(x_j)O_k^{1/2}(x_k)} dx_j &< \infty, \end{aligned}$$

where $\mathcal{X}_k = \{x_k \mid (x_0, \dots, x_d) \in \mathcal{X} \text{ for some values of } (x_l : l \neq k)\}$ and $\mathcal{X}_{j(x_k)} = \{x_j \mid (x_0, \dots, x_d) \in \mathcal{X} \text{ for some values of } (x_l : l \notin \{j, k\})\}$.

Note that assumptions A1-A4 are standard in kernel smoothing theory. In Assumptions A5 and A6 we only assume that the two-dimensional marginal densities of p are bounded from above and bounded away from 0, but we do not make the assumption that the one-dimensional marginal densities have this property. This allows that the support of a two-dimensional marginal density O_{jk} has a triangle shape $\{(x_j, x_k) : x_j + x_k \leq c; x_j, x_k \geq 0\}$ for some constant $c > 0$. This can be easily seen. Suppose for simplicity that O_{jk} is the uniform density on the triangle. Then $O_j(x_j) = 2c^{-2}(c - x_j)_+$ and $O_k(x_k) = 2c^{-2}(c - x_k)_+$ and we have

$$\int \frac{1}{O_j(x_j)O_k(x_k)} dx_j dx_k = \int_{x_j+x_k \leq c; x_j, x_k \geq 0} \frac{2}{c^2} \frac{1}{(c - x_j)(c - x_k)} dx_j dx_k < \infty.$$

Thus, our assumption A5 on one-dimensional marginals is fulfilled. One can easily verify that also A8 holds for this example. This discussion can be extended to other shapes of two-dimensional marginals that differ from rectangle supports.

The estimators $\hat{\alpha}_0, \dots, \hat{\alpha}_d$ of (3.8) can be rewritten as solutions of

$$\int_{\mathcal{X}_{x_k}} \hat{O}(x) dx_{-k} - \int_{\mathcal{X}_{x_k}} \hat{\alpha}^* \prod_j \hat{\alpha}_j(x_j) \hat{E}(x) dx_{-k} = 0, \quad k = 0, \dots, d.$$

Since, $\int_{\mathcal{X}_{x_k}} O(x) dx_{-k} - \int_{\mathcal{X}_{x_k}} \alpha^* \prod_j \alpha_j(x_j) E(x) dx_{-k} = 0$, the difference of those two terms is zero as well, and we have

$$\begin{aligned} 0 &= \widehat{\Delta}_k(x_k) - \int_{\mathcal{X}_{x_k}} \left\{ \widehat{\alpha}^* \prod_j \widehat{\alpha}_j(x_j) - \alpha^* \prod_{j=0}^d \alpha_j(x_j) \right\} \widehat{E}(x) dx_{-k} \\ &= \widehat{\Delta}_k(x_k) - \int_{\mathcal{X}_{x_k}} \left[(1 + \widehat{\delta}^*) \prod_{j=0}^d \{1 + \widehat{\delta}_j(x_j)\} - 1 \right] \prod_{j=0}^d \alpha_j(x_j) \widehat{E}(x) dx_{-k}, \end{aligned} \quad (3.13)$$

where

$$\begin{aligned} \widehat{\Delta}_k(x_k) &= \int_{\mathcal{X}_{x_k}} \{ \widehat{O}(x) - O(x) \} dx_{-k} + \int_{\mathcal{X}_{x_k}} \alpha^* \prod_{j=0}^d \alpha_j(x_j) \{ \widehat{E}(x) - E(x) \} dx_{-k}, \\ \widehat{\delta}_j(x_j) &= \frac{\widehat{\alpha}_j(x_j) - \alpha_j(x_j)}{\alpha_j(x_j)}, \\ \widehat{\delta}^* &= \frac{\widehat{\alpha}^* - \alpha^*}{\alpha^*}. \end{aligned}$$

Note that $\widehat{\delta}$ is defined as root of a non-linear operator. Motivated by (3.13), we define an approximation, $\bar{\delta}^*$ and $\bar{\delta}_j(x_j)$ ($0 \leq j \leq d$), as solution of the linear equation

$$\int_{\mathcal{X}_{x_k}} \left[\bar{\delta}^* + \sum_{j=0}^d \bar{\delta}_j(x_j) \right] \alpha^* \prod_{j=0}^d \alpha_j(x_j) \widehat{E}(x) dx_{-k} = \widehat{\Delta}_k(x_k) \quad (3.14)$$

under the constraint

$$\int \bar{\delta}_k(x_k) \left[\int \prod_{j=0}^d \alpha_j(x_j) \widehat{E}(x) dx_{-k} \right] dx_k = 0,$$

where

$$\bar{\delta}^* = \frac{\int_{\mathcal{X}} \{ \widehat{O}(x) - O(x) \} dx + \int_{\mathcal{X}} \alpha^* \prod_{j=0}^d \alpha_j(x_j) \{ \widehat{E}(x) - E(x) \} dx}{\int_{\mathcal{X}} \alpha^* \prod_{j=0}^d \alpha_j(x_j) \widehat{E}(x) dx}.$$

Note that the constraint is identical to (3.7) for the choice $w_k(x_k) = \int \prod_{j \neq k} \alpha_j(x_j) \widehat{E}(x) dx_{-k}$ if we replace the right hand side of (3.7) by $\int \prod_{j=0}^d \alpha_j(x_j) \widehat{E}(x) dx$. This norming cannot be used in practice because α is unknown but it will simplify the theoretical discussion and the results can be carried over to feasible weighting.

This can be rewritten to an integral equation of the second kind

$$\bar{\delta}_k(x_k) + \sum_{j \neq k} \int_{\mathcal{X}_{j(x_k)}} \widehat{\pi}_{k,j}(x_k, x_j) \bar{\delta}_j(x_j) dx_j = \widehat{\mu}_k(x_k) - \bar{\delta}^*,$$

with

$$\begin{aligned}
 \tilde{O}(x) &= \alpha^* \prod_{j=0}^d \alpha_j(x_j) \hat{E}(x), \\
 \tilde{O}_{j,k}(x_j, x_k) &= \int \tilde{O}(x) \, dx_{-(j,k)} \\
 \tilde{O}_k(x_k) &= \int \tilde{O}(x) \, dx_{-k} \\
 \hat{\pi}_{k,j}(x_k, x_j) &= \frac{\tilde{O}_{j,k}(x_j, x_k)}{\tilde{O}_k(x_k)}, \\
 \hat{\mu}_k(x_k) &= \frac{\hat{\Delta}_k(x_k)}{\tilde{O}_k(x_k)}.
 \end{aligned}$$

Note that all these functions depend on n . The integral equation can also be simply written as $\bar{\delta} + \hat{\pi}\bar{\delta} = \hat{\mu} - \bar{\delta}^*$, where $\hat{\pi}$ is the integral operator with kernel $\hat{\pi}_{k,j}$, see Mammen, Støve, and Tjøstheim (2009) and Mammen and Yu (2009). We will show that $\bar{\delta}$ approximates $\hat{\delta}$. Before we come to this point we state a proposition that gives the asymptotics for $\bar{\delta}$.

For the next results we need some conditions on the estimators $\hat{E}$ and $\hat{O}$. We decompose $\hat{\mu}_k$ into three terms $\hat{\mu}_k = \hat{\mu}_k^A + \hat{\mu}_k^B + \hat{\mu}_k^C$, that depend on n . For some deterministic functions $O^*(x)$ and $E^*(x)$ these terms are defined as:

$$\begin{aligned}
 \hat{\mu}_k^A(x_k) &= \frac{\int_{\mathcal{X}_{x_k}} \prod_{j=0}^d \alpha_j(x_j) \left\{ \hat{E}(x) - E^*(x) \right\} dx_{-k} + \int_{\mathcal{X}_{x_k}} \left\{ \hat{O}(x) - O^*(x) \right\} dx_{-k}}{\tilde{O}_k(x_k)}, \\
 \hat{\mu}_k^B(x_k) &= \frac{\int_{\mathcal{X}_{x_k}} \prod_{j=0}^d \alpha_j(x_j) \left\{ E^*(x) - E(x) \right\} dx_{-k} + \int_{\mathcal{X}_{x_k}} \left\{ O^*(x) - O(x) \right\} dx_{-k}}{O_k(x_k)}, \\
 \hat{\mu}_k^C(x_k) &= \left[\frac{O_k(x_k)}{\tilde{O}_k(x_k)} - 1 \right] \hat{\mu}_k^B(x_k),
 \end{aligned}$$

such that with

$$\pi_{k,j}(x_k, x_j) = \frac{\int \prod_{j=0}^d \alpha_j(x_j) E(x) dx_{-(k,j)}}{\int \prod_{j=0}^d \alpha_j(x_j) E(x) dx_{-k}}$$

and $\bar{\delta}^{*,r} = \int \hat{\mu}_k^r(x_k) \tilde{O}_k(x_k) \, dx_k$ for $r \in \{A, B, C\}$ the following assumptions hold:

B1 It holds that $\int \tilde{O}(x)^2 dx = O_P(1)$ and

$$\tilde{O}_{j,k}(x_j, x_k) - O_{j,k}(x_j, x_k) = o_P((\log n)^{-1/2})$$

uniformly over $0 \leq j < k \leq d$ and x_j, x_k , where $O_{j,k}(x_j, x_k) = \int O(x) dx_{-(j,k)}$.

B2

$$\sup_{x_j} |O_j^{1/2}(x_j) \hat{\mu}_j^A(x_j)| = O_P((\log n)^{1/2} n^{-2/5})$$

for $0 \leq j \leq d$, where $O_j(x_j) = \int O(x) dx_{-j}$.

B3 For x_j with $O_j(x_j) > 0$ it holds that

$$n^{2/5} \hat{\mu}_j^A(x_j) \rightarrow N(0, \sigma_j^2(x_j))$$

for $0 \leq j \leq d$ with some function $\sigma_j^2(x_j) > 0$.

B4

$$\int \hat{\mu}_j^A(x_j)^2 O_j(x_j) dx_j = O_P(n^{-4/5})$$

and

$$\int \hat{\mu}_j^B(x_j)^2 O_j(x_j) dx_j = O(n^{-4/5})$$

for $0 \leq j \leq d$.

B5 It holds that

$$\sup_{x_j \in \mathcal{X}_j} O_j^{1/2}(x_j) \int_{\mathcal{X}_{k(x_j)}} \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)} \hat{\mu}_k^A(x_k) dx_k = o_P(n^{-2/5}).$$

Proposition 3.1. *Make the assumptions [A1]–[A8], [B1]–[B5]. Then the function $\bar{\delta} = (\bar{\delta}_0, \dots, \bar{\delta}_d)$, introduced in (3.14), exists and is uniquely defined, with probability tending to one. Moreover, it has the following expansion:*

$$\left\| \bar{\delta} - \hat{\mu}^A - (I - \pi)^{-1}(\hat{\mu}^B - \bar{\delta}^{B,*}) \right\|_{O, \infty} = o_P(n^{-2/5}),$$

where, for a function $f(x) = (f_0(x_0), \dots, f_d(x_d))^\top$, we define

$$\|f\|_{O, \infty} = \sup_{x \in \mathcal{X}} \max_{0 \leq j \leq d} |O_j^{1/2}(x_j) f_j(x_j)|.$$

From the proposition we get as a corollary the asymptotic distribution of $\bar{\delta}_j(x_j)$.

Proposition 3.2. *Make the assumptions [A1]–[A8], [B1]–[B5]. Then for x_j ($0 \leq j \leq d$) with $O_j(x_j) > 0$ it holds that*

$$n^{2/5}\{\bar{\delta}_j(x_j) - [(I - \pi)^{-1}(\hat{\mu}^B - \bar{\delta}^{B,*})]_j(x_j)\} \rightarrow N(0, \sigma_j^2(x_j)),$$

in distribution. Under the additional assumption $\hat{\mu}_j^B(x_j) = O(n^{-2/5})$ we have that the bias $[(I - \pi)^{-1}(\hat{\mu}^B - \bar{\delta}^{B,})]_j(x_j)$ is of order $O(n^{-2/5})$.*

Equation (3.13) can be rewritten as

$$\hat{\mathcal{F}}(\hat{\delta}^*, \hat{\delta}_0, \dots, \hat{\delta}_d) = 0,$$

where

$$\begin{aligned} \hat{\mathcal{F}}(f^*, f_0, \dots, f_d)(x) = & \left(\int_{\mathcal{X}_{x_k}} \left[(1 + f^*) \prod_{j=0}^d \{1 + f_j(x_j)\} - 1 \right] \right. \\ & \left. \times \prod_{j=0}^d \alpha_j(x_j) \hat{E}(x) dx_{-k} - \hat{\Delta}_k(x_k) \right)_{k=0, \dots, d}. \end{aligned}$$

The following theorem states that $\bar{\delta}$ is indeed a good approximation of the relative estimation error $\hat{\delta}$.

Theorem 3.3. *Under assumptions [A1]–[A8], [B1]–[B5] it holds that with probability tending to one there exists a solution $\hat{\delta}^*$ and $\hat{\delta} = (\hat{\delta}_0, \dots, \hat{\delta}_d)$ of the equation $\hat{\mathcal{F}}(f^*, f_0, \dots, f_d) = 0$ with*

$$\begin{aligned} \|\hat{\delta} - \bar{\delta}\|_{O, \infty} &= o_p(n^{-2/5}), \\ \hat{\delta}^* - \bar{\delta}^* &= o_p(n^{-2/5}). \end{aligned}$$

For this solution we get that

$$n^{2/5}\{(\hat{\alpha}_j - \alpha_j)(x_j) - \alpha_j(x_j)[(I - \pi)^{-1}(\hat{\mu}^B - \bar{\delta}^{B,*})]_j(x_j)\} \rightarrow N(0, \alpha_j^2(x_j)\sigma_j^2(x_j)),$$

in distribution, for x_j ($0 \leq j \leq d$) with $O_j(x_j) > 0$.

3.5 Application: Outstanding loss liabilities

In order to illustrate the practical aspects of the proposed approach, we analyze the reported claims from a motor business line in Cyprus.

This is exactly the same data set as in Hiabu et al. (2016). In fact the one driving motivation of this paper was to generalize the approach in Hiabu et al. (2016) by using relaxed assumptions.

The data we are considering consist of the number of claims reported between 2004 and 2013. During these 10 years, $n = 58180$ claims were reported. The data is given as $\{(T_1, Z_1), \dots, (T_n, Z_n)\}$, where T_i denotes the underwriting date of claim i , and Z_i the time between underwriting date and the date of report of a claim in days, also called reporting delay. The data, therefore, exist on a triangle, with $T_i + Z_i \leq 31 \text{ December } 2013 = R_0$, which is a subset of the full support $\mathcal{R} = [0, R_0]^2$ ($0 = 1 \text{ January } 2004$). Our aim is to forecast the number of future claims from contracts written in the past which have not been reported yet. Hereby it is implicitly assumed that the maximum reporting delay of a claim is 10 years. Actuaries call this assumption that the triangle is fully run off. In our data set, this is a reasonable assumption, see also Figure 3.1.

To estimate the number of outstanding claims we would like to estimate the conditional hazard given the underwriting date, $\alpha_z(t) = \alpha_1(t)\alpha_2(z)$.

While Hiabu et al. (2016) assume that T and Z are independent, we do not impose such a strong restriction, but only the multiplicativity of the conditional hazard.

To justify their independence assumption, Hiabu et al. (2016) plotted Figure 3.2. The points in the plots are derived by first transforming the data into a triangle with dimension 3654×3654 ,

$$\mathcal{N}_{r^*, s^*} = \sum_{i=1}^n I(X_i = r, Y_i = s), \quad (r^*, s^*) \in \{1, 2, \dots, 3654\}^2,$$

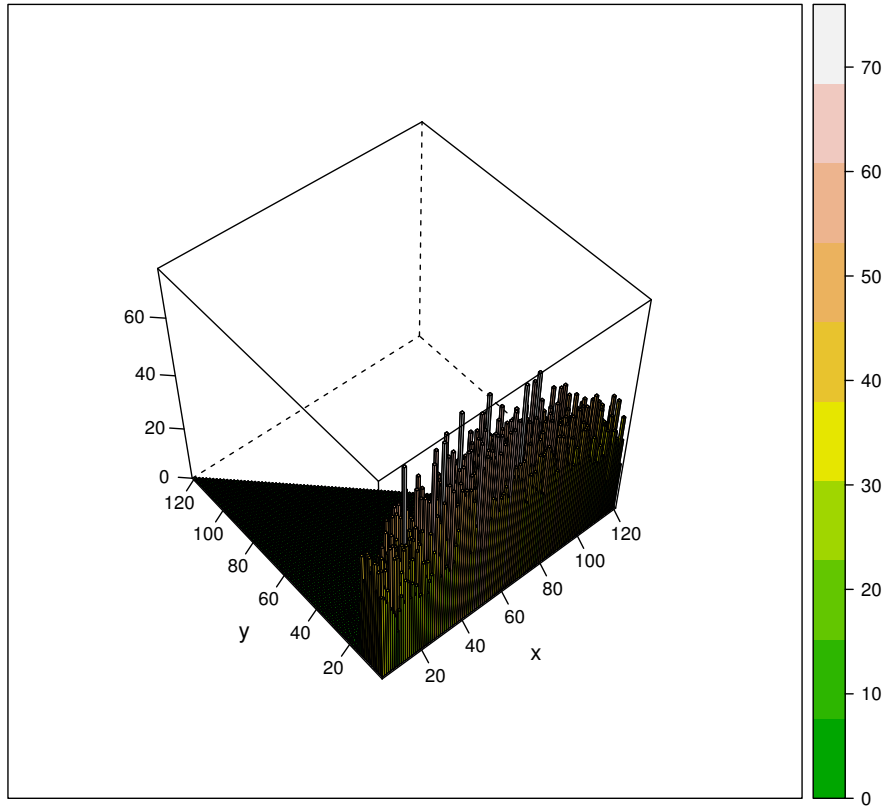
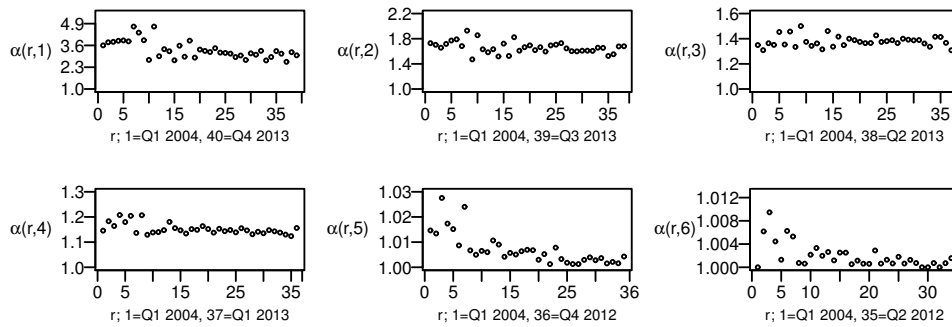


FIGURE 3.1: Histogram of claim numbers of a motor business line between 2004 and 2013.

FIGURE 3.2: Development factors of the first six quarter for individual underwriting quarter.



and then aggregating the data into a quarterly triangle, $(\mathcal{N}_{r,s}^Q)$, with dimension 40×40 . Then, for $k = 1, \dots, 6$, one derives $\bar{\alpha}(r, s) = \sum_{l=1}^{s+1} \mathcal{N}_{r,l} / \sum_{l=1}^s \mathcal{N}_{r,l}$, which are known as development factors under actuaries. The values are displayed in Figure 3.2.

If the multiplicativity assumption would be satisfied the points should lie around a horizontal line in each plot. In Hiabu et al. (2016) it was argued that the multiplicativity assumption might only be violated in the 5th and 6th development quarter because constancy does not show up in the 5th and 6th plot. It was then argued that those quarters do not have great impact on the reserve in order to justify their approach.

In this paper, however, it seems that we are able to catch the dependency structure given in the data set. The development factors plotted are in nature very similar to the hazard we want to estimate and the downward drift seems indeed in all 6 plots to be multiplicative of similar size. Note that it can be argued that the drift in the first 4 plots is not detectable since the multiplicative drift part is so small that it is hidden by the greater noise in those first figures.

We continue with the task of estimating the hazard function. We can not apply our theory directly, since we only observe T if and only if $T \leq R_0 - Z$ which is a right truncation and thus does not fit directly into the model of the previous section. This problem is also considered in Hiabu et al. (2016), and a solution is to transform the random variable T to $T^R = R_0 - T$. This has the result that the right truncation becomes a left truncation, $T^R \geq Z$. Thus, considering the random variable T^R as our variable of interest, we are in the framework of Section 3.2.1 in the previous section. In the notation of the example we now have $T = T^R, d = 1, Z = Z, \delta = 1, \mathcal{I} = \{(t, z) \in \mathcal{R} | 0 \leq z \leq t\}$. We conclude that the counting process $N_i(t) = \mathbb{1}\{T_i^R \leq t\}$, satisfies Aalen's multiplicative intensity model with respect to the filtration given in Section 3.2.1 and

$$\alpha_z(t) = \alpha(t, z) = \lim_{h \downarrow 0} h^{-1} \Pr\{T^R \in [t, t+h) | T^R \geq t, Z(t) = z\},$$

$$Y_i(t) = \mathbb{1}\{(t, Z_i(t)) \in \mathcal{I}, t \leq T_i^{R,*}\}.$$

Therefore we can estimate the unstructured hazard as described in Section 3.3.1. Since estimating the optimal bandwidth via cross-validation, see below, turned out too computationally expensive, we aggregated the triangle $\mathcal{N}_{r^*,s^*}$ into bins of two days, see also

TABLE 3.1: Number of claims forecasts in the real data application. In quarters; 1 = 2014 Q1, 39 = 2022 Q3. We compare the backfitting approach of this paper (MH), the classic chain ladder method (CLM) and the approach of Hiabu et al. (2016).

Future quarter	1	2	3	4	5	6	7	8	9	10	11	12 – 39	Tot.
Hiabu et al. 2016	970	684	422	166	14	5	3	2	1	1	1	0	2270
CLM	948	651	387	148	12	5	3	2	1	1	1	0	2160
MH	872	621	400	130	53	7	4	3	2	1	1	1	2193

Appendix 3.C.

To derive the structured estimators, we also set $\mathcal{X} = \mathcal{I} \setminus \{(0, R_0), (R_0, 0)\}$. Note that that this suffices assumption [A5]. However, this also means that the projections in assumption [A7] do not include the corner points $\{(0, R_0), (R_0, 0)\}$. But since we will also assume [A2] which ensures the continuity of α , the identification on the whole square, including the boundary, will still hold. The components of the multiplicative conditional hazard are then computed as in (3.9). We used a cross-validation method to derive a bandwidth estimate. Further details are given in the Appendix 3.A. After several trials we run the minimization over the set $b_2 \in \{2, 3, 4, 5\}$ and $b_1 \in \{1300, 1400, 1500, 1600, 1700, 1800\}$, and found the cross-validated minimum to be $b_2 = 3$, $b_1 = 1600$ (unit= 2days).

The results of the estimation procedure are given in Figure 3.4 and 3.3. The first figure shows the components of the structured estimator, and the latter one shows the difference, $\tilde{\alpha}(x) - \hat{\alpha}(x)$, of the structured and unstructured estimator.

Finally the reserve can be estimated as

$$R = \sum_{i=1}^n \frac{\int_{T-Z_i}^T \hat{f}_{Z_i}(t) dt}{\int_0^{T-Z_i} \hat{f}_{Z_i}(t) dt}, \quad \hat{f}_z(t) = \hat{\alpha}_1(T-t) \hat{\alpha}_2(z) \exp \left\{ - \int_0^{T-t} \hat{\alpha}_1(s) \hat{\alpha}_2(z) ds \right\}.$$

If one is interested in the 'cash-flow' of the next periods, one can decompose the reserve further. If the future is divided into M periods, each with length $\delta = R_0/M$, then the amount of claims forthcoming in the a th ($a = 1, 2, \dots, M$) period can be then estimated by

$$R(a) = \sum_{i=1}^n \frac{\int_{(T-Z_i+a\delta-1)\wedge T}^{(T-Z_i+a\delta)\wedge T} \hat{f}_{Z_i}(t) dt}{\int_0^{T-Z_i} \hat{f}_{Z_i}(t) dt}.$$

In Table 3.1, we have estimated the number of claims arising in the next quarters. We compare the approach of this paper with the results derived in Hiabu et al. (2016), and

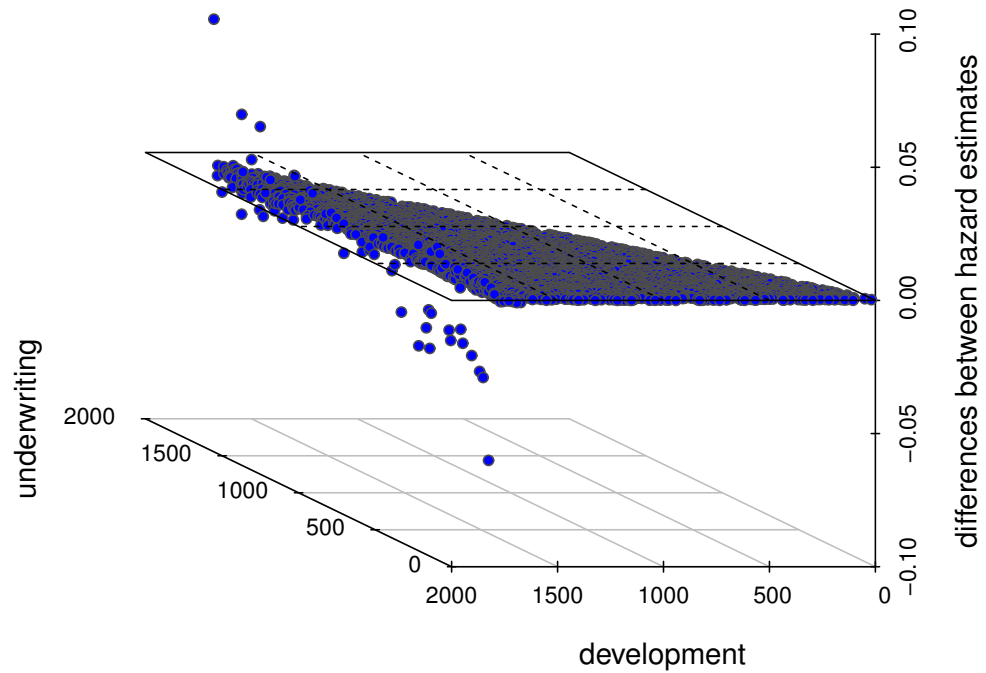


FIGURE 3.3: Difference between structured and unstructured hazard estimator

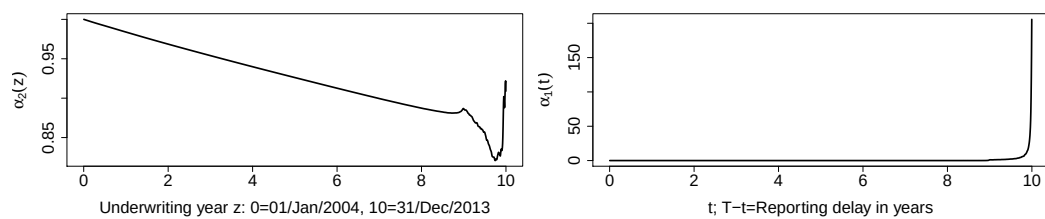


FIGURE 3.4: The estimated hazard components

with the traditional chain ladder method. The two latter approaches have in common that they assume independence between underwriting date, T , and reporting delay, Z . We see that while all approaches estimate a similar total claim number those three approaches have a very different distributions around the quarters than our method. It seems that the obvious violation of the independence assumption has not a big influence on the total claim number size, since it balances the different development pattern arising from different periods out. However, the problem becomes quite serious if one is interested in more detailed estimates like the cash flow.

3.A Bandwidth selection

A crucial part in practice is finding the right amount of smoothing when using non parametric approaches. For our application we will stick to the maybe most straight forward way in estimating the optimal bandwidth - the data-driven cross-validation method.

The data-driven cross-validation method in density estimation goes back to Rudemo (1982) and Bowman (1984). Nowadays, a slightly modified version (see Hall (1983)) is used which aims to minimize the integrated squared error. In our framework, the cross-validation bandwidth has been proposed in Nielsen and Linton (1995). For the practical purpose, in contrast to the previous chapters, we will allow the bandwidth to be different in each direction. Cross-validation arises from the idea to minimize the integrated squared error

$$n^{-1} \sum_{i=1}^n \int_0^{R_0} [\hat{\alpha}\{X_i(s)\} - \alpha\{X_i(s)\}]^2 Y_i(s) ds.$$

By expanding the square, only two of the three summands depend on the bandwidth and are thus considered. While $\int \hat{\alpha}(X_i(s))^2 ds$ is feasible, we have to estimate $\sum_i \int \hat{\alpha}(X_i(s)) \alpha(X_i(s)) Y_i(s) ds$. In cross-validation this is done by the unbiased leave one out estimator

$$\int \hat{\alpha}^{[i]} \{X_i(s)\} dN_i(s),$$

where $\hat{\alpha}^{[i]}$ is the leave one out version which arises from the definition of structured estimator $\hat{\alpha}$ by setting $N_i = 0$. Concluding, we define the cross validation bandwidth

b_{CV} as

$$b_{CV}(K) = \arg \min_b \sum_{i=1}^n \int \hat{\alpha}(X_i(s))^2 ds - 2 \sum_{i=1}^n \int \hat{\alpha}^{[i]} \{X_i(s)\} dN_i(s).$$

Theoretical properties of cross validation in hazard estimation in the one dimensional case are derived in Mammen, Martínez-Miranda, and Nielsen (2015). To our knowledge there is no theoretical analysis of cross-validation in the multivariate hazard case of this paper. An extensive simulation study of the multivariate case can be found in Gámiz et al. (2013).

3.B Proofs

3.B.1 Proof of Proposition 3.1

The proof of this proposition follows the lines of the proof of Theorem 1 in Mammen, Linton, and Nielsen (1999) but it needs some modifications in the last step of the proof because we have weaker assumptions than the ones assumed in the latter theorem. We outline that the first part of the proof in Mammen, Linton, and Nielsen (1999) also goes through under our weaker assumptions and we show how additional arguments can be used in the last part.

Note that under our assumptions [A5], [A6] we get that $\int O_{jk}(x_j, x_k)^2 O_j(x_j)^{-1} O_k(x_k)^{-1} dx_j dx_k < \infty$. As in Lemma 1 in Mammen, Linton, and Nielsen (1999) this implies that for some constants $c, C > 0$

$$c \max_{0 \leq j \leq d} \|\delta_j\| \leq \|\delta_0 + \dots + \delta_d\| \leq C \max_{0 \leq j \leq d} \|\delta_j\| \quad (3.15)$$

for $\delta_j \in \mathcal{L}_j = \{\delta_j : \mathcal{X}_j \rightarrow \mathbb{R} : \int_{\mathcal{X}_j} \delta_j^2(x_j) O_j(x_j) dx_j < \infty, \int_{\mathcal{X}_j} \delta_j(x_j) O_j(x_j) dx_j = 0\}$ where $\|\dots\|$ denotes the norm $\|m(x)\|^2 = \int m(x)^2 O(x) dx$ and $\mathcal{X}_j = \{x_j : x \in \mathcal{X}\}$. Furthermore, one gets that $\|T\| = \sup\{\|T(\delta_0 + \dots + \delta_d)\| : \delta_j \in \mathcal{L}_j \text{ with } \|\delta_0 + \dots + \delta_d\| < 1\} < 1$, where here T is the operator $T = \Psi_d \cdot \dots \cdot \Psi_0$ with

$$\begin{aligned} \Psi_j(\delta_0 + \dots + \delta_d)(x) &= \delta_0(x_0) + \dots + \delta_{j-1}(x_{j-1}) + \delta_j^*(x_j) + \delta_{j+1}(x_{j+1}) + \dots + \delta_d(x_d), \\ \delta_j^*(x_j) &= - \sum_{k \neq j} \int \delta_k(x_k) \pi_{j,k}(x_j, x_k) dx_k. \end{aligned}$$

Furthermore, note that for $j \neq k$ it holds that

$$\begin{aligned} \left\| \frac{\widehat{O}_j(x_j) - O_j(x_j)}{O_j(x_j)} \right\| &= o_P(1), \\ \left\| \frac{\widehat{O}_{j,k}(x_j, x_k)}{O_j(x_j)O_k(x_k)} - \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)O_k(x_k)} \right\| &= o_P(1), \\ \left\| \frac{\widehat{O}_{j,k}(x_j, x_k)}{\widehat{O}_j(x_j)O_k(x_k)} - \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)O_k(x_k)} \right\| &= o_P(1). \end{aligned}$$

These equations follow from [A5], [A6] and [B1]. Note that [A6] and [B1] imply that, uniformly for x_j, x_k it holds that $\widehat{O}_{j,k}(x_j, x_k) - O_{j,k}(x_j, x_k) = o_P(1)O_{j,k}(x_j, x_k)$. This gives that $[\widehat{O}_j(x_j)/O_j(x_j)] - 1 = o_P(1)$, uniformly for $x_j \in \mathcal{X}_j$ and $0 \leq j \leq d$. Together with [A5] and [B1], this implies the three equations. As in Lemma 2 in Mammen, Linton, and Nielsen (1999) we conclude from these equations that

$$\|\widehat{T}\|_n < \gamma$$

for some $\gamma < 1$ with probability tending to one and

$$\|\widehat{T} - T\|_n = o_P(1), \quad \|\widehat{\Psi}_j - \Psi_j\|_n = o_P(1) \quad (0 \leq j \leq d).$$

Here, we define $\widehat{T}$, $\|\dots\|_n$, $\mathcal{X}_{n,j}$, $\widehat{\Psi}_j$ as T , $\|\dots\|$, $\mathcal{X}_j$, Ψ_j but with O_j, π_{jk} replaced by $\widehat{O}_j, \widehat{\pi}_{jk}$ ($0 \leq j, k \leq d; j \neq k$). Arguing as in the first part of Lemma 3 in Mammen, Linton, and Nielsen (1999) this gives that $\bar{\delta}(x) = \bar{\delta}^A(x) + \bar{\delta}^B(x) + \bar{\delta}^C(x)$, where for $r \in \{A, B, C\}$

$$\bar{\delta}^r(x) = \sum_{l=0}^s \widehat{T}^l \widehat{\tau}^r(x) + R^{r,[s]}(x)$$

with $\|R^{r,[s]}\| \leq C\gamma^s$ with probability tending to one for some constant $C > 0$. Here we put

$$\widehat{\tau}^r = \widehat{\Psi}_d \cdot \dots \cdot \widehat{\Psi}_1 (\widehat{\mu}_0^r - \delta^{*,r}) + \dots + \widehat{\Psi}_d (\widehat{\mu}_{d-1}^r - \delta^{*,r}) + (\widehat{\mu}_d^r - \delta^{*,r}).$$

Up to this point we followed closely the arguments in the proof of Theorem 1 in Mammen, Linton, and Nielsen (1999). The arguments of the further parts of the proof of the latter theorem would need that, in our notation,

$$\sup_{x_j \in \mathcal{X}_j} \int_{\mathcal{X}_{k(x_j)}} \frac{\widehat{O}_{j,k}^2(x_j, x_k)}{\widehat{O}_j^2(x_j)O_k(x_k)} dx_k \quad (3.16)$$

is bounded by a constant, with probability tending to one. This would imply that with probability tending to one for some constant $C > 0$ for all functions $g : \mathcal{X}_{k(x_j)} \rightarrow \mathbb{R}$

$$\sup_{x_j \in \mathcal{X}_j} \left| \int_{\mathcal{X}_{k(x_j)}} \frac{\hat{O}_{j,k}(x_j, x_k)}{\hat{O}_j(x_j)} g(x_k) dx_k \right| \leq C \|g\|, \quad (3.17)$$

as can be seen by application of the Cauchy-Schwarz inequality. The proof of Theorem 1 in Mammen, Linton, and Nielsen (1999) shows that this can be used to show that $\sup_{x \in \mathcal{X}, 0 \leq j \leq d} |R_j^{r,[s]}(x)| \leq C\gamma^s$ with probability tending to one for some constant $C > 0$. Furthermore, it implies that

$$\sup_{x \in \mathcal{X}} \max_{0 \leq j \leq d} \left| \left(\bar{\delta} - \hat{\mu}^A - (I - \pi)^{-1}(\hat{\mu}^B - \bar{\delta}^{B,*}) \right)_j(x_j) \right| = o_p(n^{-2/5}).$$

Unfortunately in our setting (3.16) does not hold and thus we cannot follow that (3.17) holds in our setting. One can also check that in general (3.17) does not hold under our assumptions. Thus we do not have that T and $\hat{T}$ map a function with bounded L_2 -norm into a function with bounded L_∞ -norm. This also does not hold if we choose our weighted norm $\|\cdot\|_{O,\infty}$ as L_∞ -norm. We now argue that after twice application of T or $\hat{T}$ a function with bounded $\|\cdot\|$ -norm is transformed into a function with bounded $\|\cdot\|_{O,\infty}$ -norm. This follows from the following two estimates for functions $g : \mathcal{X}_k \rightarrow \mathbb{R}$

$$\int_{\mathcal{X}_{x_j}} \left(\int_{\mathcal{X}_{k(x_j)}} \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)} g(x_k) dx_k \right)^2 O_j^{1-\delta}(x_j) dx_j \leq C \int_{\mathcal{X}_k} O_k(x_k) g^2(x_k) dx_k, \quad (3.18)$$

$$\sup_{x_j \in \mathcal{X}_j} O_j^{1/2}(x_j) \left| \int_{\mathcal{X}_{k(x_j)}} \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)} g(x_k) dx_k \right| \leq C \left(\int_{\mathcal{X}_k} O_k^{1-\delta}(x_k) g^2(x_k) dx_k \right)^{1/2} \quad (3.19)$$

with some constant $C > 0$. Using these bounds one can proceed as in Mammen, Linton, and Nielsen (1999) by using similar arguments as used there. One needs to bound one further term in the above expansion of $\bar{\delta}^r(x)$ because we can bound $\|\cdot\|_{O,\infty}$ -norms only after a double application of T or $\hat{T}$. To bound this term one uses that a function with bounded $\|\cdot\|_{O,\infty}$ -norm is mapped by T and $\hat{T}$ into a function with bounded $\|\cdot\|_{O,\infty}$ -norm. This follows from

$$\sup_{x_j \in \mathcal{X}_j} O_j^{1/2}(x_j) \left| \int_{\mathcal{X}_{k(x_j)}} \frac{O_{j,k}(x_j, x_k)}{O_j(x_j)} g(x_k) dx_k \right| \leq C^* \sup_{x_k \in \mathcal{X}_k} O_k^{1/2}(x_k) |g(x_k)|. \quad (3.20)$$

with some constant $C^* > 0$. For the proof of Proposition 1 it remains to show (3.18)–(3.20). The bound (3.20) follows directly from the last inequality in Condition B5. For

the proof of (3.18) note that the left hand side of (3.18) can be bounded by a constant times

$$\int_{\mathcal{X}_{x_j, x_k}} \frac{1}{O_j^{1+\delta}(x_j)O_k(x_k)} dx_j dx_k \int_{\mathcal{X}_k} O_k(x_k)g^2(x_k)dx_k.$$

Thus, (3.18) follows by application of the first inequality in Condition B5. For the proof of (3.19) note that the left hand side of (3.19) can be bounded by a constant times

$$\left(\sup_{x_k \in \mathcal{X}_k} \int_{\mathcal{X}_{j(x_k)}} \frac{1}{O_j^{1-\delta}(x_j)O_k(x_k)} dx_j \int_{\mathcal{X}_k} O_k^{1-\delta}(x_k)g^2(x_k)dx_k \right)^{1/2}.$$

Here, (3.19) follows by application of the second inequality in Condition B5.

3.B.2 Proof of Proposition 3.2

The statement of Proposition 3.2 follows immediately from (B3) and Proposition 1.

3.B.3 Proof of Theorem 3.3

The main tool to prove this theorem is the Newton-Kantorovich theorem, see for example Deimling (1985). Since this theorem is central in our considerations we will state it here.

Theorem 3.4 (Newton-Kantorovich theorem). *Consider Banach spaces X, Y and a continuous differentiable map $F : B_r(x_0) \subset X \mapsto Y$. Also assume that the following conditions are satisfied*

- (a) $\|F'(x_0)^{-1}F(x_0)\| \leq \gamma,$
- (b) $\|F'(x_0)^{-1}\| \leq \beta,$
- (c) $\|F'(x) - F'(x^*)\| \leq l\|x - x^*\|$ for all $x, x^* \in B_r(x_0),$
- (d) $2\gamma\beta l < 1$ and $2\gamma < r.$

Then the equation

$$F(x) = 0$$

has a unique solution x^* in $\overline{B}_{2r}(x_0)$ and furthermore, x^* can be approximated by Newtons iterative method

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k),$$

and it holds that

$$\|x_k - x^*\| \leq \frac{\gamma}{2^{k-1}} q^{2^k-1}, \quad \text{with } q = 2\gamma\beta l < 1.$$

We come now to the proof of Theorem 3.3.

Proof of Theorem 3.3. We define the deterministic operator $\mathcal{F}$

$$\mathcal{F}(f_0, \dots, f_d)(x) = \left(\int_{\mathcal{X}_{x_k}} \left[\prod_{j=0}^d \{1 + f_j(x_j)\} - 1 \right] \prod_{j=0}^d \alpha_j(x_j) E(x) dx_{-k} \right)_{k=0, \dots, d}.$$

Note that $\mathcal{F}(0) = 0$. The Fréchet derivatives of $\widehat{\mathcal{F}}$ and $\mathcal{F}$ in 0 are

$$\begin{aligned} \widehat{\mathcal{F}}'(0)(f) &= \left(\int_{\mathcal{X}_{x_k}} \sum_j f_j(x_j) \alpha(x) \widehat{E}(x) dx_{-k} \right)_{k=0, \dots, d}, \\ \mathcal{F}'(0)(f) &= \left(\int_{\mathcal{X}_{x_k}} \sum_j f_j(x_j) \alpha(x) E(x) dx_{-k} \right)_{k=0, \dots, d}. \end{aligned}$$

The main idea of our proof is to apply the Newton-Kantorovich theorem, Theorem 3.4, with the mapping $F = \widehat{\mathcal{F}}$, norm $\|\dots\|_{O, \infty}$ and the starting point $x_0 = \bar{\delta}$. We will show that

$$\|\widehat{\mathcal{F}}(\bar{\delta})\|_{O, \infty} = O_p(n^{-4/5}), \quad (3.21)$$

and that $\widehat{\mathcal{F}}'$ is locally Lipschitz around 0, i.e., that there exist constants r^*, C such that with probability tending to one

$$\|\widehat{\mathcal{F}}'(g)(f) - \widehat{\mathcal{F}}'(g^*)(f)\|_{O, \infty} \leq C \|g - g^*\|_{O, \infty} \|f\|_{O, \infty} \quad \text{for all } g, g^* \in B_{r^*}(0) \quad (3.22)$$

Furthermore, we will show, that

$$\mathcal{F}'(0) \text{ is invertible, with } \|\mathcal{F}'(0)^{-1}\|_{O, \infty} < C^*, \quad \text{for some } C^* > 0. \quad (3.23)$$

We now argue that by application of the Newton-Kantorovich theorem (3.21)-(3.23) imply

$$\|\bar{\delta} - \hat{\delta}\|_{O,\infty} = O_p(n^{-4/5}). \quad (3.24)$$

This implies the statement of the theorem.

We now show that (3.21)-(3.23) imply (3.24). Since $\|\bar{\delta}\|_{p,\infty} = o_P(1)$, the inequality (3.22) also holds with a constant r for all $g, g^* \in B_r(\bar{\delta})$ with probability tending to one. This gives condition 3. of the Newton-Kantorovich theorem.

Furthermore, note that [A5] and [B1] imply that $[\hat{O}_j(x_j)/O_j(x_j)] - 1 = o_P(1)$, uniformly for $x_j \in \mathcal{X}_j$ and $0 \leq j \leq d$, as shown in the proof of Proposition 1. This gives $\|\hat{\mathcal{F}}'(0) - \mathcal{F}'(0)\|_{O,\infty} = o_P(1)$. This together with $\|\bar{\delta}\|_{O,\infty} = o_P(1)$ and (3.22) gives

$$\|\hat{\mathcal{F}}'(\bar{\delta}) - \mathcal{F}'(0)\|_{O,\infty} = o_P(1).$$

Therefore with probability tending to one, condition (3.23) also holds if $\mathcal{F}'(0)$ is replaced by $\hat{\mathcal{F}}'(\bar{\delta})$. Thus, (3.21)-(3.23) that conditions 1. - 4. of the Newton-Kantorovich theorem are satisfied with probability tending to one, with $\gamma = C|\hat{\mathcal{F}}(\bar{\delta})|$. This shows (3.24).

It remains to show (3.21), (3.22) and (3.23).

Proof of (3.22). First note that the Fréchet derivative of $\hat{\mathcal{F}}$ in $(g_0, \dots, g_d)$ is given as

$$\begin{aligned} \left(\hat{\mathcal{F}}'(g_0, \dots, g_d)(f_0, \dots, f_d)(x) \right)_k &= \sum_{l=0}^d \int_{\mathcal{X}_{x_k}} f_l(x_l) \prod_{j \neq l} \{1 + g_j(x_j)\} \alpha(x) \hat{E}(x) dx_{-k} \\ &= \sum_{l=0}^d \int_{\mathcal{X}_{x_k}} f_l(x_l) \left[\sum_{\substack{\nu \in \{0,1\}^{d+1} \\ \nu_l=0}} \prod_{j=0}^d g_j(x_j)^{\nu_j} \right] \alpha(x) \hat{E}(x) dx_{-k}. \end{aligned}$$

Claim (3.22) follows by application of Cauchy-Schwarz inequality and Conditions A5, B1.

Proof of (3.23). We have to show that $\mathcal{F}'(0)$ is invertible. For the proof of this claim we start by showing that it is bijective. For the proof of injectivity, assume that $\mathcal{F}'(0)(f) =$

0. We will show that this implies that $f = 0$. It holds that

$$\int_{\mathcal{X}_{x_k}} \sum_{j=0}^d f_j(x_j) \alpha(x) E(x) dx_{-k} = 0, \quad \text{for all } k = 0, \dots, d.$$

Hence,

$$\int_{\mathcal{X}} f_k(x_k) \sum_{j=0}^d f_j(x_j) \alpha(x) E(x) dx = 0, \quad \text{for all } k = 0, \dots, d.$$

Then by summing up over k , we conclude

$$\int_{\mathcal{X}} \left\{ \sum_{j=0}^d f_j(x_j) \right\}^2 \alpha(x) E(x) dx = 0,$$

which implies

$$\sum_{j=0}^d f_j(x_j) = 0, \quad \text{a.e. on } \mathcal{X}.$$

By application of (3.15) this implies that $f = 0$.

Now we check that $\mathcal{F}'(0)$ is surjective. Consider a function $g = (g_0, \dots, g_d)$ with $g_k : \mathcal{X}_{x_k} \mapsto \mathbb{R}$, $k = 0, \dots, d$ such that $\langle \mathcal{F}'(0)(f), g \rangle = 0$ for all $f = (f_0, \dots, f_d)$ with $f_k : \mathcal{X}_{x_k} \mapsto \mathbb{R}$. Since $\mathcal{F}'(0)$ is linear, it is sufficient to show that $g = 0$. By choosing $f = g$, we deduce that

$$\int_{\mathcal{X}} g_k(x_k) \sum_{j=0}^d g_j(x_j) \alpha(x) E(x) dx = 0, \quad \text{for all } k = 0, \dots, d,$$

and with exactly the same arguments as for the injectivity we conclude that $g = 0$. Thus, we have shown that $\mathcal{F}'(0)$ is invertible.

It remains to show that $\mathcal{F}'(0)^{-1}$ is bounded, but this follows directly from the bounded inverse theorem since $\mathcal{F}'(0)$ is bounded.

Proof of (3.21). Since $\|\bar{\delta}\|_{O, \infty} = O_p(\zeta)$ and $\widehat{\mathcal{F}}'$ is Lipschitz a first order Taylor expansion yields

$$\widehat{\mathcal{F}}(\bar{\delta}) = \widehat{\mathcal{F}}(0) + \widehat{\mathcal{F}}'(0)(\bar{\delta}) + O_p(\zeta^2).$$

Equation (3.21) follows from $\widehat{\mathcal{F}}(0) + \widehat{\mathcal{F}}'(0)(\bar{\delta}) = -\widehat{\Delta} + \widehat{\Delta} = 0$.

□

3.C Discrete data

Data is given as $\mathbf{N}_{r',r}$, with $(r', r) \in \mathcal{I}_{disc}$, $\mathcal{I}_{disc} = \{(r', r) \mid r' = 1, \dots, T_0; r = 0, \dots, T_0 - 1 \text{ and } r' \leq r\}$. We define occurrence $O_{r',r}$ and exposure $E_{r',r}$.

$$O_{r',r} = \sum_{j=1}^{n_{r'}} \int_r^{r+1} dN_{r',j}(s) = \mathbf{N}_{r',(T_0-r)},$$

$$E_{r',r} = \sum_{j=1}^{n_{r'}} \int_{r-0.5}^{r+0.5} Y_{r',j}(s) ds = Y_{r',j}(r+0.5) = \sum_{k \leq (T_0-r)} \mathbf{N}_{r',k}.$$

Then the local linear hazar estimator $\tilde{\alpha}$ becomes

$$\tilde{\alpha}(x) = \frac{\sum_{r',r \in \mathcal{I}_{disc}} \{1 - (x - (r + 0.5, r'))D_{disc}(x)^{-1}c_{1,disc}(x)\} K_b(x - (r + 0.5, r'))O_{r',r}}{\sum_{r',r \in \mathcal{I}_{disc}} \{1 - (x - (r + 0.5, r'))D_{disc}(x)^{-1}c_{1,disc}(x)\} K_b(x - (r + 0.5, r'))E_{r',r}},$$

where D_{disc} and $c_{1,disc}$ are the discrete versions of D and c_1 , respectively:

$$c_{11,disc}(x) = n^{-1} \sum_{r',r \in \mathcal{I}_{disc}} K_b(x - (r + 0.5, r'))(t - r + 0.5)E_{r',r},$$

$$c_{12,disc}(x) = n^{-1} \sum_{r',r \in \mathcal{I}_{disc}} K_b(x - (r + 0.5, r'))(t - r')E_{r',r},$$

$$d_{00,disc}(x) = \sum_{r',r \in \mathcal{I}_{disc}} K_b(x - (r + 0.5, r'))(t - r + 0.5)^2 E_{r',r},$$

$$d_{01,disc}(x) = \sum_{r',r \in \mathcal{I}_{disc}} K_b(x - (r + 0.5, r'))(t - r + 0.5)(z - r')E_{r',r},$$

$$d_{11,disc}(x) = \sum_{r',r \in \mathcal{I}_{disc}} K_b(x - (r + 0.5, r'))(z - r')^2 E_{r',r}.$$

The cross validation criteria is then

$$Q(b) = n^{-1} \sum_{r',r \in \mathcal{I}_{disc}} \{\hat{\alpha}(r, r') - \alpha(r, r')\}^2 E_{r',r},$$

and thus

$$\hat{Q}_b = n^{-1} \sum_{r',r \in \mathcal{I}_{disc}} \{\hat{\alpha}(r, r')\}^2 E_{r',r} - 2 \sum_{r',r \in \mathcal{I}_{disc}} \hat{\alpha}^{[r,r']}(r, r')O_{r',r}.$$

Finally,

$$\hat{f}(t, z) = \hat{\alpha}_1(R_0 - t)\hat{\alpha}_2(z) \exp \left\{ - \int_0^{T_0-t} \hat{\alpha}_1(s)\hat{\alpha}_2(z)ds \right\}.$$

References

- Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Beran, R. (1981). *Nonparametric regression with randomly censored survival data*. Tech. rep. Dept. Statist., Univ. California, Berkeley.
- Bowman, A. W. (1984). “An Alternative Method of Cross-Validation for the Smoothing of Density Estimates”. In: *Biometrika* 71, pp. 353–360.
- Dabrowska, D. M. (1987). “Non-parametric Regression with Censored Survival Time Data”. In: *Scand. Actuar. J.* 14, pp. 181–197.
- Deimling, K. (1985). *Nonlinear functional analysis*. Berlin: Springer.
- Gámiz, M. L., L. Janys, M. D. Martínez-Miranda, and J. P. Nielsen (2013). “Bandwidth selection in marker dependent kernel hazard estimation”. In: *Comput. Stat. Data An.* 68, pp. 155–169.
- Hall, P. (1983). “Large sample optimality of least squares cross-validation in density estimation”. In: *Ann. Stat.* 11, pp. 1156–1174.
- Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.
- Hiabu, M., E. Mammen, M. D. Martínez-Miranda, and J. P. Nielsen (2016). “In-sample forecasting with local linear survival densities”. In: *Biometrika* Forthcoming.
- Linton, O. B. (1997). “Efficient estimation of additive nonparametric regression models”. In: *Econometric Theory* 16, pp. 502–52373.
- (2000). “Efficient Estimation of Generalized Additive Nonparametric Regression Models”. In: *Biometrika* 84, pp. 469–473.
- Linton, O. B. and J. P. Nielsen (1995). “A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration”. In: *Biometrika* 82, pp. 93–100.
- Linton, O. B., J. P. Nielsen, and S. Van de Geer (2003). “Estimating Multiplicative and Additive Hazard Functions by Kernel Methods”. In: *Ann. Stat.* 31, pp. 464–492.

- Mammen, E., O. B. Linton, and J. P. Nielsen (1999). “The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions”. In: *Ann. Stat* 27, pp. 1443–1490.
- Mammen, E., M. D. Martínez-Miranda, and J. P. Nielsen (2015). “In-sample forecasting applied to reserving and mesothelioma”. In: *Insurance Math. Econom.* 61, pp. 76–86.
- Mammen, E., B. Støve, and D. Tjøstheim (2009). “Nonparametric additive models for panels of time series”. In: *Econometric Theory* 25, pp. 442–481.
- Mammen, E. and K. Yu (2009). “Nonparametric estimation of noisy integral equations of the second kind”. In: *J. Korean Statist. Soc.* 38, pp. 99–110.
- Marron, J. S. (1994). “Visual understanding of higher-order kernels”. In: *J. Comput. Graph. Statist.* 3, pp. 447–458.
- Marron, J. S. and M. P. Wand (1992). “Exact mean integrated squared error”. In: *Ann. Stat.* 20, pp. 712–736.
- McKeague, I. W. and K. J. Utikal (1990). “Inference for a Nonlinear Counting Process Regression Model”. In: *Ann. Stat.* 18, pp. 1172–1187.
- Nielsen, J. P. (1998). “Marker dependent kernel hazard estimation from local linear estimation”. In: *Scand. Actuar. J.* 1998, pp. 113–124.
- Nielsen, J. P. and O. B. Linton (1995). “Kernel estimation in a non-parametric marker dependent hazard model”. In: *Ann. Stat.* 23, pp. 1735–1748.
- Rudemo, M. (1982). “Empirical Choice of Histograms and Kernel Density Estimators”. In: *Scand. J. Stat.* 9, pp. 65–78.
- Stone, C. J. (1982). “Optimal Global Rates of Convergence for Nonparametric Regression”. In: *Ann. Stat.* 10, pp. 1040–1053.
- (1985). “Additive regression and other nonparametric models”. In: *Ann. Stat.* 13, pp. 689–705.

4

On the relationship between classical chain ladder and granular reserving

This chapter has been accepted for publication in *Scandinavian Actuarial Journal*. As of today, there is no publication date. A version of this chapter is also available on the Social Science Research Network (SSRN): <http://ssrn.com/abstract=2782320>

On the relationship between classical chain ladder and granular reserving

M. Hiabu

Cass Business School, City, University of London, United Kingdom

Abstract

We connect classical chain ladder to the continuous chain ladder model of Martínez-Miranda et al. (2013). This is done by defining explicitly how the classical run-off triangles are generated from *iid* observations in continuous time. One important result is that the development factors have a one to one correspondence to a histogram estimator of a hazard running in reversed development time. A second result is that chain ladder has a systematic bias if the row effect has not the same distribution when conditioned on any of the aggregated periods. This means that the chain ladder assumptions on one level of aggregation, say yearly, are different from the chain ladder assumptions when aggregated in quarters and the optimal level of aggregation is a classical bias variance trade-off depending on the data-set. We introduce smooth development factors arising from non-parametric hazard kernel smoother improving the estimation significantly.

Keywords: Chain Ladder, Granular Reserving, Development Factors, Solvency II, Non-Life Insurance.

4.1 Introduction

Reserving is the process behind setting capital reserves for outstanding liabilities in non-life insurance. Insurance companies are obligated to account for claims that have been reported but not settled yet and also for incurred claims which have not even been reported. The reserve is often the major part of a non-life insurers balance sheet. Accurate estimation is necessary for pricing future policies and also for the assessment of solvency and net worth of the company. This in turn plays a major role in decisions for financial investments and also for sales or acquisitions of insurances. Finally, wrong assessment can lead to bankruptcy of major companies, with consequences for the whole economic system; for example in the UK, the non-life insurance market accounts for 5% of the gross national product. These considerations come in hand with a growing sense that the reserving process has to be done more rigorous including accurate point forecast and discussions about its uncertainty around.

In practice, actuaries usually use the chain ladder method to calculate the reserve. The method is based on historical data aggregated as run off triangles, i.e., paid claims, claims counts, or incurred claims. For the sake of simplicity of the mathematical arguments we only consider claim counts. Chain ladders development factors (see (4.6)) are hereby the central object. One expression of their importance is maybe its many names: CL (chain ladder) factor, link-ratio, age to age factor, or forward factor. But despite its central role and intuitive appeal, as of today, practitioners and also academics are struggling with the understanding of development factors in terms of classical mathematical statistics. This might have let the author in Schmidt (2012) saying:

“[...] loss reserving is an art of which statistics is, although important, just a part.”

This goes in hand with England and Verrall (2002) remarking on the usual reserving practise that

”very often, the chain ladder technique is the first method to be applied, followed by manual smoothing of the resultant development factors, then adjustment of the results in line with expert opinion combined with additional information”.

With these statements in mind, the reserving task remains, by its very nature, a statistical problem. Hence, a better statistical understanding of those practices and the reserving problem is necessary not only to get reasonable point estimates and to quantify the risk and uncertainty in a reproducible way but also for understanding the underlying assumptions under which these results hold.

A main result of this paper is that when the classical run-off triangles are modeled as arising from observations in continuous time, then there is a quite easy understanding of the development factors in terms of mathematical statistics. We will show that there is a one to one correspondence to a histogram estimator of a hazard function (also known as force of mortality in the actuarial branch of longevity) in reversed development time. In Section 4.3, we show that

$$\hat{\lambda}_j = \{1 - \hat{\alpha}^H(T - x_j)\}^{-1},$$

where $\hat{\lambda}$ are the development factors defined in (4.6) and $\hat{\alpha}^H$ is the histogram estimator of the hazard function, see (4.8). This translates the estimation problem of development factors to the well known estimation problem of a hazard function in survival analysis. In this survival analysis framework it is possible to relax classical assumptions of chain ladder, for instance by allowing calendar time effects. A possibility of extension is adding covariates when estimating the hazard. Both possibilities transform the one dimensional hazard to the multivariate case. In this paper we improve chain ladder with a third possibility and make the maybe most easy improvement in estimating the development factors. We replace the histogram estimator of the hazard by more efficient non-parametric kernel smoother of the hazard, see also Hiabu et al. (2016) (Chapter 2). The one to one correspondence then leads to non-parametric kernel smoothed development factors.

Modeling the complete data generating process leads to another discovery in this paper. An underlying assumption of any stochastic model describing the classical chain ladder method is the independence of underwriting date (row) effect and delay (column) effect, since the development factors do not depend on the underwriting date. In Proposition 4.4 below, it is shown that if this holds on the individual level, then chain ladder in its aggregated form is only consistent if the underwriting effect is identically distributed within a period. Already the simple example of a continuously linear increasing trend

in the book-size will make chain ladder in-consistent by adding a systematic bias; more precisely the reserve will be overestimated in that case. Hence, if one does not see the aggregation in classical chain ladder as a smoothing step where the aggregation level converges to zero with growing sample size, then this should indeed be seen as the underlying assumptions of chain ladder.

There has been a lot of literature aimed at building a statistical model around the chain ladder method; see Kremer (1982), Verrall (1991), Mack (1993), Renshaw and Verrall (1998), Wüthrich, Merz, and Bühlmann (2008), and Kuang, Nielsen, and Nielsen (2009) among others. In the last years there is a growing sense in the industry that aggregated data or macro data is not accurate enough and maybe outdated in times of big data. This argumentation is not completely correct, since it is the very aim of statistics to compress information into a single number or function. Therefore, aggregation should be seen as a statistical pre-smoothing step. The problem then, however, is that there is a) little discussion about the optimal level of aggregation, which of course varies with the data at hand, and b) no discussion about the underlying individual data which justifies this type of aggregation. While discussing the underlying model of chain ladder, the papers mentioned before do especially not discuss the data generating process or sampling scheme which make it hard to understand and justify the implicit assumptions. Renshaw and Verrall (1998), for instance, say that the chain ladder method assumes “stationarity of the reporting process” without further defining what this process is and how the actual data arises from this process. Finally, granular methods are necessary if one is interested in a more detailed cashflow.

Models on individual data and continuous time have been developed by Arjas (1989) and Norberg (1993), where the individual claim development is modeled as a marked point process. These more theoretical contributions have been made more applicable through the work of Antonio and Plat (2014). A different semi-parametric approach based on copulas is given in Zhao, Zhou, and Wang (2009) and Zhao and Zhou (2010). A comparison of an individual model and chain ladder estimates derived from its aggregation are discussed in Huang, Wu, and Zhou (2016).

In recent literature there have also been developed models based on individual data assumptions but where the traditional run-off triangle data structure from chain ladder is kept. The idea is to keep the triangular structure and thus do not completely throw away

existing reserving theory and practice. Verrall, Nielsen, and Jessen (2010), Martínez-Miranda, Nielsen, and Verrall (2012), Hiabu et al. (2015), and Schiegl (2015) have assumptions on the individual data but work entirely with aggregated observations. This makes it hard to check the underlying assumptions on the individual data. Drieskens et al. (2012), Rosenlund (2012), Pigeon, Antonio, and Denuit (2013), and Godecharle and Antonio (2015) rely on individual data but work on aggregated time. Martínez-Miranda et al. (2013) formulated a continuous chain ladder model which keeps the traditional run-off triangle structure of classical chain ladder but considers individual data in continuous time.

This paper aims to connect and compare the continuous chain ladder model of Martínez-Miranda et al. (2013) and classical chain ladder. In practice one can imagine the continuous model being based on the classical data by defining a period as a second instead of, say, a year, which results in a triangle of only 0's and 1's. An important result of this paper is that chain ladder's estimation techniques corresponds to survival analysis techniques when the development time is reversed. With the time reversal one does not need exposure data to estimate the quantities of interest which is also the case in the classical chain ladder method. This is different to the individual data approaches based on Arjas (1989), Norberg (1993), and Antonio and Plat (2014), and will be explained in more detail in the next section.

4.2 The continuous model

4.2.1 Model formulation

In this chapter we will formulate the stochastic model of continuous chain ladder (Martínez-Miranda et al., 2013), and afterwards embed it into a counting process framework. For a better understanding in what follows it is helpful to be familiar with the classical chain ladder method, in particular the run-off triangle, see for example Taylor (1986) and Wüthrich and Merz (2008). The idea of continuous chain ladder is that claims are point observations on the usual run-off triangle rather than being aggregated into bins as is assumed in the classical chain ladder method. With the counting process formulation we will then define a hazard function as part of the counting process intensity. In chapter

4.3, we will show that the hazard in reversed development time is the continuous version of the well known development factors of the chain ladder method.

The data in the classical chain ladder method are given as a run-off triangle and are one half of the square including the future claims which are needed to be estimated. We consider the probability space, $(\mathcal{S}, \mathcal{B}(\mathcal{S}), P)$, where $\mathcal{S}$ is the square $\{(x, y) : 0 \leq x, y \leq T\}$. The underwriting date, Y , and the reporting delay, X , of a claim are hence random with probability measure P , which describes how likely it is to see a claim on a certain position on the square. Since $\mathcal{S}$ is bounded by T , we implicitly assume that firstly all claims are reported within a maximum delay of T from their underwriting date and secondly that we have T time units of observed underwriting dates. Generally to avoid extrapolation, which we are not doing here, the maximum delay of a claim must be smaller than the range of observed underwriting dates. The theory described here would, as chain ladder does, work if X and Y would be bounded by T_1, T_2 , with $T_1 \leq T_2$. To simplify the notation, we have assumed $T = T_1 = T_2$. If this is not the case in practise, i.e. $T_1 < T_2$, the remaining columns can just be filled with zeroes to obtain the same results.

We will assume that the density with respect to the Lebesgue measure, $f = dP/d\lambda$, is well defined and multiplicative, i.e., $f(x, y) = f_1(x)f_2(y)$. Hence, we assume that the components X and Y are independent. This assumption can be checked by usual independence tests, see Tsai (1990), Mandel and Betensky (2007), and Addona, Atherton, and Wolfson (2012). A more pragmatic solution is plotting the individual development factors and checking whether they lie on a horizontal line, see Hiabu et al. (2016) (Chapter 2).

We further assume that observations are only sampled on a subset of the full support of the density f . The truncated density is supported on the triangle, $\mathcal{I} = \{(x, y) : 0 \leq x, y \leq T, x + y \leq T\}$ - the well known run-off triangle. In this case, we consider observations of n independent and identically distributed claims, $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, with $X_i \leq T - Y_i$, or equivalently $Y_i \leq T - X_i$, where T is the calendar time the data are collected. Note that (X_1, Y_1) is not distributed according to P and does not have density f , since we already know that it is on the upper triangle. Hence, its density is given by $f(x, y)/\int_{\mathcal{I}} f(x, y) dx dy$. The observation schemes, $X_i \leq T - Y_i$, and $Y_i \leq T - X_i$ can be understood as random right-truncation when targeting only X or Y , respectively.

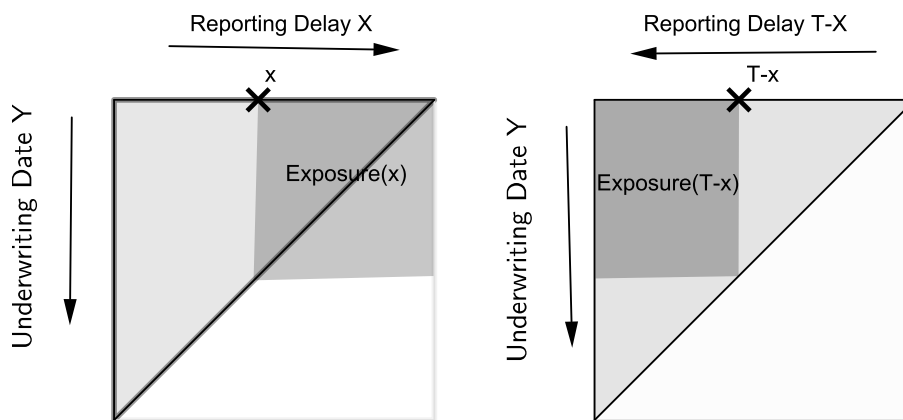


FIGURE 4.1: The exposure in forward moving time (left) and reversed time (right). Only in reversed time we observe the exposure.

The well established method to make inference on such observation schemes is to reformulate the problem into a counting process framework, see for example Andersen et al. (1993). In the following we will focus on inference on the reporting delay X . Due to symmetry all the results can be easily adapted for the random variable Y . The development factors in chain ladder only correspond to estimates depending on X . To this end, we define a counting processes indicating the occurrences of X_i , $i = 1, \dots, n$. A crucial point here is that right-truncation is not tractable as such, since the exposure is not observable: In the counting process model, one needs to know at every point in time how many individuals are at risk. Assume that we move T years forward in time and hence know about every claim on the square. Exposure in x is then the amount of claims having a greater reporting delay than x but could have been observed already at point of data collection if the delay would have been exactly x . This amount is not known at time of data collection, see Figure 4.1.

By reversing the time of the counting process, however, the right-truncation becomes a left-truncation, see for example Ware and DeMets (1976) and Lagakos, Barraj, and De Gruttola (1988), and exposure is observable, since all past claims are known. Note that in the models of Arjas (1989), Norberg (1993), and Antonio and Plat (2014) time is not reversed, and hence extra exposure data is needed to calibrate their model.

We define the time reversed counting processes as

$$N_i(t) = I(T - X_i \leq t), \quad (i = 1, \dots, n),$$

where I denotes the indicator function, with respect to the filtration

$$\mathcal{F}_t^i = \sigma \left(\left\{ (T - X_i) \leq s : s \leq t \right\} \cup \left\{ (Y_i) \leq s : s \leq t \right\} \cup \mathcal{N} \right),$$

satisfying the usual conditions, and where $\mathcal{N} = \{A : A \subseteq B, \text{ with } B \in \mathcal{B}(\mathcal{S}), P(B) = 0\}$. Adding the null set, $\mathcal{N}$, to the filtration guarantees its completeness. This is a technically useful construction, but is not strictly necessary, since the subsequent results also hold if one does not assume completeness of the filtration, see Jacod (1979) and Jacod and Shiryaev (1987).

The random intensity of N_i, ν_i , is well-defined since X is absolutely continuous. It can be described, almost surely, through $\nu_i(t) = \lim_{h \downarrow 0} h^{-1} E [N_i \{(t+h)-\} - N_i(t-)| \mathcal{F}_{t-}^i]$. Straightforward computations lead to Aalen's multiplicative intensity model (Aalen, 1978):

$$\nu_i(t) = \alpha(t)Z_i(t),$$

where the hazard ratio α , and the predictable filtering process (individual exposure), Z_i , are

$$\alpha(t) = \lim_{h \downarrow 0} h^{-1} pr \{ (T - X) \in [t, t+h) \mid (T - X) \geq t \} = \frac{f_1(T-t)}{F_1(T-t)} = \frac{f_1^R(t)}{S_1^R(t)},$$

$$Z_i(t) = I\{Y_i < t \leq (T - X_i)\},$$

and $F_1 = \int_0^\cdot f_1(x)dx$ is the cumulative distribution function. The crucial point in Aalen's multiplicative intensity model is that the hazard function, α , does not depend on Y . In chapter 4.3 we will show that the hazard in reversed development time, α , is the continuous version of the well known development factors λ of the chain ladder method. Before finishing this chapter, we introduce the notation $N(t) = \sum N_i(t)$ and the exposure $Z(t) = \sum Z_i(t)$.

4.2.2 Estimation in the continuous framework

In this section we briefly introduce three nonparametric estimators of the hazard function α in the continuous time framework: The histogram estimator, the local constant

estimator and the local linear estimator. The local linear and the local constant estimator are well studied in the statistical literature of kernel smoothing, and we will only state the results and properties of the estimator for people not familiar with smoothing theory. The histogram estimator is known from applied fields as in age-period-cohort models of demographic problems.

An alternative to estimate the hazard function α would be to assume a parametric form on the intensity ν_i , see Borgan (1984) and Andersen et al. (1993). We chose not to do so in this paper, since a nonparametric estimation technique is more in the spirit of the chain ladder technique.

For the asymptotic properties we consider the following assumptions.

Assumption (S)

- S1. *The bandwidth $h = h(n)$ satisfies $h \rightarrow 0$ and $nh \rightarrow \infty$ for $n \rightarrow \infty$.*
- S2. *The hazard function α is strictly positive and it holds that $\alpha \in C_2([0, T])$.*
- S3. *The kernel K is symmetric, has bounded support and has a second moment.*

Assumptions (S1) - (S3) are standard regularity assumptions in smoothing theory (Silverman, 1986; Simonoff, 1998). Note that under assumption (S2), the asymptotic relative exposure $\gamma(t) = pr(Z_1(t) = 1)$ is continuous and from empirical process theory it is known that

$$\sup_{s \in [0, T]} |Z(s)/n - \gamma(s)| = o_p(1). \quad (4.1)$$

4.2.2.1 The histogram estimator of the hazard

The maybe simplest way to derive an estimator of the hazard function, α , is the histogram estimator. Let's assume that a parameter, $h > 0$, as bin width is given. A histogram estimator of α on equally sized bins, with bin size h is derived by dividing the number of observations - relative to the bin width - in one bin by the number of

exposure at that bin. For t in the bin $[c1, c2)$, that is

$$\hat{\alpha}_h^H(t) = \frac{h^{-1} \sum_{i=1}^n \int_{c1}^{c2} dN_i(s)}{\sum_{i=1}^n \int_{c1}^{c2} Z_i(s) ds} = \frac{O^H(t)}{E^H(t)}. \quad (4.2)$$

In Hoem (1969) optimality of the histogram estimator is proven if the true hazard, α , is constant on the bins. The following proposition shows the asymptotic properties when local constancy is not assumed. The proof can be found in the Appendix 4.C.

Proposition 4.1. *Assume that assumptions (S1)-(S2) are satisfied. The histogram estimator has asymptotically a bias of order h and variance of order $(nh)^{-1}$. More precisely, the following pointwise asymptotics holds for $t \in (0, T)$:*

$$(nh)^{1/2} \left\{ \hat{\alpha}_h^H(t) - \alpha(t) - B(t) \right\} \xrightarrow{\mathcal{D}} N \left\{ 0, \sigma^2(t) \right\},$$

where

$$B(t) = \alpha'(t)h^{-1} \int_{c1}^{c2} (t-s) ds + o(h), \quad \sigma^2(t) = \alpha(t)\gamma(c2)^{-1}.$$

4.2.2.2 Local polynomial estimator of the hazard

The idea of local polynomial fitting is quite old and might originate from early time series analysis, see Macaulay (1931). It has been adapted to the regression case in Stone (1977) and Cleveland (1979). A general overview of local polynomial fitting can be found in Fan and Gijbels (1996). The local constant estimator has a reduced convergence rate at boundaries. This is not the case for polynomials of order $p \geq 1$. In general, a higher order reduces bias but increases variance. But variance only increases when the order changes from odd to even. In this paper we will only consider the cases $p = 0, 1$, that is the local constant and the local linear estimator of the hazard function.

We define the local constant estimator, $\hat{\alpha}_{h,K}^{LC}(t)$ of $\alpha(t)$, as the minimizer, $\hat{\Theta}_0$, in the equation

$$\hat{\Theta}_0 = \arg \min_{\Theta_0 \in \mathbb{R}} \sum_{i=1}^n \left[\int K_h(t-s) \Theta_0^2 Z_i(s) ds - 2 \int K_h(t-s) \Theta_0 dN_i(s) \right], \quad (4.3)$$

where for a given kernel, K , and a bandwidth, h , $K_h(t) = h^{-1}K(t/h)$. The definition of the local constant estimator as the minimizer of (4.3) can be motivated by the fact that

its minimizer equals the least square criteria,

$$\arg \min_{\Theta_0 \in \mathbb{R}} \left(\lim_{\varepsilon \rightarrow 0} \sum_{i=1}^n \int \left[\left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} dN_i(u) - \Theta_0 \right\}^2 - \xi(\varepsilon) \right] \times K_h(t-s) Z^i(s) ds \right),$$

where $\xi(\varepsilon) = \{\varepsilon^{-1} \int_{s-\varepsilon}^s dN^i(u)\}^{-2}$ is a just a vertical shift which is added to make the expression well-defined. The solution of (4.3), see also Nielsen and Tanggaard (2001), is given by

$$\hat{\alpha}_{h,K}^{LC}(t) = \frac{\sum_{i=1}^n \int K_h(t-s) dN_i(s)}{\sum_{i=1}^n \int K_h(t-s) Z_i(s) ds} = \frac{O^{LC}(t)}{E^{LC}(t)}.$$

For every Kernel K we define

$$\mu_i(K) = \int s^i K(s) ds, \quad R(K) = \int K^2(s) ds.$$

The following proposition states that the local constant estimator is efficient in optimal rate sense.

Proposition 4.2 (Hjort, West, and Leurgans (1992)). *Assume that assumption (S) is satisfied. Then, the following pointwise asymptotics holds for $t \in (0, T)$:*

$$(nh)^{1/2} \left\{ \hat{\alpha}_{h,K}^{LC}(t) - \alpha(t) - B(t) \right\} \xrightarrow{\mathcal{D}} N \left\{ 0, \sigma^2(t) \right\},$$

where

$$B(t) = \mu_2(K) h^2 \left\{ \frac{1}{2} \alpha''(t) + \alpha'(t) \gamma'(t) \gamma(t)^{-1} \right\} + o(h^2), \quad \sigma^2(t) = R(K) \alpha(t) \gamma(t)^{-1}.$$

Note that this result only holds in the interior of the support $[0, T]$. Following the proof one can easily see that the bias is of order b in the boundary region, i.e., the intervals $[0, b)$ and $(T - b, T]$. There have been several estimators proposed to derive convergence in the full support. Due to its simplicity but also convincing properties the local linear estimator became the maybe most popular kernel smoother. Similarly to the local constant estimator, we define the local linear estimator (Nielsen, 1998), $\hat{\alpha}_{h,K}^{LL}(t)$ of

$\alpha(t)$, as the minimizer, $\hat{\Theta}_0$, in the equation

$$\begin{pmatrix} \hat{\Theta}_0 \\ \hat{\Theta}_1 \end{pmatrix} = \arg \min_{\Theta_0, \Theta_1 \in \mathbb{R}} \sum_{i=1}^n \left[\int K_h(t-s) \{\Theta_0 - \Theta_1(t-s)\}^2 Z_i(s) ds \right. \\ \left. - 2 \int K_h(t-s) \{\Theta_0 - \Theta_1(t-s)\} Z_i(s) dN_i(s) \right]. \quad (4.4)$$

With solution

$$\begin{aligned} \hat{\alpha}_{h,K}^{LL}(t) &= n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) dN_i(s) \\ &= \frac{\sum_{i=1}^n \int K_h(t-s) \{a_2(t) - a_1(t)(t-s)\} dN_i(s)}{\sum_{i=1}^n \int K_h(t-s) \{a_2(t) - a_1(t)(t-s)\} Z_i(s) ds} \\ &= \frac{O^{LL,\delta}(t)}{E^{LL,\delta}(t)}, \end{aligned} \quad (4.5)$$

where

$$\bar{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2} K_h(t-s),$$

and

$$a_j(t) = n^{-1} \int K_h(t-s)(t-s)^j Z(s) ds \quad (j = 0, 1, 2).$$

The notation of $\bar{K}_{t,h}$ is chosen because it is indeed, given (S3), a second order kernel with respect to the measure $Z(s)ds$:

$$n^{-1} \int \bar{K}_{t,h}(t-s) Z(s) ds = 1, \quad n^{-1} \int \bar{K}_{t,h}(t-s)(t-s) Z(s) ds = 0,$$

$$n^{-1} \int \bar{K}_{t,h}(t-s)(t-s)^2 Z(s) ds > 0.$$

Furthermore, Nielsen and Tanggaard (2001) showed that $\bar{K}_{t,h}(t-s)$ is asymptotically equivalent to $k_{t,h}(t-s)Z^{-1}(s)$, with

$$k_{t,h}(t-s) = \frac{c_2(t) - c_1(t)(t-s)}{c_0(t)c_2(t) - \{c_1(t)\}^2} K_h(t-s), \quad c_j(t) = n^{-1} \int K_h(t-s)(t-s)^j ds,$$

which in turn pointwise equals $K(t-s)$, for n large enough. This considerations make it not surprising that the local linear estimator has similar point-wise asymptotics as the local constant estimator.

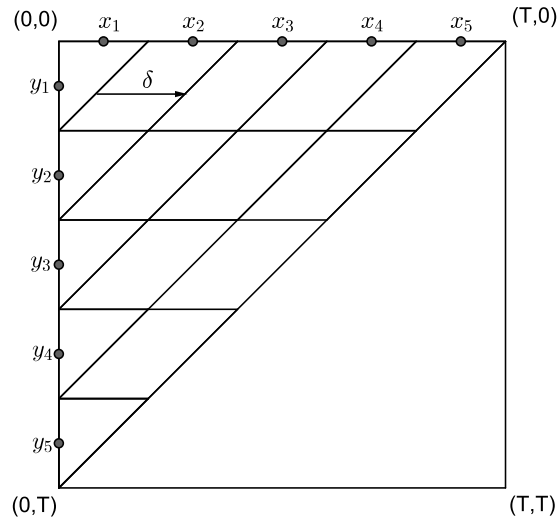


FIGURE 4.2: The usual aggregation of a triangle in the chain ladder method. The bin-width δ represents the length of a period.

Proposition 4.3 (Nielsen (1998)). *Assume that assumption (S) is satisfied. Then, the following asymptotics holds for $t \in (0, T)$:*

$$(nh)^{1/2} \left\{ \hat{\alpha}_{h,K}^{LL}(t) - \alpha(t) - B(t) \right\} \xrightarrow{\mathcal{D}} N \left\{ 0, \sigma^2(t) \right\},$$

where

$$B(t) = \frac{1}{2} \mu_2(K) \alpha''(t) h^2 + o(h^2), \quad \sigma^2(t) = R(K) \alpha(t) \gamma(t)^{-1}.$$

4.3 Discretization of the continuous model

4.3.1 The model

In the previous section we have defined several estimators of the hazard function, given observations in continuous time. In the non-life insurance context, data are usually aggregated in so called yearly or quarterly run-off triangles. This is done by aggregating the continuous triangle, $\mathcal{I}$, into a grid of parallelograms, see Figure 4.2.

The observation scheme is very similar to those in a Lexis diagram known from age-period-cohort models in demography, and aggregation of the same parallelograms is there

known as the first principle set (Hoem, 1969; Keiding, 1990). In the language of age-period-cohort models, the form of a parallelogram arises because while data are collected with respect to cohort (underwriting date) and year (claim delay), the aggregation is done with respect to cohort and period (calendar time). While aggregation into squares would make many things easier, the triangular observation scheme would then imply that the number of observation changes with different aggregation level, and in particular forecasting would not be possible for the last underwriting period.

Let δ be the grid width with integer valued inverse. The individual data of independent, identically, distributed data (X_i, Y_i) are aggregated to observations, (X_i^δ, Y_i^δ) , with support on

$$\mathcal{I}^\delta = \{(x_j, y_k) = ((j + 0.5)\delta, (k + 0.5)\delta) : j, k = 0, 1 \dots, T\delta^{-1} - 1, j + k \leq T\}.$$

The discrete observations are then described via

$$(X_i^\delta, Y_i^\delta) = (x_j, y_k) \Leftrightarrow Y_i \in [k\delta, (k + 1)\delta) \text{ and } X_i + Y_i \in [(j + k)\delta, (j + k + 1)\delta)$$

Note that this implies that

$$X_i \in [(j - 1)\delta \vee 0, (j + 1)\delta).$$

The parallelogram aggregation adds a non-trivial dependency structure between the components X^δ and Y^δ , even though X and Y might be independent. The chain ladder method implicitly assumes independence between underwriting date and development delay, since the development factors do not change for different underwriting dates, see also various paper discussing the underlying model of chain ladder, e.g., Mack (1993) and Renshaw and Verrall (1998).

It is then important to note that the necessary independence of the components of (X^δ, Y^δ) does generally not follow from the independence of X and Y . Consider the following assumptions.

Assumption (D)

D(i) The density function f_2 of Y is multiplicatively separable in the sense that there exist functions g_1 and g_2 so that for every $y \in [y_j - 0.5\delta, y_j + 0.5\delta)$, it holds that $f_2(y) = g_1(y_j)g_2(y - y_j)$.

D(ii) The random variables X^δ and Y^δ are independent.

Assumption $D(i)$ basically says that the random variable Y needs to have the same distribution when conditioned on any bin on the grid.

Sufficient conditions are for instance local constancy, $f_2(y) = g(y_k)$, as most often assumed in age-period-cohort literature, or an exponential growth $f_2(y) = c \exp(y)$, where c is a norming constant.

Note that those assumptions can not be checked if one has only access to the aggregated data, also whether assumption D(i) or D(ii) are satisfied or by how much they violated does depend on the level of aggregation.

Proposition 4.4. *If X and Y have a multiplicative separable density, i.e., $f(x, y) = f_1(x)f_2(y)$, and f_1 is not further specified, then assumption $D(i)$ and $D(ii)$ are equivalent.*

The proof can be found in the Appendix 4.C.

The implication of this proposition is that if assumption D(i) is not satisfied, then the level of aggregation in chain ladder is a bias-variance trade off. The optimal level of aggregation should then be derived via a cross-validation method which needs to be developed.

The classical run-off triangle data is given in the form $(N_{r,s})$, $r, s = 1, \dots, T$, $r+s \leq T+1$, and are the total numbers of claims of insurance incurred in period (most often a year or quarter) r which have been reported in period $r+s$, i.e., with s periods delay from r . The triangle is derived from the random variables (X^δ, Y^δ) , via

$$\mathcal{N}_{r,s} = \#\{i : Y_i^\delta = r, X_i^\delta = s\} = \sum_{i=1}^n I\{Y_i^\delta = r, X_i^\delta = s\}.$$

Since we are assuming that there is no so called tail, i.e., all claims are reported within T periods, forecasts are obtained by estimating and summing up the values $(N_{r,s})$,

$r, s = 1, \dots, T$, $r + s \geq T + 2$. For this, chain ladder estimates development factors,

$$\hat{\lambda}_s = \frac{\sum_{k=1}^{T-s+1} \sum_{l=1}^s N_{k,l}}{\sum_{k=1}^{T-s+1} \sum_{l=1}^{s-1} N_{k,l}}, \quad (s = 2, \dots, T), \quad (4.6)$$

and forecasts are derived as $\hat{N}_{r,s} = N_{r,T-r+1} \prod_{l=T-r+1}^s \lambda_l$.

For deriving estimators in the discrete framework, we want to use the already developed theory in the continuous time case and introduce f_1^δ as the density of X^δ with respect to the counting measure $\mu(A) = \delta \# \{j \mid (j+0.5)\delta \in A, j = 0, 1, \dots, T\delta^{-1} - 1\}$, $A \in \mathcal{B}([0, 1])$.

$$f_1^\delta(t) = \begin{cases} 0 & \text{if } t \neq (j + 0.5)\delta \\ \delta^{-1} \int_{j\delta}^{(j+1)\delta} f_1(t) dx & \text{if } t = (j + 0.5)\delta. \end{cases}$$

We define the time reversed counting processes as

$$N_i^\delta(t) = I(T - X_i^\delta \leq t), \quad (i = 1, \dots, n),$$

with respect to the filtration

$$\mathcal{F}_t^{i,\delta} = \sigma\left(\left\{(T - X_i^\delta) \leq s : s \leq t\right\} \cup \mathcal{N}\right).$$

Similar to the continuous case one derives that the intensity of the counting process is

$$\nu_i^\delta(t) = \alpha^\delta(t) Z_i^\delta(t),$$

where the hazard ratio α , and the predictable filtering process, Z_i^δ , are

$$\alpha^\delta(t) = \frac{f_1^\delta(T - t)}{F_1^\delta(T - t)} = \frac{f_1^{\delta,R}(t)}{S_1^{\delta,R}(t)}, \quad Z_i^\delta(t) = I\{Y_i^\delta \leq t \leq (T - X_i^\delta)\}.$$

This means that also the discrete observation can be translated into Aalen's multiplicative intensity model. The main difference to the continuous case is that the Lebesgue measure is replaced by the counting measure, μ , which lives on a grid according to the aggregation level of the data. For the development of the theoretical properties of the discrete estimators in the next section we introduce the following functions for

$$t \in [0.5\delta, T - 0.5\delta],$$

$$\bar{\alpha}^\delta(t) = \frac{\delta^{-1} \int_{t-0.5\delta}^{t+0.5\delta} f(s) ds}{\int_{t-0.5\delta}^1 f(s) ds}, \quad \bar{Z}_i^\delta(t) = I\{Y_i - 0.5\delta \leq t \leq (T - X_i) + 0.5\delta\}.$$

Note that $\bar{\alpha}^\delta(x_j) = \alpha^\delta(x_j)$ and $\bar{Z}_i^\delta(x_j) = Z_i^\delta(x_j)$. For δ converging to zero, we have that

$$\sup_{s \in [0, T]} |\bar{Z}^\delta(s)/n - \gamma(s)| = o_p(1). \quad (4.7)$$

4.3.2 The histogram estimator and chain ladders development factors

Let us assume that one chooses a bandwidth $h = c\delta$, $c = 1, 2, \dots$, as bin width. Then, for t in the bin $[c_1, c_2)$, with width h , the histogram estimator of the previous section translates to

$$\hat{\alpha}_h^{H,\delta}(t) = \frac{h^{-1} \sum_{i=1}^n \int_{c_1}^{c_2} dN_i^\delta(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i^\delta(s) d\mu(s)} = \frac{O^{H,\delta}(t)}{E^{H,\delta}(t)}. \quad (4.8)$$

We will suppress the subscript h in $\hat{\alpha}_h^{H,\delta}$, if $h = \delta$. Note that while the nominator equals the one from the continuous estimator in (4.2), the denominators are different. The reason is that when considering discrete observations, the exposure in the considered bins is not observed and hence needs to be estimated, see also Hoem (1969).

Proposition 4.5. *Assume that assumptions (S) and (D) hold, then for $h = \delta$, $\hat{\alpha}_h^{H,\delta}(x_j)$ is an unbiased estimator of $\alpha^\delta(x_j)$. The estimation error is asymptotically normal with variance $(nh)^{-1} \gamma^{-1}(t_j) \alpha(t_j)$.*

For $h = c\delta$, $c = 2, 3, \dots$, it holds that

$$(nh)^{1/2} \left\{ \hat{\alpha}_h^H(x_j) - \alpha(x_j) - B(x_j) \right\} \xrightarrow{\mathcal{D}} N \left\{ 0, \sigma^2(x_j) \right\},$$

where

$$B(t) = \frac{1}{24} f^{R''}(x_j) \{S^R(t_j)\}^{-1} \delta^2 + (\bar{\alpha}^\delta)'(x_j) \left\{ (x_j - c_1) - \frac{1}{2}h \right\} + o(h + \delta^2),$$

$$\sigma^2(t) = \bar{\alpha}^\delta(x_j) \gamma(x_j)^{-1}.$$

The proof can be found in the Appendix 4.C.

We now discuss the relationship between chain ladder's development factors and the discrete histogram estimator of the hazard, $\hat{\alpha}_h^{H,\delta}$, when one chooses that the bin-width h equals the discretization δ . Note that the development factor, (4.6), can be rewritten as

$$\hat{\lambda}_j = \frac{E^{H,\delta}(x_j)}{E^{H,\delta}(x_j) - \delta O^{H,\delta}(x_j)}. \quad (4.9)$$

Theorem 4.6. *Assume that λ_j is the j -th development factor derived from the chain ladder algorithm. It holds that*

$$\hat{\lambda}_j = \frac{1}{1 - \delta \hat{\alpha}^{H,\delta}(T - x_j)}. \quad (4.10)$$

Furthermore, it holds that

$$\hat{\lambda}(x_j) = 1 + \delta \hat{\alpha}^{H,\delta}(T - x_j) + O_p(\delta^2). \quad (4.11)$$

Proof. This follows directly from (4.8) and (4.9). □

Equation (4.10) tells us that there is an exact and deterministic relationship between the histogram estimator and the development factor. Equation (4.11) even gives asymptotic equality when the development factors are subtracted by 1.

We conclude the following. In continuous time, chain ladders development factors and a histogram estimator of the hazard in reversed time are the same entity. Or in other words the development factors aim to estimate a hazard in reversed time via a histogram approach. To make this clear, we introduce the new notation

$$\hat{\lambda}^{H,\delta}(x_j) = \hat{\lambda}_j = \frac{1}{1 - \delta \hat{\alpha}^{H,\delta}(T - x_j)}. \quad (4.12)$$

When working in the continuous setting, or say daily level, those classical development factors will be too noisy. Thus, one will need to increase, h or equivalently δ , to increase performance. A better alternative might be to replace the classical development-factors by kernel smoothed versions.

In the next section we introduce discrete versions of the local constant and local linear kernel estimator.

4.3.3 Local polynomial estimator

It is straight forward to see that the solutions of the discrete versions of (4.3) and (5.3) are given by

$$\hat{\alpha}_{h,K}^{LC,\delta}(t) = \frac{h^{-1} \sum_{i=1}^n \int K_h(t-s) dN_i^\delta(s)}{\sum_{i=1}^n \int K_h(t-s) Z_i^\delta(s) d\mu(s)} = \frac{O^{LC,\delta}(t)}{E^{LC,\delta}(t)}.$$

and

$$\hat{\alpha}_{h,K}^{LL,\delta}(t) = \frac{\sum_{i=1}^n \int K_h(t-s) \left\{ a_2^\delta(t) - a_1^\delta(t)(t-s) \right\} dN_i^\delta(s)}{\sum_{i=1}^n \int K_h(t-s) \left\{ a_2^\delta(t) - a_1^\delta(t)(t-s) \right\} Z_i^\delta(s) d\mu(s)} = \frac{O^{LL,\delta}(t)}{E^{LL,\delta}(t)},$$

where

$$a_j^\delta(t) = n^{-1} \int K_h(t-s) (t-s)^j Z^\delta(s) d\mu(s) \quad (j = 0, 1, 2).$$

Proposition 4.7. *Assume that assumption (S) and (D) are satisfied. Then, the following asymptotics holds for $t \in (0, T)$:*

$$\begin{aligned} (nh)^{1/2} \left\{ \hat{\alpha}_{h,K}^{LC}(t) - \alpha(t) - B_{LC,\delta}(t) \right\} &\xrightarrow{\mathcal{D}} N \left\{ 0, \sigma_{LC,\delta}^2(t) \right\}, \\ (nh)^{1/2} \left\{ \hat{\alpha}_{h,K}^{LL}(t) - \alpha(t) - B_{LL,\delta}(t) \right\} &\xrightarrow{\mathcal{D}} N \left\{ 0, \sigma_{LL,\delta}^2(t) \right\}, \end{aligned}$$

where

$$\begin{aligned} B_{LC,\delta}(x) &= \frac{1}{24} f^{R''}(x_j) \{S^R(x_j)\}^{-1} \delta^2 + \mu_2(K) h^2 \left\{ (\bar{\alpha}^\delta)'(x_j) \gamma'(x_j) \gamma^{-1}(x_j) + \frac{1}{2} (\bar{\alpha}^\delta)''(x_j) \right\} + o(\delta^2 + h^2), \\ B_{LL,\delta}(x) &= \frac{1}{24} f^{R''}(x_j) \{S^R(x_j)\}^{-1} \delta^2 + \frac{1}{2} \mu_2(K) h^2 (\bar{\alpha}^\delta)''(x_j) + o(\delta^2 + h^2), \\ \sigma_{LC,\delta}^2(x) &= \sigma_{LL,\delta}^2(x) = R(K) \alpha^\delta(t) \gamma(x)^{-1}. \end{aligned}$$

The proof can be found in the Appendix 4.C.

We now define the local constant and local linear development factors which can be used in the chain ladder approach.

$$\hat{\lambda}^{LC,\delta}(x_j) = \frac{1}{1 - \delta \hat{\alpha}_h^{LC,\delta}(T - x_j)}, \quad \hat{\lambda}^{LL,\delta}(x_j) = \frac{1}{1 - \delta \hat{\alpha}_h^{LL,\delta}(T - x_j)}. \quad (4.13)$$

4.4 Simulation study

To illustrate the finite sample performance, we simulated three models assuming independent underwriting and delay components. For simplicity we set $T = 1$. For the development component, in model 1 and 2, we chose that $X \sim \text{Beta}(2, 5)$, and in the third model we chose a more steep development pattern with $X \sim \text{Exponential}(5)$. For the underwriting variable, Y , in the first model, we assume a uniform distribution and hence the chain ladder assumptions are satisfied for every aggregation. In the second and third model the density of Y is linearly increasing, i.e., $f_2(y) = 2y$. This means that the aggregated approaches will estimate a biased reserve. We also tried many other distributions but they did not change the conclusions we derived from those three models presented here.

We have run 500 repetitions with sample-sizes of $n = 200, 1000, 5000, 10000$ to estimate the relative error, $\text{error} = (E[R] - \hat{R})/E[R]$, where $\hat{R}$ is the reserve estimate derived by the chain ladder algorithm using the development factors in (4.12) and (4.13).

This is done by calculating the chain ladder development factors for aggregation levels $\delta \in \{0.01, 0.02, 0.04, 0.1, 0.2\}$. For $\delta = 0.01$ we also calculated the local linear and the local constant versions. The discretization $\delta = 0.01$ should approximate the continuous model well enough for the smaller sample sizes ($n=200, 1000$). For the greater sample sizes ($n=5000, 10000$), the performance of the local polynomial estimators could have been improved with a smaller δ , according to the asymptotic theory in the previous section. The way the code is implemented, computation time depends on the aggregation level, δ , and is unaffected by the sample size, n . More details are given in the Appendix.

This paper does not discuss the problem of how to choose a bandwidth b . This issue needs to be addressed separately where a cross-validation procedure is developed and assessed. Depending on the estimation purpose, i.e. if one interested in the full sum of the lower triangle or in the diagonal sums of the lower triangle, one will have different loss functions with different optimal bandwidths. In this simulation study, we have used the bandwidth optimal for the given loss function in each simulation step. This choice is infeasible in practice. To give an idea about the robustness of the estimators with respect to the bandwidth, for the local constant estimator, we also included a bandwidth which is randomly picked in every simulation step from a quite wide range depending on

model and sample size. An eye picked or cross-validated bandwidth is then expected to have a performance in between the optimal and random choice. For the local polynomial estimators we have used the Epanechnikov kernel, as kernel K .

In Table 4.1 and Figure 4.3, we see that the histogram estimator becomes better the more one aggregates. This is consistent with the theory, since there is no bias in the estimation and accuracy can hence be reduced by aggregation via reduction of variance. But even when aggregated to a triangle with only 5 periods ($\delta = 0.2$) the kernel estimators are competitive, and the local constant estimator is even favorable. A change in sample size does not seem to alter the conclusion but improves the estimators uniformly.

The results of the second simulation study are presented in Table 4.2. Here, one can see that the choice of the aggregation for the chain ladder approach is a classical bias variance trade off. The results are also visualized in via boxplots in Figure 4.4. Also in this model the conclusion is to prefer the kernel estimators.

Similar results are given in the third model, Table 4.3 and Figure 4.5, which indicates that the results are independent of the distribution choice and also hold in a harder estimation problem with a sharp decay of mass.

An interesting result is that independent of the models, extreme estimation errors are always overestimating the reserve. This is independent of the true distributions or the way the development factors are estimated but seems to be a feature of the chain ladder technique.

In both models the local linear estimator performed surprisingly bad compared to the local constant estimator, but might be better in other scenarios.

4.5 Concluding remarks

In this paper we connected classical chain ladder to the continuous chain ladder model of Martínez-Miranda et al. (2013). We derive a one to one connection between the development factors and a histogram hazard estimator and then improve this histogram estimator by more efficient kernel smoothers. However, the hazard interpretation also

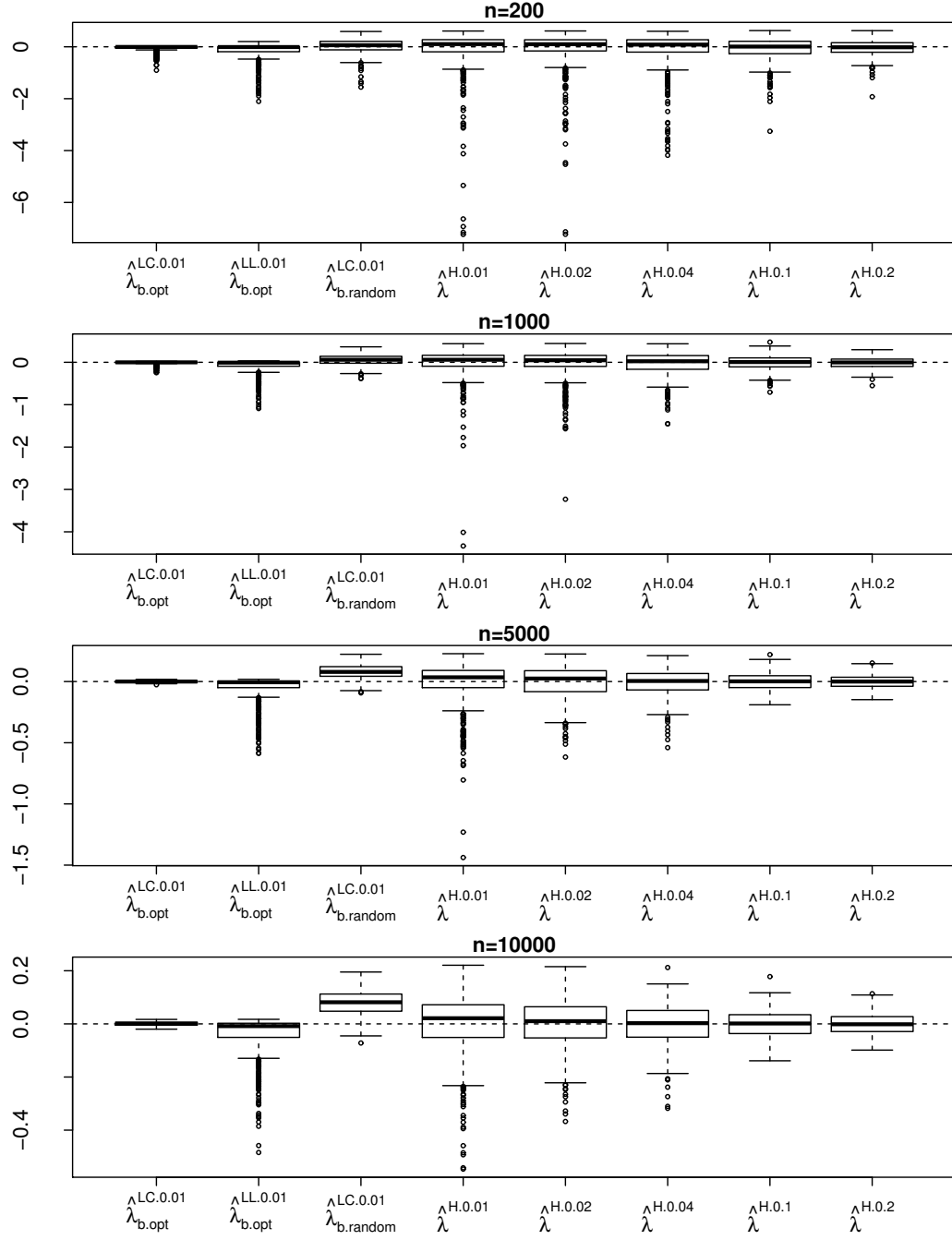


FIGURE 4.3: Boxplot results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a Beta distribution with parameters $(2, 5)$, and the underwriting date density, Y , is uniformly distributed. Sample size is $n = 200, 1000, 5000, 10000$. For $n = 200, 1000$: $b.random \in [0.05, 0.3]$, for $n = 1000, 5000$: $b.random \in [0.05, 0.25]$

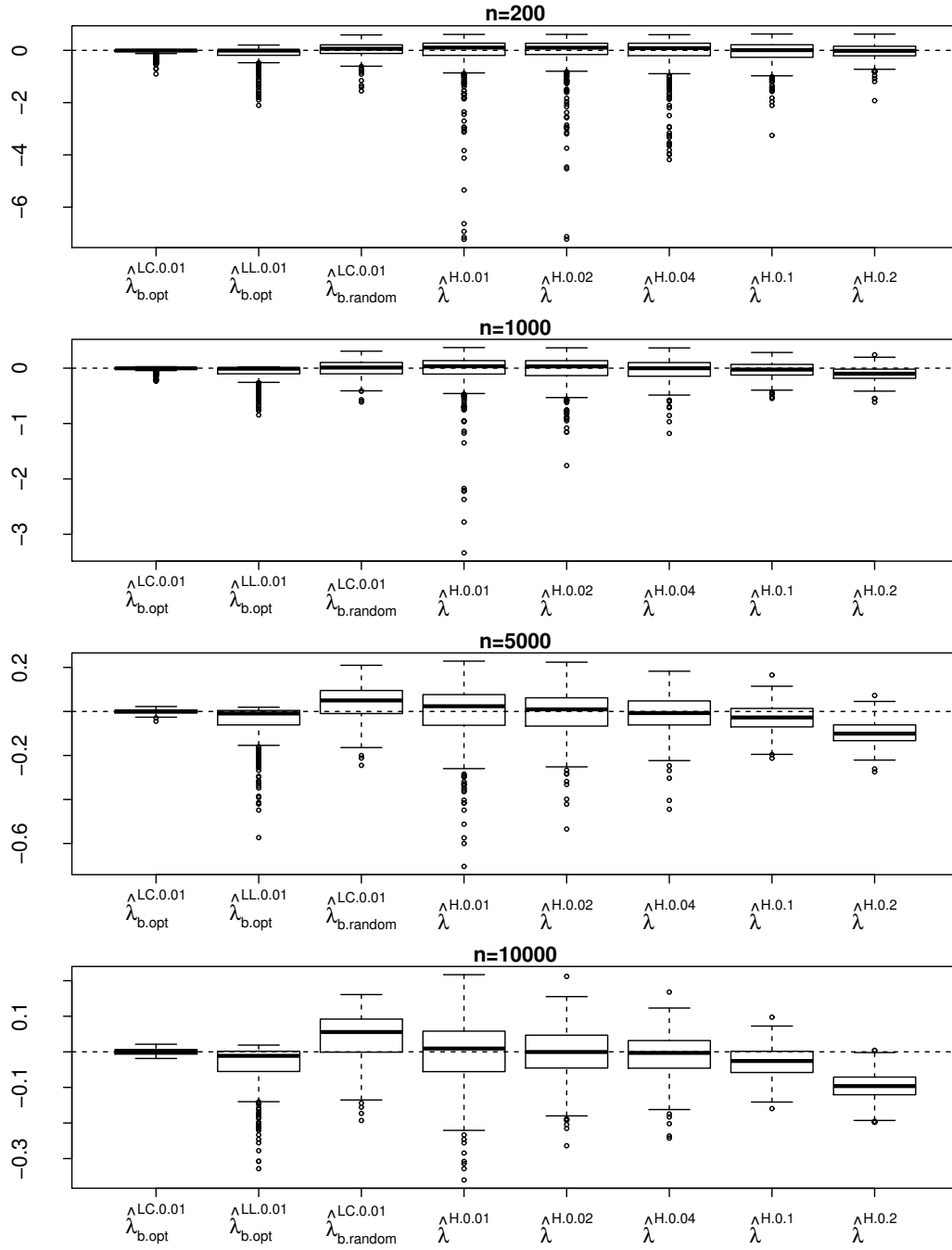


FIGURE 4.4: Boxplot results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a Beta distribution with parameters $(2, 5)$, and the underwriting date density, Y , is linear increasing, $f_2(y) = 2y$. For $n = 200, 1000$: $b.random \in [0.05, 0.3]$, for $n = 1000, 5000$: $b.random \in [0.05, 0.25]$

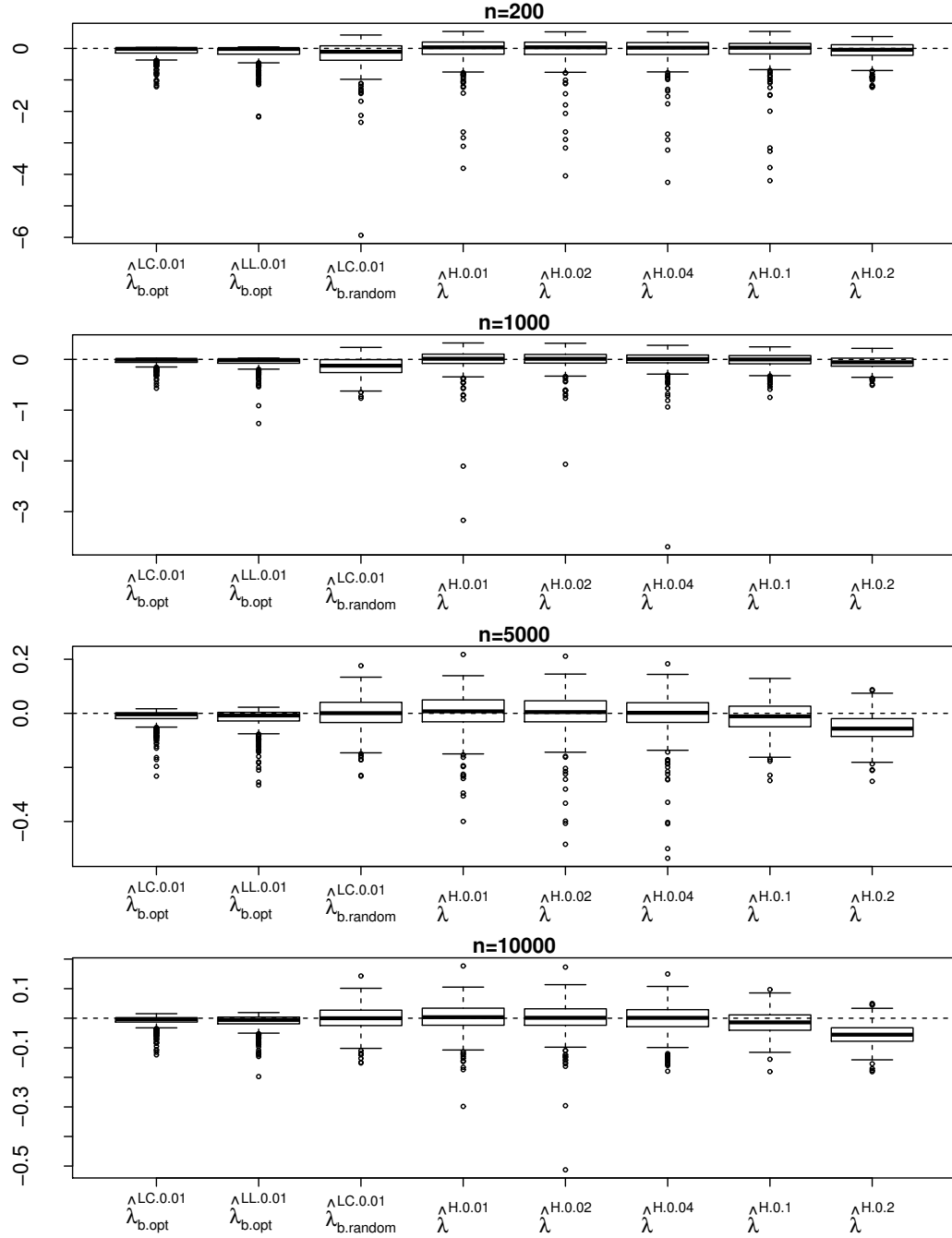


FIGURE 4.5: Boxplot results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a exponential distribution with rate 5, and the underwriting date density Y , is linear increasing, $f_2(y) = 2y$. For $n = 200, 1000$: $b.random \in [0.05, 0.25]$, for $n = 1000, 5000$: $b.random \in [0.01, 0.1]$

n		$\hat{\lambda}_{b,opt}^{LL,0.01}$	$\hat{\lambda}_{b,opt}^{LC,0.01}$	$\hat{\lambda}_{b,random}^{LC,0.01}$	$\hat{\lambda}^{H,0.01}$	$\hat{\lambda}^{H,0.02}$	$\hat{\alpha}^{H,0.04}$	$\hat{\lambda}^{H,0.1}$	$\hat{\lambda}^{H,0.2}$
200	Mean	-1.746	-0.590	0.217	-1.311	-1.146	-1.217	-0.765	-0.474
	Median	-0.089	-0.053	0.607	1.042	0.999	0.848	0.082	-0.178
	SD	3.674	1.289	2.825	9.033	8.177	7.498	4.528	2.957
1000	Mean	-0.906	-0.105	0.536	-0.234	-0.304	-0.313	-0.146	-0.088
	Median	-0.082	-0.010	0.601	0.601	0.479	0.243	0.059	-0.029
	SD	1.850	0.364	1.218	3.845	3.390	2.625	1.631	1.261
5000	Mean	-0.514	-0.004	0.801	-0.115	-0.076	-0.090	-0.022	-0.004
	Median	-0.059	-0.001	0.788	0.348	0.246	0.046	0.015	-0.001
	SD	1.090	0.083	0.559	1.785	1.335	1.088	0.709	0.544
10000	Mean	-0.429	0.005	0.804	-0.045	-0.012	-0.032	-0.007	0.000
	Median	-0.084	0.003	0.814	0.214	0.102	0.027	0.015	-0.014
	SD	0.813	0.079	0.446	1.163	0.932	0.748	0.511	0.388

TABLE 4.1: Simulation results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a Beta distribution with parameters (2, 5), and the underwriting date density, Y , is uniformly distributed. Sample size is $n = 200, 1000, 5000, 10000$. For $n = 200, 1000$: $b.random \in [0.05, 0.3]$, for $n = 1000, 5000$: $b.random \in [0.05, 0.25]$

n		$\hat{\lambda}_{b,opt}^{LL,0.01}$	$\hat{\lambda}_{b,opt}^{LC,0.01}$	$\hat{\lambda}_{b,random}^{LC,0.01}$	$\hat{\lambda}^{H,0.01}$	$\hat{\lambda}^{H,0.02}$	$\hat{\alpha}^{H,0.04}$	$\hat{\lambda}^{H,0.1}$	$\hat{\lambda}^{H,0.2}$
200	Mean	-1.746	-0.590	0.217	-1.311	-1.146	-1.217	-0.765	-0.474
	Median	-0.089	-0.053	0.607	1.042	0.999	0.848	0.082	-0.178
	SD	3.674	1.289	2.825	9.033	8.177	7.498	4.528	2.957
1000	Mean	-0.824	-0.125	-0.062	-0.401	-0.374	-0.373	-0.352	-1.043
	Median	-0.094	-0.030	0.106	0.343	0.321	-0.017	-0.258	-0.999
	SD	1.546	0.365	1.500	3.604	2.622	2.051	1.414	1.271
5000	Mean	-0.464	-0.000	0.400	-0.084	-0.059	-0.120	-0.286	-0.975
	Median	-0.092	-0.001	0.500	0.237	0.094	-0.072	-0.273	-1.008
	SD	0.867	0.093	0.743	1.264	1.020	0.840	0.602	0.534
10000	Mean	-0.351	0.000	0.410	-0.053	-0.033	-0.092	-0.275	-0.958
	Median	-0.112	0.005	0.556	0.092	-0.005	-0.029	-0.258	-0.963
	SD	0.584	0.087	0.665	0.846	0.705	0.584	0.423	0.369

TABLE 4.2: Simulation results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a Beta distribution with parameters (2, 5), and the underwriting date density, Y , is linear increasing, $f_2(y) = 2y$. For $n = 200, 1000$: $b.random \in [0.05, 0.3]$, for $n = 1000, 5000$: $b.random \in [0.05, 0.25]$

allows for straight forward generalisations to more flexible models allowing for calendar time effects and covariates with specific claim informations. This can be done by extending the univariate hazard estimation case to the multivariate case.

Another point is that we only considered claim counts. A generalisation suitable for claim amounts is explained in Chapter 5. The assumption necessary hereby is that the influences of development delay and underwriting date on the claim severity are independent to each other. If this holds, then everything done in this paper can be done in the same way with claim amounts. The uncertainty will, however, depend on the

n		$\hat{\lambda}_{b.opt}^{LL,0.01}$	$\hat{\lambda}_{b.opt}^{LC,0.01}$	$\hat{\lambda}_{b.random}^{LC,0.01}$	$\hat{\lambda}^{H,0.01}$	$\hat{\lambda}^{H,0.02}$	$\hat{\alpha}^{H,0.04}$	$\hat{\lambda}^{H,0.1}$	$\hat{\lambda}^{H,0.2}$
200	Mean	-1.348	-1.033	-1.772	-0.437	-0.454	-0.539	-0.609	-0.759
	Median	-0.170	-0.126	-1.061	0.326	0.334	0.238	0.218	-0.434
	SD	2.518	1.946	4.328	4.101	4.208	4.194	4.353	2.744
1000	Mean	-0.553	-0.426	-1.400	-0.111	-0.055	-0.148	-0.200	-0.620
	Median	-0.120	-0.073	-1.242	0.125	0.113	0.062	-0.017	-0.567
	SD	1.117	0.800	1.790	2.242	1.731	2.209	1.364	1.198
5000	Mean	-0.188	-0.143	0.005	0.032	0.021	-0.032	-0.133	-0.539
	Median	-0.080	-0.037	0.006	0.073	0.046	0.019	-0.113	-0.565
	SD	0.375	0.298	0.582	0.670	0.708	0.740	0.554	0.506
10000	Mean	-0.123	-0.094	-0.001	0.022	0.008	-0.023	-0.142	-0.543
	Median	-0.058	-0.031	-0.001	0.034	0.015	0.012	-0.140	-0.556
	SD	0.255	0.199	0.422	0.475	0.522	0.468	0.382	0.352

TABLE 4.3: Simulation results over 500 repetitions for the relative estimation error of the reserve. The development delay, X , has a exponential distribution with rate 5, and the underwriting date density Y , is linear increasing, $f_2(y) = 2y$. For $n = 200, 1000$: $b.random \in [0.05, 0.25]$, for $n = 1000, 5000$: $b.random \in [0.01, 0.1]$

severity distribution. With this generalisation it would be interesting to compare the methods in this paper with other individual reserving models like Martínez-Miranda, Nielsen, and Verrall (2012) or Antonio and Plat (2014).

In this the paper, we have derived asymptotic results for the estimation uncertainty of the hazard/development factors. Uncertainty of the reserve or estimated sum of the lower triangle is not discussed in this paper. An analytic derivation seems not to be straightforward, since even if the true development factors are known, chain ladder uses the observed values to project into the lower triangle. However, since we are in a full statistical model, one could develop and implement a bootstrap approach which can also include parameter uncertainty. This would also be possible in the more general framework of Chapter 5 which is suitable for claim amounts.

4.A Computational complexity

In this section we give a brief and not so scientific outline about the computational cost involved in the chain ladder algorithm implemented for this paper. The complexity does hereby not depend on the sample size but only on the dimension of the triangle, i.e, the number $(T\delta^{-1})$. In Table 4.4 we provide an idea of the computational complexity of the algorithm running in a standard computer (Intel(R) Core(TM) i5-4590S with 3.00 GHz

and 8.00 GB-RAM with R working under Windows 7-64 bit). Specifically we have evaluated the run-time of one arbitrary simulated sample with $(T\delta^{-1}) = 100, 1000, 10000$. We have hereby split the computation time in the three different categories. Firstly, the aggregation from a triangle of size $T\delta^{-1}$ to a smaller triangle (Aggregation); note that it does hereby not matter to which size the triangle is aggregated. Secondly the calculation of the development factors via the different methods $(\hat{\lambda}^H, \hat{\lambda}^{LC}, \hat{\lambda}^{LL})$ and lastly the chain ladder algorithm when the development factors are given (CL algorithm).

$T\delta^{-1}$	Aggregation	CL algorithm	$\hat{\lambda}^H$	$\hat{\lambda}^{LC}$	$\hat{\lambda}^{LL}$
100	0.011	0.036	0.001	0.106	0.1109
1000	1.223	3.382	1.032	8.976	6.274
10000	117 (2min)	1351 (23min)	4659 (78min)	5176 (86min)	4933 (82min)
Complexity	$(T\delta^{-1})^2$	$(T\delta^{-1})^3$	$(T\delta^{-1})^3$	$(T\delta^{-1})^3\mathcal{B}$	$(T\delta^{-1})^3\mathcal{B}$

TABLE 4.4: Computation time in seconds and complexity for the aggregation, the chain ladder algorithm and the development factor estimators. The local polynomial estimators also depend on the number of bandwidths $\mathcal{B}$. The running time for the LC and LL estimators are given for a choice with $\mathcal{B} = 50$.

4.B A martingale CLT

In Ramlau-Hansen (1983), the author presented a central limit theorem for the martingale $M(t) = N(t) - \int \alpha(t)Z(t)dt$. This is essential to derive asymptotic normality of the kernel and also the histogram estimator of the hazard function α . As mentioned in that paper this central limit theorem is only a special case of Corollary 2 in Liptser and Shirayev (1981) which also covers the discrete setting of chapter 4.3 in this paper. The result can be stated as follows.

Theorem 4.8. *For the continuous case: Consider a predictable process $W_n(t)$ and assume that for some $\sigma^2 \geq 0$ the following conditions are satisfied:*

$$\begin{aligned} \int W_n^2(t)Z(t)\alpha(t)dt &= \sigma^2 + o_p(1), \\ \int W_n^2(t)I\{W_n^2(t) > \varepsilon\}Z(t)\alpha(t)dt &= o_p(1) \quad \text{for all } \varepsilon > 0. \end{aligned}$$

Then, it holds that $\int W_n(u)dM(u) \rightarrow N(0, \sigma^2)$, in distribution. For the discrete case: Consider a predictable process $H_n(t)$ and assume that for some $(\sigma^\delta)^2 \geq 0$ the following

conditions are satisfied:

$$\begin{aligned} \int H_n^2(t) Z^\delta(t) \alpha^\delta(t) d\mu(t) &= \sigma^2 + o_P(1) \\ \int H_n^2(t) I\{H^2(t) > \varepsilon\} Z^\delta(t) \alpha^\delta(t) d\mu(t) &= o_P(1) \quad \text{for all } \varepsilon > 0, \end{aligned}$$

then it holds that

$$\int H_n(t) dM^\delta(t) \rightarrow N(0, (\sigma^\delta)^2),$$

where $M^\delta(t) = N^\delta(t) - \int \alpha^\delta(t) Z^\delta(t) d\mu(t)$.

4.C Proofs

4.C.1 Proof of Proposition 4.1

By defining

$$\alpha^*(t) = \frac{\sum_{i=1}^n \int_{c_1}^{c_2} d\Lambda_i(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i(s) ds},$$

we divide the estimation error $\hat{\alpha}_h^H(t) - \alpha(t)$ into a deterministic part, $\alpha^*(t) - \alpha(t)$, and a variable part, $\hat{\alpha}_h^H(t) - \alpha^*(t)$. By a first order Taylor expansion we get for the deterministic part that

$$\alpha^*(t) - \alpha(t) = \frac{\sum_{i=1}^n \int_{c_1}^{c_2} \{\alpha(s) - \alpha(t)\} Z_i(s) ds}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i(s) ds} = \alpha'(t) h^{-1} \int_{c_1}^{c_2} (t - s) ds + o(h).$$

For the variable part we have

$$\hat{\alpha}_h^H(t) - \alpha^*(t) = \frac{\sum_{i=1}^n \int_{c_1}^{c_2} dM_i(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i(s) ds}.$$

From (4.1) it directly follows that the second condition of Theorem 4.8 in the Appendix is satisfied for $W(s) = (nh)^{1/2} I(s \in [c_1, c_2]) \{\int_{c_1}^{c_2} Z(s) ds\}^{-1}$. To calculate the asymptotic variance several first order Taylor expansions of $\gamma(s)$ and $\alpha(s)$ yield

$$\int W^2(s) \alpha(s) Z(s) ds = \alpha(t) \gamma(t)^{-1} + o(1),$$

where here and below, the integral, $\int$, with no limits denotes integration over the whole support, that is $\int_0^T$. We deduce that $\hat{\alpha}_h^H(t) - \alpha^*(t)$ is centered and asymptotically normal with variance $\sigma^2(t)$.

4.C.2 Proof of Proposition 4.5

By defining

$$\alpha^*(t_j) = \frac{\sum_{i=1}^n \int_{c_1}^{c_2} d\Lambda_i^\delta(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i^\delta(s) d\mu(s)},$$

we divide the estimation error $\hat{\alpha}_h^H(t_j) - \alpha(t_j)$ into a deterministic part, $\alpha^*(t_j) - \alpha(t_j)$, and a variable part, $\hat{\alpha}_h^H(t_j) - \alpha^*(t_j)$. By a first order Taylor expansion we get for the deterministic part that

$$\begin{aligned} \alpha^*(t_j) - \alpha(t_j) &= \frac{\sum_{i=1}^n \int_{c_1}^{c_2} \{\alpha^\delta(s) - \alpha(t_j)\} Z_i^\delta(s) d\mu(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i^\delta(s) d\mu(s)} \\ &= \alpha^\delta(t_j) - \alpha(t_j) + (\bar{\alpha}^\delta)'(t_j) h^{-1} \int_{c_1}^{c_2} (t_j - s) ds + o(h) \\ &= \frac{1}{24} f^{R''}(t_j) \{S^R(t_j)\}^{-1} \delta^2 + (\bar{\alpha}^\delta)'(t_j) \left\{ (t_j - c_1) - \frac{1}{2}h \right\} + o(h + \delta^2). \end{aligned}$$

For the variable part we have

$$\hat{\alpha}_h^{H,\delta}(t_j) - \alpha^*(t_j) = \frac{\sum_{i=1}^n \int_{c_1}^{c_2} dM_i^\delta(s)}{\sum_{i=1}^n \int_{c_1}^{c_2} Z_i^\delta(s) ds}.$$

From (4.7) it directly follows that the second condition of Theorem 4.8 in the Appendix is satisfied for $H(s) = (nh)^{1/2} I(s \in [c_1, c_2]) \{\int_{c_1}^{c_2} Z^\delta(s) d\mu(s)\}^{-1}$. To calculate the asymptotic variance several first order Taylor expansions of $\gamma(s)$ and $\bar{\alpha}^\delta(s)$ yield

$$\int H^2(s) \bar{\alpha}^\delta(s) Z^\delta(s) d\mu(s) = \bar{\alpha}^\delta(t_j) \gamma(t_j)^{-1} + o(1).$$

We deduce that $\hat{\alpha}_h^{H,\delta}(t) - \alpha^*(t)$ is centered and asymptotically normal with variance $\sigma^2(t)$.

4.C.3 Proof of Proposition 4.7

We only show the result for local constant estimator. The case for the local linear estimator is proved in the same way after the kernel $\bar{K}_{t,h}(t - s)$ is replaced by $K(t -$

$s)Z^{-1}(s)$. In Nielsen and Tanggaard (2001) it was shown that this can be done when studying pointwise first order asymptotics. We define

$$\alpha^*(t_j) = \frac{\sum_{i=1}^n \int K_h(t_j - s) d\Lambda_i^\delta(s)}{\sum_{i=1}^n \int K_h(t_j - s) Z_i^\delta(s) d\mu(s)}.$$

The estimation error can then be divided into into a deterministic part, $\hat{\alpha}_h^H(t_j) - \alpha(t_j)$, $\alpha^*(t_j) - \alpha(t_j)$, and a variable part, $\hat{\alpha}_h^H(t_j) - \alpha^*(t_j)$. By a second order Taylor expansion we get for the deterministic part that

$$\begin{aligned} \alpha^*(t_j) - \alpha(t_j) &= \frac{\int K_h(t_j - s) \{\alpha^\delta(s) - \alpha(t_j)\} Z_i^\delta(s) d\mu(s)}{\int K_h(t_j - s) Z_i^\delta(s) d\mu(s)} \\ &= \alpha^\delta(t_j) - \alpha(t_j) + (\bar{\alpha}^\delta)'(t_j) \frac{\int K_h(t_j - s)(t - s) Z(s) d\mu(s)}{\int K_h(t_j - s) Z(s) d\mu(S)} \\ &\quad + \frac{1}{2} (\bar{\alpha}^\delta)''(t_j) \frac{\int K_h(t_j - s)(t - s)^2 Z(s) d\mu(s)}{\int K_h(t_j - s) Z(s) d\mu(S)} + o(h^2) \\ &= \frac{1}{24} f^{R''}(t_j) \{S^R(t_j)\}^{-1} \delta^2 \\ &\quad + \mu_2(K) h^2 \left\{ (\bar{\alpha}^\delta)'(t_j) \gamma'(t_j) \gamma^{-1}(t_j) + \frac{1}{2} (\bar{\alpha}^\delta)''(t_j) \right\} + o(\delta^2 + h^2) \end{aligned}$$

For the variable part we have

$$\hat{\alpha}_h^{H,\delta}(t_j) - \alpha^*(t_j) = \frac{\sum_{i=1}^n \int K_h(t_j - s) dM_i^\delta(s)}{\sum_{i=1}^n \int K_h(t_j - s) Z_i^\delta(s) ds}.$$

From (4.7) it directly follows that that the second condition of Theorem 4.8 in the Appendix is satisfied for $H(s) = (nh)^{1/2} \{\int K_h(t_j - s) Z^\delta(s) d\mu(s)\}^{-1}$. To calculate the asymptotic variance Taylor expansions of $\gamma(s)$ and $\bar{\alpha}^\delta(s)$ yield

$$\int H^2(s) \bar{\alpha}^\delta(s) Z^\delta(s) d\mu(s) = \bar{\alpha}^\delta(t_j) \gamma(t_j)^{-1} + o(1).$$

We deduce that $\hat{\alpha}_h^{H,\delta}(t) - \alpha^*(t)$ is centered and asymptotically normal with variance $\sigma^2(t)$.

4.C.4 Proof of Proposition 4.4

We have to show that Assumption D is a necessary and sufficient condition so that $P(X^\delta = x_j, Y^\delta = y_k)$ factorizes for $(x_j, y_k) \in \mathcal{I}^\delta$. We will only show that Assumption D is a necessary condition the sufficiency is easily shown by plugging in the solution in

a similar manner. Consider the case for $x_j = 0.5\delta$. It holds that

$$pr(X^\delta = 0.5\delta)pr(Y^\delta = y_k) = \sum_{l=0}^{T\delta-1} \int_0^\delta f_1(x) \int_{l\delta}^{(l+1)\delta-x} f_2(y) dy dx \int_{k\delta}^{(k+1)\delta} f_2(y) dy.$$

We also have that

$$pr(X^\delta = 0.5\delta, Y^\delta = y_k) = \int_0^\delta f_1(x) \int_{k\delta}^{(k+1)\delta-x} f_2(y) dy dx$$

Without further restrictions on f_1 those two terms can only be equal if for almost every $x \in [0, \delta]$, and every k

$$\sum_{l=0}^{T\delta-1} \int_{l\delta}^{(l+1)\delta-x} f_2(y) dy \int_{k\delta}^{(k+1)\delta} f_2(y) dy = \int_{k\delta}^{(k+1)\delta-x} f_2(y) dy$$

We assume without loss of generality that there is a k where $\int_{k\delta-x}^{(k+1)\delta} f_2(y) dy$ is not zero for all $x \in [0, \delta]$ (otherwise restrict the range of x). Fixing this k we conclude that

$$\sum_{l=0}^{T\delta-1} \int_{l\delta}^{(l+1)\delta-x} f_2(y) dy / \int_{k\delta}^{(k+1)\delta-x} f_2(y) dy$$

does not depend on x , which shows the necessity.

References

- Aalen, O. O. (1978). “Non-parametric inference for a family of counting processes”. In: *Ann. Stat.* 6, pp. 701–726.
- Addona, V., J. Atherton, and D. B. Wolfson (2012). “Testing the assumption of independence of truncation time and failure time”. In: *Int. J. Biostat.* 8.
- Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Antonio, K. and R. Plat (2014). “Micro-level stochastic loss reserving for general insurance”. In: *Scand. Actuar. J* 2014, pp. 649–669.
- Arjas, E. (1989). “The claims reserving problem in non-life insurance: Some structural ideas”. In: *Astin Bulletin* 19, pp. 139–152.
- Borgan, Ø. (1984). “Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data”. In: *Scand. J. Stat.*, pp. 1–16.

- Cleveland, W. S. (1979). “Robust locally weighted regression and smoothing scatter-plots”. In: *J. Am. Stat. Assoc.* 74, pp. 829–836.
- Drieskens, D., M Henry, J.-F. Walhin, and J. Wielandts (2012). “Stochastic projection for large individual losses”. In: *Scand. Actuar. J.* 2012, pp. 1–39.
- England, P. D. and R. J. Verrall (2002). “Stochastic Claims Reserving In General Insurance”. In: *British Actuarial Journal* 8, pp. 443–544.
- Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.
- Godecharle, E. and K. Antonio (2015). “Reserving by Conditioning on Markers of Individual Claims: A Case Study Using Historical Simulation”. In: *North American Actuarial Journal* 19, pp. 273–288.
- Hiabu, M., E. Mammen, M. D. Martínez-Miranda, and J. P. Nielsen (2016). “In-sample forecasting with local linear survival densities”. In: *Biometrika* Forthcoming.
- Hiabu, M., C. Margraf, M. D. Martínez-Miranda, and J. P. Nielsen (2015). “Cash flow generalisations of non-life insurance expert systems estimating outstanding liabilities”. In: *Expert Systems With Applications* 45, pp. 400–409.
- Hjort, N. L., M. West, and S. Leurgans (1992). “Semiparametric Estimation Of Parametric Hazard Rates”. In: *Survival Analysis: State of the Art*. Ed. by J. P. Klein and P. Goel. Vol. 211. Nato Science. Springer, pp. 211–236.
- Hoem, J. M. (1969). “Fertility rates and reproduction rates in a probabilistic setting”. In: *Biométrie-Praximétrie* 10, pp. 38–66.
- Huang, J., X. Wu, and X. Zhou (2016). “Asymptotic behaviors of stochastic reserving: Aggregate versus individual models”. In: *European J. Oper. Res.* 249, pp. 657–666.
- Jacod, J. (1979). *Calcul stochastique et problemes de martingales*. Berlin: Springer.
- Jacod, J. and A. N. Shiryaev (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer.
- Keiding, N. (1990). “Statistical inference in the Lexis diagram”. In: *PHILOS T ROY SOC A* 332, pp. 487–509.
- Kremer, E. (1982). “IBNR-claims and the two-way model of ANOVA”. In: *Scand. Actuar. J.* 1982, pp. 47–55.
- Kuang, D., B. Nielsen, and J. P. Nielsen (2009). “Chain-ladder as maximum likelihood revisited”. In: *Ann. Actuar. Sci* 4, pp. 105–121.
- Lagakos, S. W., L. M. Barraj, and V. De Gruttola (1988). “Nonparametric Analysis of Truncated Survival Data, with Application to AIDS”. In: *Biometrika* 75, pp. 515–523.

- Liptser, R. S. and A. N. Shirayev (1981). “A functional central limit theorem for semi-martingales”. In: *Theory Probab. Appl.* 25, pp. 667–688.
- Macaulay, F. R. (1931). *The smoothing of time series*. New York: National Bureau of Economic Research.
- Mack, T. (1993). “Distribution-free calculation of the standard error of chain ladder reserve estimates”. In: *Astin Bulletin* 23, pp. 213–225.
- Mandel, M. and R. A. Betensky (2007). “Testing goodness of fit of a uniform truncation model”. In: *Biometrics* 63, pp. 405–412.
- Martínez-Miranda, M. D., J. P. Nielsen, S. Sperlich, and R. Verrall (2013). “Continuous Chain Ladder: Reformulating and generalising a classical insurance problem”. In: *Expert. Syst. Appl.* 40, pp. 5588–5603.
- Martínez-Miranda, M. D., J. P. Nielsen, and R. Verrall (2012). “Double Chain Ladder”. In: *Astin Bull.* 42, pp. 59–76.
- Nielsen, J. P. (1998). “Marker dependent kernel hazard estimation from local linear estimation”. In: *Scand. Actuar. J.* 1998, pp. 113–124.
- Nielsen, J. P. and C. Tanggaard (2001). “Boundary and bias correction in kernel hazard estimation”. In: *Scand. J. Stat.* 28, pp. 675–698.
- Norberg, R. (1993). “Prediction of Outstanding Liabilities in Non-Life Insurance”. In: *Astin Bull.* 23, pp. 95–115.
- Pigeon, M., K. Antonio, and M. Denuit (2013). “Individual loss reserving with the multivariate skew normal framework”. In: *Astin Bulletin* 43, pp. 399–428.
- Ramlau-Hansen, H. (1983). “Smoothing counting process intensities by means of kernel functions”. In: *Ann. Stat.* 11, pp. 453–466.
- Renshaw, A. E. and R. J. Verrall (1998). “A stochastic model underlying the chain-ladder technique”. In: *British Actuarial Journal* 4, pp. 903–923.
- Rosenlund, S. (2012). “Bootstrapping individual claim histories”. In: *Astin Bull.* 42, p. 291.
- Schiegl, M. (2015). “A model study about the applicability of the Chain Ladder method”. In: *Scand. Actuar. J.* 2015, pp. 482–499.
- Schmidt, K. D. (2012). “Loss prediction based on run-off triangles”. In: *AStA Adv. Stat. Anal.* 96, pp. 265–310.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.
- Simonoff, J. S. (1998). *Smoothing methods in statistics*. New York: Springer.

- Stone, C. J. (1977). “Consistent nonparametric regression”. In: *Ann. Stat.* 5, pp. 595–620.
- Taylor, G. C. (1986). *Claims reserving in non-life insurance*. Amsterdam: Elsevier Science Ltd.
- Tsai, W.-Y (1990). “Testing the assumption of independence of truncation time and failure time”. In: *Biometrika* 77, pp. 169–177.
- Verrall, R. J. (1991). “Chain ladder and maximum likelihood”. In: *J. Inst. Actuar.* 118, pp. 489–499.
- Verrall, R., J. P. Nielsen, and A. Jessen (2010). “Including Count Data in Claims Reserving”. In: *Astin Bull.* 40, pp. 871–887.
- Ware, J. H. and D. L. DeMets (1976). “Reanalysis of some baboon descent data”. In: *Biometrics* 32, pp. 459–463.
- Wüthrich, M. V. and M. Merz (2008). *Stochastic claims reserving methods in insurance*. West Sussex: John Wiley & Sons.
- Wüthrich, M. V., M. Merz, and H. Bühlmann (2008). “Bounds on the estimation error in the chain ladder method”. In: *Scand. J. Stat.* 2008, pp. 283–300.
- Zhao, X. B. and X. Zhou (2010). “Applying copula models to individual claim loss reserving methods”. In: *Insurance Math. Econom* 46, pp. 290–299.
- Zhao, X. B., X. Zhou, and J. L. Wang (2009). “Semiparametric model for prediction of individual claim loss reserving”. In: *Insurance Math. Econom* 45, pp. 1–8.

5

Continuous chain-ladder with paid data

This chapter is a working paper. A version of this chapter is also available on the Social Science Research Network (SSRN): <http://ssrn.com/abstract=2782364>

Continuous chain-ladder with paid data

M. Hiabu

Cass Business School, City, University of London, United Kingdom

Abstract

In survival analysis one is usually interested in making inference on the transition time. Recent literature explains how claims data in the non-life insurance context can be embedded in this framework and it is used to calculate future claim numbers. However, when reserves are to be calculated there is a cost associated to every claim, depending on the transition time. We introduce a local polynomial estimator of the cost weighted density of the survival time. This enables one to forecast the cost of future claims. This is done without the use of more complex marked point process theory. Consistency and a central limit theorems for the normalized estimation errors are provided.

5.1 Introduction

For centuries, researchers have been interested in estimating and predicting demographic quantities. Even before the establishment of mathematical statistics, they have collected data in life tables to investigate size and distribution of the population. Around 1870, the Lexis diagram emerged with an attempt of leading demographers to formalize those data in a useful and coherent manner. It can be best explained as a two-way ANOVA arrangement, where data is organized on a two dimensional plane of (calendar time, age) counting the number of people dead in those aggregated cells, with the purpose of making inference on the three dimensional system of (calendar time, age, cohort), where cohort is given by the diagonals. In the actuarial discipline of ‘reserving in non-life insurance’, data is arranged in so-called run-off triangles. The appearance is similar to Lexis diagrams, but the plane is given as (cohort, age), where cohort is the accident data of a claim, and age the time from that date to a payment. Developed at least in the beginning of the last century, the chain-ladder method is still the industry standard for estimating the future cost for outstanding liabilities from those run-off triangles. It is a deterministic algorithm which often gives reasonable point estimates, however, the estimator does not specify the assumptions that it is based on, nor the uncertainty of the estimation.

Stochastic models around the chain-ladder method have been developed in Kremer (1982), Verrall (1991), Mack (1993), Renshaw and Verrall (1998), and Kuang, Nielsen, and Nielsen (2009) and many others. A comprehensive summary can be found by England and Verrall (2002). The drawback of those papers is that they do not discuss how the data arises as aggregation from individual data. This is needed when one wants to truly understand the underlying assumptions of the model. (Taylor, 1986) coins those models as macro-models which are in direct contrast to micro-models which begin on the individual level.

Recent literature addresses this gap and connects the chain ladder method and its data to counting process theory in survival analysis. In Hiabu (2016) (Chapter 4), we introduce a full statistical model including the data generating process which is built on the continuous model of Martínez-Miranda et al. (2013). The authors explain that the estimation and sampling technique of the chain-ladder method is different from other

sampling techniques used in classical (bio-)statistical literature: Individuals or policies are only followed if a failure, i.e., a claim occurs. This has the advantage that less data is required than in classical survival data, and censoring does not occur. Truncation occurs when $cohort + age$ is greater than the date of data collection. However when all failures are observed, inference on the two dimensional random variable (cohort, age) on the unobserved area is still possible via survival analysis techniques as explained in Hiabu et al. (2016) (Chapter 2).

However, in contrast to the life tables in demographic data, the data in the run-off triangles are usually not the aggregation of events, but events with their associated cost. In other words, claim numbers are not summarized, but claim amounts. Martínez-Miranda et al. (2013) explain how the classical survival data in the chain ladder method can be understood as arising from continuous data and how the estimators can be understood as histogram estimators. However, the authors also point out that theory is limited to the case where one is interested in the event times rather than the claim cost.

In this paper, we introduce a cost weighted density estimator based on a local polynomial least squares minimisation principle, which is known from regression (Stone, 1977) and translated to the survival density setting in (Nielsen, Tanggaard, and Jones, 2009). We do so by introducing a mark representing the cost associated to the jump-observations of the counting process and elaborate under which assumptions non-parametric estimation is possible. Consistency and a central limit theorems for the normalized estimation errors are provided. An application for the estimation of outstanding liabilities can be found in Martínez-Miranda et al. (2013).

There also exist other micro-models for estimating outstanding liabilities in non-life insurance. Arjas (1989) and Norberg (1993) formulated models in a classical bio-statistical setup via marked-point processes. The problem with their models is that one is not interested in full inference on the marked point process, i.e. for instance the distribution of the mark/cost. This distribution is not necessary to derive an estimate of the outstanding liabilities. As mentioned before, those approaches also require information about the exposure (i.e. information about the number of policies in the portfolio), which does not carry information about the cost of the single claims and might be quite volatile.

5.2 Model formulation

We now formulate the model under a quite general counting process framework. The special case of estimating outstanding liabilities (reserving) is explained in the next section. Consider a probability space $(\Omega, \mathcal{F}, P)$. When observing n individuals, let $N_i = I(t \geq X_i)$ be a $\{0, 1\}$ -valued counting process, which observes the failure of the i th individual in the time interval $[0, 1]$. The process N_i is adapted to an increasing, right-continuous, complete filtration, $\mathcal{F}_t^i \subset \mathcal{F}, t \in [0, 1]$. We further observe the $\{0, 1\}$ -valued $\mathcal{F}_i$ -predictable process, Y_i , which equals unity when the i th individual is at risk. Finally we observe a covariate Z_i , which is given X_i independent to $\mathcal{F}_i$, and is the cost of the occurred failure, zero if no failure occurred. Assuming independence between the individuals, we thus have independent identically distributed observations of triples (N_i, Y_i, Z_i) ($i = 1, \dots, n$).

We assume that the random variable (X, Z) has density f with respect to the Lebesgue measure and has support $\{[0, 1] \times \mathbb{R}_+\}$. The filtered observation (X_1, Z_1) then has density f^* , which differs from f due to the incomplete observation described via the exposure process Y_1 . We assume the following relationship between Y_1 and N_1 .

Assumption 1 [*Aalen's multiplicative intensity model*] The intensity of the counting process N_1 exists and can be decomposed as

$$\lambda_1(t) = \lim_{h \downarrow 0} h^{-1} E \left[N_1 \{ (t+h)- \} - N_1(t-) \mid \mathcal{F}_{t-}^1 \right] = \alpha(t) Y_1(t), \quad (5.1)$$

where α is a continuous function.

The most prominent example of an observation scheme satisfying Aalen's multiplicative intensity model is left truncation. In cross-sectional observations for example, one starts following individuals from a specific point in time. This means one observes triplets (U_i, X_i, Z_i) ($i = 1, \dots, n$) where U_i is age at which an individual enters the study, X_i is the age at which an event happens. Hence, $U_i \leq X_i$, and the counting process formulation is $N_i(t) = I(X_i \leq t)$ and $Y_i(t) = I(U_i \leq t < X_i)$. Assumption 1 is satisfied if U and X are independent. Examples, without the observations of Z , of prevalent cohort data, nursing-home data and AIDS blood transfusion data are given in Wang (1989), see also Andersen et al. (1993).

The observation of (N_i, Z_i) can be interpreted as observing a marked point process, see e.g. Jacobsen (2006). But we are not interested in making inference on the marked point process as such. We want to estimate the cost weighted density,

$$\tilde{f}^X(t) = \frac{E[Z|X=t]}{E[Z]} f^X(t) = \frac{E[Z|X=t]}{E[Z]} \alpha(t) \exp \left\{ - \int_0^t \alpha(s) ds \right\}. \quad (5.2)$$

Note that the conditional expectations to point events with probability zero here and below are well defined through the marginals of f and f^* . To be able to estimate this quantity non-parametrically we assume

Assumption 2 The random variable Z_1 is uniformly integrable and

$$\frac{E[Z_1 | \Delta N_1(t) = 1]}{E[Z_1 | Y_1(t) = 1]} = \frac{E[Z|X=t]}{E[Z|X \geq t]},$$

where $\Delta N_i(t) = \lim_{h \downarrow 0} N_i\{(t+h)-\} - N_i(t-)$.

Under the the left truncation observation scheme, i.e., $Y_i(t) = I(U_i \leq t < X_i)$, we will show that this is, under mild assumptions, equivalent to the following assumption.

Assumption 2* The conditional expectation of the cost Z_1 , given (X_1, U_1) is multiplicatively separable, i.e., it can be written as $E[Z_1 | X_1, U_1] = g_1(X_1)g_2(U_1)$, with two functions g_1, g_2 .

Proposition 5.1. *Assume $Y_i(t) = I(U_i \leq t < X_i)$. If the random variable (U, X, Z) has density g , continuous and bounded from above and below, with respect to the Lebesgue measure, then Assumption 2 is equivalent to Assumption 2*.*

5.3 Reserving and In-sample forecasting

In-sample forecasting has been introduced in Martínez-Miranda et al. (2013), and has been further developed and generalized in Mammen, Martínez-Miranda, and Nielsen (2015), Lee et al. (2015), and Hiabu et al. (2016) (Chapter 2). The data in in-sample forecasting can be seen as incomplete observations due to right-truncation. Hence, given that the truncation is independent of the survival time, Assumption 1 can be

fulfilled by reversing the time of the counting process, which turns the right truncation to a left truncation. We have $\tilde{X}_1 = 1 - X_1$, $N_1 = I(t \geq \tilde{X}_1)$, see Ware and DeMets (1976) and Hiabu et al. (2016). On a deeper glance it is a different sampling technique compared to classical survival data, in that only but all failures are observed and there is no information (needed) about the amount of individuals under risk. We observe (U_i, X_i, Z_i) ($i = 1, \dots, n$), where X describes the time from origin until a specific event, and U is the calendar time of origin. Observations are then truncated if $U + X$ is larger than the date of data collection. Note that under Assumption 1, U and X are independent.

In non-life insurance, outstanding liabilities are traditionally estimated using the chain ladder method. The method is applied on so called run-off triangle of historical claim amounts which are aggregated on a two dimensional grid of underwriting date of the claim's underlying policy and the time between this date and the payment. On an individual basis this readily translates into the above model as follows. Given n observations of independent and identically distributed historical claims, let the random variable U_i describe the underwriting date of the underlying policy, and let the random variable X_i be the time between this date and the payment. The mean of the outstanding claim amount is then given as

$$\tau \int_0^1 \int_{1-u}^1 \tilde{f}^X(x) \tilde{f}^U(u) dx du,$$

where $(\tilde{f}^X(x) \tilde{f}^U(u))$ is the cost weighted density of (U, X) (cf. (5.2)) and

$$\tau = n \left\{ \int_0^1 \int_0^{1-u} \tilde{f}^X(x) \tilde{f}^U(u) dx du \right\}^{-1}.$$

Due to symmetry, the components $\tilde{f}^X(x)$ and $\tilde{f}^U(u)$ can be estimated separately via the approach described in the next section. This approach generalises the theory described in the previous chapter since it allows to estimate outstanding claim amounts instead of only claim counts.

5.4 Local polynomial estimation

We first define the cost-weighted Kaplan–Meier product-limit estimator of the survival function $\tilde{S}(t) = \int_0^t \tilde{f}^X(s) ds = \{E[Z_1 | X_1 \geq t] / E[Z_1]\} \int_0^t f^X(s) ds$,

$$\hat{\tilde{S}}(t) = \prod_{s \leq t} \{1 - \Delta \hat{A}(s)\},$$

where $\hat{A}(t) = \sum_{i=1}^n Z_i \left\{ \sum_{j \neq i} Z_j Y_j(s) \right\}^{-1} dN_i(s)$ is motivated by the Aalen estimator, estimating, $\tilde{A}(t) = \int_0^t E[Z_1 | X = s] \{E[Z_1 | X \geq s]\}^{-1} \alpha(s) ds$. Let $q_p(z) = \sum_i^p \theta_i z^i$ denote a polynomial of degree p . We define the local polynomial estimator of degree p , $\widehat{\tilde{f}}_{p,h,K}^X(t)$ of $\tilde{f}^X(t)$ as the minimizer $\hat{\theta}_0$ in the equation

$$\begin{aligned} \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} = \arg \min_{\theta_0, \theta_1 \in \mathbb{R}} \sum_{i=1}^n \left[\int K_h(t-s) \{q_p(t-s)\}^2 Z_i Y_i(s) W(s) ds \right. \\ \left. - 2 \int K_h(t-s) q_p(t-s) \frac{Z_i^2}{\sum_j Z_j Y_j(s)} \hat{\tilde{S}}(s) Y_i(s) W(s) dN_i(s) \right] \end{aligned} \quad (5.3)$$

Here and below, an integral $\int$ with no limits denotes integration over the whole support, i.e., $\int_0^T$. In addition, for kernel K and bandwidth h , $K_h(t) = h^{-1} K(t/h)$. The definition of the local polynomial estimator as the minimizer of (5.3) can be motivated by the fact that the sum on the right hand side of (5.3) equals the limit of

$$\sum_{i=1}^n \int \left[\left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \hat{\tilde{S}}(u) dN_i(u) - q_p(t-s) \right\}^2 - \xi(\varepsilon) \right] K_h(t-s) Z_i Y_i(s) W(s) ds,$$

for ε converging to zero. Here, $\xi(\varepsilon) = \{\varepsilon^{-1} \int_s^{s+\varepsilon} (\sum_j Z_j Y_j(u))^{-1} Z_i \hat{\tilde{S}}(u) dN_i(u)\}^{-2}$ is a vertical shift subtracted to make the expression well-defined. Because $\xi(\varepsilon)$ does not depend on q_p , $\hat{\theta}_0$ is defined by a local weighted least squares criterion. The function, W , is an arbitrary predictable weight function. There exist two popular weightings: the first being the natural unit weighting, $W(s) = 1$, while the second is the Ramlau–Hansen weighting, $W(s) = \{n/Y(s)\} I\{Y(s) > 0\}$. The latter becomes the classical kernel density estimator in the simple unfiltered case. However, in the framework of filtered observations the natural unit weighting, $W(s) = 1$, tends to be more robust (Nielsen, Tanggaard, and Jones, 2009), so we use it.

In the sequel we will only consider the cases $p = 0, 1$, i.e., the local constant and local linear case. While a higher degree in conjunction with higher order kernels improves the asymptotic properties, finite sample studies show that improvements are only visible with unrealistically big sample sizes. In the local constant case of (5.3) we derive the first order condition

$$2 \sum_{i=1}^n K_h(t-s) Z_i Y_i(s) ds = 2 \sum_{i=1}^n K_h(t-s) Z_i Y_i(s) dN_i(s),$$

and conclude the local constant estimator

$$\hat{f}_{0,h,K}(t) = \frac{\sum_{i=1}^n \int K_h(t-s) \hat{\hat{S}}(s) Z_i dN_i(s)}{\sum_{i=1}^n \int K_h(t-s) Z_i Y_i(s) ds}$$

We make the following assumptions.

- S1. The bandwidth $h = h(n)$ satisfies $h \rightarrow 0$ and $n^{1/4}h \rightarrow \infty$ for $n \rightarrow \infty$.
- S2. The density f^X is strictly positive and two times continuously differentiable.
- S3. The kernel K is symmetric, has bounded support and has finite second moment.
- S4. There is a strictly positive and continuous function γ with $\sup_{s \in [0,1]} |\sum_{i=1}^n Y_i(s)/n - \gamma(s)| = o_p(1)$, for $n \rightarrow \infty$.
- S5. The function $l(t) = E[Z_1 | Y_1(t) = 1]$ is continuously differentiable.

We introduce the following notation. For every kernel, K , let

$$\mu_j(K) = \int s^j K(s) ds, \quad R(K) = \int K^2(s) ds, \quad \bar{K}^*(u) = \frac{\mu_2(K) - \mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2} K(u).$$

Proposition 5.2. Under Assumption 1, 2 and (S1)–(S5), for $t \in (0, T)$, $n \rightarrow \infty$,

$$(nh)^{1/2} \left\{ \hat{f}_{0,k,h}(t) - \tilde{f}^X(t) - B_0(t) \right\} \rightarrow N \left\{ 0, \sigma_0^2(t) \right\},$$

in distribution, where

$$B_0(t) = \frac{1}{2} h^2 \mu_2(K) \left[\tilde{f}''(t) h^2 + \tilde{f}'(t) \frac{\{l(t)\gamma(t)\}'}{l(t)\gamma(t)} \right],$$

$$\sigma_0^2(t) = \left\{ \frac{E[Z_1 | X = t]}{E[Z_1]} \right\}^2 R(K) f(t) S(t) \gamma(t)^{-1}.$$

For the local linear case, we introduce the following quantities.

$$G_j(t) = \sum_{i=1}^n \int K_h(t-s)(t-s)^j Z_i dN(s) \quad (j = 0, 1).$$

$$a_j(t) = \sum_{i=1}^n \int K_h(t-s)(t-s)^j Z_i Y_i(s) ds \quad (j = 0, 1, 2).$$

The first order condition for $p = 1$ then reads

$$G_0(t) = \theta_0 a_0 + \theta_1 a_1,$$

$$G_1(t) = \theta_0 a_1 + \theta_1 a_2.$$

Hence the solution θ_0 is given by

$$\hat{f}_{1,h,K}(t) = n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) \hat{\tilde{S}}(s) Z_i dN_i(s), \quad (5.4)$$

where

$$\bar{K}_{t,h}(t-s) = n \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2} K_h(t-s).$$

If K is a second-order kernel, then $n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s) Z_i Y_i(s) ds = 1$, $n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s)(t-s) Z_i Y_i(s) ds = 0$, $n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s)(t-s)^2 Z_i Y_i(s) ds > 0$, so that $\bar{K}_{t,h}$ can be interpreted as a second-order kernel with respect to the measure, μ , where $d\mu(s) = n^{-1} \sum_{i=1}^n Z_i Y_i(s) ds$.

Since

$$\sup_{t \in [h, 1-h]} |a_j(t) - h^j \mu_j(K) g(t) \gamma(t)| = o_p(1) \quad (j = 1, 2, 3), \quad (5.5)$$

one can easily verify that $n^{-1} \sum_i \bar{K}_{t,h}(t-s) Z_i Y_i(s)$ converges locally uniform almost surely to $\bar{K}_h^*(t-s)$, where $\bar{K}_h^*$ arises from $\bar{K}^*$ by replacing u and $K(u)$ with the local versions $h^{-1}u, h^{-1}K(u/h)$; see also Nielsen and Tanggaard (2001). Furthermore, if K is symmetric, then $\bar{K}^*(t) = K(t)$.

Proposition 5.3. *Under Assumption 1, 2 and (S1)–(S5), for $t \in (0, T)$, $n \rightarrow \infty$,*

$$(nh)^{1/2} \left\{ \hat{f}_{1,k,h}(t) - \tilde{f}^X(t) - B_1(t) \right\} \rightarrow N \left\{ 0, \sigma_1^2(t) \right\},$$

in distribution, where

$$B_1(t) = \frac{1}{2}h^2\mu_2(K)\tilde{f}''(t)h^2,$$

$$\sigma_1^2(t) = \left\{ \frac{E[Z_1 | X = t]}{E[Z_1]} \right\}^2 R(K)f(t)S(t)\gamma(t)^{-1}.$$

5.5 Concluding remarks

In this paper, we have introduced a local constant and a local linear estimator for a mark weighted survival density. In the context of reserving in non-life insurance, this extends the theory of continuous chain ladder, which is described in the previous chapters, see also Martínez-Miranda et al. (2013), from handling claim counts to now also handling claim amounts. It turns out that one can use the same estimator in both cases. If claim amounts are estimated, asymptotic bias and variance will additionally depend on the conditional mean severity of a claim. The fact that the same estimator can be used is not so surprising, since the traditional chain-ladder method is also applied on both claim counts and claim amounts. However, the estimation of claim amounts comes with the cost of additional assumptions. Assumption 2* dictates that the influence of the payment delay, X , and the underwriting date, U , on the claim's severity, Z , must act independently. An application or simulation study is not presented in this paper. An application can be found in Martínez-Miranda et al. (2013), which was done without discussing the underlying theory presented here. For future research, it would also be interesting to examine continuous chain ladder for claim counts and claim amounts acting together, as is done in the double chain ladder framework (Martínez-Miranda, Nielsen, and Verrall, 2012) for aggregated run-off triangles.

5.A Proofs

5.A.1 Proof of Proposition 5.1

First note that

$$\begin{aligned} \frac{E[Z_1 | \Delta N_1(t) = 1]}{E[Z_1 | Y_1(t) = 1]} &= \frac{E[Z_1 | X = t, U \leq t]}{E[Z_1 | X > t, U \leq t]} \\ &= \frac{\int_0^\infty \int_0^t z g(u, t, z) du dz \int_t^1 \int_0^\infty \int_0^t g(u, s, z) du dz ds}{\int_t^1 \int_0^\infty \int_0^t z g(u, s, z) du dz ds \int_0^\infty \int_0^t g(u, t, z) du dz}. \end{aligned}$$

Now, since X and U are independent,

$$\begin{aligned} &\int_t^1 \int_0^\infty \int_0^t g(u, s, z) du dz ds / \int_0^\infty \int_0^t g(u, t, z) du dz \\ &= \int_t^1 \int_0^\infty \int_0^1 g(u, s, z) du dz ds / \int_0^\infty \int_0^1 g(u, t, z) du dz = \alpha^{-1}(t). \end{aligned}$$

Hence Assumption 2 is equivalent to

$$\frac{\int_0^t \int_0^\infty z g(u, t, z) dz du}{\int_t^1 \int_0^t \int_0^\infty z g(u, s, z) dz du ds} = \frac{\int_0^1 \int_0^\infty z g(u, t, z) dz du}{\int_t^1 \int_0^1 \int_0^\infty z g(u, s, z) dz du ds}.$$

With continuity arguments this holds if and only if $\int_0^\infty z g(u, s, z) dz$ is multiplicatively separable in s and u . This completes the proof with the independence of X and U .

5.A.2 Estimation of the weighted survival function

We first analyse the process $\hat{A}_1(t) = \int_0^t Z_1 / \{\sum_{j \neq 1} Z_j Y_j(s)\} dN_1(s)$, where the integral can be understood pathwise in Lebesgue-Stieltjes sense. From (5.1) we conclude that

$$\begin{aligned} &\lim_{h \downarrow 0} h^{-1} E \left[\hat{A}_1 \{(t+h)-\} - \hat{A}_1(t-) \mid \mathcal{F}_{t-}^1 \right] \\ &= \lim_{h \downarrow 0} h^{-1} E \left[\frac{Z_1}{\sum_{j \neq 1} Z_j Y_j(X_1)} \mid X_1 \in [t, t+h) \right] E \left[N_1 \{(t+h)-\} - N_1(t-) \mid \mathcal{F}_{t-}^1 \right] \\ &= \frac{E[Z_1 | \Delta N_1(t) = 1]}{(n-1)E[Z_1 | Y_1(t) = 1] \gamma(t)} \alpha(t) Y_1(t) \end{aligned}$$

Hence,

$$\tilde{\Lambda}_i(t) = \frac{1}{(n-1)} \int_0^t \frac{E[Z_1 | \Delta N_1(s) = 1]}{E[Z_1 | Y_1(s) = 1] \gamma(s)} \alpha(s) Y_i(s) ds, \quad (i = 1, \dots, n),$$

is a compensator of the uniformly integrable submartingale $\hat{A}_i$. We denote the resulting martingale by $\widetilde{M}_i = \hat{A}_i - \widetilde{\Lambda}_i$. Since M is cadlag with finite variation, the quadratic variation equals the sum of square differences:

$$[\widetilde{M}_1(t)] = \sum_{0 \leq s \leq t} (\Delta \widetilde{M}_1(s))^2 = \int_0^t \left\{ \frac{Z_1}{\{\sum_{j \neq 1} Z_j Y_j(s)\}} \right\}^2 dN(s).$$

And by similar arguments as before we can calculate its compensator to derive the predictable variation process

$$\langle \widetilde{M}_1(t) \rangle = \int_0^t \left\{ \frac{E[Z_1 | \Delta N_1(s) = 1]}{(n-1)E[Z_1 | Y_1(s) = 1]\gamma(s)} \right\}^2 \alpha(s) Y_1(s) ds$$

Proposition 5.4. *Under Assumption 1-2, S1-S4 it holds that*

$$n^{1/2} \sum_i \widetilde{M}_i \rightarrow U(\sigma^2), \quad \sigma^2 = \int_0^t \left\{ \frac{E[Z_1 | \Delta N_1(s) = 1]}{E[Z_1 | Y_1(s) = 1]} \right\}^2 \alpha(s) \gamma^{-1}(s) ds,$$

in distribution in Skorohod topology sense, where U is a zero mean Gaussian martingale with covariance, $\text{Cov}\{U(s), U(t)\} = \sigma^2(s \wedge t)$.

Proof. This follows from a martingale central limit theorem in Rebolledo (1980), see also Andersen et al. (1993)[p.83]. For the assumptions to be satisfied, we verify that

$$\langle \sum_i \widetilde{M}_i(t) \rangle = n \sum_i \int_0^t \left\{ \frac{E[Z_1 | \Delta N_1(s) = 1]}{(n-1)E[Z_1 | Y_1(s) = 1]\gamma(s)} \right\}^2 \alpha(s) Y_i(s) ds \rightarrow \sigma^2,$$

where we have used that $\langle \widetilde{M}_i, \widetilde{M}_j \rangle = 0$ for $i \neq j$. The Lindenberg condition follows from $n^{-1/2} Z_1 \rightarrow 0$, and the fact that jumps happen at the same time with zero probability. $\square$

Corollary 5.5. *Under Assumption 1-2, S1-S4 it holds that*

$$n^{1/2} \sup_t |\hat{\hat{S}}(t) - \widetilde{S}(t)| = O_p(1)$$

Proof. This directly follows from applying the functional delta method on Proposition 5.4, since $\hat{\hat{S}}$ and $\widetilde{S}$ are functionals of $\hat{A}$ and $\widetilde{A}$, respectively. $\square$

5.A.3 Proof Proposition 5.2

We first split the estimation error into a stable part and a martingale part, $\widehat{f}_0 - \widetilde{f}^X = B_0 + V_0$, via

$$B_0 = \widetilde{f}_0^* - \widetilde{f}^X, \quad V_0 = \widehat{f}_0 - \widetilde{f}_0^*,$$

where

$$\widetilde{f}_0^* = \frac{\sum_{i=1}^n \int K_h(t-s) \widehat{S}(s) E[Z_1 | \Delta N_1(s) = 1] Y_i(s) \alpha(s) ds}{\sum_{i=1}^n \int K_h(t-s) Z_i Y_i(s) ds}.$$

We now discuss the asymptotics of B and V separately, and conclude the proof by showing that that $B_0(t) = \frac{1}{2} \mu_2(\overline{K}^*) f''(t) h^2 + o(h^2)$, and then that

$$(nh)^{1/2} V_0(t) \rightarrow N \left\{ 0, \{E[Z_1 | X = t] / E[Z_1]\}^2 R(K) f(t) S(t) \gamma(t)^{-1} \right\}.$$

We start with V . The main tool is the following Lemma. We define

$$\overline{M}_i = \int Z_i dN_i(s) - \int \{E[Z_1 | \Delta N_1(s) = 1] \alpha(s) Y_i(s) ds.$$

Under Assumption 1,2 and S1-S4, one can show that

$$n^{1/2} \sum_i \overline{M}_i \rightarrow U(\sigma^2), \quad \sigma^2 = \int_0^t \{E[Z_1 | \Delta N_1(s) = 1]\}^2 \alpha(s) \gamma^{-1}(s) ds,$$

in distribution in Skorohod topology sense, where U is a zero mean Gaussian martingale with covariance, $\text{Cov}\{U(s), U(t)\} = \sigma^2(s \wedge t)$. With Proposition (survival function), S4 and the central limit theorem stated above we conclude that $(nh)^{1/2} V_0 \rightarrow N(0, \sigma_0^2)$, with

$$\sigma_0^2 = h \frac{\int K_h^2(t-s) \widetilde{S}^2(s) E^2[Z_1 | \Delta N_1(s) = 1] \alpha(s) \gamma(s) ds}{\left\{ \int K_h(t-s) E[Z_1 | Y_1(s) = 1] \gamma(s) ds \right\}^2}$$

A first order Taylor expansion in the numerator as well as denominator then gives the desired result. We continue with the asymptotics for B_0 . After reshuffling, and replacing $\widehat{S}(s)$ by $\widetilde{S}(s)$, which we can do by arguing with Proposition 5.4, we have that

$$B_0(t) = \frac{\sum_{i=1}^n \int K_h(t-s) Y_i(s) \{ \widetilde{f}^X(s) E[Z_1 | Y_1(s) = 1] - Z_i \widetilde{f}^X(t) \} ds}{\sum_{i=1}^n \int K_h(t-s) Z_i Y_i(s) ds} + o(h^2).$$

From assumption (S4) we can further use that $n^{-1} \sum_i Z_i Y_i(s)$ converges uniformly to $E[Z_1|Y_1 = 1]\gamma(s)$. Hence,

$$B_0(t) = \frac{\int K_h(t-s) E[Z_1|Y_1(s) = 1] \gamma(s) \{ \tilde{f}^X(s) - \tilde{f}^X(t) \} ds}{\int K_h(t-s) E[Z_1|Y_1(s) = 1] \gamma(s) ds} + o(h^2)$$

The proof is concluded by a Taylor expansion in the numerator and denominator and using that K is a second order kernel.

5.A.4 Proof of Proposition 5.3

From (5.5), S3 and Proposition 5.4, we conclude that it is enough to consider the asymptotic behavior of

$$n^{-1} \sum_{i=1}^n \int K_h(t-s) \frac{Z_i \tilde{S}(s)}{\sum_i Z_i Y_i(s)} dN_i(s).$$

Analog to the local constant case, we split the estimation error into a stable and a martingale part

$$B_1 = \tilde{f}_1^* - \tilde{f}^X + o_p(n^{-1/2}), \quad V_1 = \widehat{\tilde{f}_1} - \tilde{f}_1^* + o_p(n^{-1/2}),$$

where

$$\tilde{f}_1^* = \int K_h(t-s) \tilde{f}^X(s) ds.$$

The asymptotic limit of the bias part, B_1 , is now easily derived via a second order Taylor expansion. The martingale part can be concluded with similar arguments as in Appendix 5.A.2.

References

- Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Arjas, E. (1989). “The claims reserving problem in non-life insurance: Some structural ideas”. In: *Astin Bulletin* 19, pp. 139–152.
- England, P. D. and R. J. Verrall (2002). “Stochastic Claims Reserving In General Insurance”. In: *British Actuarial Journal* 8, pp. 443–544.

- Hiabu, M. (2016). “On the relationship between classical chain ladder and granular reserving”. In: *Scand. Actuar. J.* Forthcoming.
- Hiabu, M., E. Mammen, M. D. Martínez-Miranda, and J. P. Nielsen (2016). “In-sample forecasting with local linear survival densities”. In: *Biometrika* Forthcoming.
- Jacobsen, M. (2006). *Point process theory and applications: marked point and piecewise deterministic processes*. Birkhäuser.
- Kremer, E. (1982). “IBNR-claims and the two-way model of ANOVA”. In: *Scand. Actuar. J.* 1982, pp. 47–55.
- Kuang, D., B. Nielsen, and J. P. Nielsen (2009). “Chain-ladder as maximum likelihood revisited”. In: *Ann. Actuar. Sci* 4, pp. 105–121.
- Lee, Y. K., E. Mammen, J. P. Nielsen, and B. Park (2015). “Asymptotics for In-Sample Density Forecasting”. In: *Ann. Stat.* 43, pp. 620–651.
- Mack, T. (1993). “Distribution-free calculation of the standard error of chain ladder reserve estimates”. In: *Astin Bulletin* 23, pp. 213–225.
- Mammen, E., M. D. Martínez-Miranda, and J. P. Nielsen (2015). “In-sample forecasting applied to reserving and mesothelioma”. In: *Insurance Math. Econom.* 61, pp. 76–86.
- Martínez-Miranda, M. D., J. P. Nielsen, S. Sperlich, and R. Verrall (2013). “Continuous Chain Ladder: Reformulating and generalising a classical insurance problem”. In: *Expert. Syst. Appl.* 40, pp. 5588–5603.
- Martínez-Miranda, M. D., J. P. Nielsen, and R. Verrall (2012). “Double Chain Ladder”. In: *Astin Bull.* 42, pp. 59–76.
- Nielsen, J. P. and C. Tanggaard (2001). “Boundary and bias correction in kernel hazard estimation”. In: *Scand. J. Stat.* 28, pp. 675–698.
- Nielsen, J. P., C. Tanggaard, and M. C. Jones (2009). “Local linear density estimation for filtered survival data”. In: *Statistics* 43, pp. 176–186.
- Norberg, R. (1993). “Prediction of Outstanding Liabilities in Non-Life Insurance”. In: *Astin Bull.* 23, pp. 95–115.
- Rebolledo, R. (1980). “Central limit theorems for local martingales”. In: *Probab. Theory Related Fields* 51, pp. 269–286.
- Renshaw, A. E. and R. J. Verrall (1998). “A stochastic model underlying the chain-ladder technique”. In: *British Actuarial Journal* 4, pp. 903–923.
- Stone, C. J. (1977). “Consistent nonparametric regression”. In: *Ann. Stat.* 5, pp. 595–620.

- Taylor, G. C. (1986). *Claims reserving in non-life insurance*. Amsterdam: Elsevier Science Ltd.
- Verrall, R. J. (1991). “Chain ladder and maximum likelihood”. In: *J. Inst. Actuar.* 118, pp. 489–499.
- Wang, M.–C (1989). “A Semiparametric Model for Randomly Truncated Data”. In: *J. Am. Stat. Assoc.* 84, pp. 742–748.
- Ware, J. H. and D. L. DeMets (1976). “Reanalysis of some baboon descent data”. In: *Biometrics* 32, pp. 459–463.

