European Actuarial Journal (2022) 12:485–502
https://doi.org/10.1007/s13385-021-00291-0

ORIGINAL RESEARCH PAPER

A nonparametric sequential learning procedure for estimating the pure premium

Jun Hu[1] · Liang Hong[2]

Received: 18 December 2020 / Revised: 15 April 2021 / Accepted: 14 July 2021 /
Published online: 2 August 2021

Abstract
With the advent of the "big" data era, large-sample properties of a statistical learning method are becoming more and more important in an actuary's daily work. For a fixed sample size, regardless of how large it is, the variance of an estimator can be larger than a pre-assigned level to an arbitrary extent. In this paper, we propose a nonparametric sequential learning procedure for estimating the pure premium. Our method not only provides an accurate estimate of the pure premium but also guarantees that the mean of our random sample sizes is close to the unobservable optimal fixed sample size and the variance of our estimator is close to all small pre-determined levels. In addition, our method is nonparametric and applicable to any claims distribution; hence it avoids potential issues associated with a parametric model such as model misspecification risk and the effect of selection.

Keywords Large-sample properties · Nonlife insurance · Nonparametric methods · Point estimation · Property and casualty insurance · Sequential methods · Statistical learning

1 Introduction

Accurate estimation of the pure premium is one of the most important tasks in an actuary's daily work because it affects several key aspects of an insurer's business operations such as product pricing, reserves estimation, regulatory compliance, and

✉ Liang Hong
liang.hong@utdallas.edu

Jun Hu
junhu@oakland.edu

1 Department of Mathematics and Statistics, Oakland University, 146 Library Drive, Rochester, MI 48309, USA

2 Department of Mathematical Sciences, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, USA

risk management. Since the underlying claims distribution is unknown, there are two common approaches to estimating the pure premium. In the first approach, the actuary first proposes a parametric model, then applies one of those model-selection tools such as AIC or BIC to pick a "best" model, and finally takes the mean of the chosen claims model to be the estimate of the pure premium. The second approach adopts a model-free setup and estimates the pure premium using a nonparametric estimator. While the first approach seems very reasonable, there are two potential issues lurking behind the scene. First, using data to choose a model via a model selection tool might adversely affect the outcome of statistical inference. This phenomenon is called the *effect of selection* (e.g. Refs. [3, 17, 29]). Second, even if the actuary correctly applies a model selection procedure, there is no guarantee that the chosen model is correct, subjecting estimation to model misspecification biases. Since the second approach avoids these two issues, it is the preferable choice when it comes to estimation of the pure premium. With the arrival of the "big" data era, the insurer has easy access to insurance data sets of large sample sizes. As a result, good asymptotic properties of an estimator become more relevant. Practically speaking, when the actuary applies a method to estimate the pure premium, it is reasonable to expect estimation to be more accurate when the actuary has more data. This should be a minimum requirement for a good estimator. In other words, the actuary should only use an asymptotically unbiased estimator with an asymptotically vanishing variance.

When claims data for a homogeneous group of policyholders is available, the sample mean is a natural estimator of the pure premium. This estimator has a number of appealing properties such as strong consistency, intuitive interpretability, computational simplicity, and model-independence. Therefore, it seems to be a perfect candidate for estimating the pure premium. However, as claims data is random, it is also essential to consider the variance of an estimator. The variance of the sample mean vanishes when the sample size goes to infinity. But, all real-world applications only involve finite samples. This means the variance of any estimator cannot be zero. From a practical point of view, it is desirable that the variance of the estimator should be below or at least close to a small pre-assigned level. For a fixed sample size, this is impossible to achieve, even if the sample mean has an asymptotically vanishing variance; see Sect. 2 for more details. In this paper, we propose a sequential learning method that provides accurate estimation of the pure premium and, at the same time, guarantees that (i) the mean of our random sample sizes is close to the unknown optimal fixed sample size; and (ii) the variance of our estimator is close to all small pre-determined levels. To our knowledge, the insurance literature contains very few methods from sequential analysis. The only exceptions we found are [2, 25, 26], where the former two papers employ stochastic approximation to evaluate the credibility formula for distributions outside the exponential family, and the latter paper applies sequential testing to decide whether full credibility should be taken in the limited fluctuation credibility approach. The techniques of sequential analysis are also rarely discussed in popular textbooks or monographs on loss modeling, predictive modeling, and statistical learning (e.g. Refs. [10, 16, 22]). Therefore, our paper is the first one in the insurance literature that gives a full-blown sequential analysis for an estimation problem.

The remainder of the paper is organized as follows. In Sect. 2, we provide readers with some background of sequential estimation. In particular, we use a concrete example to illustrate the fact that the variance of a good estimator, such as the sample mean, may be arbitrarily larger than a pre-assigned level, even if that estimator has an asymptotically vanishing variance. Next, in Sect. 3, we give a detailed description of our proposal and establish several desirable properties of our method. In Sect. 4, we give three numerical examples, one based on simulated data and two based on real data, to demonstrate the excellent performance of our method. Finally, we provide some concluding remarks in Sect. 5. The Appendix contains proofs of all theorems.

2 Background

2.1 Why a sequential learning procedure?

To give readers a flavor for why sequential estimation is important and useful, we start with a well-known toy example. Suppose $X_1, \dots, X_n$ is a random sample from $\mathsf{N}(\theta, \sigma^2)$, where θ and σ^2 are both unknown. A data scientist wants to estimate θ and make the variance of his/her estimator fall below a pre-determined level, say, $b > 0$. Since the sample mean $\bar{X}_n = \sum_{k=1}^{n} X_k/n$ is evidently a sufficient statistic for θ, the Lehmann–Scheffé theorem (e.g. Ref. [36], Theorem 5.5) immediately implies that $\bar{X}_n$ is a minimum variance unbiased estimator (MVUE). However, since σ^2 is unknown and can be arbitrarily large, there is no way the data analyst can bound $\mathsf{V}(\bar{X}_n) = \sigma^2/n$ below any pre-decided $b > 0$.

To circumvent this difficulty, Stein [41, 42] propose the following two-stage sequential learning scheme:

The data analyst first obtains a pilot sample $\mathbf{X}^m = (X_1, \dots, X_m)$ where the pilot sample size $m > \max\{2, 1/b + 3\}$, computes the sample variance $S_m^2 = \sum_{i=1}^{m}(X_i - \bar{X}_m)^2/m$, and puts $k = k(\mathbf{X}^m) = 1 + \lfloor mS_m^2 \rfloor = 1 + \lfloor \sum_{i=1}^{m}(X_i - \bar{X}_m)^2 \rfloor$ where $\lfloor a \rfloor$ denotes the greatest integer less than or equal to $a \in \mathbb{R}$. Note that k is random since it depends on the pilot sample $\mathbf{X}^m$. Now take an additional sample of size k and let $n = m + k$ be the terminal sample size. The intuition of this two-stage sampling scheme is quite clear: the data analyst first assesses the magnitude of σ^2 based upon the pilot sample $\mathbf{X}^m$. If S_m^2 is relatively large, then σ^2 is likely to be large. Hence, relatively more samples will be needed in the second stage. On the other hand, if S_m^2 turns out to be relatively small, then relatively fewer samples will be taken in the second stage. This two-stage procedure is summarized in Algorithm 1.

Algorithm 1: Sequential learning procedure for estimating a normal mean

1 Take a pilot sample $X_1, \ldots, X_m$ where $m > \max\{2, 1/b + 3\}$;
2 Compute the corresponding sample variance $S_m^2 = \sum_{i=1}^m (X_i - \bar{X}_m)^2/m$;
3 Set $k = k(\mathbf{X}^m) = 1 + \lfloor mS_m^2 \rfloor$ where $\lfloor a \rfloor$ denotes the greatest integer less than or equal to $a \in \mathbb{R}$;
4 Take an additional sample of size k and let $n = m + k$ be the terminal sample size;
5 Return $\bar{X}_n$ as the estimator of θ.

It turns out that $\bar{X}_n$ produced by Algorithm 1 is an unbiased estimator of θ and $\mathsf{V}(\bar{X}_n) < b$. To see this, we first notice that the iid normal assumption implies that $\bar{X}_m$ and S_m^2 are independent, which further implies the independence of k and $\sum_{i=1}^m X_i$. By the iterated expectation formula, we have

$$\begin{aligned} \mathsf{E}(\bar{X}_n) &= \mathsf{E}[\mathsf{E}(\bar{X}_n \mid \mathbf{X}^m)] \\ &= \mathsf{E}\left[\theta + \left(\sum_{i=1}^m X_i - m\theta\right)/(m+k)\right] \\ &= \theta + \mathsf{E}\left(\sum_{i=1}^m X_i - m\theta\right)\mathsf{E}\left(\frac{1}{m+k}\right) \\ &= \theta. \end{aligned}$$

Similarly, $\mathsf{E}(\bar{X}_n^2) = \theta^2 + \mathsf{E}[\sigma^2/(m+k)]$. In view of our iid normal assumption, $\sum_{i=1}^m (X_i - \bar{X}_m)^2/\sigma^2 \sim \chi_{m-1}^2$ where χ_{m-1}^2 denotes the Chi-square distribution with $m-1$ degrees of freedom. It follows that

$$\begin{aligned} \mathsf{V}(\bar{X}_n) &= \mathsf{E}[\sigma^2/(m+k)] \\ &\le \mathsf{E}[1/(S_m^2/\sigma^2)] \\ &= 1/(m-3) < b. \end{aligned}$$

References [34, 39, 40] further study sequential estimation of the mean for a normal distribution. However, these results all depend heavily on the normality assumption of the sample. In most insurance applications, the claims distribution is supported on $\mathbb{R}_+ = [0, \infty)$; hence, the normality assumption is clearly inappropriate. Starr and Woodroofe [40] investigates the same problem for a gamma distribution. But, even if an actuary chooses a positive distribution (e.g. gamma or Pareto) using a model selection tool such AIC or BIC, a parametric assumption can still subject the actuary to model misspecification risk [19, 21]. In addition, recent research shows that using a goodness-of-fit test to choose a model may have adverse effects on statistical inference (e.g., Refs. [3, 17, 29]). To avoid these potential issues, we will take a nonparametric approach in this paper.

Throughout, we assume that insurance claims $X_1, \ldots, X_n$ are iid from an unknown distribution function F supported on $\mathbb{R}_+$. Our main quantity of interest is the pure premium $\mu = \mathsf{E}(X_1)$. In the nonparametric setting, the claims distribution does not take a parametric form. Therefore, there is no way one can bound the variance of a reasonable estimator below all small pre-determine levels. Thus, our goal is to find a procedure that provides an accurate estimation of μ and, at the same time, guarantees that the variance of the estimator is close to a small pre-assigned level $b > 0$. Note that a small b indicates a small variance of an estimator which is certainly desirable. For accurate estimation of μ, a natural estimator is the sample mean $\hat{\mu} = \bar{X}_n = \sum_{i=1}^{n} X_i/n$. This estimator has several appealing properties: (i) it is strongly consistent; (ii) it is intuitive to interpret and simple to calculate; (iii) it is nonparametric and model-independent; and (iv) it has an asymptotically vanishing variance. However, in any practical application, the sample size n is finite. Following the same line of reasoning as in the aforementioned motivating example, we know that it is impossible for the insurer to control the variance of $\hat{\mu}$ within a small range of a pre-assigned level using a fixed sample size n, no matter how large n is. Therefore, we must resort to a sequential learning procedure.

2.2 Some key components of sequential estimation

Sequential estimation is called for when a fixed sample scheme cannot meet the goal of the estimation as in the above example. This usually happens when part of the goal can be achieved only if an unobservable parameter $\theta = \theta(F)$ was known. In such a case, a typical trick is first to use a pilot sample $X_1, \dots, X_m$ to estimate that parameter θ based on some statistic $\hat{\theta} = g(X_1, \dots, X_m)$ where g is an m-ary real-valued function and then proceed as if $\hat{\theta}$ is the true parameter θ. In the above toy example, $\theta = \sigma^2$ and $\hat{\theta} = S_m^2$.

In a sequential procedure, we do not fix the sample size n in advance. Therefore, the convention is to let $(X_1, X_2, \dots)$ denote the claims data for the sake of convenience. In practice, we cannot keep sampling forever. For this reason, we need to specify a rule for deciding when to stop taking additional samples. Such a rule is called a *stopping time*. Apparently, a stopping time cannot be arbitrary. At a minimum, a stopping time should stipulate when to stop taking further samples, based on previously obtained samples. Therefore, we define a stopping time N to be a function from the X-space $\mathbb{R}_+$ to $\overline{\mathbb{Z}}_+ = \{0, 1, 2, \dots, \} \cup \{+\infty\}$ such that the event $[N = j]$ is independent of $X_{j+1}, X_{j+2}, \dots$ for all non-negative integers j. Since a stopping time is data-dependent by definition, the sample size N in a sequential procedure is random. Note that N is allowed to take the value $+\infty$ in which case sampling never stops. Therefore, it is crucial to verify that $\mathbb{P}(N < +\infty) = 1$ in all applications. In the above example, the stopping time is given by

$$N = \inf\{n \geq m \,:\, n \geq m + 1 + mS_m^2\} = m + 1 + \lfloor mS_m^2 \rfloor. \tag{1}$$

Since each normal random variable is trivially finite with probability one, we immediately have $\mathbb{P}(N < +\infty) = 1$.

Finally, once we stop taking more samples according to a stopping time N, we must have an estimator h for estimating our quantity of interest (e.g., μ), based on our sample $X_1, \ldots, X_N$. In the above example, our estimator of μ is $\bar{X}_N = \sum_{i=1}^{N} X_i/N$ where N is given by Eq. (1). There is no hard-and-fast rule for setting a stopping time N and choosing the final estimator h. But a rule of thumb is that these two quantities should be set in accordance with the goal of the statistical inference and they should be relatively easy to implement numerically. For a textbook treatment of sequential estimation, we refer to [14, 15, 31, 37, 38].

3 Nonparametric sequential estimation of the pure premium

3.1 Our proposal

It is natural to ask whether the sequential estimation scheme in the above toy example can be carried over for a nonparametric setup. Unfortunately, the answer is negative. As a matter of fact, the argument in that example depends on two properties of iid normal random variables: (1) $\bar{X}_m$ and S_m^2 are independent and (2) $\sum_{i=1}^{m}(X_i - \bar{X}_m)^2/\sigma^2 \sim \chi^2_{m-1}$. In the nonparametric setting we are considering, these two properties fail and the argument will not go through. Therefore, we need to seek a different path. Towards this, we notice that if σ^2 were known, then the optimal fixed sample size $n^\star$ for achieving

$$b \geq \mathsf{V}(\bar{X}_n) = \sigma^2/n,$$

would be

$$n^\star = \sigma^2/b. \tag{2}$$

Here and throughout, we follow the notational convention of sequential analysis to tacitly disregard the fact that $n^\star$ may not be an integer. Since σ^2 is unknown in a real-world application, we follow the typical trick in sequential estimation to first take a pilot sample $X_1, \ldots, X_m$ (where $m \geq 2$) and estimate σ^2 as $S_m^2 = \sum_{i=1}^{m}(X_i - \bar{X}_m)^2/m$. Then we take additional samples one by one according to the following stopping time:

$$N = \inf\{n \geq m : n \geq S_n^2/b\}, \tag{3}$$

where $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/n$ denotes the updated sample variance. Finally, our estimator of μ will be $\hat{\mu} = \bar{X}_N$ where $\bar{X}_N$ is the sample mean of the whole sample $X_1, \ldots, X_N$. The proposed sequential procedure for estimating μ is summarized in Algorithm 2.

Algorithm 2: Nonparametric sequential learning procedure for estimating the mean

Take a pilot sample $X_1, \ldots, X_m$ where $m \geq 2$;
Compute the corresponding sample variance $S_m^2 = \sum_{i=1}^{m}(X_i - \bar{X}_m)^2/m$;
for $m \leq n < S_n^2/b$ **do**
  Take one additional sample X_{n+1};
  Compute the corresponding sample variance $S_{n+1}^2 = \sum_{i=1}^{n+1}(X_i - \bar{X}_{n+1})^2/(n+1)$;
  Update n as $n+1 \to n$;
end
Return $\bar{X}_n$ as the estimator of μ.

3.2 Properties

The optimal fixed sample size $n^\star$ defined by Eq. (2) is a deterministic number, but N given by Eq. (3) is a random variable. Since $S_n^2 \to \sigma^2$ with probability one as $n \to +\infty$ by the law of large numbers, we have $\mathbb{P}(N < +\infty) = 1$. That is, Algorithm 2 will stop at some point with probability one. The next three theorems depict the asymptotic behavior of N as $b \to 0$ under some mild moment conditions. Theorem 3.1 indicates that $\mathsf{E}(N)$ will not overshoot $n^\star$ too much; Theorems 3.2 and 3.3 ensure that $\mathsf{E}(N)$ will be close to the optimal fixed sample size $n^\star$.

Theorem 3.1 *Let $b > 0$ and $m \equiv m(b)$ (which depends on b). If $\delta b^{-1/2} \leq m(b) = o(b^{-1})$ as $b \to 0$ for some $\delta > 0$ and $\mathsf{E}|X_1|^2 < +\infty$, then*

$$\mathsf{E}(N) \leq n^\star + 1 + o(1) \text{ as } b \to 0. \tag{4}$$

Proof See the Appendix. □

Theorem 3.2 *Let $b > 0$ and $m \equiv m(b)$. If $\delta b^{-1/2} \leq m(b) = o(b^{-1})$ as $b \to 0$ for some $\delta > 0$ and $\mathsf{E}|X_1|^4 < +\infty$, then*

$$\mathsf{E}(N) = n^\star + O(1) \text{ as } b \to 0. \tag{5}$$

Proof See the Appendix. □

Theorem 3.3 *Let $b > 0$ and $m \equiv m(b)$. If $\delta b^{-1/2} \leq m(b) = o(b^{-1})$ as $b \to 0$ for some $\delta > 0$ and $\mathsf{E}|X_1|^{6+\epsilon} < +\infty$ for some $\epsilon > 0$, then*

$$\mathsf{E}(N) = n^\star + \sigma^{-2}\mathsf{E}(U) - \mathsf{V}[(X_1 - \mu)^2] - 1 + o(1) \text{ as } b \to 0, \tag{6}$$

where U is a random variable following the same distribution as the first hitting time of the random walk $2n - \sum_{i=1}^{n} X_i^2$ at the boundary $y = 0$.

Proof See the Appendix. □

In addition to the above three nice properties of N, Anscombe's [1] random central limit theorem implies asymptotic normality of $\bar{X}_N$. That is,

$$\sqrt{N}(\bar{X}_N - \mu) \to N(0, \sigma^2) \text{ in distribution as } b \to 0. \tag{7}$$

This guarantees that our method will provide an accurate estimate of the pure premium μ. Since a small value of $b > 0$ is desirable here, it is also necessary to assess the performance of our estimator for all small values of $b > 0$. The next theorem shows that the variance of our proposed estimator is close to all small pre-assigned levels.

Theorem 3.4 *Let $b > 0$ and $m \equiv m(b)$. If $\delta b^{-1/2} \leq m(b) = o(b^{-1})$ as $b \to 0$ for some $\delta > 0$ and $\mathsf{E}|X_1|^{4+\epsilon} < +\infty$ for some $\epsilon > 0$, then*

$$\mathsf{V}(\bar{X}_N) = b + o(b) \text{ as } b \to 0. \tag{8}$$

Proof See the Appendix. □

As mentioned above, our stopping time N is random. Thus, bounding $\mathsf{V}(\bar{X}_N)$ is more challenging than bounding $\mathsf{V}(\bar{X}_n)$ where n is a fixed sample size. Since the underlying distribution of the claims $X_1, X_2, \ldots$ is unknown, no method can allow the actuary to bound $\mathsf{V}(\bar{X}_N)$ from above by b for all small $b > 0$. Therefore, the best one can hope is that $\mathsf{V}(\bar{X}_N)$ is close to a pre-assigned small b. Theorem 3.4 shows that this is exactly what our method is able to achieve.

3.3 Choice of the pilot sample size *m*

In the above Algorithm 2, the pilot sample size m is required to be at least 2 so that the pilot sample variance S_m^2 is well-defined. The upside of a large pilot sample size is that Algorithm 2 will converge relatively quickly; but the downside is that the pilot sample size can potentially exceed the optimal fixed sample size $n^\star$ (oversampling). On the other hand, a small pilot sample size will usually increase computational time but it is unlikely to exceed $n^\star$. In practice, an actuary will need to choose the pilot sample size based upon the desirable value of b, the size of the given dataset, and other prior information. If b is relatively small, the given data is large, oversampling is acceptable, or the prior information indicates the true distribution might have a heavy tail, then a relatively large pilot sample size might be appropriate. But, if b is not excessively small, the given dataset is small or moderate, oversampling is costly, or the prior information suggests the true distribution has a light tail, the actuary might want to take a small pilot sample size. In general, our recommendation is that the actuary take a relatively conservative pilot sample size.

4 Examples

4.1 Simulated data

Here we perform a simulation study and apply our nonparametric sequential learning method to estimate the pure premiums of three claims distributions:

1. The gamma distribution with shape parameter 3 and scale parameter 3, i.e., the gamma distribution with the density function:

$$f(x) = \frac{1}{54}x^2 e^{-x/3}, \quad x > 0.$$

2. The type-II Pareto distribution with the distribution function:

$$G(x) = 1 - \left(\frac{30}{x+30}\right)^7, \quad x > 0.$$

3. The folded-t distribution (introduced to the insurance literature by Brazauskas and Kleefeld [4, 5]) with parameters $s = 5$ and $\nu = 8$, i.e., the folded-t distribution with the density function:

$$h(x) = 0.4h_{t(8)}(x/5), \quad x > 0,$$

where $h_{t(8)}$ is the density function of the central Student's t distribution with 8 degrees of freedom. It follows from Psaraki and Panaretoes [32] that the mean and variance of h are

Table 1 Optimal fixed sample size $n^\star$, terminal sample size N, the ratio of the pure premium estimate to the pure premium $\hat{\mu}/\mu$, and the ratio of the estimated variance of the estimator to the pre-assigned level $\hat{\mathbb{V}}(\bar{X}_N)/b$, with varying values of b, for the three distributions in Sect. 4.1

	b	$n^\star$	N	$\hat{\mu}/\mu$	$\hat{\mathbb{V}}(\bar{X}_N)/b$
Gamma	0.10	270	267	0.9990	1.0007
	0.08	338	336	0.9996	1.0006
	0.05	540	539	0.9997	1.0003
	0.03	900	901	1.0001	1.0002
	0.01	2700	2700	0.9999	1.0001
Pareto	0.10	350	329	0.9873	1.0003
	0.08	438	417	0.9916	1.0004
	0.05	700	677	0.9943	1.0002
	0.03	1167	1149	0.9976	1.0001
	0.01	3500	3484	0.9993	1.0000
Folded-t	0.10	139	135	0.9908	0.9895
	0.08	173	167	0.9913	1.0005
	0.05	276	271	0.9951	1.0008
	0.03	460	455	0.9978	1.0003
	0.01	1381	1375	0.9992	1.0002

$$2s\sqrt{\nu/\pi}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)(\nu-1)} = 4.4194$$

and

$$s^2\left[\frac{\nu}{\nu-2} - \frac{4\nu}{\pi(\nu-1)^2}\left(\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\right)^2\right] = 13.8021,$$

respectively.

For each distribution, we generate $M = 1,000$ random samples. For each random sample, we run Algorithm 2 on the sample with a pilot sample size $m = 100$, record the stopping time N given by Eq. (3), and calculate the ratio of our estimate to the true mean $\hat{\mu}/\mu$ as well as the ratio of the estimated variance of our estimator to the pre-assigned level $\hat{\mathsf{V}}(\bar{X}_N)/b$. The optimal fixed sample size $n^\star$ and the average values of N, $\hat{\mu}/\mu$, and $\hat{\mathsf{V}}(\bar{X}_N)/b$ are reported in Table 1.

It is clear that our methods provide accurate estimation of the pure premium μ. In addition, for each of the five given values of b, the value of $\hat{\mathsf{V}}(\bar{X}_N)$ is close to the value of b, across three different claims distributions. In general, the smaller the value of b is, the larger the value of N will be. As b decreases, our estimator of the pure premium becomes more accurate and $\mathsf{V}(\bar{X}_N)$ gets closer to b. These observations are all consistent with the theorems in Sect. 3.2.

4.2 1991 Norwegian fire claims data

The 1991 Norwegian fire claims data has 624 entries of fire claims in thousand Norwegian krones, ranging from 500 to 49,692. It is available from http://lstat.kuleuven.be/Wiley/ or the R package `CASdatasets`. We first re-scale the data by dividing each entry by 500 so that the range of the scaled data is [1, 99.384]. Table 2 provides the summary statistics for the scaled data.

Since the original data is in the ascending order, we randomly permutate the data and then apply our sequential method to the scaled data with $m = 100$ for $b = 0.10$, 0.05, and 0.01. Table 3 provides the values of the stopping time N, our estimate of

Table 2 Summary statistics for scaled 1991 Norwegian fire claims data

n	Min	1st Qu	Median	Mean	3rd Qu	Max	Variance
624	1.000	1.506	2.197	3.641	3.818	99.384	36.468

Table 3 Terminal sample size n, pure premium estimate $\hat{\mu}$, and the estimated variance of the estimator $\hat{\mathsf{V}}(\bar{X}_N)$, for the 1991 Norwegian fire claims data

b	N	$\hat{\mu}$	$\hat{\mathsf{V}}(\bar{X}_N)$
0.10	100	2.9851	0.0581
0.05	122	3.0726	0.0496
0.01	≥ 625	NA	NA

the pure premium $\hat{\mu}$, and the estimated variance of our estimator $\hat{\mathsf{V}}(\bar{X}_N)$. In Table 3, NA stands for "not available".

We make the following observations. For $b = 0.10$, a pilot sample of size 100 is sufficiently large to ensure that $\mathsf{V}(\bar{X}_N) \leq b$. When $b = 0.05$, a terminal sample of 122 is required to guarantee that $\mathsf{V}(\bar{X}_N)$ is close to b. As for $b = 0.01$, our method shows that the given sample (whose size is 624) is not large enough to guarantee that $\mathsf{V}(\bar{X}_N)$ is close to b and more samples are needed to achieve that goal; therefore, the corresponding values of $\hat{\mu}$ and $\hat{\mathsf{V}}(\bar{X}_N)/b$ are not available. This is no surprise. Since Table 2 indicates that the true distribution has a large variance relative to its mean and has a heavy tail, it is reasonable that a small b like 0.01 requires more samples for Algorithm 2 to converge.

In this case, we know neither the true mean μ nor the true variance σ^2. However, based on the above simulation study, we are inclined to believe that our method performs satisfactorily: $\hat{\mu} = \bar{X}_N$ gives accurate estimation of the pure premium μ, N is close to the unknown optimal sample size $n^\star$, and $\mathsf{V}(\bar{X}_N)$ is close to the pre-determined level b.

4.3 Automobile insurance claims data

Finally, we examine a data set of claims from private passenger vehicle policies written by a large US non-life insurer. This data set contains settled amounts on 6673 closed claims with four covariates: age, gender, class, and state. It is available from the companion website of [9]: https://instruction.bus.wisc.edu/jfrees/jfreesbooks. We first apply the recursive binary splitting (e.g. Ref. [16], Sect. 9.2) to segment the data into relatively homogeneous groups, and then estimate the pure premium for each group. Since the state code is randomly assigned, we do not include the covariate "state" in our analysis. For the covariate "class", the code starts with either "C" or "F" and is followed by a numerical number. These numerical numbers do not seem to be ordinal; hence, we simply classify "class" into "C" and "F". The recursive binary splitting segments the original data into four groups: (i) age > 81.5, (ii) age < 81.5 and class = F, (iii) age < 81.5, class = C, and gender = F(Female),

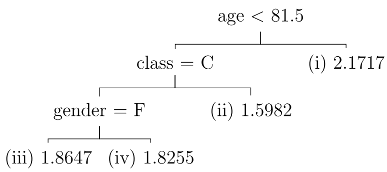

Fig. 1 Regression tree for the automobile insurance claims data

Table 4 Sample size n, terminal sample size N, pure premium estimate $\hat{\mu}$, and the estimated variance of the estimator $\hat{\mathsf{V}}(\bar{X}_N)$, for the four groups, based on the automobile insurance claims data

Group	n	N	$\hat{\mu}$	$\hat{\mathsf{V}}(\bar{X}_N)$
(i)	484	449	2.1843	0.0299
(ii)	316	181	1.7743	0.0299
(iii)	2316	124	1.7404	0.0297
(iv)	3657	216	1.9046	0.0298

and (iv) age < 81.5, class $=$ C, and gender $=$ M(Male). Figure 1 depicts the corresponding tree, which has three internal nodes and four terminal nodes/leaves. At a given internal node, the label indicates the left-hand branch emanating from that split. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to age < 81.5, and the right-hand branch corresponds to age ≥ 81.5. The number in each leaf is the mean of the response for the observations that fall there.

For each of the four resultant groups, we re-scale the claims by diving 1000 and run our sequential method on the scaled claims with $b = 0.05$ and $m = 100$. For each group, Table 4 gives the sample size, values of the stopping time N, our estimate of the pure premium $\hat{\mu}$, and the estimated variance of our estimator $\hat{\mathsf{V}}(\bar{X}_N)$. A few observations are in order. First, the degree of segmentation has some nontrivial impact on the estimate of the pure premium as anticipated. Second, the terminal sample size is also affected by the degree of segmentation. Note that Group (i) lies in the first layer of the tree, Group (ii) is in the second layer, and Group (iii) and (iv) belong to the last layer. In general, the deeper the layer is, the more homogeneous the group is. Hence, for the same b and m, the terminal sample size tends to be smaller relative to the actual sample size for a group in a deeper layer. Finally, the estimated variance of the estimator $\hat{\mathsf{V}}(\bar{X}_N)$ does not seem to be affected by the degree of segmentation in this case; this is no surprise since terminal sample size is less than the sample size of each group in this case.

5 Concluding remarks

Recent insurance literature has witnessed a surge of applications of novel machine/statistical learning methods to various insurance problems (e.g. Refs. [11–13, 18, 20, 27, 28, 33, 43]). This paper is also one such effort: it employs a full-fedged sequential learning method for an estimation problem in insurance. In a fixed sample scheme, the variance of an estimator of the pure premium could be larger than a pre-determined level to an arbitrary extent. We have proposed a nonparametric sequential method for learning the pure premium. Our method gives accurate estimation of the pure premium and, at the same time, provides provable guarantee that the average of our stopping time is close to the unknown optimal fixed sample size and that the variance of our estimator is close to all small pre-assigned levels. Though the pure premium is one key quantity in an actuary's daily work, the actuary also needs to learn other features of the claims distribution, such as higher moments, Value-at-Risk, expected shortfall, to make informed decisions. Therefore, one important problem for future research is to find a method which can accurately estimate the density function of the claims distribution and, at the same time, ensures that the corresponding error is close to all small pre-determined levels.

Appendix

Preliminaries

Throughout this appendix, we assume without loss of generality that $\mu = 0$ and $\sigma = 1$. We let $T_n = \sum_{i=1}^{n} X_i$ denote the partial sum. The following preliminary results will be used in the proofs below. Equation (12) follows from Theorem IV-3 of [44], and others follow from Lemmas 2, 4 and 5 of [8].

$$\mathsf{E}|X_1|^{2p} < +\infty, p \geq 1 \Rightarrow \left\{(b^{-1}N)^p \,:\, 0 < b \leq 1\right\} \text{ is uniformly integrable.} \tag{9}$$

$$\mathsf{E}|X_1|^{2} < +\infty \Rightarrow \left\{(bN^{-1})^p \,:\, 0 < b \leq 1\right\} \text{ is uniformly integrable for all } p > 0. \tag{10}$$

$$\mathsf{E}|X_1|^{2} < +\infty, 0 < \gamma < 1 \Rightarrow \mathbb{P}(N \leq \gamma b^{-1}) = O(b^p) \text{ as } b \to 0 \text{ for all } p > 0. \tag{11}$$

$$\mathsf{E}|X_1|^{2p} < +\infty, p \geq 2 \Rightarrow \left\{|b^{1/2}(N - b^{-1})|^p \,:\, 0 < b \leq 1\right\} \text{ is uniformly integrable.} \tag{12}$$

$$\mathsf{E}|X_1|^{2p} < +\infty, p \geq 1 \Rightarrow \left\{(b^{1/2}T_N)^{2p} \,:\, 0 < b \leq 1\right\} \text{ is uniformly integrable.} \tag{13}$$

$$\mathsf{E}|X_1|^{2p} < +\infty, p \geq 2 \Rightarrow \left\{|b^{1/2}(\sum_{i=1}^{N} X_i^2 - N)|^p \,:\, 0 < b \leq 1\right\} \text{ is uniformly integrable.} \tag{14}$$

Proof of Theorem 3.1

From the definition of the terminal sample size N in Eq. (3), we know that, on the set $\{N > m\}$,

$$(N-1)^2 \leq b^{-1}\sum_{i=1}^{N-1}(X_i - \bar{X}_{N-1})^2 \leq b^{-1}\sum_{i=1}^{N-1}(X_i - \mu)^2 = b^{-1}\sum_{i=1}^{N-1} X_i^2.$$

Jensen's inequality and Wald's first identity (e.g. Refs. [35], Theorem 3.3.2; [15], Theorem 2.4.4) imply that

$$\mathsf{E}^2[N-1] \leq \mathsf{E}[(N-1)^2] \leq b\mathsf{E}[N-1] + m^2\mathbb{P}(N = m). \tag{15}$$

In view of Eq. (9), we know that $\mathsf{E}[b(N-1)] < \infty$ for any $b > 0$. Next, we derive the order of $m^2\mathbb{P}(N = m)$. Note that as $b \to 0$,

$$\begin{aligned}\mathbb{P}(N = m) \leq \mathbb{P}(N \leq m) &\leq \mathbb{P}(N \leq \gamma b^{-1}), \text{ for some } 0 < \gamma < 1, \\ &= O(b^r), \text{ for all } r > 0,\end{aligned}$$

where the last equality follows from Eq. (11). This and Eq. (15) imply that

$$\begin{aligned}\mathsf{E}[N-1] \leq & b^{-1} + m^2\mathbb{P}(N=m)\mathsf{E}^{-1}[N-1] \leq b^{-1} + m^2\mathbb{P}(N=m)(m-1)^{-1} \\ \leq & b^{-1} + o(b^{-2})O(b^r)o(b^{-1}) = b^{-1} + o(1), \text{ as } b \to 0,\end{aligned}$$

for an appropriately chosen $r > 3$. In view of Eq. (2), we have $\mathsf{E}(N) \leq n^\star + 1 + o(1)$ as $b \to 0$.

Proof of Theorem 3.2

It suffices to show that $\mathsf{E}[N] \geq n^\star + O(1)$ as $b \to 0$, since the reverse inequality follows from Theorem 3.1. Let $\mathbb{I}(E)$ denote the indicator function of an event E. Note that as $b \to 0$,

$$\begin{aligned}\mathsf{E}[N-n^*] &\geq b^{-1}\mathsf{E}\left[S_N^2\mathbb{I}(N>m)-1\right] + m\mathbb{P}(N=m) \\ &= b^{-1}\mathsf{E}\left[N^{-1}\left(\sum_{i=1}^N X_i^2 - N\right)\right] - b^{-1}\mathsf{E}[\bar{X}_N^2] + O(1) \\ &= \mathsf{E}\left[N^{-1}(b^{-1}-N)\left(\sum_{i=1}^N X_i^2 - N\right)\right] - b^{-1}\mathsf{E}[\bar{X}_N^2] + O(1),\end{aligned}$$

where the last equality comes from the fact that $\mathsf{E}\left[\sum_{i=1}^N X_i^2\right] = \mathsf{E}[N]$.

Equations (10) and (13) imply that $\{(bN)^{-r} : 0 < b \leq 1\}$ is uniformly integrable for all $r > 0$ and that $\left\{b^2T_N^4 : 0 < b \leq 1\right\}$ is uniformly integrable. Therefore, Cauchy-Schwarz's inequality implies that, as $b \to 0$,

$$\begin{aligned}b^{-1}\mathsf{E}[\bar{X}_N^2] &= \mathsf{E}\left[(bN)^{-2}(bT_N^2)\right] \\ &\leq \mathsf{E}^{1/2}\left[(bN)^{-4}\right]\mathsf{E}^{1/2}\left[(b^2T_N^4)\right] \\ &= O(1)\cdot O(1) = O(1).\end{aligned}$$

Next, it is easy to see from Eqs. (12) and (14) that $\left\{b(N-b^{-1})^2 : 0 < b \leq 1\right\}$ and $\left\{b\left(\sum_{i=1}^N X_i^2 - N\right)^2 : 0 < b \leq 1\right\}$ are both uniformly integrable. Since

$$b^{-1}N^{-2}(N-b^{-1})^2\mathbb{I}\left(N \geq \frac{1}{2}b^{-1}\right) \leq 4b(N-b^{-1})^2,$$

we know that $\left\{b^{-1}N^{-2}(N-b^{-1})^2\mathbb{I}\left(N \geq \frac{1}{2}b^{-1}\right) : 0 < b \leq 1\right\}$ is also uniformly integrable. Therefore, Eq. (11) shows that, as $b \to 0$,

$$
\begin{aligned}
&\mathsf{E}\left[b^{-1}N^{-2}(N-b^{-1})^2\right] \\
&\quad = \mathsf{E}\left[b^{-1}N^{-2}(N-b^{-1})^2\mathbb{I}\left(N \geq \frac{1}{2}b^{-1}\right)\right] + \mathsf{E}\left[b^{-1}N^{-2}(N-b^{-1})^2\mathbb{I}\left(N < \frac{1}{2}b^{-1}\right)\right] \\
&\quad \leq O(1) + m^{-2}b^{-3}\mathbb{P}\left(N < \frac{1}{2}b^{-1}\right) = O(1) + o(b^2)O(b^{-3})O(b^r) = O(1),
\end{aligned}
$$

for an appropriately chosen $r > 1$.

Finally, we apply Cauchy–Schwarz's inequality to obtain

$$
\begin{aligned}
&\mathsf{E}\left[N^{-1}(b^{-1}-N)\left(\sum_{i=1}^{N} X_i^2 - N\right)\right] \\
&\quad \leq \mathsf{E}^{1/2}\left[b^{-1}N^{-2}(N-b^{-1})^2\right] \cdot \mathsf{E}^{1/2}\left[b\left(\sum_{i=1}^{N} X_i^2 - N\right)^2\right] = O(1),
\end{aligned}
$$

as $b \to 0$. It follows that $\mathsf{E}[N - n^\star] = O(1)$ as $b \to 0$.

Proof of Theorem 3.3

Define $U_b = S_N^{-2}N - b^{-1}$, which is the overshoot in the sense of [23, 24]. By Taylor's theorem, we have

$$
U_b = N - b^{-1} - \left(\sum_{i=1}^{N} X_i^2 - N\right) + N\bar{X}_N^2 + \lambda_N^{-3}N(S_N^2 - 1)^2,
$$

where λ_N is a random variable such that $\min\left\{1, S_N^2\right\} \leq \lambda_N \leq \max\left\{1, S_N^2\right\}$. In the spirits of Chang and Hsiung [6] and [30],

$$
U_b \to U \text{ in distribution, as } b \to 0,
$$

where U is a random variable following the same distribution as the first hitting time of the random walk

$$
n - \left(\sum_{i=1}^{n} X_i^2 - n\right) = 2n - \sum_{i=1}^{n} X_i^2
$$

at the boundary $y = 0$, and $\mathsf{E}(U)$ exists. Now Theorem 3.2 follows from the Main Theorem of Chang and Hsiung [6].

Proof of Theorem 3.4

Observe that

$$b^{-1}\left[\mathsf{V}(\bar{X}_N) - b\right] = \mathsf{E}\left[b^{-1}\bar{X}_N^2\right] - 1 = \mathsf{E}\left[b^{-1}N^{-2}T_N^2 - bT_N^2\right] + \mathsf{E}[bT_N^2] - 1$$
$$= \mathsf{E}\left[T_N^2\left(b^{-1}N^{-2} - b\right)\right] + \mathsf{E}[bN] - 1,$$

where the last equality follows from Theorem 2 of [7]. Equation (9) implies the uniform integrability of $\{bN : 0 < b \le 1\}$. Also, it is easy to see from Eq. (3) that $bN \to 1$ with probability one as $b \to 0$. Therefore, $\mathsf{E}[bN] = 1 + o(1)$ as $b \to 0$. Now it remains to show that as $b \to 0$,

$$\mathsf{E}\left[T_N^2\left(b^{-1}N^{-2} - b\right)\right] = o(1). \tag{16}$$

To this end, we rewrite $\mathsf{E}\left[T_N^2\left(b^{-1}N^{-2} - b\right)\right]$ as follows:

$$\mathsf{E}\left[T_N^2\left(b^{-1}N^{-2} - b\right)\right] = \mathsf{E}\left[T_N^2(b^{-1}N^{-2} - bS_N^{-4})\right] + b\mathsf{E}\left[T_N^2(S_N^{-4} - 1)\right]. \tag{17}$$

For the first term on the right-hand side of Eq. (17), we have

$$T_N^2(b^{-1}N^{-2} - bS_N^{-4}) = -b(N^{-1}T_N^2)U(b^{-1}N^{-1} + S_N^{-2}),$$

where $b^{-1}N^{-1} + S_N^{-2} \to 2$ with probability one as $b \to 0$. Following the Lemma on page 830 of [30], we know that $N^{-1}T_N^2$ and U are asymptotically independent as $b \to 0$ and that $N^{-1}T_N^2$ is asymptotically distributed as the Chi-square distribution with one degree of freedom. In view of Eqs. (10), (13), and (2.8) of [30], we conclude that as $b \to 0$,

$$\mathsf{E}\left[T_N^2(b^{-1}N^{-2} - bS_N^{-4})\right] = -2b\mathsf{E}[U] + o(b) = O(b).$$

The second term on the right-hand side of Eq. (17) can be expressed as

$$\mathsf{E}\left[(N^{-1}T_N^2)(Nb)(S_N^{-2} - 1)(S_N^{-2} + 1)\right],$$

where $(Nb)(S_N^{-2} + 1) \to 2$ with probability one as $b \to 0$. Applying Taylor's theorem, we arrive at

$$\mathsf{E}\left[(N^{-1}T_N^2)(Nb)(S_N^{-2} - 1)(S_N^{-2} + 1)\right]$$
$$= \mathsf{E}\left[(N^{-1}T_N^2)\left\{-(S_N^2 - 1) + \kappa_N^{-3}(S_N^2 - 1)^2\right\}\{2 + o(1)\}\right],$$

as $b \to 0$, where κ_N is a random variable such that $\min\left\{1, S_N^2\right\} \le \kappa_N \le \max\left\{1, S_N^2\right\}$. By Hölder's inequality, Eq. (14), and the condition that $\mathsf{E}|X_1|^{4+\epsilon} < \infty$, we know that as $b \to 0$,

$$\mathsf{E}\left|(N^{-1}T_N^2)\left\{-(S_N^2 - 1)\right\}\right| \le \mathsf{E}^{1/2}\left[N^{-1}T_N^2\right]^2 \cdot \mathsf{E}^{1/2}\left[S_N^2 - 1\right]^2$$
$$= O(1)O(b^{1/2}) = O(b^{1/2}),$$

and

$$\mathsf{E}\left[(N^{-1}T_N^2)\left\{\kappa_N^{-3}(S_N^2-1)^2\right\}\right] \le \mathsf{E}^{\frac{2+\epsilon}{4+\epsilon}}\left[N^{-1}T_N^2\right]^{\frac{4+\epsilon}{2+\epsilon}} \cdot \mathsf{E}^{\frac{2}{4+\epsilon}}\left[\kappa_N^{-3}(S_N^2-1)^2\right]^{\frac{4+\epsilon}{2}}$$
$$= O(b^{(4+\epsilon)(2+\epsilon)/8}).$$

Therefore, Eq. (16) holds.

Acknowledgements We thank the Associate Editor and two anonymous reviewers for many useful comments and suggestions.

References

1. Anscombe FJ (1952) Large sample theory of sequential estimation. Math Proc Camb Philos Soc 48(4):600–607
2. Baron M, Xu R (2016) Sequential testing for full credibility. Variance 10(2):227–239
3. Berk R, Brown L, Buja A, Zhuang K, Zhao L (2013) Valid post-selection inference. Ann Stat 41:802–837
4. Brazauskas V, Kleefeld A (2011) Folded and log-folded-t distributions as models for insurance loss data. Scand Actuar J 1:59–74
5. Brazauskas V, Kleefeld A (2016) Modeling severity and measuring tail risk of Norwegian fire claims. N Am Actuar J 20(1):1–16
6. Chang IS, Hisung CA (1979) Approximations to the expected sample size of certain sequential procedures. In: Proceedings of the conference on recent developments in statistical methods and applications, Institution of Mathematics Academia Sinica, pp 71–82
7. Chow YS, Robbins H, Teicher H (1965) Moments of randomly stopped times. Ann Math Stat 36:789–799
8. Chow YS, Yu KF (1981) The performance of sequential procedure for the estimation of the mean. Ann Stat 9:184–189
9. Frees EW (2010) Regression modeling with actuarial and financial applications. Cambridge University Press, Cambridge
10. Frees EW, Derrig RA, Meyers G (2014) Predictive modeling applications in actuarial science, vol. I: predictive modeling techniques. Cambridge University Press, Cambridge
11. Gan G (2013) Application of data clustering and machine learning in variable annuity valuation. Insur Math Econ 53(3):795–801
12. Gan G, Valdez EA (2019) Data clustering with actuarial applications. N Am Actuar J. https://doi.org/10.1080/10920277.2019.1575242
13. Ghahari A, Newlands NK, Lyubchich V, Gel YR (2019) Deep learning at the interface of agricultural insurance risk and spatio-temporal uncertainty in weather extremes. N Am Actuar J 23(4):535–550
14. Ghosh BK, Sen PK (1990) Handbook of sequential analysis. Dekker, New York
15. Ghosh M, Mukhopadhyay N, Sen PK (1997) Sequential estimation. Wiley, New York
16. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York
17. Hong L, Kuffner T, Martin R (2018) On prediction of future insurance claims when the model is uncertain. Variance 12(1):90–99
18. Hong L, Martin R (2017) A flexible Bayesian nonparametric model for predicting future insurance claims. N Am Actuar J 21:228–241
19. Hong L, Martin R (2020) Model misspecification, Bayesian versus credibility estimation, and Gibbs posteriors. Scand Actuar J 2020(7):634–649
20. Hong L, Martin R (2021) Valid model-free prediction of future insurance claims. N Am Actuar J. https://doi.org/10.1080/10920277.2020.1802599 (**to appear**)
21. Kleijn BJK, van der Vaart AW (2012) The Bernstein-Von-Mises theorem under misspecification. Electron J Stat 6:354–381

22. Klugman SA, Panjer HH, Willmot GE (2008) Loss models: from data to decisions, 3rd edn. Wiley, Hoboken
23. Lai TL, Siegmund D (1977) A nonlinear renewal theory with applications to sequential analysis I. Ann Stat 5(5):946–954
24. Lai TL, Siegmund D (1979) A nonlinear renewal theory with applications to sequential analysis II. Ann Stat 7(1):60–76
25. Landsman Z, Makov UE (1999a) Sequential credibility evaluation for symmetric location claim distributions. Insur Math Econ 24(3):291–300
26. Landsman Z, Makov UE (1999b) On stochastic approximation and credibility. Scand Actuar J 1:15–31
27. Lee SCK, Lin XS (2018) Delta boosting machine with application to general insurance. N Am Actuar J 22(3):405–425
28. Liu K, Tan KS (2020) Real-time valuation of large variable annuity portfolios: a green mesh approach. N Am Actuar J. https://doi.org/10.1080/10920277.2019.1697707 (**to appear**)
29. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the lasso. Ann Stat 42(2):413–468
30. Martinsek AT (1983) Second order approximation to the risk of a sequential procedure. Ann Stat 11:827–836
31. Mukhapadhyay N, de Silva BM (2009) Sequential methods and their applications. CRC Press, Boca Raton
32. Psaraki S, Panaretoes J (1990) The folded-t distribution. Commun Stat Theory Methods 19(7):2717–2734
33. Risk J, Ludkovski M (2016) Statistical emulators for pricing and hedging longevity risk products. Insur Math Econ 68:45–60
34. Robbins H (1959) Sequential estimation of the mean of a normal population. In: Grenander U (ed) Probability and statistics, H. Crámer volume. Almquist and Wiksell, Uppsala, pp 235–245
35. Ross SM (1996) Stochastic processes, 2nd edn. Wiley, Hoboken
36. Schervish MJ (1995) Theory of statistics. Springer, New York
37. Sen PK (1981) Sequential nonparametrics: invariance principles and statistical inferences. Wiley, New York
38. Sen PK (1984) Nonparametric sequential estimation. Handb Stat 4:487–514
39. Starr N (1966) On the asymptotic efficiency of a sequential procedure for estimating the mean. Ann Math Stat 37:1173–1185
40. Starr N, Woodroofe M (1969) Remarks on sequential point estimation. Proc Natl Acad Sci 63:285–288
41. Stein C (1945) A two sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat 16:243–258
42. Stein C (1949) Some problems in sequential estimation. Econometrics 17:77–78
43. Wüthrich M (2018) Neural networks applied to chain-ladder reserving. Eur Actuar J 8:407–436
44. Yu KF (1978) Limit theorems for a class of sequential procedures. Doctoral Dissertation. Columbia University