SORT

Statistics and Operations Research Transactions

Volume
46

Number 1, January-June 2022

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 46, Number 1, January-June 2022

ISSN: 1696-2281
eISSN: 2013-8830

9 771696 228009 46001

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) -formerly *Qüestiió*- is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society and the Catalan Statistical Society. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
www.idescat.cat/sort/

ISSN: 1696-2281
eISSN: 2013-8830
SORT 46 (1) January-June (2022)

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística

Generalitat
de Catalunya
Institut d'Estadística
de Catalunya

SORT

Volume 46
Number 1
January-June 2022
ISSN: 1696-2281
eISSN: 2013-8830

Invited article

Articles

SORT 46 (1) January-June 2022, 3-38 DOI: 10.2436/20.8080.02.116

Fifty years later: new directions in Hawkes processes

John Worrall[1,2], Raiha Browning[1,2], Paul Wu[1,2] and Kerrie Mengersen[1,2]

Abstract

The Hawkes process is a self-exciting Poisson point process, characterised by a conditional intensity function. Since its introduction fifty years ago, it has been the subject of numerous research directions and continues to inspire new methodological and theoretical developments as well as new applications. This paper marks half a century of interest in Hawkes processes by presenting a snapshot of four state-of-the-art research directions, categorised as frequentist and Bayesian methods, other modelling approaches and notable theoretical developments. A particular focus is on nonparametric approaches, with advances in kernel estimation and computational efficiencies. A survey of real world applications is provided to illustrate the breadth of application of this remarkable approach.

MSC: *60G55, 62G05.*

Keywords: *Hawkes process, point process, nonparametric.*

1. Introduction

Events occur in the world with frequencies fluctuating over time and space, but often these events are not isolated and their occurrence increases the likelihood of further events. A mathematical model introduced by Hawkes (1971) describes the sequential arrival of these events as a non-Markovian process with a self-exciting nature. The Hawkes process (HP) has wide application in areas such as seismology (Ogata, 1981; Rasmussen, 2013); crime analysis (Yang et al., 2018; Zhuang and Mateu, 2019); traffic incidents (Kalair, Connaughton and Di Loro, 2021; Li, Cui and Chen, 2018); terrorism (Porter and White, 2010; White, Porter and Mazerolle, 2012); finance (Bacry, Mastromatteo and Muzy, 2015); infectious diseases (Kelly et al., 2019; Browning et al., 2021); and social media trends (Hall and Willett, 2016; Zhang, Walder and Rizoiu, 2020b).

[1] School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.

[2] QUT Centre for Data Science.

Received: May 2022.

In a HP, the self-exciting nature of the data is modelled through the conditional intensity function which governs the expected arrival rate of events. An important characteristic of this intensity function is the triggering kernel. There has been much research in learning these triggering kernels, including investigation of the underlying assumptions defined through simple parametric functions such as power laws, multiple exponential distributions, and Gaussian, Rayleigh and Weibull functions (Chen, Hawkes and Scalas, 2021; Chiang, Liu and Mohler, 2021). Further studies consider approaches to adapting these parametric models (Kobayashi and Lambiotte, 2016; Du et al., 2016), extending to the multidimensional setting and improving scalability of estimation by low-rank approximation (Zhou, Zha and Song, 2013; Bacry et al., 2020) and mean-field theory (Bacry et al., 2016a).

More flexible approaches to learning the kernel include representing the function as piecewise constant on a finite grid. Seminal work by (Lewis and Mohler (2011); Bacry, Dayri and Muzy (2012) provides a nonparametric framework for estimation that is also being actively explored. Bayesian nonparametric approaches are also emerging, with leading work in the area including (Donnet, Rivoirard and Rousseau, 2020; Zhou et al., 2020b; Zhang et al., 2020b). However, relaxing assumptions and increasing the expressiveness of functions comes at a cost: there is a requirement either for discretisation of the input domain or improved computational requirements to meet increasing practical demands.

These requirements have motivated new research into efficient algorithms for the analysis of HPs, and concomitant investigation of the characteristics of these algorithms. For example Achab et al. (2018) encodes causality of a multivariate process via a moment matching method fitting to second and third order cumulants. Work by Zhang et al. (2020b) takes advantage of latent branching structure and stationarity assumptions to reduce computational complexity and to efficiently infer a flexible representation of the kernel using Gaussian processes. In another very promising direction, Yang et al. (2017) focus on sequential (online) learning by approximating the function in a reproducing kernel Hilbert space. New Bayesian perspectives are lending themselves readily to handling the sheer volume and scalability in online learning (Broderick et al., 2013; Chérief-Abdellatif, Alquier and Khan, 2019; Markwick, 2020).

Another direction for Bayesian nonparametric approaches is in extending HPs to also cluster events via Dirichlet processes (Du et al., 2015). In these examples the form of the triggering kernel is generally parametric, and interest lies in the clustering of the events themselves.

Other recent directions of research into HPs arise from the perspective of graphs (Liu, Yan and Chen, 2018), stochastic differential equations (SDEs) (Lee, Lim and Ong, 2016; Kanazawa and Sornette, 2020) and neural networks (Zhang et al., 2020a; Du et al., 2016). These frameworks aim to provide more flexibility and less bias, while taking advantage of the techniques made available from a rapidly growing statistical data science community.

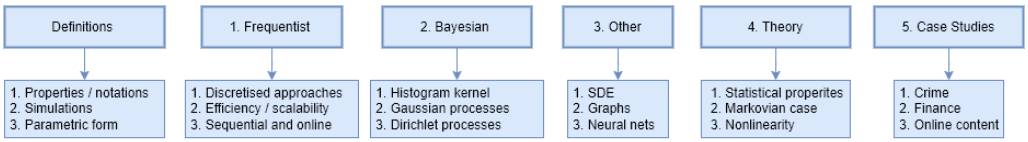


Figure 1. *Structure of paper.*

In addition, there has been considerable and significant research in theoretical properties and guarantees of HPs. Recent bodies of work include advances in estimating higher order statistical properties (Jovanović, Hertz and Rotter, 2015; Cui, Hawkes and Yi, 2020), asymptotic properties of the Markovian class of HPs (Gao and Zhu, 2018b; Zhu, 2015) and developments around nonlinear generalisation (Torrisi, 2016; Gao and Zhu, 2018a; Sulem, Rivoirard and Rousseau, 2021).

This paper aims to provide a review of these new directions in the modelling and analysis of HPs, with an emphasis on nonparametric Bayesian approaches and brief reference to the underpinning theory. We preface the review with a brief overview of notation, definitions and properties of HPs, and close the paper by presenting a survey of recent applications and some substantive applications in crime, finance and social media. The structure of the paper is illustrated in Fig 1.

1.1. Definitions and basic properties

This section provides a brief summary of mathematical definitions, properties and the general form of the HP which will be used throughout the remaining sections.

Definition 1.1 (Poisson process)**.**
A *nonhomogeneous Poisson process* with time varying arrival rate $\lambda(t)$ is defined as a counting process, $N(t) : t \geq 0$ which satisfies $t \in \mathbb{R}^+$, with associated history $\mathcal{H}_t : t \geq 0$, such that probability is given by

$$\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}_t) = \begin{cases} \lambda(t)h + o(h) & \text{m} = 1 \\ o(h) & \text{m} > 1 \\ 1 - \lambda(t)h + o(h) & \text{m} = 0. \end{cases} \tag{1}$$

Of particular interest in the study of nonhomogeneous Poisson processes is the HP $N(t)$ where $\lambda(t) : \mathbb{R}^+ \to \mathbb{R}^+$.

The time intervals between events (shown in Fig 2 as $t_1, t_2 \ldots, t_7$) are described as inter-arrival event times (Rasmussen, 2018). The point process can be characterised by the distribution function of the next arrival time conditioned on the past. Thus the conditional cumulative density function $F(t|\mathcal{H}_\mu)$ of next arrival time T_{k+1} can be expressed in terms of the conditional density function $f(s|\mathcal{H}_\mu)$,

$$F(t|\mathcal{H}_\mu) = \int_\mu^t \mathbb{P}(T_{k+1} \in [s, s+ds]|\mathcal{H}_\mu)\, ds = \int_\mu^t f(s|\mathcal{H}_\mu) ds.$$

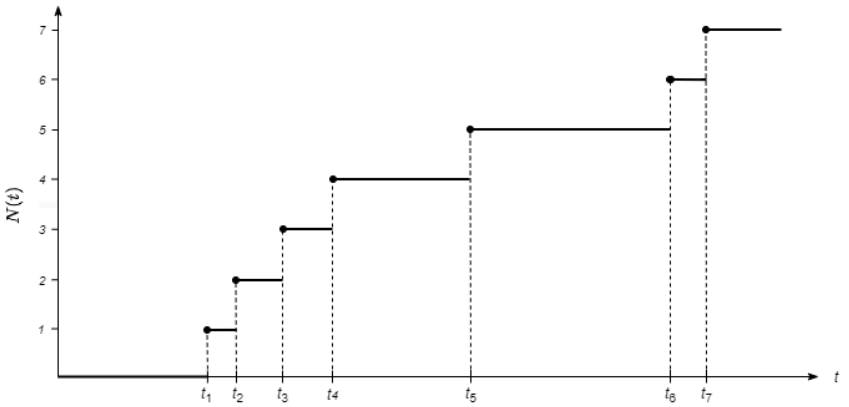


Figure 2. *Point process with stochastic realisation* $\{t_1, t_2 \ldots\}$ *and counting process N(t).*

where $\mathcal{H}_\mu$ is the history of the process until the last arrival (Ozaki, 1979). Where the conditional distribution is given using the law of total probabilities,

$$f(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} f(t_i | \mathcal{H}_\mu)).$$

Definition 1.2 (Conditional intensity function)**.**
Let the conditional density be $f(t|\mathcal{H}_{t_n})$ and the corresponding cumulative distribution function $F(t|\mathcal{H}_{t_n})$ for any $t > t_n$. Then $\lambda^*(t)$ is the conditional intensity or hazard function (Cox, 1955). The notation $*$ borrowed from (Daley and Vere-Jones, 2003) is used to represent conditioning on the history up to time t. A more intuitive definition of the conditional intensity function (Daley and Vere-Jones, 2003) is its expected rate of arrivals conditioned on the associated history,

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)} = \lim_{h \to 0} \frac{\mathbb{E}[N(t+h) - N(t) | \mathcal{H}_t]}{h}.$$

Hawkes (Hawkes, 1971) introduced a class of self-exciting process to model contagious processes, characterised by this conditional intensity function.

Definition 1.3 (HP)**.**
Let $D \in \mathbb{N}^+$ and $\{(t_i^j)\}_{j=1,\ldots,D}$ be a D-dimensional point process, with associated counting processes $N_t = (N_t^1, \ldots, N_t^D)$. A multidimensional Hawkes process (MHP) is defined with intensities $\lambda_i^*(t), i = 1, \ldots, D$ given by

$$\lambda_i^*(t) = \mu_i + \sum_{j=1}^{D} \int_0^t \phi_{ij}(t-s) dN_j(s) \tag{2}$$

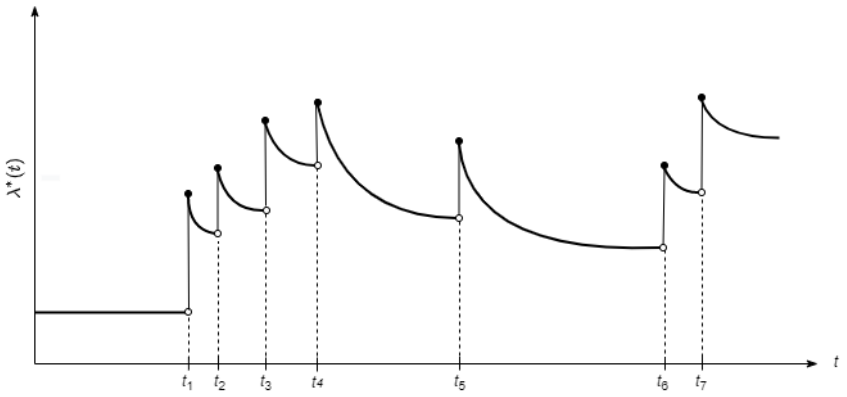


Figure 3. *Conditional intensity function of the HP, with exponential decay.*

where $\mu_i > 0$ is the non-negative background intensity of process i and $\phi_{ij}(\cdot) : (0,\infty) \to (0,\infty]$ is the excitation function from process j onto process i. When D=1, the univariate HP is expressed as

$$\lambda^*(t) = \mu + \int_0^t \phi\,(t-s)\,dN(s). \tag{3}$$

The self-excitation term within the expression of the HP is designed to capture the influences of all previous events in the current conditional intensity value. In multidimensional cases the self-exciting and mutually-exciting terms are, respectively, $\phi_{ii}(\cdot)$ and $\phi_{ij}(\cdot)$, $i \neq j$. A popular kernel choice is the exponential decay,

$$\phi_{ij}(t-s) = \left(\alpha_{ij} e^{-\beta_{ij}(t-s)}\right)_{i,j=1,\ldots,D} \tag{4}$$

where each arrival in the system instantaneously increases the arrival intensity by α_{ij} and over time the arrivals influence the decay at rate β_{ij} (Fig 3.)

The standard temporal HP can be extended to include spatial dependence, thereby capturing the clustering behaviours of the process through time and space. The process has an analogous description to the temporal HP.

Definition 1.4 (Spatio-temporal HP)**.**
Let $D \in \mathbb{N}^+$. Let $\{(t_i^j),(x_i^j),(y_i^j)\}_{j=1,\ldots,D}$ be a D-dimensional point process, with an associated counting process $N_t = (N_t^1,\ldots,N_t^D)$. A multidimensional spatio-temporal Hawkes process is defined with intensities $\lambda_i^*(\cdot), i = 1,\ldots,D$ given by

$$\lambda_i^*(t,x,y) = \mu_i + \sum_{j=1}^{D} \int_0^t \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi_{ij}(t-s,x-u,y-v) dN_j(s) dN_j(u) dN_j(v). \tag{5}$$

Further generalisations

In the past fifty years, there have been several popular types of generalisations of the conditional density. Three more common approaches are described in reference to the following univariate case equation,

$$g\big(\lambda^*(t)\big) = \mu(t) + \sum_{t>s} \phi(t-s, \xi_i) \tag{6}$$

1. Generalisations of the baseline process, $\mu(t)$, as a function of time effects on exogenous activity;
2. The Marked HP, where marks (ξ_i) associated to events (t_i) have different effects on intensity;
3. Nonlinear processes, where $g\big(\lambda^*(t)\big)$ is a nonlinear function with support in $\mathbb{R}^+$.

Regardless of the assumed background and triggering function forms, the fitness of the HP model is typically measured via the likelihood (Daley and Vere-Jones, 2003).

Definition 1.5 (Likelihood of HP)**.**
Let $N(\cdot)$ be a regular point process on $[0,T]$ for some finite positive T, and let $t_1,...,t_n$ denote a realisation of $N(\cdot)$ over $[0,T]$. Then, the likelihood function L is expressible in the form

$$L(t_1 \ldots, t_n \mid \mu, \phi) = \left[\prod_{i=1}^{n} \lambda^*(t_i)\right] \exp\left(-\int_0^T \lambda^*(u)\, du\right). \tag{7}$$

A condition in ensuring the estimated model is stable and has access to most properties of the HP is stationarity.

Definition 1.6 (Stationarity of HP)**.**
Let $N(\cdot)$ be a multivariate HP on $[0,T]$ for some finite positive T, where $N(\cdot)$ is stationary if a translation in time does not change its distribution. Let Φ be a $D \times D$ matrix with entries given by,

$$\Phi_{ij} = \int_0^\infty \phi_{ij}(u) du.$$

A sufficient condition for stationarity is that $\rho(\Phi) < 1$, where $\rho(\Phi)$ is spectral radius of Φ given as

$$\rho(\Phi) = \max_{x \in \mathcal{S}(\Phi)} |x| \tag{8}$$

where $\mathcal{S}(\Phi)$ is a set of all eigenvalues of Φ.

A number of simulation procedures are available for ensuring stationarity and other stochastic properties of the HP. The generation of synthetic data sets from these methods ensures statistical equivalence to the real population of interest and is an invaluable tool in supporting model design and development.

1.2. Simulating a HP

Concerning the experimental aspects of a self-exciting process, two synthetic generation algorithms are popular.

The first of these is the thinning method, a standard approach to producing nonhomogeneous Poisson processes. The intuition of the algorithm is to combine two generated homogeneous Poisson processes of different rates and to remove points probabilistically, so the remaining points satisfy a time-varying intensity $\lambda(\cdot)$. For the Ogata modified algorithm Ogata (1981), the intensity has no asymptotic upper bound, although it is common to set non-increasing periods without any arrival.

Simulation of a HP may also be represented as an immigration-birth process, leading to a branching simulation procedure (Fig 4). Here immigrants are generated via a homogeneous Poisson rate λ, conditioned on K immigrants with arrival times uniformly i.i.d in time window $(0,T]$. Each immigrant descendant forms a nonhomogeneous Poisson process with intensity (α/β) giving arrival times $[I_i+E_1, I_i+E_2, \ldots, I_i+E_{D_i}]$.

Algorithm 1 Simulating univariate HP by thinning

Require: $(\lambda^*(\cdot), T)$
  Initialisation $P \leftarrow []$, $t \leftarrow 0$, $\varepsilon \leftarrow 10^{-10}$
  while $t < T$ **do**
    Set upper bound $M \leftarrow \lambda^*(t+\varepsilon)$
    Generate candidate point $E \leftarrow$ Exp(M)
    $t \leftarrow t+\varepsilon$
    Set with probability $U \sim$ Unif(0,1)
    if $t < T$ and $U \leq \lambda^*(t)$ **then**
      $P \leftarrow [P,t]$
    end if
  end while
  return P

1.3. Parametric models for HPs

There is a rich literature on parametric methods for modelling HPs. These approaches have many uses, particularly when the parametric form of the process is obvious. They are often simple to implement and can provide useful insights into the behaviour of these processes. A brief summary of parametric methods is provided here, covering some of the most popular forms for the triggering kernel and inference techniques for these models.

Algorithm 2 Simulating univariate HP by clusters

Require: $(T, \lambda, \alpha, \beta)$
  Initialisation $P \leftarrow [], i \leftarrow 1$
  Generate immigrants $K \sim \text{Pois}(\lambda T)$
  $I_1, I_2, .., I_K \sim \text{Unif}(0, T)$
  Generate descendants $D_1, D_2, .., D_K \sim \text{Pois}(\alpha/\beta)$
  while $i < K$ **do**
    if $D_i > 0$ **then**
      $E_1, E_2, .., E_{D_i} \sim \text{Exp}(\beta)$
      $P \leftarrow P \cup [I_i + E_1, I_i + E_2, .., I_i + E_{D_i}]$
    end if
    Set i = i + 1
  end while
  Remove invalid descendants (0,T] $P \leftarrow (P_i : P_i \in P, P_i \leq T)$
  Add immigrants $P \leftarrow \text{Sort}(P \cup [I_1, I_2, .., I_n])$
  return P

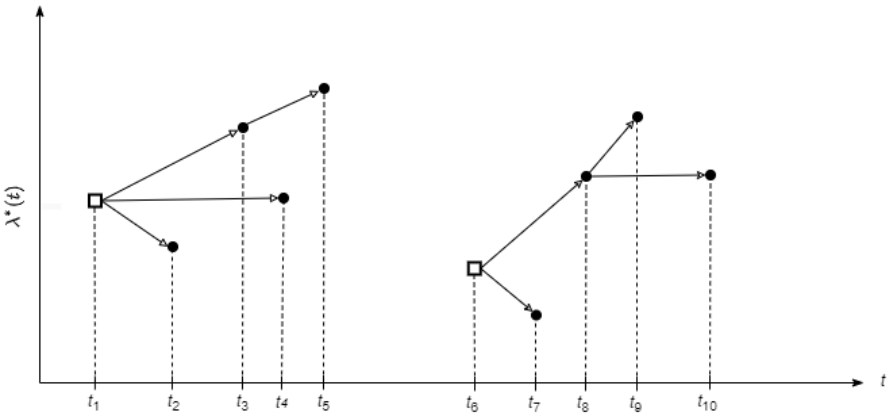


Figure 4. *HP immigrant-birth representation (squares indicate immigrants and circles indicate offspring/descendants).*

1.3.1. Choice of triggering kernel

Although the structure of the conditional intensity function is quite flexible, the most common triggering kernel is parameterised as an exponential decay

$$\phi(t-s) = \alpha e^{-\beta(t-s)}.$$

Here α represents the overall strength of excitation and β denotes the influence decay rate of the arrivals. Hawkes (1971) used this parametric form to derive theoretical properties of the covariance density function and Bartlett spectrum, via the frequency

domain. The Laplace transform is given as

$$\mathcal{L}\{\cdot\}(s) = \frac{\alpha\mu(2\beta-\alpha)}{2(\beta-\alpha)(s+\beta-\alpha)} \tag{9}$$

where $s \in \mathbb{C}$. Evaluating the power spectral density (defined in terms of the covariance density) of a HP provides a set of useful tools in discriminating and fitting between models and access to other techniques from the spectral theory field.

In addition, the exponential decay has several other advantageous properties, such as straightforward computation of the expected value of an arbitrary function on $N(t)$, direct simulation, and efficient computation of the likelihood. Most of these properties descend from the Markov property, where the intensity and the pair $(\lambda(t), N(t))$ are Markovian, in the following form,

$$d\lambda(t) = -\beta\lambda(t)dt + \alpha\beta dN(t) \tag{10}$$

Several other parametric kernel forms have also become popular. These include the power law, sinusoidal, Gaussian and rectangular functions supporting different types of interactions among events. In almost all realistic applications, however, it is not obvious which parametric form of the excitation function for HPs is the most appropriate. This has generated a great deal of recent interest in nonparametric specification of the kernel function. Under this representation, traditional assumptions about the triggering kernel can be relaxed to capture the complexities and subtleties of the excitation effects retrieved from the data. Before moving to a more comprehensive discussion of nonparametric directions in Sections 2 and 3, we complete the introduction to HPs with an overview of spatio-temporal approaches and matters of inference.

1.3.2. Spatio-temporal approaches

A number of authors propose spatio-temporal self-exciting processes. Generally, the triggering kernel is constructed in a separable fashion, where the temporal and spatial dependence can be decomposed (Mohler et al., 2011; Schoenberg, 2016; Reinhart, 2018). A popular parameterisation for the respective kernels is exponential decay in time and Gaussian decay in space. Several Bayesian approaches have also been introduced to model spatio-temporal HPs. These include Holbrook, Ji and Suchard (2022), who model the outbreak of Ebola in West Africa and extend the standard spatio-temporal Hawkes model to learn the evolutionary history of the virus which informs the characteristics for each variant of the virus. Holbrook et al. (2020) also account for uncertainty in the location of events by placing a prior on the spatial position of events.

A popular case of the spatio-temporal HP, originally introduced as a marked, purely temporal process, is an adaption of the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata, 1988) to incorporate spatial dynamics (Ogata, 1998). The spatial ETAS model was introduced in the context of modelling earthquakes through the baseline parameter, and their corresponding aftershocks represented by the triggering kernel. The

marks are denoted by m and correspond to the magnitude of the earthquake. Thus, the intensity function can be written in the form,

$$\lambda_m^*(t) = \mu(x,y) + \sum_{t>s} \phi_m(t-s, x-u, y-v) \tag{11}$$

$$\phi_m(t,x,y) = \kappa_m \times \frac{(p-1)c^{p-1}}{(t+c)^p} \times \left(\frac{1}{\pi\sigma_m} \cdot f(\frac{x^2+y^2}{\sigma_m}) \right), \tag{12}$$

where κ_m is the expected number of aftershocks for an earthquake of magnitude m, σ_m is a scale factor, f is a function such that $\int_0^\infty f(x)dx = 1$ holds, and p and c are global constants. The second and third terms of ϕ represent the temporal and spatial decay functions respectively.

1.4. Inference

A range of inference approaches have been used for estimating the parameters of these parametric models. A common procedure is that of maximum likelihood estimation, where the likelihood function is maximised to obtain the set of parameter values that produce the highest likelihood.

Other approaches are based on the branching representation of the HP which allows the likelihood to be decomposed into conditionally independent immigrant and offspring processes. Due to this latent structure, inference methods such as Expectation-Maximisation (EM) and Variational Inference (VI) can be used to integrate over this latent space. A detailed explanation of the EM algorithm for HPs is provided in Laub, Lee and Taimre (2021) and a similar construction is used when performing VI for these models. These inference techniques that utilise the latent structure of HPs are discussed further in the subsequent sections of this review. Efficient Gibbs samplers have also been developed, using the decomposition of the likelihood and placing conjugate priors on the parameters of the model.

We turn now to four general directions of research that illustrate current activity in HPs. These include frequentist nonparametric kernel adaptation and presentation, Bayesian nonparametric approaches, other approaches (stochastic differential equations, graphs and neural networks) and theoretical aspects of HPs. This is intended to be a canvas rather than an exhaustive review of all research directions.

2. Direction 1: Frequentist nonparametric kernel adaptation and estimation

There is now a large literature on various directions of research into nonparametric kernels for HPs. The following discussion focuses on a selection of these directions, based on their novelty, currency and interest to the authors. The focus is initially on several frequentist approaches, followed by efficient estimation methods and finally sequential or online models.

2.1. Discretised scheme

A frequentist approach to estimating the excitation and/or baseline function is defined by approximating the function as a binned grid, where function values are piecewise constant within each bin and the width of each bin is selected optimally to model local variations of the excitation.

2.1.1. Stochastic declustering

Early work by Zhuang, Ogata and Vere-Jones (2002) supports this approach by attempting to differentiate between 'true' background events and triggered events. Such differentiation using the probability for background events, p_{ii}, is called stochastic declustering. Motivated by this work, Model Independent Stochastic Declustering (MISD) was introduced as a nonparametric HP with homogeneous background rate (Marsan and Lengliné, 2008) and later extended for the more general case of varying $\mu(t)$ (Lewis and Mohler, 2011). This method makes use of the branching structure to reduce both baseline and triggering kernel into a density estimation problem. The augmented likelihood of observations D and branching structure B with two independent components is then given by

$$p(D,B|u(t),\phi(\tau)) = \underbrace{\prod_{i=1}^{N} u(t_i)^{b_{ii}} \exp(-uT)}_{u(t)} \cdot \underbrace{\prod_{i=2}^{N}\prod_{j=1}^{i-1} \phi(t_i - t_j)^{b_{ij}} \prod_{i=1}^{N} \exp\left(-\int_0^T \phi(\tau)d\tau\right)}_{\phi(t)} \tag{13}$$

The recovered parameters are then updated via an expectation step, where b_{ij} is replaced by the expectation $\mathbb{E}[b_{ij}] = p_{ij}$, representing the probability that event i is caused by event j

$$p_{ij}^k = \frac{\phi^k(t_i - t_j)}{u^k + \sum_{j=1}^{i-1} \phi^k(t_i - t_j)}, \quad p_{ii}^k = \frac{u^k}{u^k + \sum_{j=1}^{i-1} \phi^k(t_i - t_j)}. \tag{14}$$

This allows for the construction of a matrix P^k, giving events caused by the background rate (diagonal elements) or another event (non-diagonal elements). The maximisation step then updates parameters given the current matrix of probabilities such that,

$$u^{k+1} = \frac{1}{T}\sum_{j=1}^{n} p_{ii}^k, \qquad \phi_m{}^{k+1} = \frac{1}{\delta t}\sum_{i>j\in A_m} p_{ij}^k \tag{15}$$

where δt is a discretisation parameter controlling the bin grid and A_m is the set of pairs of events.

In examining the MISD model, we illustrate in Fig 5 a synthetic exponential kernel (red) compared with the estimated kernel (blue) from the MISD model with varying discretisation parameters.

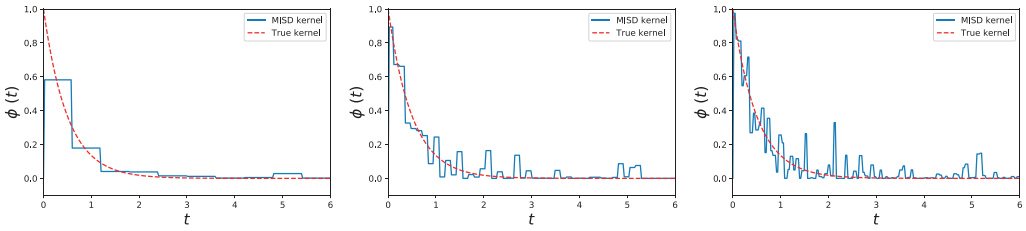


Figure 5. *Kernel estimate (blue) on synthetic exponential kernel, increased bin size 10,50,100 (left to right).*

Results highlight empirically the sensitivity of the chosen discretisation parameter δ to the structure of the kernel ϕ. Incorrect choice of the number of bins leads to underfitting (left) and overfitting (right). This motivates future work that improves on bandwidth choice and boundary effects, which are unavoidable topics of kernel estimation.

Another approach to defining the excitation function on a grid or set of grids is through exploiting relations in the model in the frequency domain between second order statistics and the triggering kernel.

2.1.2. Wiener-Hopf integral

Bacry and Muzy (2016) showed that the kernel matrix of a MHP can be estimated by relating the jump correlation matrix of event processes to a series of Wiener-Hopf equations. This relationship between the first and second order characterisation properties, triggering kernel and background rate of a HP is exploited in the frequency domain to satisfy a unique causal solution. Given this unique solution, the unknown kernel may be solved by a discretised system of linear equations via quadrature and inversion. The triggering kernel matrix function and conditional expectation $g(t)$ satisfy the following Wiener-Hopf equation,

$$g(t) \ = \Phi(t) + \Phi(t) * g(t) \ , \forall t > 0 \tag{16}$$

where $*$ represents convolution (Bacry et al., 2015).

The numerical approximation requires selecting a grid and quadrature scheme for $g(t)$, computing first the estimated $\widetilde{g}$ (Jovanović et al., 2015). Considering the univariate case, where N_t jumps are all size 1 and stationary, the first order property (mean event rate) is

$$\Lambda dt \ = \mathbb{E}(dN_t) = \frac{\mu}{1 - \int \phi \ (\tau) d\tau} dt$$

with second order statistic are summed up by infinitesimal covariances,

$$\mathrm{Cov}\,(dN_{t1}, dN_{t2}) = \mathbb{E}\,(dN_{t1}\ dN_{t2}) - \mathbb{E}\,(dN_{t1})\,\mathbb{E}\,(dN_{t2})$$

under assumption N_t has stationary increments, $\mathrm{Cov}\,(dN_{t1}, dN_{t2})$ only depends on $\tau = t_2 - t_1$ this part of this covariance is can be written as

$$v(\tau)d\tau = \mathbb{E}\,(dN_0\, dN_\tau) - \mathbb{E}\,(dN_0)\ \mathbb{E}\,(dN_\tau)\,.$$

where the second order statistic can be rewritten in terms of conditional expectations,

$$g(\tau)dt\ = v\,(\tau)\,d\tau\ /\Lambda\ = \mathbb{E}\,(dN_\tau | dN_0 = 1)\ -\ \Lambda d\tau\ .$$

details of proof in Bacry and Muzy (2016).

Approximation of the equation is commonly given via the Gaussian quadrature method for discretised Wiener-Hopf systems on the interval $[t_{\min},\ t_{\max}]$ and is shown as

$$\widetilde{g}_{ij}(t_n) = \widetilde{\phi}_{ij}(t_n) + \sum_{l=1}^{D}\sum_{k=1}^{K} w_k\ \widetilde{g}_{il}\,(t_n - t_k) w_k \widetilde{\phi}_{ij}(t_n)\ ,\ \forall n \in [0,K], i,j \in [0,D]. \tag{17}$$

Inverting the obtained linear systems results in estimation of the matrix kernel at quadrature points $\widetilde{\phi}_{ij}$ and lastly estimating μ using the first order cumulant.

In the example below we again consider a simulated exponential decay kernel and approximate Wiener-Hopf equation with optimal bandwidth given by the MSE with respect to grid size.

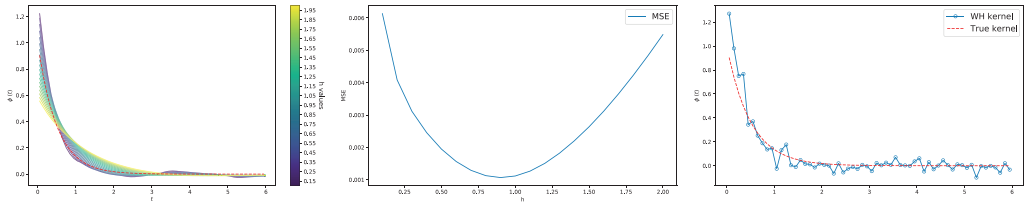


Figure 6. *Kernel estimates with actual in red (left) and mean square error (MSE, middle) with varying width(h). Optimum kernel estimate (right).*

The more expressive bin grid approach compared to parametric methods requires a larger sample size and is restricted to non-Markovian regimes, thus a larger computational cost. This has led to a body of work focusing on computational efficiencies.

2.2. Improved estimation scalability and efficiency

Achab et al. (2018) decreases computational costs by replacing estimation of kernels through matching cumulants (or moments). This strategy relates the branching structure of an MHP to Granger causality, estimating cumulative values to quantify the causal relationship among each node by estimating the matrix,

$$\int \Phi(t)\ dt = \int_0^\infty \phi_{ij}(t)\ dt\ \geq 0\ \text{ for } 1 \leq\ i,j \leq d. \tag{18}$$

It first computes from sequences moments, up to the third estimates, $\hat{\mathcal{M}}$ and minimises the L^2 error between these estimates and actual moments $\mathcal{M}^{\text{true}}$ (uniquely determined from $||\Phi(t)||$) where

$$||\hat{\Phi}(t)|| = \arg\min_{||\Phi(t)||} ||M(||\Phi(t)||) - \hat{M}||^2, \tag{19}$$

where the matrix $\mathcal{R}$ is given as

$$\mathcal{R} = (\mathbb{I}^d - \hat{\Phi}())^{-1}.$$

The explicit relationships between the matrix and cumulants are then defined as the following identities Λ, C, K. Estimation is given from general formulae for the integral of cumulants of an MHP in Jovanović et al. (2015), where 3rd order statistics are connected to skewness of N_t (Achab et al., 2018), shown as

$$\begin{aligned} K^{ijk}dt = \mathbb{E}\left(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right) \\ - dt\ \Lambda^i \mathbb{E}\left((\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right) \end{aligned} \tag{20}$$

where $\Delta_H N_t^i = N_{t+H}^i - N_{t-H}^i$ with first and second moments given as

$$\begin{aligned} \Lambda^i dt &= \mathbb{E}(dN_t^i) = \lim_{n\to\infty} \frac{1}{T}\ \lambda_i{}^*(t) = (\mathbb{I} - ||\Phi||)^{-1}\mu_i \\ C^{ij}dt &= \mathbb{E}(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)) \end{aligned}$$

and with $\hat{\Lambda}, \hat{C}\hat{K}$ to be incorporated in the estimator $\hat{\mathcal{R}} = \arg\min_{\mathcal{R}} L(\mathcal{R}))$ such that

$$L(\mathcal{R}) = (1-k)||K^c(\mathcal{R} - \hat{K}^c)||_2^2 + k||C(\mathcal{R}) - \hat{C}||_2^2,$$

where K^c is the tensor contraction of tensor K, and the coefficient k is used to scale the two terms

$$k = \frac{||\hat{K}^c||_2^2}{||\hat{K}^c||_2^2 + ||\hat{C}||_2^2}.$$

Inverting (19) leads to the recovered matrix

$$\hat{\Phi}(t) = (\mathbb{I}^d - \hat{\mathcal{R}})^{-1}.$$

In the univariate case, the $\hat{\Phi}$ can be estimated from the second order statistics, whereas in higher dimensions the third order or skewness is required for unique $\hat{\Phi}(t)$.

The nonparametric cumulant method outperforms the previously discussed MISD and Wiener-Hopf algorithms, given its reduced complexity. The recovered matrix also provides a quantifiable degree of endogeneity in a system and the causality structure of a network.

Another approach to improving the computational bin grid process is by updating parameters in a single pass.

2.3. Sequential and online approaches

Yang et al. (2017) proposed an online procedure where the triggering function belongs to a Reproducing Kernel Hilbert Space (RKHS). Assume that there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that there is a positive definite kernel,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

where $n \in \mathbb{N}, c \in \mathbb{R}$ and that for any $x \in \mathcal{X}$, the evaluation functional is bounded as

$$f(x) = \langle f, K(x,\cdot)\rangle_H \leq C||f||_H$$

for some constant C. Suppose that $f(x)$ satisfies the decreasing tail property with tail function $\varepsilon_f(\cdot)$ if

$$\sum_{k=m}^{\infty}(t_k - t_{k-1}) \sup_{x\in(t_k,t_{k-1}]} |f(x)| \leq \varepsilon_f(t_{m-1}), \forall m > 0, \tag{21}$$

where $\varepsilon_f(\cdot)$ is a bounded and continuous function such that $\lim_{t\to\infty}\varepsilon_f(t) = 0$. Then the assumed triggering function belongs to a RKHS where similarities among high-dimensional and complex distributions are mapped onto lower-dimensional ones. The process then takes the log-likelihood function and optimises over a discretised version,

$$\begin{aligned} L_i(\lambda) &= \sum_{d=1}^{D}\sum_{k=1}^{M(t)}\int_{\chi_{k-1}}^{\chi_k} \lambda_d(s)ds - y_{d,k}\, log\lambda_d(t_k) \\ &= \sum_{d=1}^{D}\Delta L_{d,i}(\lambda_d) \end{aligned}$$

where partitioning $\{0, \chi_1 ... \chi_{M(t)}\}$ on the interval $[0, T]$ is defined as

$$\chi_{k+1} = \min_{t_i > \chi_k}\{\iota * \lfloor \chi_k/\iota \rfloor + \iota, t_i\}$$

for some small $\iota > 0$. The discretised version is then expressed as

$$\begin{aligned} L(\lambda) &= \sum_{d=1}^{D}\sum_{k=1}^{M(t)}\int_{\chi_{k-1}}^{\chi_k} (\chi_k - \chi_{k-1}) - \lambda_d(\chi_k) - y_{d,k}\log(\lambda_i(\chi_k)) \\ &= \sum_{d=1}^{D}\Delta L_{d,t}(\lambda_d). \end{aligned} \tag{22}$$

To perform fast evaluation, the optimisation algorithm processes each partition and employs the following three properties. The first is a truncation of the intensity function that considers arrivals within a recent window $[t - z, t)$ as

$$\lambda_i^z(t) = \mu_i \sum_{i=1}^{p}\int_0^t \mathbb{1}\{t - \tau < z\} f_{ij}(t - \tau) dN_j(\tau). \tag{23}$$

The second is Tikhonov regularisation over the baseline and triggering kernel, adding weight terms to the loss function and keeping the resultant values small. The third is enforcing positivity through the projection steps in the optimised triggering function part. By reducing complexity through the RKHS and by exploiting inter-arrival MHP properties, the resulting nonparametric algorithm recovers estimates over a single pass with comparable computation cost of alternative parametric online learning algorithms.

2.4. Summary

Nonparametric HPs comprise a major direction of current research. Notwithstanding the advantages of a nonparametric framework, such as the enablement of a more expressive triggering function, the approach induces a number of challenges. Firstly, the discretisation approach employed in fitting the nonparametric model requires a larger sample size compared to more traditional parametric methods that fit better on shorter and fewer arrival sequences. However, they may underfit on longer sequences. This is easily seen when relating the bin division grid concept to a histogram of inter-arrival times. Secondly, computational loads for estimation and inference become much larger, given the above-mentioned sample size requirements and as the binning process is a sequential process that cannot take account of the Markovian property given in the exponential function.

Several research directions to address these challenges have been discussed. The first is an improved computational estimation method in matching cumulants. The second is a reduction in computational complexity though some assumptions on the kernel (belonging to RKHS) to estimate parameters given a single pass on some discretised time domain.

3. Direction 2: Bayesian nonparametric approaches

Bayesian approaches inherit the usual benefits of more flexible hierarchical modelling and probabilistic inference. A number of Bayesian nonparametric approaches to modelling HPs have been proposed in the literature. In particular, the majority of methods discussed in this section estimate either the baseline rate, triggering kernel, or both, using either a nonparametric histogram kernel or Gaussian processes. Also of interest in the Bayesian nonparametric literature for HP is using the self-exciting properties of HPs to determine the clustering of events using Dirichlet processes.

3.1. Histogram kernel

A drawback of the binning based methods for estimating a histogram function discussed in Section 2 is that they require *a priori* selection of the grid size. This often leads to models that are either overfitted or underfitted. Donnet et al. (2020) propose a Bayesian nonparametric approach for modelling HP that eases this choice.

The authors derive posterior concentration rates for HPs, and exemplify these results through a nonparametric histogram representation of the triggering kernel. This form

of kernel is motivated in the neuroscience context, mimicking the behaviour of action potentials to model the interaction between neurons in the brain. The resulting histogram kernel defined on the compact set $(0,A)$ with J components, change points at $s = (s_0 = 0, s_1, ..., s_{J-1}, s_J = A)$ and respective heights $w = (w_1, ..., w_J)$ such that $\sum_{j=1}^{J} w_j = 1$, has the form,

$$\phi(t|J,w,s) = \delta \sum_{j=1}^{J} \frac{w_j}{s_j - s_{j-1}} \mathbb{1}_{t\in(s_{j-1},s_j)}. \tag{24}$$

where $\delta \sim \text{Bern}(p)$ is an indicator variable that determines whether the histogram function is active with probability p, or whether the heights for all components is zero. The parameters of this model are inferred using Reversible-jump Markov chain Monte Carlo (RJMCMC), which proves to be a costly procedure. RJMCMC (Green, 1995) is a trans-dimensional approach to Bayesian inference that allows the model to move between different parameter spaces. In this example the various parameter spaces are determined by the possible values of J. A drawback of RJMCMC, as found in this study, is that it is computationally expensive and experiences slower mixing that more standard approaches.

3.2. Gaussian processes

To circumvent the issue of slow inference, several authors have proposed efficient algorithms by instead estimating the model parameters as flexible functions using Gaussian processes. A common feature in all of these approaches is the augmentation of the branching structure to decompose the likelihood function into conditionally independent processes.

Zhang et al. (2019) suggest a practical direction for improvement by proposing a flexible triggering kernel represented as a quadratic transformation of a Gaussian process $f(\cdot)$ given by,

$$\phi(t) = \frac{1}{2} f^2(t). \tag{25}$$

This form is selected as it has certain analytical advantages. With a conjugate Gamma prior on the baseline parameter μ, the adapted Laplace method (Walder and Bishop, 2017) approximates the posterior conditioning on the branching structure, resulting in scalable estimates of theoretical linear time complexity, $O(n)$. An approximated sampling structure (Halpin, 2013) is used to reduce computation by considering only high-probability triggering relationships; this is achieved by assuming that the probability for extremely unlikely triggering relationships is very close to zero. The model is estimated through an EM implementation of both a block Gibbs sampler and MAP estimator. The approach maintains conjugacy relationships due to the decomposition of the likelihood to facilitate a closed form in computation of sequential updates to the model.

In a similar style, Zhou et al. (2021) uses Gaussian processes to represent both the baseline rate and triggering kernel. They also perform a quadratic transformation of a

Gaussian process and further employ a sparse GP approximation (Titsias, 2009) to reduce complexity and avoid costly optimisation procedures. These authors also decouple the baseline rate and triggering kernel by augmenting the branching structure, thereby introducing a fast EM style mean-field variational Bayes algorithm.

Zhou et al. (2020a) instead adopt a sigmoid transformation of a Gaussian process for the baseline rate and triggering kernel, again using a sparse GP approximation. These functions have the form,

$$\mu(t) = \lambda_\mu^* \sigma(f_1(t)), \quad \phi(t) = \lambda_\phi^* \sigma(f_2(t)) \tag{26}$$

where λ_μ^* and λ_ϕ^* are upper bounds of μ and ϕ respectively, $\sigma(\cdot)$ is the sigmoid function and $f_1(\cdot)$ and $f_2(\cdot)$ are generated from a Gaussian process. In addition to augmenting the branching structure, several other processes are introduced to allow for conjugate inference. Several efficient inference schemes are also proposed, namely a Gibbs sampler, an EM algorithm and a mean-field variational inference algorithm. In this experiment all three algorithms performed comparatively well. In this work the sigmoid function is defined as a Gaussian representation, with Polya-Gamma augmentation (Polson, Scott and Windle, 2013),

$$\sigma(z) = \frac{e^{z/2}}{2\,\cosh(z/2)} = \int_0^\infty e^{h(\omega,z)} p_{PG}(\omega|b,0)\, d\omega \tag{27}$$

where $h(\omega,z) = z/2 - z^2\omega/2 - \log 2$ and PG is the Polya-Gamma distribution.

Data augmentation provides a mechanism that eliminates the need to evaluate the high dimensional integral, allowing for efficient conjugate inference. This augmentation strategy with the polya-gamma technique is an interesting development as the likelihood becomes conjugate for the GP prior, thereby allowing for speed compared to other augmentation techniques, and it is an effective method for posterior inference. Malemshinitski, Ojeda and Opper (2022) extend the process further by allowing nonlinear and inhibitory effects in the kernel, ensuring that the intensity is non-negative via a sigmoidal link function. This approach is also computationally efficient, given the new likelihood form and mean-field variational inference; however, it does not rely on the commonly used branching structure.

3.3. Dirichlet approaches

Yet another popular direction Bayesian nonparametric modelling is in incorporating the dynamics of HPs into Dirichlet processes to inform event clustering. This enables capture of the diversity of event types, while the self-exciting process describes the temporal dynamics. The framework of the Dirichlet process means that the number of clusters grows as the complexity of the data increases.

An example of this is the Dirichlet-Hawkes (DHP) model proposed by Du et al. (2015). The authors cluster streams of data, such as news articles and social media content, using a Dirichlet process augmented with a temporal HP to determine the intensity

of arrivals. The overarching idea of this work is to determine related actions of media platforms given a particular occurrence of a highly impactful event, through word content and time of occurrence.

The model is a generalisation of the Dirichlet process. Generally, the probability of joining an existing cluster or a new cluster is proportional to the number of observations currently in each cluster or the concentration parameter respectively. The authors instead specify these probabilities as counts that are temporally weighted by the triggering kernel $\phi_{\theta_k}(t-t_i)$ for each existing cluster, or the baseline rate μ for new customers. Hence the baseline rate acts as the concentration parameter in the Dirichlet process.

Let θ_k be the parameters of the bag-of-words document content model for the kth cluster and w_n^v be the vth word in the nth document. Then the model is given by,

$$\begin{aligned} w_n^v &\sim \text{Multinomial}(\theta_k) \\ \theta_k &\sim DHP(\mu, G_0) \\ G_0 &\sim \text{Dirichlet}(\theta_0) \end{aligned}$$

where θ_0 is the concentration parameter for the base distribution in the Dirichlet process.

The choice of algorithm for parameter inference is motivated by the streaming context. A Sequential Monte Carlo framework is used which allows the authors to reuse previous samples. When necessary, duplicate timestamps are resampled as this is a violation of the assumptions of a point process. A Gibbs sampler similar to Neal (2000) is embedded within this framework to sample the cluster labels in the following way. For event at time t_n with cluster allocation s_n,

- Remove t_n from cluster s_n.
- Calculate the probability of t_n belonging to cluster j,

$$p(s_n = j | t_n, \text{rest}) = \begin{cases} \frac{\phi_{\theta_k}(t_n - t_i)}{\mu + \sum_{t_n > t_i} \phi_{\theta_k}(t_n - t_i)} & \text{if } j \text{ occupied,} \\ \frac{\mu}{\mu + \sum_{t_n > t_i} \phi_{\theta_k}(t_n - t_i)} & \text{otherwise.} \end{cases} \tag{28}$$

- Sample cluster allocation for t_n from (28). If j is unoccupied draw θ_j from G_0.

Blundell, Heller and Beck (2012) present another extension on Dirichlet processes for HP. The authors combine HPs with the infinite relational model (IRM) (Xu et al., 2006; Kemp et al., 2006), a graph based approach to modelling the relationship between entities given previously declared relationships. In this model events are represented as vertices on a graph and are clustered according to a Chinese restaurant process (CRP) (Aldous, 1985). Each pair of clusters (in both directions) has a corresponding HP with a parametric form for the conditional intensity function. Let V be the set of events or vertices, π denote the partition of events, and n_j be the number of immigrant events in

cluster n_j. For $\lambda^*_{pq}(t)$ the model then has the form,

$$\begin{aligned}
\pi &\sim CRP(\alpha) \\
\lambda^*_{pq}(t) &= \mu_{pq} n_p n_q + \int_0^t \phi_{pq}(t-s) dN_j(s) \; \forall p,q \in \text{range}(\pi) \\
N_{pq}(\cdot) &\sim \text{HP}(\lambda^*_{pq}(\cdot)) \\
N_{uv}(\cdot) &\sim \text{Thinning}(N_{\pi(u)\pi(v)}(\cdot))
\end{aligned}$$

where α is the concentration parameter of the CRP and the thinning process $N_{uv}(\cdot)$ determines the edges of the directed graph by thinning, or distributing, all events in both clusters among the edges such that $N_{pq} = \sum_{u,v} N_{uv}(\cdot)$. Parameter inference is performed using Markov chain Monte Carlo methods as there is no conjugate prior available for this likelihood. The partition of the individuals in the model is updated via a Gibbs sampler, in a similar fashion to Du et al. (2015) with modifications for their model. The remaining model parameters are updated using a slice sampler.

A natural extension of the above approaches is the hierarchical Dirichlet. The inclusion of hierarchies facilitates description of a wide range of phenomena in the data and system under inspection. For example, a hierarchical Dirichlet HP proposed by Markwick (2020) is applied to 5 minute foreign exchange trade data, that is grouped daily for individual day HPs whilst allowing pooling of information where there is less data. The model is able to learn seasonality in trading events simultaneously, with the nonparametric background rate shown as,

$$\begin{aligned}
\mu_d(t) &\sim \mu_0 \cdot f_D(t) \; , \\
f_D(t) &\sim \int k(t|\theta) dG_D(\theta), \\
G_D &\sim DP(\alpha_D, G_0) \\
G_0 &\sim DP(\nu, H)
\end{aligned} \tag{29}$$

where μ_0 and f_D are the amplitude and density of controlling events on a day, respectively, with individual days d grouped by Days, D. The individual Dirichlet process model G_D with base measure for mixing kernel k (beta distributions) is,

$$k(y_i|\theta) = \text{Beta}(y_i|\mu, \upsilon, T) = \frac{y_i^{\frac{\mu\upsilon}{T}-1}(T-y_i)^{\upsilon(1-\frac{\mu}{T}-1)}}{B(\frac{\mu}{T}, \upsilon(1-\frac{\mu}{T}))T^{\upsilon-1}}$$

with non-conjugate prior for the mixture kernel,

$$G_0(\mu, \upsilon|T, \alpha_0, \beta_0) = U(\mu|[0,T])\text{Inv-Gamma}(\upsilon|\alpha_0, \beta_0)$$

and the global Dirichlet process learnt from the data. Augmenting the latent structure and selecting conjugate priors for the model parameters lead to a fully-Gibbs sampling algorithm. The model benefits from the ability of trades being updated in real time (online) and modelling the days of week's trades whilst sharing data amongst all groups with dynamic forecasts.

3.4. Summary

A number of non-parametric Bayesian inference procedures for the HP were reviewed. Computational improvements were highlighted with approximation and estimation strategies giving linear time complexity. Other approaches provided important improvements in flexibility and uncertainty in modelling the kernel. Finally, a scalable online clustering method in Dirichlet Process allows for the number of samples to grow with the HP, while the hierarchical approach supports pooling information and aiding where limited data size is available.

4. Direction 3: Other approaches

This section presents a brief review of three other directions in HP research. These include stochastic differential equations, graphs and neural networks.

In the first direction, Lee et al. (2016) extended the HP model to include randomness of the triggering kernel and introduced contagion parameters to control the levels of excitation. Each level of the excitation function is a stochastic process and is solved using a stochastic differential equation that follows a Geometric Brownian Motion and Exponential Langevin dynamics, inferred through Bayesian methods. The model attempts to better approximate applications where self-excitation intensities are accelerated with correlated levels of contagion.

The second direction points to graph-based approaches. This allows the user to determine the interaction between components within multivariate HPs by recovering the latent network structure. Generally, this is achieved by estimating the infectivity matrix, for which the ijth element describes the expected number of offspring events expected in dimension i given an event in dimension j. Several authors have introduced sparse and low-rank approximations to the matrix to control interactions within the network and improve computational efficiency.

An early example is given by Linderman and Adams (2014). The authors combine HPs with random graph models by decomposing the infectivity matrix into a binary adjacency matrix representing network sparsity, and a weight matrix to model interaction strength. A parallelisable Gibbs sampler is used to infer the model parameters. Guo et al. (2015) augment this approach for uncovering the latent network with a new Bayesian language model to study the evolution of dialogue within a social network over time. Linderman, Wang and Blei (2017) focus on inferring the latent structure of a social network when the data is not fully observed, where several types of missing data are considered. A new sequential Monte Carlo approach is proposed to recover the missing data. Liu et al. (2018) exploit MHP spatio-temporal properties by introducing a graph regularisation method, in which a penalisation term from the proximity of the infectivity matrix to a spatial connection matrix learns the influence among MHP characteristics.

Bacry et al. (2015) provide an extension by introducing a sparsity and low-rank induced penalisation, resulting in an excitation matrix of few non-zero and independent rows. This enhances scalability and improves estimation of the kernel. In a similar vein,

Zhou et al. (2013); Bacry et al. (2020) perform inference for higher dimensional HPs by modelling the excitation function as a low-rank approximation with regularised objective functions. The sparsity introduced in the infectivity matrix ensures that individuals are only impacted by a small number of users in the network while a small fraction of hubs can have wide-spread influence. Similarly, the Mean-Field Hypothesis as described by Bacry et al. (2016a) improves computational efficiency when recovering parameters in higher dimensions, given fluctuations of stochastic intensity are small.

In a third direction, the nonlinearity of the intensity function can be modelled as a neural network. Recurrent neural networks encode sequences of input states and output states, where each state is determined by the preceding state and the hidden state captures other past states. The parameters are fitted by an optimisation procedure on a nonlinear function, such as a sigmoidal or hyperbolic tangent.

Improving on the recurrent neural network issues, long short-term memory (LSTM) architecture mitigates the vanishing gradient problem, extending memory by modelling HP intensities of multiple events trained through 'forget gates' to control influences of past events on the current state (Mei and Eisner, 2017). Some other neural network approaches are the self-attentive/transformer models (Zuo et al., 2020; Zhang et al., 2020a) and graph convolution networks (Shang and Sun, 2019), showing computational efficiencies and improved prediction accuracy.

Several approaches have also been proposed to model spatio-temporal HPs using neural networks. Okawa et al. (2021) construct the intensity function for HPs to accept images as input by combining convolutional neural networks with continuous convolution kernels to output a multiplicative factor that influences the process in addition to the standard temporal and spatial triggering kernels. An alternative model that relies on neural networks to approximate the conditional intensity function of the HP is proposed by Du et al. (2021). The authors introduce a framework that learns the graph structure of the process which is then combined with temporal and spatial information. There are numerous other variations in neural network approaches as this is a significant body of active research in this area.

4.1. Summary

In this section we presented and discussed three further approaches, namely stochastic differential equations, graphs and neural networks. Although these related fields do not conveniently fit in the previous sections, they highlight the breadth of HPs in different research areas and show significant recent growth.

5. Direction 4: Theoretical guarantees and statistical properties

There is an emerging deep literature on theoretical aspects of HPs. Here we touch on three of these, namely developments in statistical properties, the special case of Markovian HPs, and nonlinear representation of self exciting processes.

With respect to developments in statistical properties, definitions past the first and second order statistics are possible given weakly stationary state conditions (Daley and Vere-Jones, 2003), but they become less intuitive and tractable as their statistical order increases. By introducing a combinatorial formula, Jovanović et al. (2015) allow the integral of cumulants (and consequent moments) to be calculated of arbitrary order for HPs. Specifically, given a set of $s \in \{1,...,D\}$ components and one of times $t_s = \{t_1,...,t_{|s|}\}$, the cumulant density of a HP is defined as

$$k(N^s) = dt^{-|s|} \sum_{\pi} (|\pi| - 1!)(-1)^{|\pi|-1} \prod_{B \in \pi} \left\langle \prod_{i \in B} dN_t^i \right\rangle, \tag{30}$$

where the sum runs over all partitions $|\pi|$ in s, $|\cdot|$ denotes the number of blocks and B labels individually the blocks of π. Moments in terms of cumulants are expressed as

$$\left\langle \prod_{i \in s} dN_{t_i}^i \right\rangle dt^{-|s|} = \sum_{\pi} \prod_{B \in \pi} k(N^B). \tag{31}$$

Jovanović et al. (2015) represent HPs as a cluster process, showing how to express (30) as a sum of integral terms by enumerating all possible rooted trees. The contribution of enumerating these 'family trees' that represent the complex interactions between point events, can be performed systematically and thus ease computational costs.

In an alternative approach to finding moments, Cui et al. (2020) used elementary derivations of self-exciting processes, setting the objective function to evaluate probabilistic arguments that yield a differential equation for the required moment.

Some other progress made in the direction of asymptotic results is in the study of a special class of HPs that is Markovian. For instance, when the exciting function is exponential, the joint process $(N_t, \lambda t)$ is then Markovian (10). In the paper by Gao and Zhu (2018b), the functional law of large numbers and central limit theorems are derived for the linear HP where the initial intensity and time are large, defined as

$$\lambda(t) := \int_{-\infty}^{t} = \alpha e^{-\beta(t-s)} dN(s) = \lambda_0 \cdot e^{-\beta t} + \int_0^t \alpha e^{-\beta(t-s)} dN(s) \text{ as } \lambda_0 = n \to \infty$$

where the process λ is Markovian given $d\lambda(t) = -\beta\lambda(t)dt + \alpha dN(t)$. Such limit theorems (details in Gao and Zhu (2018b)) provide insight into macroscopic behavior of large initial intensity asymptotics of HPs. Furthering the Markovian HPs towards the nonlinear case, a proof for large deviation was obtained by Zhu (2015), where the exciting function is both exponential and a sum of exponentials. More recent work by Kanazawa and Sornette (2020) provides a theoretical framework to embed non-Markovian kernels as Markovian, with the aim of tackling more general and complex derived HP models. This process of introducing auxiliary field variables via a master equation provides a formulation in terms of linear stochastic partial differential equations that are Markovian.

Another direction in theoretical work is the study of nonlinear HPs. For example, Torrisi (2016, 2017) derive explicit bounds in the Gaussian and Poisson approximations

on nonlinear HPs using Stein's method and Malliavin calculus. Gao and Zhu (2018a) present a study of a new asymptotic regime and its relation to the mean field limit for higher dimensions. Finally, from the perspective of asymptotic frequentist properties of Bayesian estimators, Donnet et al. (2020) consider nonparametric MHP posterior concentration rate ε_t around the true parameter θ^*,

$$\mathbb{E}_{\theta^*}\left(\prod(d(\theta,\theta^*) > \varepsilon_t | N_t)\right) = o(1) \text{ as } T \to \infty \tag{32}$$

in understanding influential features of the prior. The prior models are defined as a piecewise constant function and a mixture of Beta distributions that is given by,

$$\phi_{ij}(\cdot) = \rho_{ij}\left(\int_0^1 g_{\alpha_{ij}\varepsilon} dM_{ij}(\varepsilon)\right),\ g_{\alpha\varepsilon}(x) = \frac{\Gamma(\alpha/\left(\varepsilon(1-\varepsilon)\right))}{\Gamma(\alpha/\varepsilon)\,\Gamma(\alpha/(1-\varepsilon))} x^{\frac{\alpha}{1-\varepsilon}-1}(1-x)^{\frac{\alpha}{\varepsilon}-1} \tag{33}$$

where M_{ij} are bounded signed measures on $[0,1]$ such that $|M_{kl}| = 1$. The asymptotic posterior concentration rates are derived in stochastic terms and $\mathbb{L}_1$ distances $d(\theta,\theta^*)$. Sulem et al. (2021) furthers theoretical guarantees on estimation methods by considering nonlinear and inhibition effects of MHPs, obtaining the concentration rates of the posterior distribution on the parameters.

5.1. Summary

The theory of HPs is extensive with numerous areas of development. It is not our goal to give a detail account here, rather to provide the reader with three interesting current challenges that researchers are tackling. First we show approaches to nth order cumulant density formula derived in terms of Poisson cluster processes, secondly a number of derived theorems from a special class of the HPs (Markovian) and finally explicit bounds and posterior concentration rates for nonlinear HPs.

6. Real-world cameos

As noted in the Introduction, a key aspect of HP modelling is its suitability to real world applications. Many of the papers discussed in this review motivated and illustrated their methods with substantial examples. Tables 1 and 2 provide a scan of these applications and the corresponding findings, categorised by the estimation and numerical methods described in previous sections. A small number of cameos are described in further detail below.

6.1. Cameo 1: Crime

The issue of refining the parametric form of the triggering kernel is circumvented by a nonparametric approach to parameter estimation. Mohler et al. (2011) introduce a

spatio-temporal model for burglaries in Los Angeles. The model, inspired by the ETAS model developed to model seismic activity, is given by

$$\lambda(t,x,y) = \mu_t(t)\mu_b(x,y) + \int_{-\infty}^{t}\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(t-s,x-u,y-v)dN(s)dN(u)dN(v) \qquad (34)$$

where $\mu_t(t)$ and $\mu_b(x,y)$ are temporal and spatial baseline functions, respectively. Model parameters are estimated via variable-bandwidth Kernel Density Estimation (KDE).

A recent extension of this work is the semi-parametric spatiotemporal model employed by Zhuang and Mateu (2019), which describes complexities of criminal behaviors by incorporating their biological clock and periodic social activity. The conditional intensity is defined as

$$\begin{aligned}\lambda(t,x,y) = {} & \mu_0\mu_t(t)\mu_d(t)\mu_w(t)\mu_b(x,y) + \\ & A\int_{-\infty}^{t}\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi_1(t-s)\phi_2(x-u,y-v)dN(s)dN(u)dN(v) \qquad (35)\end{aligned}$$

where relaxation coefficients A and μ_0 stabilise the estimation process via maximisation likelihood, giving the model a semiparametric component. The other terms extend the nonparametric MISD model, where the baseline periodicity is estimated via residual analysis with daily/weekly terms μ_d and μ_w, average trend μ_t and spatial background $\mu_b(x,y)$ all normalised to 1. The triggering kernels, both temporal ϕ_1 and spatial ϕ_2, are then normalised as density functions. The introduction of periodic terms and estimation of their relative contributions is used to model crime rates in Castellon, Spain. In addition to uncovering daily and weekly patterns in robberies, the authors' analysis reveals the high influence of the background rate compared to the clustering effect which explains roughly 3% of the overall intensity.

6.2. Cameo 2: Finance

Kirchner (2017) shows the close relation of HPs to an Integer Auto-Regression (INAR) where the distribution of the resulting bin count sequence is approximated as a multivariate INAR(p). Fitting a mutually exciting bivariate HP to trades and limit orders on S&P 500, Kirchner (2017) determines an asymmetric relationship between both incoming orders exciting limit order and market orders, and finds that market order has barely an effect on incoming limit order.

In further support of high frequency applications, the Hawkes Graphs approach by Embrechts and Kirchner (2018) efficiently fits dozens of event streams. This method also provides a natural approach to studying connectivity and causality.

The suitability of the nonparametric HP method to very large datasets was also demonstrated by Bacry, Jaisson and Muzy (2016b). In the approach taken by these authors, a series of Wiener-Hopf equations is solved by Gaussian quadrature to estimate the kernel matrix, where market orders of two future assets on EUREX were shown to closely fit a power law function. Rambaldi, Bacry and Lillo (2017) couples this nonparametric kernel estimation with a MHP to successfully show the complex interactions

between time of arrival of orders in limit order books (LOB) and their size. Their work highlights the fact that high frequency orders on EUREX exchange are not suitable to be described with a simple model assuming independence between volume and time.

6.3. Cameo 3: Online content

The model proposed by Du et al. (2015), summarised in Section 3.3, was applied by the authors to a stream of news articles for a 35 day period at the beginning of 2011. The aim of the model is to identify emerging news stories by clustering related news articles based on the terms used in each article.

To determine the words included in the vocabulary of the model, named entities are identified and words that do not add information to the text are pruned, leaving a vocabulary of terms consisting largely of named entities, nouns, verbs and adjectives. The triggering kernel is made up of a linear combination of known radial basis function kernels. These kernels assign mass to the excitation function based on the distance between particular reference time points and the time elapsed for a pair of events. In this study the reference time points range from 30 minutes to 168 hours, capturing a range of both short and long time excitation effects. A number of meaningful news stories were identified as clusters, including the 2011 shooting in Tuscan, the release of the film 'Dark Knight Rises', the space shuttle Endeavour's last mission and cyclone Yasi in Queensland, Australia. A key outcome for this work is the ability to track the trend of each of these stories through examining both the form of the triggering kernel and the level of overall excitation.

‘

Table 1. *Recent research applications in nonparametric HPs. i.e. traffic incidents, financial markets, crime, memes and epidemiology.*

Researchers	Date	Application	Type/Method	Results
∗ Zhou F, Luo S, Li Z et al.	2021	NY vehicle incidents	HP, MISD, EM, VI	Flexible baseline and kernel avoids overfitting and improves on vehicle collision predictions.
∗∗ Kalair K, Connaughton C and Di Loro P	2021	UK traffic incidents	HP, SP, MISD	Self-excitation shown to account for 6-7% of observed secondary incidents.
∗ Markwick D	2020	FX trades	HP, EM, MCMC	Hierarchical model accounts natural daily trading, with real time updating and dynamic forecasting.
∗ Donnet S, Rivoirard V and Rousseau J	2020	Neuronal	MHP, MCMC	Recovers interaction graphs of neurons activity, simulating action potentials.
∗∗ Park J, Chaffeea A, Harrigan R, Schoenberg F	2020	Ebola spread west Africa	HP, MISD	Utility of HPs as alternative approach in forecasting epidemic spread, showing improved RMSE on SEIR models.
∗∗ Zhuang J and Mateu J	2019	Spain’s crime	HP, SP, MISD	3% of crime explained through clustering and daily/weekly periodicity activity.
∗ Zhang R, Walder C, Rizoiu M and Xie L	2019	Twitter meme	HP, MCMC, EM	Captures and compares categories of contents given decaying tweets and measure diffusion in linear time complexity.
∗∗ Schoenberg F, Gordon J and Harrigan R	2018	US plague spread	HP, MISD	Estimated contagion time 0-7.5 days with fitted model improving on computation and kernel estimates.
∗∗ Nichols K, Trevino E, Ikeda N et al.	2018	California earthquakes	HP, EM, MISD, ETAS	Adapted model for smaller magnitude earthquakes, accounting for more varied spatial and temporal features.
∗∗ Achab M, Bacry E, Gaiffas S et al.	2018	Meme tracking	MHP, MO	Shows an improvement on existing methods in relative error (7%), rank correlation and computing time.
∗∗ Embrechts P and Kirchner M	2018	FX trades	MHP, AR	The Hawkes Graph models supports multi-type high frequency streaming data types and causality analysis.
∗ Seonwoo Y, Park S and Oh A	2018	New York Times articles	HP, SMC	Texts in news articles are reconstructed for narratives and thread structures from the New York times, highlighting algorithm’s performance efficiency.

Intensity functions categories
∗ Bayesian nonparametric
∗∗ Frequentist nonparametric

Type/Method Acronyms
(HP)Univariate Hawkes process, (MHP)Multivariate Hawkes process, (SP)Spatial,
(MISD)Model Independent Stochastic Declustering, (WH)Wiener-Hopf,
(ETAS) Epidemic Type Aftershock Shock, (AR)Auto Regressive, (MO)Moments,
(GD)Gradient descent, (MAP)Maximum a posterior, (EM)Expectation Maximisation,
(VI)Variational Inference, (MCMC) Markov chain Monte Carlo, (SMC)Sequential Monte Carlo

Table 2. *Continued. Recent research applications in nonparametric HPs.*

Researchers	Date	Application	Type/Method	Results
∗∗ Yang Y, Etesami J, He N and Kiyavash N	2017	Meme tracking	MHP, RKHS	Popular phrases on news agencies are model in a scalable and online learning efficient algorithm, O(log T).
∗∗ Kirchner M	2017	FX trades	MHP, AR	Fits bivariate HPs on trades and limit order to LOB E-mini S&P 500, asymmetric relation shown between components.
∗∗ Rambaldi M, Bacry E and Lillo F	2017	FX trades	MHP, WH	Analysing interaction between time and arrival of orders (in LOB) and their size, shows unsuitability of independence between volume and time with simpler models.
∗ Xu H and Zha H	2017	ICU signal analysis	MHP, EM, MCMC	HP sequence clustering on pattern recognition and signal processing for clinical decision-making in the intensive care unit.
∗∗ Bacry E, Jaisson T and Muzy J	2016b	FX trades	MHP, WH	Models market orders arrivals on EUREX exchange showing power-law like shape and suitability to larger datasets.
∗∗ Hall E and Willett R	2016	Meme tracking	MHP, GD	Estimates underlying network of posts with scalable and online algorithm of $O(n^2)$ complexity.
∗ Mavroforakis C, Valera I and Gomez-Rodriguez M	2016	Online user activity	HP, SMC	Learns users patterns on Stack Overflow through tracing online activity, grouped streaming data with hierarchical Dirichlet HPs.
∗ Fox EW, Schoenberg, FP and Gordon JS	2016	Japan's earthquakes	HP, SP, ETAS	Incorporated histogram estimator and variable bandwidth kernel improves seismicity model and provides support as a powerful exploratory tool.
∗ Linderman S and Adams R	2015	Chicago gang violence	HP, SP, MCMC	Spatial Gaussian mixture model to predict gang related homicide that exhibit mutually exciting properties.
∗ Du N, Farajtabar M, Ahmed A et al.	2015	News	HP, MCMC, MAP	Uncovering topic specific clusters with learned lantent temporal dynamics showing an intuitive way in tracking trends online.
∗ Blundell C, Heller K and Beck J	2012	Social networks	MHP, MCMC	Inferred social network structure based on email interaction threads.

Intensity functions categories
∗ Bayesian nonparametric
∗∗ Frequentist nonparametric

Type/Method Acronyms
(HP)Univariate Hawkes process, (MHP)Multivariate Hawkes process, (SP)Spatial,
(MISD)Model Independent Stochastic Declustering, (WH)Wiener-Hopf,
(ETAS) Epidemic Type Aftershock Shock, (AR)Auto Regressive, (MO)Moments,
(GD)Gradient descent, (MAP)Maximum a posterior, (EM)Expectation Maximisation,
(VI)Variational Inference, (MCMC) Markov chain Monte Carlo, (SMC)Sequential Monte Carlo

7. Conclusion and Challenges

The past fifty years has seen the HP embedded as a staple methodology in the statistical literature. The growth in research directions inspired by the HP is itself a HP! Even after half a century, this pursuit continues through new theoretical, methodological and computational developments and new applications. The papers referenced in this review were selected to highlight some of the current directions in these areas and to provide a broad overview for new readers in the field. A range of research directions, in particular parametric, nonparametric, online and Bayesian approaches, were highlighted along with a number of real-world applications. The quantity and quality of the work reviewed here, and the large body of literature that was unfortunately not included, are a portent for another fifty years of exciting research related to HPs.

Acknowledgement

This research was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and the Australian Research Council (ARC) Laureate Fellowship Program under the project 'Bayesian Learning for Decision Making in the Big Data Era' (ID: FL150100150).

References

Achab, M., Bacry, E., Gaiffas, S., Mastromatteo, I. and Muzy, J. F. (2018). Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, *18* (2016), 1–28.

Aldous, D. J. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983 1–198. Lecture Notes in Math.* 1117. Springer, Berlin.

Bacry, E., Dayri, K. and Muzy, J. F. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *European Physical Journal B*, *85* (5).

Bacry, E., Mastromatteo, I. and Muzy, J. F. (2015). Hawkes Processes in Finance. *Market Microstructure and Liquidity*, *01* (01), 1550005.

Bacry, E. and Muzy, J. F. (2016). First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62 (4), 2184–2202.

Bacry, E., Gaïffas, S., Mastromatteo, I. and Muzy, J. F. (2016a). Mean-field inference of Hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, *49* (17).

Bacry, E., Jaisson, T. and Muzy, J. F. (2016b). Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, *16* (8), 1179–1201.

Bacry, E., Bompaire, M., Gaïffas, S. and Muzy, J. F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, *21*, 1–32.

Blundell, C., Heller, K. A. and Beck, J. M. (2012). Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, *4*, 2600–2608.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C. and Jordan, M. I. (2013). Streaming variational Bayes. *Advances in Neural Information Processing Systems*, 1–9.

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V. and Rousseau, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. *PLOS ONE*, *16 (4 April),* 1–28.

Chen, J., Hawkes, A. G. and Scalas, E. (2021). A Fractional Hawkes Process. *SEMA SIMAI Springer Series*, *26*, 121–131.

Chérief-Abdellatif, B. E., Alquier, P. and Khan, M. E. (2019). A generalization bound for online variational inference. *Asian Conference on Machine Learning*, 662–677.

Chiang, W. H., Liu, X. and Mohler, G. (2021). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting*, (40).

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17* (2), 129–157.

Cui, L., Hawkes, A. and Yi, H. (2020). An elementary derivation of moments of Hawkes processes. *Advances in Applied Probability*, *52* (1), 102–137.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*. Springer-Verlag.

Donnet, S., Rivoirard, V. and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of statistics*, 2698–2727.

Du, H., Zhou, Y., Ma, Y. and Wang, S. (2021). Astrologer: Exploiting graph neural Hawkes process for event propagation prediction with spatio-temporal characteristics. *Knowledge-Based Systems*, *228*, 107247.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M. and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17 August*, 1555–1564.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J. and Song, L. (2015). Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-August*, 219–228.

Embrechts, P. and Kirchner, M. (2018). Hawkes graphs. *Theory of Probability and its Applications*, *62* (1), 132–155.

Fox, E. W., Schoenberg, F. P. and Gordon, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, *10* (3), 1725–1756.

Gao, F. and Zhu, L. (2018a). Some asymptotic results for nonlinear Hawkes processes. *Stochastic Processes and their Applications*, *128* (12), 4051–4077.

Gao, X. and Zhu, L. (2018b). Limit theorems for Markovian Hawkes processes with a large initial intensity. *Stochastic Processes and their Applications*, *128* (11), 3807–3839.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Guo, F., Blundell, C., Wallach, H. and Heller, K. (2015). The Bayesian echo chamber: Modeling social influence via linguistic accommodation. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)2015, San Diego, CA, USA. JMLR: W&CP, volume 38*.

Hall, E. C. and Willett, R. M. (2016). Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, *62* (7), 4327–4346.

Halpin, P. F. (2013). An EM algorithm for Hawkes process. *New Developments in Quantitative Psychology: Proceedings of the 77th International Meeting of the Psychometric Society*, (212).

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *33* (3), 438–443.

Holbrook, A. J., Ji, X. and Suchard, M. A. (2022). From viral evolution to spatial contagion: a biologically modulated Hawkes model. *Bioinformatics (Oxford, England)*, *38* (7), 1846–1856.

Holbrook, A. J., Loeffler, C. E., Flaxman, S. R. and Suchard, M. A. (2021). Scalable Bayesian inference for self-excitatory stochastic processes applied to big American gunfire data. *Statistics and Computing, January 2021*, *31* (4).

Jovanović, S., Hertz, J. and Rotter, S. (2015). Cumulants of Hawkes point processes. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *91* (4).

Kalair, K., Connaughton, C. and Di Loro, P. A. (2021). A non-parametric Hawkes process model of primary and secondary accidents on a UK smart motorway. *Journal of the Royal Statistical Society Series C, vol. 70* (1), 80-97.

Kanazawa, K. and Sornette, D. (2020). Field master equation theory of the self-excited Hawkes process. *Physical Review Research*, *2* (3), 33442.

Kelly, J. D., Park, J., Harrigan, R. J., Hoff, N. A., Lee, S. D., Wannier, R., Selo, B., Mossoko, M., Njoloko, B., Okitolonda-Wemakoy, E., Mbala-Kingebeni, P., Rutherford, G. W., Smith, T. B., Ahuka-Mundeke, S., Muyembe-Tamfum, J. J., Rimoin, A. W. and Schoenberg, F. P. (2019). Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. *Epidemics*, *28*, 100354.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, 381–388.

Kirchner, M. (2017). An estimation procedure for the Hawkes process. *Quantitative Finance*, *17* (4), 571–595.

Kobayashi, R. and Lambiotte, R. (2016). TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. *Tenth International AAAI Conference on Web and Social Media*.

Laub, P. J., Lee, Y. and Taimre, T. (2021). The elements of Hawkes processes. *Springer International Publishing. Cham.*

Lee, Y., Lim, K. W. and Ong, C. S. (2016). Hawkes processes with stochastic excitations. *33rd International Conference on Machine Learning, ICML 2016*, *1*, 132–145.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, *1* (1), 1–20.

Li, Z., Cui, L. and Chen, J. (2018). Traffic accident modelling via self-exciting point processes. *Reliability Engineering and System Safety*, *180* (July), 312–320.

Linderman, S. W., and Adams, R. P. (2014). Discovering latent network structure in point process data. *31st International Conference on Machine Learning, ICML 2014*, *4*, 3268–3281.

Linderman, S. W., and Adams, R. P. (2015). Scalable Bayesian inference for excitatory point process networks. *Computer Science*. 1–16.

Linderman, S. W., Wang, Y., and Blei, D. M. (2017). Bayesian inference for latent Hawkes processes. *Advances in Approximate Bayesian Inference Workshop at the 31st Conference on Neural Information Processing Systems.*

Liu, Y., Yan, T. and Chen, H. (2018). Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics. *IJCAI International Joint Conference on Artificial Intelligence*, *2018-July*, 2475–2482.

Malem-shinitski, N., Ojeda, C. and Opper, M. (2022). Variational Bayesian inference for non-linear Hawkes process with Gaussian process self-effects. *Entropy, 24(3):* 356.

Markwick, D. (2020). Bayesian nonparametric Hawkes processes with applications. Doctoral thesis (Ph.D), UCL (University College London).

Marsan, D. and Lengliné, O. (2008). Extending earthquakes' reach through cascading. *Science*, *319* (5866), 1076–1079.

Mavroforakis, C., Valera, I. and Rodriguez, M. G. (2017). Modeling the dynamics of online learning activity. *In Proceedings of the 26th International World Wide Web Conference, 2017.*

Mei, H. and Eisner, J. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Advances in Neural Information Processing Systems*, *2017- Decem* (Nips), 6755–6765.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106* (493), 100–108.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9* (2), 249–265.

Nichols, K., Trevino, E., Ikeda, N., Philo, D., Garcia, A. and Bowman, D. (2018). Interdependency amongst earthquake magnitudes in Southern California. *Journal of Applied Statistics*, *45* (4), 763–774.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, *27* (1), 23–31.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of Computational and Graphical Statistics*, *83* (401), 9–27.

Ogata, Y. (1998). Space-Time Point-Process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50* (2), 379–402.

Okawa, M., Iwata, T., Tanaka, Y., Toda, H., Kurashima, T. and Kashima, H. (2021). Dynamic Hawkes processes for discovering time-evolving communities' states behind diffusion processes. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21).*

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, *31* (1), 145–155.

Park, J., Chaffee, A. W., Harrigan, R. J. and Schoenberg, F. P. (2020). A non-parametric Hawkes model of the spread of Ebola in west Africa. *Journal of Applied Statistics, 49* (3), 621-6371.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108* (504), 1339–1349.

Porter, M. D. and White, G. (2010). Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics*, *4* (1), 106–124.

Rambaldi, M., Bacry, E. and Lillo, F. (2017). The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, *17* (7), 999–1020.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, *15* (3), 623–642.

Rasmussen, J. G. (2018). Lecture Notes: Temporal point processes and the conditional intensity function. *arXiv:1806.00221v1*.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, *33* (3), 299–318.

Schoenberg, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica*, *26*, 861-879.

Schoenberg, F. P., Gordon, J. S. and Harrigan, R. J. (2018). Analytic computation of nonparametric Marsan–Lengliné estimates for Hawkes point processes. *Journal of Nonparametric Statistics*, *30* (3), 742–757.

Seonwoo, Y., Oh, A. and Park, S. (2018). Hierarchical Dirichlet Gaussian marked Hawkes process for narrative reconstruction in continuous time domain. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3316–3325.

Shang, J. and Sun, M. (2019). Geometric Hawkes processes with graph convolutional recurrent neural networks. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 4878–4885.

Sulem, D., Rivoirard, V. and Rousseau, J. (2021). Bayesian estimation of nonlinear Hawkes process. https://arxiv.org/abs/2103.17164v2

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Artificial intelligence and statistics*, *5*, 567–574.

Torrisi, G. L. (2016). Gaussian approximation of nonlinear Hawkes processes. *Annals of Applied Probability*, *26* (4), 2106–2140.

Torrisi, G. L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *53* (2), 679–700.

Walder, C. J. and Bishop, A. N. (2017). Fast Bayesian intensity estimation for the permanental process. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5459– 5471.

White, G., Porter, M. D. and Mazerolle, L. (2012). Terrorism Risk, Resilience and Volatility: A Comparison of Terrorism Patterns in Three Southeast Asian Countries. *Journal of Quantitative Criminology*, *29* (2), 295–320.

Xu, H. and Zha, H. (2017). A Dirichlet mixture model of Hawkes processes for event sequence clustering. *Advances in Neural Information Processing Systems*, *2017-December* (Nips), 1355–1364.

Xu, Z., Tresp, V., Yu, K. and Kriegel, H.-P. (2006). Infinite hidden relational models. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 544–551.

Yang, Y., Etesami, J., He, N. and Kiyavash, N. (2017). Online learning for multivariate Hawkes processes. *Advances in Neural Information Processing Systems*, *2017-December* (2), 4938–4947.

Yang, Y., Etesami, J., He, N. and Kiyavash, N. (2018). Nonparametric Hawkes processes: online estimation and generalization bounds. (Nips 2017), 1–39.

Zhang, R., Walder, C., Rizoiu, M. A. and Xie, L. (2019). Efficient non-parametric Bayesian Hawkes processes. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-Augus*, 4299–4305.

Zhang, Q., Lipani, A., Kirnap, O. and Yilmaz, E. (2020a). Self-attentive Hawkes process. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 11117– 11127.

Zhang, R., Walder, C. and Rizoiu, M.-A. (2020b). Variational inference for sparse Gaussian process modulated Hawkes process. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34* (04), 6803–6810.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2020a). Efficient inference for nonparametric Hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, *21*, 1–31.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2020b). Fast multiresolution segmentation for nonstationary Hawkes process using cumulants. *International Journal of Data Science and Analytics*, *10* (4), 321–330.

Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2021). Efficient EM-variational inference for nonparametric Hawkes process. *Statistics and Computing*, *31* (4), 1–11.

Zhou, K., Zha, H. and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Journal of Machine Learning Research*, *31*, 641–649.

Zhu, L. (2015). Large deviations for Markovian nonlinear Hawkes processes. *Annals of Applied Probability*, *25* (2), 548–581.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97* (458), 369–380.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *182* (3), 919–942.

Zuo, S., Jiang, H., Li, Z., Zhao, T. and Zha, H. (2020). Transformer Hawkes process. *37th International Conference on Machine Learning, ICML 2020*, *PartF16814*, 11628–11638.

SORT 46 (1) January-June 2022, 39-66 DOI: 10.2436/20.8080.02.117

Unusual-event processes for count data

Wanrudee Skulpakdee[1] and Mongkol Hunkrajok[2]

Abstract

At least one unusual event appears in some count datasets. It will lead to a more concentrated (or dispersed) distribution than the Poisson, gamma, Weibull, Conway-Maxwell-Poisson (CMP), and Faddy (1997) models can accommodate. These well-known count models are based on the monotonic rates of interarrival times between successive events. Under the assumption of non-monotonic rates and independent exponential interarrival times, a new class of parametric models for unusual-event (UE) count data is proposed. These models are applied to two empirical applications, the number of births and the number of bids, and yield considerably better results to the above well-known count models.

MSC: *62J99, 62M05, 62P99.*

Keywords: *Poisson count model, Gamma count model, Weibull count model, Conway-Maxwell-Poisson count model, Faddy count model.*

1. Introduction

Count data regression analysis is a collection of statistical techniques for modeling and investigating the conditional count distributions of count response variables given sets of covariates. The conditional-variance-mean function of these distributions can be classified into two different categories: linear and non-linear.

1. If the distributions are equidispersed (variance = mean), this function is linear.
2. If the distributions are overdispersed (variance > mean), this function is either linear or non-linear.

[1] Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. wanrudee.sku@nida.ac.th

[2] Independent Researcher, Bangkok, Thailand. hunkrajokmongkol@gmail.com

Received: June 2021
Accepted: November 2021

3. If the distributions are underdispersed (variance $<$ mean), this function is either linear or non-linear.

4. If the distributions are over-, under-, and equidispersed, this function is non-linear.

A renewal process is a counting process. Its times between successive events are independent and identically distributed with a non-negative distribution (Ross 2010). The primary assumption of the Poisson model is that the times between events are exponential. It follows that the Poisson model is equidispersed, and the Poisson regression model has a linear conditional-variance-mean function. The exponential distribution replaced by a less restrictive non-negative distribution such as the gamma and Weibull distributions leads to the gamma (Winkelmann 1995) and Weibull (McShane et al. 2008) count models. They allow for both overdispersion and underdispersion. The gamma and Weibull regression models have linear conditional-variance-mean functions when the additional parameter (α) equals 1, that is, the Poisson regression model. Furthermore, they have nearly linear conditional-variance-mean functions shown in Figures 1(a) and 1(b), although α does not approach 1.

The Conway-Maxwell-Poisson (CMP) model was originally introduced by Conway and Maxwell (1962). In contrast to the above models, the CMP model is not derived from an underlying renewal process. The proof can be found in the Supplementary Material. Surprisingly, however, the graphs in Figures 1(a) and 1(c) of the conditional-variance-mean functions for the gamma and the CMP are hardly distinguishable. A plausible explanation for this similarity is the equality of their approximate variance-mean ratios. These ratios are equal to a constant $1/\alpha$ (Winkelmann 1995, p. 470; Sellers and Shmueli 2010, p. 946). Likewise the gamma and Weibull count models, the CMP model consists of the rate and dispersion parameters. Thus, it allows for both over- and underdispersion.

As previously mentioned, the conditional variance and mean of the above well-known regression models are (nearly) linearly related. In some applications, these regression models are either unsatisfactory or inappropriate when the sample relative frequency distribution is created as a mixture of distributions whose relationship between the variance and the mean is non-linear.

The common assumption that the rates of interarrival times are equal may cause a (nearly) linear conditional-variance-mean function. One potential solution to this problem is to allow the unequal rates. Faddy (1997) suggested the generalization of the Poisson process $\lambda_n = \lambda\,(b+n)^{\alpha}$, $n = 0, 1, , 2, ...$, in which the rate at which new events occur depends on the number of events. The rate sequence of the Faddy (1997) process is either non-decreasing or non-increasing. The Faddy (1997) regression model has both (nearly) linear and non-linear conditional-variance-mean functions shown in Figures 1(d), but it displays only one of over-, under-, and equidispersion. Therefore, this regression model is either unsatisfactory or inappropriate when the sample relative frequency distribution is created as a mixture of over-, under-, and equidispersed distributions whose relationship between the variance and the mean must be non-linear. Note that the conditional variances and means in Figure 1 were computed in **R**

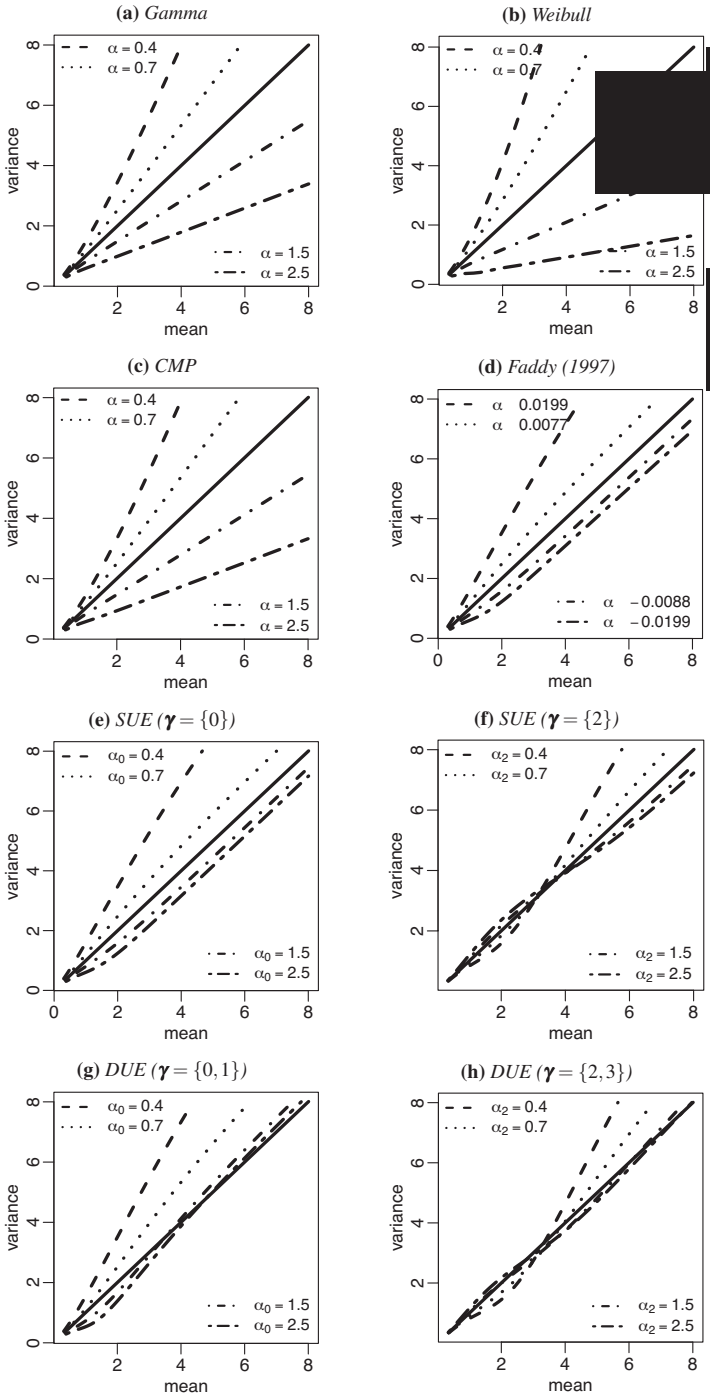


Figure 1. *Graphs showing the linear and non-linear functions of variance and mean. The Faddy (1997), DUE ($\boldsymbol{\gamma} = \{0,1\}$), and DUE ($\boldsymbol{\gamma} = \{2,3\}$) models present the cases in which b, α_1, and α_3 are $1 \times 10^{-20}, 0.687$, and 0.687, respectively.*

(R Core Team 2019) by the `dCount-conv-bi` function in the **Countr** package (Kharrat and Boshnakov 2018) for the gamma and Weibull count models, the `dcmp` function in the **COMPoissonReg** package (Sellers, Lotze and Raim, 2018) for the CMP count model, and the `Faddyprob.general` function in the **CountsEPPM** package (Smith and Faddy 2018) for the Faddy (1997) count model.

The limitation that the above regression models present only one dispersion type may be easily removed by allowing for non-monotonic rate sequences. The two examples are the single-unusual-event (SUE ($\boldsymbol{\gamma} = \{2\}$)) and double-unusual-event (DUE ($\boldsymbol{\gamma} = \{2,3\}$)) models shown in Figures 1(f) and 1(h). Their curves corresponding to $\alpha_2 \neq 1$ always cross the 45-degree (Poisson) lines. Thus, these models can fit a dataset that is a mixture of over-, under-, and equidispersion. The development and exploration of a new class of unusual-event (UE) models is the main objective of the present article. Note that the SUE ($\boldsymbol{\gamma} = \{0\}$) model is a special case of the Faddy (1997) (see Figures 1(d) and 1(e)), as described later.

The rest of this article is organized as follows. Section 2 presents the UE models and their properties, with additional details provided in Appendices A and B at the end of the paper. Section 3 discusses numerical strategies for computing UE probabilities. Section 4 provides and analyses the experimental results from the number of births and the number of bids. Finally, Section 5 concludes the paper.

2. Unusual-event models

Let $X(t)$ be a discrete random variable, representing the total number of events that occur before or at exactly time t. $\{X(t);\, t \geq 0\}$ is a pure birth process with $X(0) = 0$ and birth rates $\lambda_n (n \geq 0)$. The probabilities $P_n(t) = P\{X(t) = n \mid X(0) = 0\}$, for $n = 0, 1, 2, ...$, satisfy the Chapman-Kolmogorov forward differential equations (Cox and Miller 1965), namely

$$\begin{aligned} P_0'(t) &= -\lambda_0 P_0(t), \\ P_n'(t) &= -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t), \quad n > 0, \end{aligned} \tag{1}$$

with boundary conditions $P_0(0) = 1$ and $P_n(0) = 0,\ n > 0$.

Different distributions correlate with different birth rate sequence λ_n patterns. The simple Poisson process, which has a constant rate parameter λ, restricts that the variance equals the mean. The birth rate, which depends on the number of events, may allow for overdispersion and underdispersion. Increasing the number of parameters in the process almost always improves the goodness of fit (as assessed by the log-likelihood function), but it may cause overfitting. Thus, the rate λ_n must be a parametric function of n, as stated by Faddy and Smith (2008). Examples of pure birth processes follow below.

1. A sequence of rates

$$\lambda_n = \lambda, \quad \text{for } n = 0, 1, 2, ..., \quad \lambda > 0,$$

exhibits the Poisson distribution, which is a one-parameter count model.

2. A sequence of rates

$$\lambda_n = \begin{cases} \lambda & \text{for } n > 0 \\ \lambda_0 & \text{for } n = 0, \quad \lambda \text{ and } \lambda_0 > 0, \end{cases}$$

exhibits the Faddy (1994) distribution, which is a two-parameter count model.

3. A sequence of rates

$$\lambda_n = \begin{cases} \lambda & \text{for } n > 1 \\ \lambda_0 & \text{for } n = 0 \\ \lambda_1 & \text{for } n = 1, \quad \lambda,\ \lambda_0, \text{ and } \lambda_1 > 0, \end{cases}$$

exhibits the extended Faddy (1994) distribution, which is a three-parameter count model.

4. A sequence of rates

$$\lambda_n = \lambda\,(b+n)^{\alpha}, \quad \text{for } n = 0,1,2,..., \quad \lambda > 0,\ b > 0, \text{ and } \alpha \leq 1,$$

exhibits the Faddy (1997) distribution, which is a three-parameter count model.

The Faddy (1994), extended Faddy (1994), and Faddy (1997) models have greater flexibility than the Poisson model at the cost of additional parameters. Covariates can be incorporated into these models by setting λ as a function of the linear predictor $\beta_0 + \beta_1 x_{j1} + ... + \beta_r x_{jr}$, where $x_{jk},\ k = 1,...,r$, is the jth observation of the kth covariate, and $\beta_l,\ l = 0,...,r$, is the lth unknown parameter to be estimated. The rates $\lambda_n\,(n \geq 0)$ of the Poisson and Faddy (1997) distributions depend on the covariates, but the rates λ_0 and λ_1 of the Faddy (1994) and extended Faddy (1994) do not. One might argue that λ_0 and λ_1 can be written as a function of the linear predictor. However, the approximately doubled (Faddy (1994)) and tripled (extended Faddy (1994)) parameters comparing to the above two distributions may lead to overfitting. Perhaps the rate sequences of the Faddy (1994) and extended Faddy (1994) can be easily modified as follows:

$$\lambda_n = \begin{cases} \lambda & \text{for } n > 0 \\ \alpha_0\lambda & \text{for } n = 0, \quad \alpha_0 \text{ and } \lambda > 0, \end{cases}$$

and

$$\lambda_n = \begin{cases} \lambda & \text{for } n > 1 \\ \alpha_0\lambda & \text{for } n = 0 \\ \alpha_1\lambda & \text{for } n = 1, \quad \alpha_0,\ \alpha_1, \text{ and } \lambda > 0. \end{cases}$$

We call λ the base rate. These modified rate sequences can avoid the risk of overfitting, and the rates λ_n depend on covariates. In other words, these distributions with the fewest numbers of parameters occur when λ is a function of the linear predictor.

We call the pure birth process with this pattern of the rate sequences the unusual-event (UE) process because at least one rate differs from the base rate λ. It is defined as

$$\lambda_n = \begin{cases} \lambda & \text{for } n \notin \boldsymbol{\gamma} = \{\gamma_1, \gamma_2, ..., \gamma_m\} \\ \alpha_{\gamma_i}\lambda & \text{for } n \in \boldsymbol{\gamma}, \end{cases} \tag{2}$$

where γ_i is a non-negative integer, and $\alpha_{\gamma_i} > 0$, $i = 1, 2, ..., m$. We call α_{γ_i} the shape parameter. The UE process permits a wide range of regression models for count data, including the combinations of distributions with either one or three dispersion types. These possibilities are illustrated using the single-unusual-event (SUE) and double-unusual-event (DUE) processes.

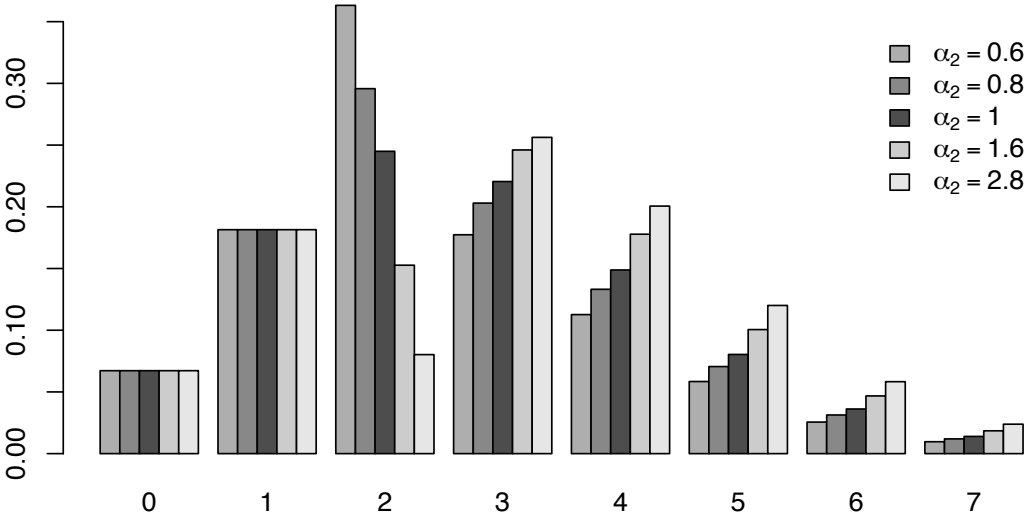


Figure 2. *SUE* ($\boldsymbol{\gamma} = \{2\}$ *and* $\lambda = 2.7$) *distributions with unequal means and dispersions.*

2.1. SUE models

Perhaps the simplest example of UE processes is a SUE process with

$$\lambda_n = \begin{cases} \lambda & \text{for } n \notin \boldsymbol{\gamma} = \{\gamma\} \\ \alpha_\gamma\lambda & \text{for } n = \gamma. \end{cases} \tag{3}$$

For $\alpha_\gamma = 1$, the SUE process simplifies to the Poisson process. It is noted that the SUE distribution is characterized by independently exponential distributed interarrival times. Figure 2 compares the probability functions of the SUE ($\boldsymbol{\gamma} = \{2\}$ and $\lambda = 2.7$) distribution for five values of α_2. It is more concentrated ($\alpha_2 < 1$) or more dispersed ($\alpha_2 > 1$) than the Poisson distribution ($\alpha_2 = 1$). The overdispersion case $\alpha_2 = 2.8$ shows a probability distribution with two distinct modes (1 and 3) referred to as a bimodal distribution.

The SUE ($\boldsymbol{\gamma} = \{2\}$) probability function is given by

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < 2 \\ (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \dfrac{((1-\alpha_2)\lambda t)^i}{(i+n)!} & \text{for } n = 2 \\ \alpha_2 (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \dfrac{((1-\alpha_2)\lambda t)^i}{(i+n)!} & \text{for } n > 2. \end{cases} \tag{4}$$

The derivations of SUE probability distributions can be found in Appendix A. For $n < 2$, the SUE ($\boldsymbol{\gamma} = \{2\}$) probability function is Poisson. It is not a function of α_2, and thus the probabilities in Figure 2 are equal at $n = 0$ and 1. For $n = 2$, the SUE ($\boldsymbol{\gamma} = \{2\}$) probability function can be simplified to $\frac{(\lambda t)^2 e^{-\lambda t}}{2!} + (1-\alpha_2)(\lambda t)^3 e^{-\lambda t} \sum_{i=0}^{\infty} \frac{((1-\alpha_2)\lambda t)^i}{(i+3)!}$. The first term is the Poisson probability function for $n = 2$. Since the second term is positive ($\alpha_2 < 1$) and negative ($\alpha_2 > 1$), the SUE ($\boldsymbol{\gamma} = \{2\}$) probability value is greater and smaller than the Poisson ($\alpha_2 = 1$) , respectively.

The Faddy (1994) process is equivalent to the SUE ($\boldsymbol{\gamma} = \{0\}$) process, but only when their regression models are not considered. Therefore, the proof of the Faddy (1994) by direct calculation that for $t > 0$,

$$Var\{X(t)\} > E\{X(t)\} \quad \text{if} \quad \frac{\lambda_0}{\lambda} = \alpha_0 < 1$$

and

$$Var\{X(t)\} < E\{X(t)\} \quad \text{if} \quad \frac{\lambda_0}{\lambda} = \alpha_0 > 1,$$

is still correct for the SUE ($\boldsymbol{\gamma} = \{0\}$) process. Alternatively, $\alpha_0 < 1$ and $\alpha_0 > 1$ result in non-decreasing and non-increasing rate sequences, which provide overdispersed and underdispersed SUE ($\boldsymbol{\gamma} = \{0\}$) distributions, respectively (see Figure 4(a)). These properties were conjectured by Faddy (1994) and proved by Ball (1995). Note that the non-increasing and non-decreasing rate sequences mean $\lambda_{n+1} \leq \lambda_n$ and $\lambda_{n+1} \geq \lambda_n$, respectively.

It is worth mentioning that the SUE ($\boldsymbol{\gamma} = \{0\}$) model is a special case of the Faddy (1997) model (see Figures 1(d) and 1(e)). Let us consider a rate sequence of the Faddy (1997) model in which the parameter b is given in the form $\sigma^{\frac{1}{|\alpha|}}$, where $0 < \sigma \leq 1$. It can be shown that

$$\lim_{\alpha \to 0^-} \lambda_n = \lim_{\alpha \to 0^-} \lambda \left(\sigma^{\frac{1}{|\alpha|}} + n\right)^{\alpha} = \begin{cases} \lambda & \text{for } n > 0 \\ \frac{\lambda}{\sigma} = \alpha_0 \lambda & \text{for } n = 0, \ \alpha_0 = \frac{1}{\sigma} \geq 1, \end{cases}$$

and

$$\lim_{\alpha \to 0^+} \lambda_n = \lim_{\alpha \to 0^+} \lambda \left(\sigma^{\frac{1}{|\alpha|}} + n\right)^{\alpha} = \begin{cases} \lambda & \text{for } n > 0 \\ \sigma\lambda = \alpha_0 \lambda & \text{for } n = 0, \ 0 < \alpha_0 = \sigma \leq 1. \end{cases}$$

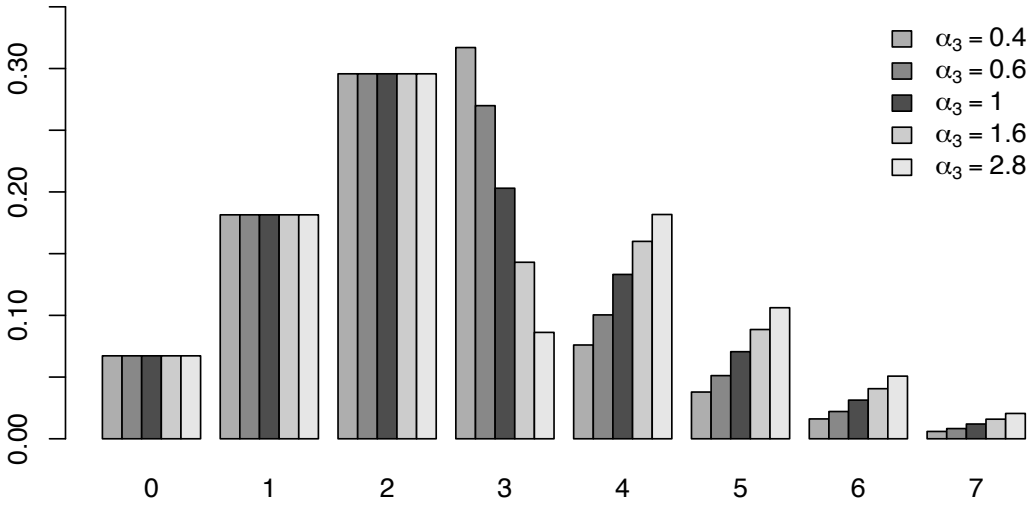


Figure 3. *DUE* ($\boldsymbol{\gamma} = \{2,3\}$, $\lambda = 2.7$, *and* $\alpha_2 = 0.8$) *distributions with unequal means and dispersions.*

Hence, the SUE ($\boldsymbol{\gamma} = \{0\}$) models with $\alpha_0 \geq 1$ and $0 < \alpha_0 \leq 1$ are the limiting cases of the Faddy (1997) model. They arise when α approaches 0 from the left and the right, respectively.

Figures 4(a)-4(d) show graphs of the variance-mean ratio with $\gamma = 0-3$ for various values of λ and α_γ. $\alpha_\gamma \neq 1$, $\gamma > 0$, results in a non-monotonic rate sequence, which provides over-, under-, and equidispersed SUE ($\boldsymbol{\gamma} = \{\gamma > 0\}$) distributions (see Figures 4(b)-4(d)). For fixed α_γ, the three dispersion types of the SUE models are defined by the base rate λ, in contrast to the gamma, Weibull, CMP, and Faddy (1997) models. We conjecture that this property holds for any non-monotonic rate sequence of SUE processes.

2.2. DUE models

Perhaps the simplest example of DUE processes is a pure birth process with

$$\lambda_n = \begin{cases} \lambda & \text{for } n \notin \boldsymbol{\gamma} = \{\gamma,\ \gamma+1\} \\ \alpha_\gamma \lambda & \text{for } n = \gamma \\ \alpha_{\gamma+1}\lambda & \text{for } n = \gamma+1. \end{cases} \tag{5}$$

For $\alpha_\gamma \neq 1$ and $\alpha_{\gamma+1} = 1$, the DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) process simplifies to the SUE ($\boldsymbol{\gamma} = \{\gamma\}$) process. We consider the DUE count model in which the rates, λ_γ and $\lambda_{\gamma+1}$, of two consecutive events are not equal to the base rate. This phenomenon appears to occur in the two empirical applications, the number of births and the number of bids, shown in Tables 4 and 5. Figure 3 compares the probability functions of the DUE ($\boldsymbol{\gamma} = \{2,3\}$, $\lambda = 2.7$, and $\alpha_2 = 0.8$) distribution for five values of α_3. It is more concentrated ($\alpha_3 < 1$)

or more dispersed ($\alpha_3 > 1$) than the SUE distribution ($\alpha_3 = 1$). The DUE ($\alpha_3 = 2.8$) distribution is a bimodal distribution whose modes are 2 and 4. The bimodal distributions of the SUE ($\boldsymbol{\gamma} = \{2\}$, $\lambda = 2.7$, and $\alpha_2 = 2.8$) and the DUE ($\boldsymbol{\gamma} = \{2,3\}$, $\lambda = 2.7$, $\alpha_2 = 0.8$, and $\alpha_3 = 2.8$) suggest that we should expect a UE distribution to be a multimodal distribution, which is a discrete probability distribution with two or more modes.

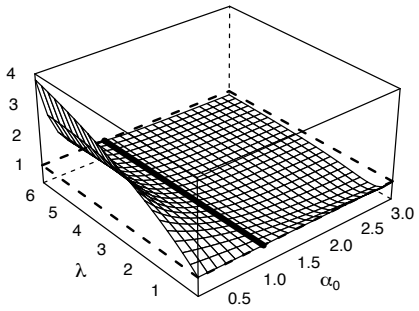
The DUE ($\boldsymbol{\gamma} = \{2,3\}$) probability function is given by

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < 2 \\ (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \dfrac{((1-\alpha_2)\lambda t)^i}{(i+n)!} & \text{for } n = 2 \\ \alpha_2 (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \dfrac{c_i (\lambda t)^i}{(i+n)!} & \text{for } n = 3 \\ \alpha_2 \alpha_3 (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \dfrac{c_i (\lambda t)^i}{(i+n)!} & \text{for } n > 3, \end{cases} \tag{6}$$

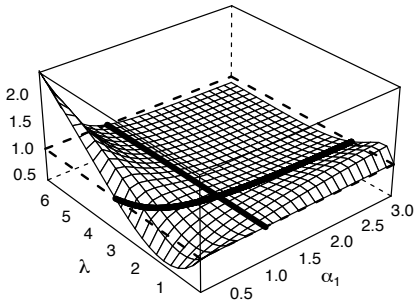
where $c_i = \sum_{k=0}^{i} (1-\alpha_2)^k (1-\alpha_3)^{i-k}$. The derivations of DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) probability distributions can be found in Appendix B. For $n < 3$, the DUE ($\boldsymbol{\gamma} = \{2,3\}$) probability functions are not dependent on α_3, and thus the probabilities in Figure 3 are equal at $n = 0$, 1, and 2. For $n = 3$, the DUE ($\boldsymbol{\gamma} = \{2,3\}$) probability function can be simplified to $\alpha_2 (\lambda t)^3 e^{-\lambda t} \sum_{i=0}^{\infty} \frac{((1-\alpha_2)\lambda t)^i}{(i+3)!} + \alpha_2(1-\alpha_3)(\lambda t)^4 e^{-\lambda t} \sum_{i=0}^{\infty} \frac{c_i(\lambda t)^i}{(i+4)!}$. The first term is the SUE ($\boldsymbol{\gamma} = \{2\}$) probability function for $n = 3$. Since the second term is positive ($\alpha_3 < 1$) and negative ($\alpha_3 > 1$), the DUE ($\boldsymbol{\gamma} = \{2,3\}$) probability value is greater and smaller than the SUE ($\boldsymbol{\gamma} = \{2\}$), respectively.

Figures 4(e)-4(h) show graphs of the variance-mean ratio with $\gamma = 0-3$ for various values of λ and α_γ. A non-decreasing rate sequence with $\alpha_0 \leq \alpha_1 < 1$ provides only overdispersed DUE ($\boldsymbol{\gamma} = \{0,1\}$) distribution, and a non-monotonic rate sequence with $\alpha_0 > \alpha_1 < 1$ produces over-, under-, and equidispersed DUE ($\boldsymbol{\gamma} = \{0,1\}$) distribution (see Figure 4(e)). $\boldsymbol{\gamma} \neq \{0,1\}$ results in a non-monotonic rate sequence, which provides over-, under-, and equidispersed DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) distributions (see Figures 4(f)-4(h)). For fixed α_γ and $\alpha_{\gamma+1}$, the three dispersion types of the DUE models are defined by the base rate λ, in contrast to the gamma, Weibull, CMP, and Faddy (1997) models. We conjecture that this property holds for any non-monotonic rate sequence of DUE processes and also UE processes.

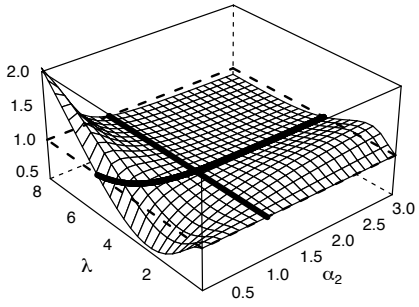
We conclude that even the simplest generalizations of the Poisson and SUE processes, the SUE and the DUE, are relatively flexible models for count data.



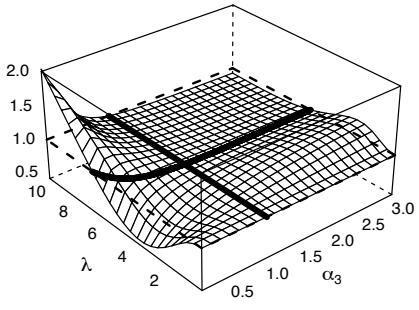
(a) *SUE* ($\boldsymbol{\gamma} = \{0\}$)



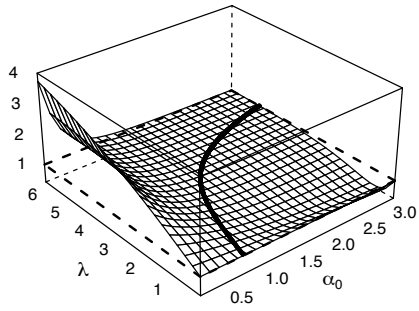
(b) *SUE* ($\boldsymbol{\gamma} = \{1\}$)



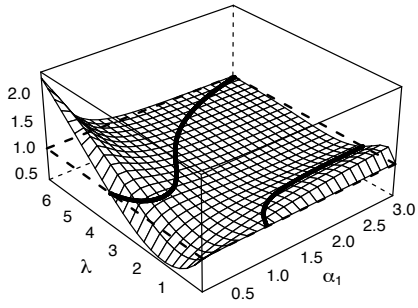
(c) *SUE* ($\boldsymbol{\gamma} = \{2\}$)



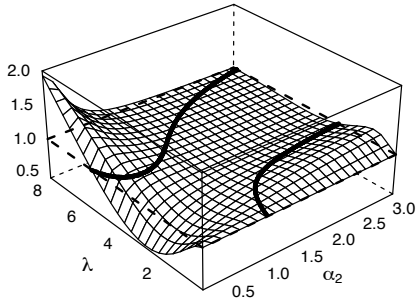
(d) *SUE* ($\boldsymbol{\gamma} = \{3\}$)



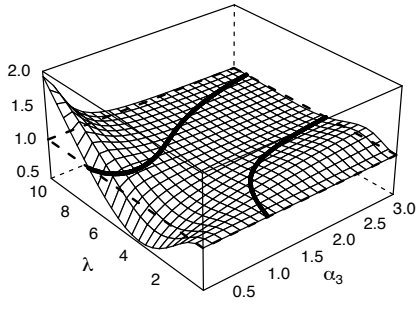
(e) *DUE* ($\boldsymbol{\gamma} = \{0,\, 1\}$, $\alpha_1 = 0.687$)



(f) *DUE* ($\boldsymbol{\gamma} = \{1,\, 2\}$, $\alpha_2 = 0.687$)



(g) *DUE* ($\boldsymbol{\gamma} = \{2,\, 3\}$, $\alpha_3 = 0.687$)



(h) *DUE* ($\boldsymbol{\gamma} = \{3,\, 4\}$, $\alpha_4 = 0.687$)

Figure 4. *Variance-mean ratios for SUE and DUE count models with* $0 < \alpha_\gamma < 3$. *Each SUE surface* ($\gamma > 0$) *contains a saddle point, which is the intersection of the straight and curved lines.*

3. Computation of UE Probabilities

The solution of the Chapman-Kolmogorov forward differential equation (1) can be written in terms of a matrix-exponential function (Cox and Miller 1965)

$$(P_0(t)\ P_1(t)\ ...\ P_n(t)) = (1\ 0\ ...\ 0)\, exp(\boldsymbol{Q}t), \tag{7}$$

where $\boldsymbol{Q}$ is the matrix of birth rates

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & 0 \\ 0 & -\lambda_1 & \lambda_1 & \cdots & 0 \\ 0 & 0 & -\lambda_2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \lambda_{n-1} \\ 0 & 0 & 0 & \cdots & -\lambda_n \end{pmatrix},$$

and an integral function (Bartlett 1978)

$$\begin{aligned} P_0(t) &= e^{-\lambda_0 t}, \\ P_n(t) &= \int_0^t \lambda_{n-1} P_{n-1}(u) e^{-\lambda_n (t-u)} du \quad \text{for } n > 0. \end{aligned} \tag{8}$$

The matrix exponentiation is the most common for computing the probabilities of pure birth processes. Researchers usually rely on this method (e.g. Faddy and Smith 2011 and Smith and Faddy 2016), perhaps because various packages for calculating the matrix exponential have been developed and made the routines available as described by Faddy and Smith (2008). The analytic solution is obtained by the integral function. It is computationally intractable (Faddy 1997; Crawford, Ho and Suchard, 2018) because there is an extremely ill-conditioned problem in the solution. A numerical solution may differ significantly from the exact solution. Therefore, the analytic solution is not appropriate for numerical computation (Podlich et al. 2004). However, this ill-conditioned problem can be solved by a Taylor series expansion. In this research, we will report the computational results from the analytic solution that was previously thought to be infeasible. The fertility and takeover bids datasets are considered in this paper. Their results are obtained by using the matrix-exponential and analytic solution approaches. We confirm here that the results from these two methods are identical.

The probability $P_n(t)$ in Equation (8) can be described as a convolution of two functions. First, the probability density function $\lambda_{n-1}P_{n-1}(u)$ is the n-fold convolution of the exponential density functions of the interarrival times between events. It presents the probability that the nth event occurs at exactly time u. Second, $e^{-\lambda_n(t-u)}$ is the survival function of the interarrival times between the nth and $(n+1)$th event. The survival function denotes the probability that the $(n+1)$th event does not occur after time u and before or at exact time t. Using Equation (8) and letting $\lambda_n = \alpha_n \lambda$, the first few probabilities of

the UE count models are obtained:

$$
\begin{aligned}
P_0(t) &= e^{-\alpha_0 \lambda t} \\
P_1(t) &= \alpha_0 e^{-\alpha_1 \lambda t} \left(\frac{e^{(\alpha_1-\alpha_0)\lambda t} - 1}{\alpha_1 - \alpha_0} \right) \\
P_2(t) &= \alpha_0 \alpha_1 e^{-\alpha_2 \lambda t} \left(\frac{(e^{(\alpha_2-\alpha_0)\lambda t} - 1)}{(\alpha_1 - \alpha_0)(\alpha_2 - \alpha_0)} + \frac{(e^{(\alpha_2-\alpha_1)\lambda t} - 1)}{(\alpha_0 - \alpha_1)(\alpha_2 - \alpha_1)} \right) \\
P_3(t) &= \alpha_0 \alpha_1 \alpha_2 e^{-\alpha_3 \lambda t} \left(\frac{(e^{(\alpha_3-\alpha_0)\lambda t} - 1)}{(\alpha_1 - \alpha_0)(\alpha_2 - \alpha_0)(\alpha_3 - \alpha_0)} \right. \\
&\quad \left. + \frac{(e^{(\alpha_3-\alpha_1)\lambda t} - 1)}{(\alpha_0 - \alpha_1)(\alpha_2 - \alpha_1)(\alpha_3 - \alpha_1)} + \frac{(e^{(\alpha_3-\alpha_2)\lambda t} - 1)}{(\alpha_0 - \alpha_2)(\alpha_1 - \alpha_2)(\alpha_3 - \alpha_2)} \right)
\end{aligned}
$$

From these equations for $P_0(t)$, $P_1(t)$, $P_2(t)$, and $P_3(t)$, one can deduce that the general UE probability function might be of the form

$$
P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \left(\prod_{i=0}^{n-1} \alpha_i \right) e^{-\alpha_n \lambda t} \sum_{i=0}^{n-1} \frac{(e^{(\alpha_n-\alpha_i)\lambda t} - 1)}{\prod_{j=0, j\neq i}^{n} (\alpha_j - \alpha_i)} & \text{for } n > 0, \end{cases} \tag{9}
$$

and this expression is similar to Bartlett (1978, eq. (9), p. 55) and Crawford et al. (2018, eq. (55), p. 13). Inserting the Taylor series expansion of $e^{(\alpha_n-\alpha_i)\lambda t}$, the UE probability distribution can be rewritten as follow:

$$
P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \left(\prod_{i=0}^{n-1} \alpha_i \right) e^{-\lambda t} \sum_{i=0}^{\infty} c_i \frac{(\lambda t)^{i+n}}{(i+n)!} & \text{for } n > 0, \end{cases} \tag{10}
$$

where

$$
c_i = \begin{cases} 1 & \text{for } i = 0 \\ \sum_{k_i=0}^{n} \sum_{k_{i-1}=0}^{k_i} \ldots \sum_{k_1=0}^{k_2} \prod_{j=k_1}^{k_i} (1 - \alpha_j) & \text{for } i > 0. \end{cases}
$$

This expression can also be obtained from Equation (7) by letting $\boldsymbol{Q} = \lambda(\boldsymbol{P} - \boldsymbol{I})$, where

$$
\boldsymbol{P} = \begin{pmatrix} 1-\alpha_0 & \alpha_0 & 0 & \cdots & 0 \\ 0 & 1-\alpha_1 & \alpha_1 & \cdots & 0 \\ 0 & 0 & 1-\alpha_2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \alpha_{n-1} \\ 0 & 0 & 0 & \cdots & 1-\alpha_n \end{pmatrix},
$$

and $\boldsymbol{I}$ denotes the identity matrix. If $\alpha_i \leq 1$ for $i = 0, ..., n$, this procedure is known as uniformization, originally introduced by Jensen (1953). Another method for obtaining Equation (10) is to use continued fractions (see Parthasarathy and Sudhesh 2006). However, the more details for computing $P_n(t)$ are intricate and cannot be discussed adequately here.

4. Experimental Results

This section shows results from two applications. The fertility data were analysed by Winkelmann (1995) and re-analysed by McShane et al. (2008), Chanialidis et al. (2018), and Kharrat et al. (2019). The takeover bids data were analysed by Jaggia and Thosar (1993) and re-analysed by Cameron and Johansson (1997), Saez-Castillo and Conde-Sanchez (2013), and Smith and Faddy (2016). For more information, the readers are referred to Winkelmann (1995) for the fertility data and Cameron and Johansson (1997) for the takeover bids data.

Experimental results obtained from the Poisson, gamma, Weibull, CMP, and Faddy (mean only) models are computed using the **stats** (R Core Team 2019), **Countr** (gamma and Weibull) (Kharrat and Boshnakov 2018), **COMPoissonReg** (Sellers et al. 2018), and **CountsEPPM** (Smith and Faddy 2018) **R** packages. The fertility and takeover bids datasets are available from the **Countr** and **mpcmp** (Fung et al. 2019) **R** packages, respectively. The UE models are implemented in **R** (R Core Team 2019) and C++. Most of the code is written in C++ via the **Rcpp** (Eddelbuettel et al. 2021) package in order to accelerate computations. The **expm** (Goulet et al. 2020) **R** package enables computation of the matrix exponential for calculating the probabilities of UE processes.

Table 1. *Shape parameter (α_γ), log-likelihood, BIC, and computation time (seconds) values of several SUE models for the fertility and takeover bids data.*

	Fertility data				Takeover bids data			
Model	α_γ	-Log-L	BIC	Time	α_γ	-Log-L	BIC	Time
Poisson	-	2101.8	4282.0	0.02	-	185.0	418.3	0.00
SUE ($\boldsymbol{\gamma} = \{0\}$)	1.46	2078.1	4241.8	0.59	2.96	171.3	395.8	0.05
SUE ($\boldsymbol{\gamma} = \{1\}$)	1.23	2096.0	4277.5	0.57	0.40	174.1	401.4	0.07
SUE ($\boldsymbol{\gamma} = \{2\}$)	0.52	2048.8	4183.1	0.66	1.00	185.0	423.1	0.00
SUE ($\boldsymbol{\gamma} = \{3\}$)	1.00	2101.8	4289.1	0.06	1.00	185.0	423.1	0.00
SUE ($\boldsymbol{\gamma} = \{4\}$)	1.00	2101.8	4289.1	0.05	1.00	185.0	423.1	0.00
SUE ($\boldsymbol{\gamma} = \{5\}$)	1.00	2101.8	4289.1	0.04	1.00	185.0	423.1	0.00

The rate between two consecutive events different from the base rate causes an excess (or a lack) of counts relative to a benchmark model such as the Poisson. This unusual event might be investigated by comparing the histogram of the sample and Poisson

Table 2. *Highest log-likelihood, lowest BIC, and computation time (seconds) values of each k-combination ($1 \leq k \leq 7$) UE regression model.*

Fertility data				Takeover bids data			
$\boldsymbol{\gamma}$	-Log-L	BIC	Time	$\boldsymbol{\gamma}$	-Log-L	BIC	Time
$\{2\}$	2048.8	4183.1	1.6	$\{0\}$	171.3	395.8	0.2
$\{2,3\}$	2040.1	4172.9	10.6	$\{1,2\}$	168.0	394.1	1.2
$\{2,3,4\}$	2037.3	4174.4	31.4	$\{1,2,3\}$	166.9	396.6	3.1
$\{2,3,4,7\}$	2035.8	4178.5	54.6	$\{1,2,3,4\}$	166.2	400.1	4.6
$\{1,2,3,4,7\}$	2034.7	4183.5	55.5	$\{1,2,3,4,5\}$	165.8	404.1	4.3
$\{1,2,3,4,6,7\}$	2034.6	4190.3	31.5	$\{1,2,3,4,5,6\}$	165.8	408.9	2.3
$\{0,1,2,3,4,6,7\}$	2034.3	4196.8	10.2	$\{1,2,3,4,5,6,7\}$	165.8	413.7	0.8

distributions. For example, Figure 5(a) contains an excess of two counts. The "excess two" phenomenon may arise in the situation that is the rate between the second and third events is less than others. In other words, the third event is unusual, and the SUE ($\boldsymbol{\gamma} = \{2\}$ and $\alpha_2 < 1$) model is preferred over other SUE models. The results in Table 1 show that the SUE ($\boldsymbol{\gamma} = \{2\}$) model has a higher log-likelihood and lower BIC values than other models. Therefore, we can conclude that the third event is the unusual event of the fertility data. Similarly, in Figure 5(b), this approach can be applied to the takeover bids data.

Visualizing the histograms can be a method for guessing unusual events, but it is hard to conclude which UE model is the best. Therefore, an exhaustive search is utilized for finding the best UE model because the number of UE models is limited. It is simple and guaranteed to find the best solution. We assume that λ_n's ($n > 7$) are equal to the base rate. For $\boldsymbol{\gamma} = \{0,1,...,7\}$, there are 255 different UE models to choose from using the combinations of all eight unusual events, 8 models (one and seven unusual events), 28 models (two and six unusual events), 56 models (three and five unusual events), and 70 models (four unusual events). Table 2 summarizes the highest log-likelihood, lowest BIC, and computation time values of the k-combination ($1 \leq k \leq 7$) models by fitting the UE regression models to the fertility and takeover bids data. For both datasets, as k increases, the log-likelihood increases monotonically. The BIC attains minimum at $k = 2$, and the DUE model is selected as the best model.

The fertility data, which consists of 10 covariates, are very slightly underdispersed with the variance-mean ratio equalling 2.328/2.384 = 0.977. The Poisson regression model is inappropriate because the mixture of conditional equidispersed distributions is always overdispersed. The gamma, Weibull, CMP, and Faddy (mean only) models display underdispersion. These regression models perhaps provide a good fit for the data because the mixture of conditional underdispersed distributions can be over-, under-, or equidispersion. The SUE ($\boldsymbol{\gamma} = \{2\}$) provides a much better fit to the data than the other

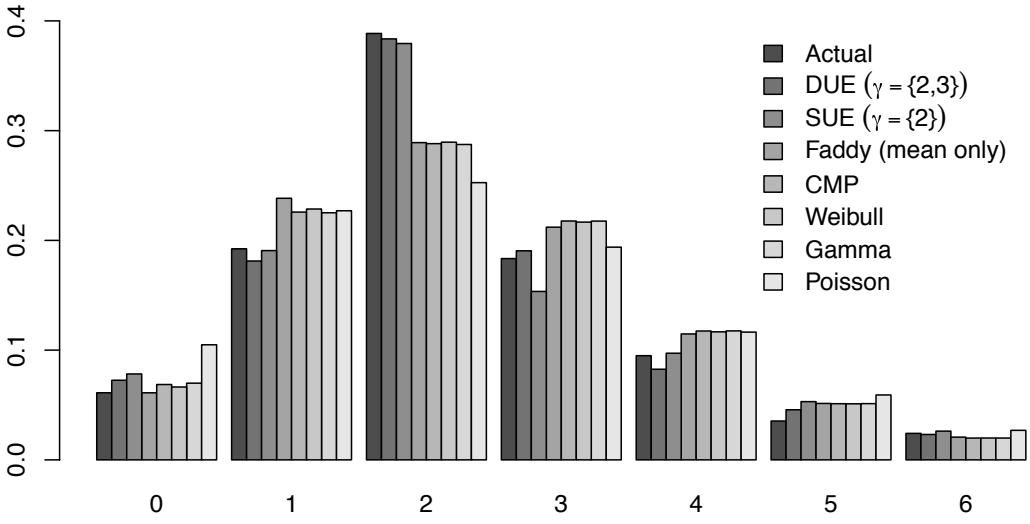
Table 3. *Variance, mean, variance-mean ratio, and BIC values of several regression models for the fertility and takeover bids data. The SUE ($\boldsymbol{\gamma} = \{2\}$) and DUE ($\boldsymbol{\gamma} = \{2,3\}$) models fit to the fertility data. The SUE ($\boldsymbol{\gamma} = \{0\}$) and DUE ($\boldsymbol{\gamma} = \{1,2\}$) models fit to the takeover bids data.*

	Fertility data				Takeover bids data			
Model	Variance	Mean	Ratio	BIC	Variance	Mean	Ratio	BIC
Sample	2.328	2.384	0.977	-	2.035	1.738	1.171	-
Poisson	2.742	2.382	1.151	4281.98	2.227	1.737	1.282	418.26
Gamma	2.175	2.383	0.913	4241.96	1.710	1.736	0.985	413.94
Weibull	2.157	2.383	0.905	4239.55	1.635	1.735	0.943	413.61
CMP	2.166	2.384	0.909	4241.25	1.657	1.738	0.954	413.92
Faddy	2.190	2.387	0.917	4244.23	1.463	1.727	0.847	401.64
SUE	2.512	2.386	1.053	4183.05	1.468	1.727	0.850	395.82
DUE	2.332	2.376	0.981	4172.87	2.142	1.740	1.231	394.12

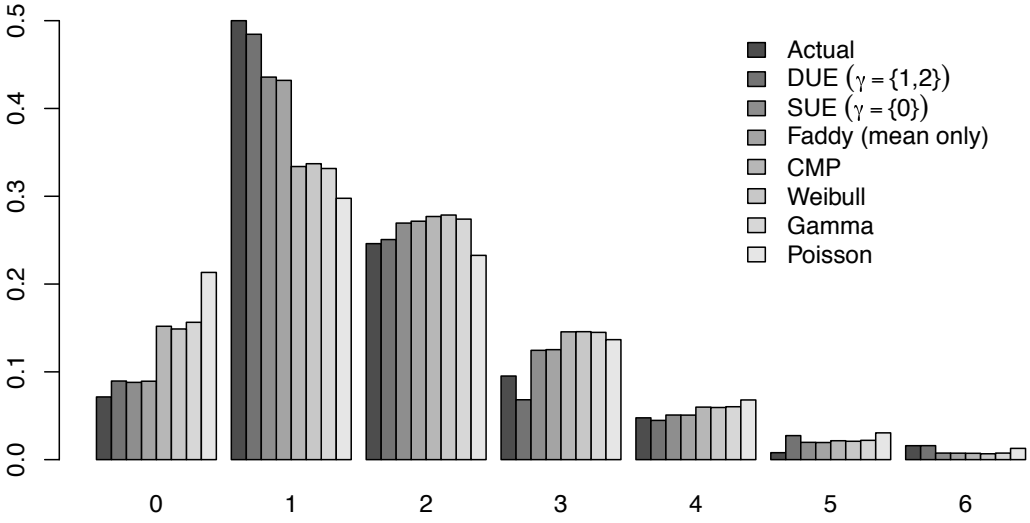
models, excluding the DUE, although its variance-mean ratio disagrees with the actual data (see Table 3). It means that the shape of the fertility data distribution resembles the SUE ($\boldsymbol{\gamma} = \{2\}$) more than the other models (see Figure 5(a)). However, the DUE ($\boldsymbol{\gamma} = \{2,3\}$) provides the best fit in terms of BIC to the data. The log-likelihood value of -2040.12 for this model with 13 parameters is much greater than -2048.77 from the SUE ($\boldsymbol{\gamma} = \{2\}$) model with 12 parameters. Because of the one additional parameter associated with a substantial increase in log-likelihood, the BIC value of 4172.87 is smaller than the SUE ($\boldsymbol{\gamma} = \{2\}$). Note that the fertility data distribution may be the combination of over-, under-, and equidispersed distributions, as described later.

For the takeover bids data, the variance-mean ratio is 2.035/1.738 = 1.171. Therefore, the data present overdispersion. The Poisson provides the worst fit in terms of BIC to the data even though it presents overdispersion as the data do (see Table 3). It interprets that the shape of the takeover bids data distribution resembles the Poisson less than the other models (see Figure 5(b)). The DUE ($\boldsymbol{\gamma} = \{1,2\}$) provides the best fit in terms of BIC to the data, and its variance-mean ratio agrees with the actual data (see Table 3). Note that the takeover bids data distribution may be the combination of over-, under-, and equidispersed distributions, as described later.

Figure 5 presents the sample and predicted probabilities evaluated at individual covariates for the Poisson, gamma, Weibull, CMP, Faddy (mean only), SUE, and DUE models. The fertility and takeover bits datasets contain an excess of two and one outcomes, respectively. It means there are more twos and ones in the two datasets than predicted by the Poisson, the gamma, etc. Figure 5(a) reveals that the models, excluding the SUE ($\boldsymbol{\gamma} = \{2\}$) and DUE ($\boldsymbol{\gamma} = \{2,3\}$) models, greatly underpredict the two outcomes because the third event is unusual. The SUE ($\boldsymbol{\gamma} = \{2\}$) model has the rate between the second and third event differs from others. However, the SUE ($\boldsymbol{\gamma} = \{2\}$) underpredicts the three outcomes because the fourth event is unusual. The DUE ($\boldsymbol{\gamma} = \{2,3\}$) has



(a) *Fertility data*



(b) *Takeover bids data*

Figure 5. *Sample and predicted relative frequency distributions.*

the consecutive rates between the second and fourth event differ from others. Thus, it leads to a considerable improvement of the predicted probabilities in the fertility case. Figure 5(b) shows that the models, excluding the DUE ($\boldsymbol{\gamma} = \{1,2\}$), underpredict the one outcome because the second and third events are unusual. The rate sequences of the SUE ($\boldsymbol{\gamma} = \{0\}$) and DUE ($\boldsymbol{\gamma} = \{1,2\}$) models are $2.962\lambda, \lambda, \lambda, \lambda, \lambda...$ and $\lambda, 0.314\lambda, 0.378\lambda, \lambda, \lambda...$, respectively. The unusual events of the SUE ($\boldsymbol{\gamma} = \{0\}$) and DUE ($\boldsymbol{\gamma} = \{1,2\}$) models disagree, but they have in common the fact that $\lambda_0 > \lambda_1$.

Tables 4 and 5 present the results from regressions for the number of children and number of bids data. The regression results from the gamma and Weibull models are produced by the "nlminb" function and the CMP, Faddy (mean only), SUE, and DUE models by the "optim" function with the "BFGS" method. The six models use the Poisson coefficients in Tables 4 and 5 as starting values of the unknown parameters $\beta_0, \beta_1, ..., \beta_r$, and the initial values of the other parameters are set to zero. The Poisson coefficients are perhaps the best initial guess for these models because the models generalize the Poisson. We note that the estimated parameters for the SUE and DUE regression models reported in Tables 4 and 5 are obtained by using the analytic solution approach. These values are identical to that produced by the matrix-exponential method, thus verifying the accuracy of the Taylor series expansion approach.

Comparing α in Table 4, these values in the gamma, Weibull, and CMP regression models are respectively 1.439, 1.236, and 1.429, which exceed one considerably, so there is an indication of underdispersion. The Faddy (mean only) also displays underdispersion because $\alpha = -0.129$. These four regression models with fixed α exhibit only one of over-, under-, and equidispersion. In other words, the dispersion types of these regression models depend only on α but not on λ. The SUE ($\boldsymbol{\gamma} = \{2\}$ and $\alpha = 0.521$) regression model displays overdispersion ($\lambda > 3.67$), underdispersion ($\lambda < 3.67$), and equidispersion ($\lambda = 3.67$) (see Figure 4(c)). The DUE ($\boldsymbol{\gamma} = \{2,3\}$, $\alpha_2 = 0.503$, and $\alpha_3 = 0.687$) regression model displays overdispersion ($\lambda > 4.31$), underdispersion ($\lambda < 4.31$), and equidispersion ($\lambda = 4.31$) (see Figure 4(g)). The dispersion types of the SUE ($\boldsymbol{\gamma} = \{2\}$) and DUE ($\boldsymbol{\gamma} = \{2,3\}$) regression models depend on α_2, α_3, and λ. It shows the flexibility of the SUE and DUE regression models to allow for over-, under-, and equidispersion, although the shape parameters are fixed. This property does not appear in the gamma, Weibull, CMP, and Faddy (1997) count models.

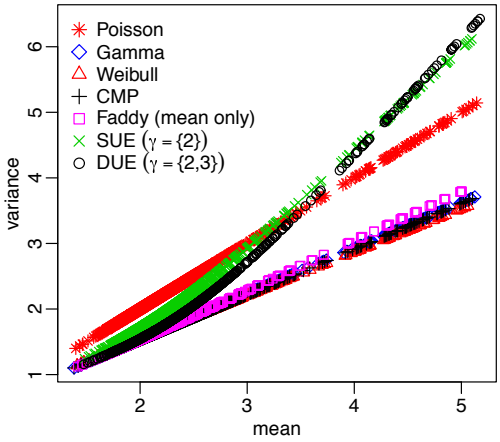
Figure 6 presents scatterplots of the fertility and takeover bids data. The dotted points are an ordered pair of the estimated mean and variance of each response variable produced by the seven models. The estimated mean and variance values of these models have to be determined numerically directly from their probability distributions using a suitable truncation (n). The points below and above the 45-degree (Poisson) line indicate underdispersion and overdispersion, respectively. In Figures 6(a) and 6(b), the SUE ($\boldsymbol{\gamma} = \{2\}$) curved (or the DUE ($\boldsymbol{\gamma} = \{2,3\}$) curved) and Poisson lines cut each other at a point, which is the estimated mean equals the estimated variance. The gamma, Weibull, CMP, and Faddy (mean only) lines are nearly coincident, indicating a similar ability of these four models to handle the fertility data. It is supported by the results in Table 4 that the log-likelihoods of these models are very similar. According to the SUE ($\boldsymbol{\gamma} = \{2\}$) and DUE ($\boldsymbol{\gamma} = \{2,3\}$) regression models, the fertility data are divided into two sets. The first set consists entirely of the underdispersed response variables, and the overdispersed response variables belong to the second. For the SUE ($\boldsymbol{\gamma} = \{2\}$), the first set (1151 members) is about 12.5 times bigger than the second set (92 members). For the DUE ($\boldsymbol{\gamma} = \{2,3\}$), the first set (1175 members) is about 17 times bigger than the second set (68 members). The gamma, Weibull, and CMP models in Figure 6(c) can

Table 4. *Regression model results for fertility data.*

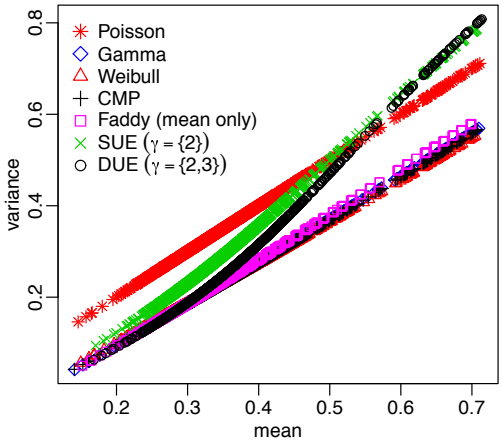
	Model													
	Poisson		Gamma		Weibull		CMP		Faddy (mean only)		SUE $\boldsymbol{\gamma} = \{2\}$		DUE $\boldsymbol{\gamma} = \{2,3\}$	
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	1.147	0.302	1.557	0.252	1.397	0.314	1.721	0.357	1.327	0.311	1.335	0.307	1.364	0.308
German	-0.200	0.072	-0.190	0.059	-0.223	0.072	-0.266	0.084	-0.197	0.073	-0.194	0.073	-0.209	0.073
Years of schooling	0.034	0.032	0.032	0.027	0.039	0.033	0.044	0.037	0.029	0.033	0.033	0.033	0.037	0.033
Vocational training	-0.153	0.044	-0.144	0.036	-0.173	0.044	-0.202	0.051	-0.177	0.044	-0.158	0.044	-0.166	0.044
University	-0.155	0.159	-0.146	0.130	-0.181	0.160	-0.207	0.182	-0.158	0.161	-0.136	0.162	-0.148	0.163
Catholic	0.218	0.071	0.206	0.058	0.242	0.070	0.289	0.082	0.241	0.071	0.212	0.071	0.221	0.072
Protestant	0.113	0.076	0.107	0.062	0.123	0.076	0.151	0.088	0.124	0.076	0.097	0.077	0.099	0.077
Muslim	0.548	0.085	0.523	0.070	0.639	0.087	0.742	0.103	0.624	0.087	0.547	0.087	0.572	0.087
Rural	0.059	0.038	0.055	0.031	0.068	0.038	0.078	0.044	0.067	0.038	0.062	0.039	0.070	0.039
Year of birth	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.000	0.002	0.001	0.002	0.001	0.002
Age at marriage	-0.030	0.007	-0.029	0.005	-0.034	0.006	-0.040	0.008	-0.035	0.007	-0.030	0.007	-0.031	0.007
ln α			0.364	0.049	0.212	0.027	0.357	0.047						
ln b									-2.317	1.600				
α									-0.129	0.064				
ln α_γ											-0.652	0.064	-0.687	0.064
ln $\alpha_{\gamma+1}$													-0.375	0.091
Log likelihood	-2101.80		-2078.23		-2077.02		-2077.88		-2075.80		-2048.77		-2040.12	
AIC (smaller is better)	4225.60		4180.45		4178.04		4179.75		4177.60		4121.54		4106.24	
BIC (smaller is better)	4281.98		4241.96		4239.55		4241.25		4244.23		4183.05		4172.87	
Time (seconds)	0.02		159.78		39.20		4.00		525.72		0.66		0.78	

Table 5. *Regression model results for takeover bids data.*

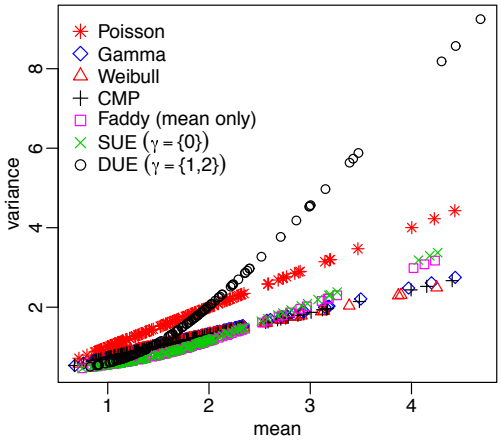
	Model													
	Poisson		Gamma		Weibull		CMP		Faddy (mean only)		SUE $\boldsymbol{\gamma} = \{0\}$		DUE $\boldsymbol{\gamma} = \{1,2\}$	
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	0.986	0.534	1.609	0.432	1.330	0.548	1.832	0.720	0.667	0.568	0.653	0.568	1.571	0.582
Leglrest	0.260	0.151	0.234	0.111	0.318	0.153	0.391	0.191	0.346	0.162	0.345	0.162	0.261	0.162
Rearest	-0.196	0.192	-0.168	0.142	-0.244	0.193	-0.307	0.240	-0.380	0.209	-0.385	0.209	-0.379	0.211
Finrest	0.074	0.217	0.072	0.160	0.043	0.217	0.113	0.268	0.015	0.229	0.016	0.229	0.000	0.231
Whtknght	0.481	0.159	0.430	0.117	0.568	0.162	0.710	0.210	0.664	0.175	0.664	0.176	0.572	0.172
Bidprem	-0.678	0.377	-0.616	0.278	-0.791	0.381	-1.009	0.477	-0.877	0.405	-0.876	0.406	-0.721	0.406
Insthold	-0.362	0.424	-0.323	0.313	-0.445	0.426	-0.546	0.521	-0.563	0.460	-0.569	0.461	-0.484	0.462
Size	0.179	0.060	0.164	0.045	0.218	0.062	0.283	0.082	0.251	0.066	0.251	0.065	0.194	0.064
Sizesq	-0.008	0.003	-0.007	0.002	-0.010	0.003	-0.012	0.004	-0.011	0.003	-0.011	0.003	-0.008	0.003
Regulatn	-0.030	0.161	-0.024	0.119	-0.042	0.160	-0.041	0.197	-0.038	0.169	-0.039	0.170	-0.029	0.172
ln α			0.544	0.161	0.331	0.093	0.551	0.152						
ln b									-29.72	42.67				
α									-0.036	0.051				
ln α_γ											1.086	0.218	-1.157	0.213
ln $\alpha_{\gamma+1}$													-0.973	0.287
Log likelihood	-184.95		-180.37		-180.21		-180.36		-171.80		-171.31		-168.04	
AIC (smaller is better)	389.90		382.74		382.41		382.72		367.61		364.62		360.08	
BIC (smaller is better)	418.26		413.94		413.61		413.92		401.64		395.82		394.12	
Time (seconds)	0.00		9.75		2.38		0.33		51.64		0.05		0.09	



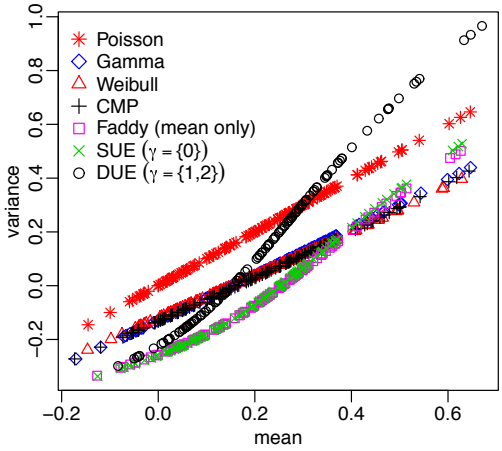
(a) *Linear scale of fertility data*



(b) *Base-10 log scale of fertility data*



(c) *Linear scale of takeover bids data*



(d) *Base-10 log scale of takeover bids data*

Figure 6. *Scatterplots of estimated variances versus estimated means.*

be interpreted similarly to Figure 6(a), but the Faddy (mean only) model is different. The Faddy (mean only) and SUE ($\boldsymbol{\gamma} = \{0\}$) curved lines lie nearly on top of each other because the rate sequence of the SUE ($\boldsymbol{\gamma} = \{0\}$) process is very similar to the Faddy (mean only) process. The rate sequences of the Faddy (mean only) and SUE ($\boldsymbol{\gamma} = \{0\}$) models are $2.92\lambda, 1.00\lambda, 0.98\lambda, 0.96\lambda, \ldots$ and $2.96\lambda, \lambda, \lambda, \lambda, \ldots$, respectively. They are non-increasing, and thus the Faddy (mean only) and SUE ($\boldsymbol{\gamma} = \{0\}$) curved lines do not cross the Poisson line. The DUE ($\boldsymbol{\gamma} = \{1,2\}$) curved line crosses the Poisson line, indicating the takeover bids data distribution is the combination of 90 underdispersed and 36 overdispersed distributions.

5. Conclusion

The Poisson, gamma, Weibull, CMP, and Faddy (1997) count models are well-known, but their underlying assumption of monotonic rate sequence limits their use in many applications. The UE count models, in contrast, are assumed that the rate sequences are non-monotonic, and the distributions of their interarrival times are exponential. One significant advantage of these new count models is the dispersion types defined by the base rate and the shape parameters. Hence, the UE count models can display over-, under-, and equidispersion, although the shape parameters are fixed numbers. In other words, the conditional variance and mean of the UE regression models must not be linearly related, allowing for a mixture of the over-, under-, and equidispersed distributions. The UE regression models are applied to the fertility and takeover bids data, and they offer significant improvements in log-likelihood compared to the above well-known regression models. For fertility data, the results show that the women's intentions to have third and fourth children, unusual events, are considerably less than other children. The behavior of these women cannot be captured by the above well-known count models with monotonic rates. Even though the UE count models offer significant improvements, future studies could improve the models for better results by replacing the exponential distribution with a non-negative distribution such as the gamma, the Weibull, etc.

Acknowledgments

The authors are grateful to the associate editor and anonymous referees for their valuable comments and suggestions.

Appendices

A. Derivations of SUE probability functions

A.1. SUE $(\boldsymbol{\gamma} = \{0\})$

Using Equations (3) and (8), the first few probabilities of the SUE $(\boldsymbol{\gamma} = \{0\})$ count model are obtained:

$$\begin{aligned} P_0(t) &= e^{-\alpha_0 \lambda t} \\ P_1(t) &= \int_0^t \alpha_0 \lambda P_0(u) e^{-\lambda(t-u)} du \\ &= \frac{\alpha_0 e^{-\lambda t}}{(1-\alpha_0)} \left(e^{(1-\alpha_0)\lambda t} - 1 \right) \end{aligned}$$

$$
\begin{aligned}
P_2(t) &= \int_0^t \lambda P_1(u) e^{-\lambda(t-u)} du \\
&= \frac{\alpha_0 e^{-\lambda t}}{(1-\alpha_0)^2} \left(e^{(1-\alpha_0)\lambda t} - 1 - (1-\alpha_0)\lambda t \right) \\
P_3(t) &= \int_0^t \lambda P_2(u) e^{-\lambda(t-u)} du \\
&= \frac{\alpha_0 e^{-\lambda t}}{(1-\alpha_0)^3} \left(e^{(1-\alpha_0)\lambda t} - 1 - (1-\alpha_0)\lambda t - \frac{((1-\alpha_0)\lambda t)^2}{2!} \right)
\end{aligned}
$$

From these equations for $P_0(t)$, $P_1(t)$, $P_2(t)$, and $P_3(t)$, one can deduce that the general SUE ($\boldsymbol{\gamma} = \{0\}$) probability function might be of the form

$$
P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \dfrac{\alpha_0 e^{-\lambda t}}{(1-\alpha_0)^n} \left(e^{(1-\alpha_0)\lambda t} - \displaystyle\sum_{i=0}^{n-1} \frac{((1-\alpha_0)\lambda t)^i}{i!} \right) & \text{for } n > 0. \end{cases} \tag{A.1}
$$

Inserting the Taylor series expansion of $e^{(1-\alpha_0)\lambda t}$, the SUE ($\boldsymbol{\gamma} = \{0\}$) probability distribution can be rewritten as

$$
P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \alpha_0 (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{((1-\alpha_0)\lambda t)^i}{(i+n)!} & \text{for } n > 0. \end{cases} \tag{A.2}
$$

A.2. SUE ($\boldsymbol{\gamma} = \{\gamma > 0\}$)

Using Equations (3) and (8), the first few probabilities of the SUE ($\boldsymbol{\gamma} = \{\gamma > 0\}$) count model are obtained:

$$
\begin{aligned}
P_0(t) &= e^{-\lambda t} \\
P_1(t) &= \int_0^t \lambda P_0(u) e^{-\lambda(t-u)} du = \lambda t e^{-\lambda t} \\
&\vdots \\
P_{\gamma-1}(t) &= \int_0^t \lambda P_{\gamma-2}(u) e^{-\lambda(t-u)} du = \frac{(\lambda t)^{\gamma-1} e^{-\lambda t}}{(\gamma-1)!} \\
P_\gamma(t) &= \int_0^t \lambda P_{\gamma-1}(u) e^{-\alpha_\gamma \lambda (t-u)} du \\
&= \frac{e^{-\lambda t}}{(1-\alpha_\gamma)^\gamma} \left(e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma-1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right)
\end{aligned}
$$

$$P_{\gamma+1}(t) = \int_0^t \alpha_\gamma \lambda P_\gamma(u) e^{-\lambda(t-u)} du$$

$$= \frac{\alpha_\gamma e^{-\lambda t}}{(1-\alpha_\gamma)^{\gamma+1}} \left(e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right)$$

From these equations for $P_0(t), P_1(t), ..., P_{\gamma-1}(t), P_\gamma(t)$, and $P_{\gamma+1}(t)$, one can deduce that the general SUE ($\boldsymbol{\gamma} = \{\gamma > 0\}$) probability function might be of the form

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < \gamma \\ \dfrac{e^{-\lambda t}}{(1-\alpha_\gamma)^n} \left(e^{(1-\alpha_\gamma)\lambda t} - \displaystyle\sum_{i=0}^{n-1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right) & \text{for } n = \gamma \\ \dfrac{\alpha_\gamma e^{-\lambda t}}{(1-\alpha_\gamma)^n} \left(e^{(1-\alpha_\gamma)\lambda t} - \displaystyle\sum_{i=0}^{n-1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right) & \text{for } n > \gamma. \end{cases} \tag{A.3}$$

Inserting the Taylor series expansion of $e^{(1-\alpha_\gamma)\lambda t}$, the SUE ($\boldsymbol{\gamma} = \{\gamma > 0\}$) probability distribution can be rewritten as

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < \gamma \\ (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{(i+n)!} & \text{for } n = \gamma \\ \alpha_\gamma (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{(i+n)!} & \text{for } n > \gamma. \end{cases} \tag{A.4}$$

B. Derivations of DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) Probability Functions

B.1. DUE ($\boldsymbol{\gamma} = \{0,1\}$)

Using Equations (5) and (8), the first few probabilities of the DUE ($\boldsymbol{\gamma} = \{0,1\}$) count model are obtained:

$$P_0(t) = e^{-\alpha_0 \lambda t}$$

$$P_1(t) = \int_0^t \alpha_0 \lambda P_0(u) e^{-\alpha_1 \lambda (t-u)} du$$

$$= \frac{\alpha_0 e^{-\lambda t}}{(\alpha_1 - \alpha_0)} \left(e^{(1-\alpha_0)\lambda t} - e^{(1-\alpha_1)\lambda t} \right)$$

$$P_2(t) = \int_0^t \alpha_1 \lambda P_1(u) e^{-\lambda(t-u)} du$$

$$= \frac{\alpha_0 \alpha_1 e^{-\lambda t}}{(\alpha_1 - \alpha_0)} \left(\frac{e^{(1-\alpha_0)\lambda t} - 1}{(1-\alpha_0)} - \frac{e^{(1-\alpha_1)\lambda t} - 1}{(1-\alpha_1)} \right)$$

$$P_3(t) = \int_0^t \lambda P_2(u) e^{-\lambda(t-u)} du$$

$$= \frac{\alpha_0 \alpha_1 e^{-\lambda t}}{(\alpha_1 - \alpha_0)} \left(\frac{e^{(1-\alpha_0)\lambda t} - 1 - (1-\alpha_0)\lambda t}{(1-\alpha_0)^2} - \frac{e^{(1-\alpha_1)\lambda t} - 1 - (1-\alpha_1)\lambda t}{(1-\alpha_1)^2} \right)$$

From these equations for $P_0(t)$, $P_1(t)$, $P_2(t)$, and $P_3(t)$, one can deduce that the general DUE ($\boldsymbol{\gamma} = \{0,1\}$) probability function might be of the form

$$P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \dfrac{\alpha_0 e^{-\lambda t}}{(\alpha_1 - \alpha_0)} \left(e^{(1-\alpha_0)\lambda t} - e^{(1-\alpha_1)\lambda t} \right) & \text{for } n = 1 \\ \dfrac{\alpha_0 \alpha_1 e^{-\lambda t}}{(\alpha_1 - \alpha_0)} \left(\dfrac{e^{(1-\alpha_0)\lambda t} - \sum_{i=0}^{n-2} \frac{((1-\alpha_0)\lambda t)^i}{i!}}{(1-\alpha_0)^{n-1}} \right. & \\ \qquad \left. - \dfrac{e^{(1-\alpha_1)\lambda t} - \sum_{i=0}^{n-2} \frac{((1-\alpha_1)\lambda t)^i}{i!}}{(1-\alpha_1)^{n-1}} \right) & \text{for } n > 1. \end{cases} \tag{B.1}$$

Inserting the Taylor series expansion of $e^{(1-\alpha_0)\lambda t}$ and $e^{(1-\alpha_1)\lambda t}$, the DUE ($\boldsymbol{\gamma} = \{0,1\}$) probability distribution can be rewritten as

$$P_n(t) = \begin{cases} e^{-\alpha_0 \lambda t} & \text{for } n = 0 \\ \alpha_0 (\lambda t) e^{-\lambda t} \sum_{i=0}^{\infty} \dfrac{c_i (\lambda t)^i}{(i+1)!} & \text{for } n = 1 \\ \alpha_0 \alpha_1 (\lambda t)^n e^{-\lambda t} \sum_{i=0}^{\infty} \dfrac{c_i (\lambda t)^i}{(i+n)!} & \text{for } n > 1, \end{cases} \tag{B.2}$$

where $c_i = \sum_{k=0}^{i} (1-\alpha_0)^k (1-\alpha_1)^{i-k}$.

B.2. DUE ($\boldsymbol{\gamma} = \{\gamma > 0, \gamma+1\}$)

Using Equations (5) and (8), the first few probabilities of the DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) count model are obtained:

$$
\begin{aligned}
P_0(t) &= e^{-\lambda t} \\
P_1(t) &= \int_0^t \lambda P_0(u) e^{-\lambda(t-u)} du = \lambda t e^{-\lambda t} \\
P_2(t) &= \int_0^t \lambda P_1(u) e^{-\lambda(t-u)} du = \frac{(\lambda t)^2 e^{-\lambda t}}{2!} \\
&\vdots \\
P_{\gamma-1}(t) &= \int_0^t \lambda P_{\gamma-2}(u) e^{-\lambda(t-u)} du = \frac{(\lambda t)^{\gamma-1} e^{-\lambda t}}{(\gamma-1)!} \\
P_{\gamma}(t) &= \int_0^t \lambda P_{\gamma-1}(u) e^{-\alpha_\gamma \lambda(t-u)} du \\
&= \frac{e^{-\lambda t}}{(1-\alpha_\gamma)^\gamma} \left(e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma-1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right) \\
P_{\gamma+1}(t) &= \int_0^t \alpha_\gamma \lambda P_\gamma(u) e^{-\alpha_{\gamma+1}\lambda(t-u)} du \\
&= \frac{\alpha_\gamma e^{-\lambda t}}{(\alpha_{\gamma+1}-\alpha_\gamma)} \left(\frac{e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma-1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!}}{(1-\alpha_\gamma)^\gamma} \right. \\
&\qquad \left. - \frac{e^{(1-\alpha_{\gamma+1})\lambda t} - \sum_{i=0}^{\gamma-1} \frac{\left((1-\alpha_{\gamma+1})\lambda t\right)^i}{i!}}{(1-\alpha_{\gamma+1})^\gamma} \right) \\
P_{\gamma+2}(t) &= \int_0^t \alpha_{\gamma+1} \lambda P_{\gamma+1}(u) e^{-\lambda(t-u)} du \\
&= \frac{\alpha_\gamma \alpha_{\gamma+1} e^{-\lambda t}}{(\alpha_{\gamma+1}-\alpha_\gamma)} \left(\frac{e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!}}{(1-\alpha_\gamma)^{\gamma+1}} \right. \\
&\qquad \left. - \frac{e^{(1-\alpha_{\gamma+1})\lambda t} - \sum_{i=0}^{\gamma} \frac{\left((1-\alpha_{\gamma+1})\lambda t\right)^i}{i!}}{(1-\alpha_{\gamma+1})^{\gamma+1}} \right)
\end{aligned}
$$

$$P_{\gamma+3}(t) = \int_0^t \lambda P_{\gamma+2}(u) e^{-\lambda(t-u)} du$$

$$= \frac{\alpha_\gamma \alpha_{\gamma+1} e^{-\lambda t}}{(\alpha_{\gamma+1} - \alpha_\gamma)} \left(\frac{e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{\gamma+1} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!}}{(1-\alpha_\gamma)^{\gamma+2}} \right.$$

$$\left. - \frac{e^{(1-\alpha_{\gamma+1})\lambda t} - \sum_{i=0}^{\gamma+1} \frac{\left((1-\alpha_{\gamma+1})\lambda t\right)^i}{i!}}{(1-\alpha_{\gamma+1})^{\gamma+2}} \right)$$

From these equations for $P_0(t)$, $P_1(t)$, ..., $P_{\gamma+1}(t)$, $P_{\gamma+2}(t)$, and $P_{\gamma+3}(t)$, one can deduce that the general DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) probability function might be of the form

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < \gamma \\ \dfrac{e^{-\lambda t}}{(1-\alpha_\gamma)^n} \left(e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{n-1} \dfrac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!} \right) & \text{for } n = \gamma \\ \dfrac{\alpha_\gamma e^{-\lambda t}}{(\alpha_{\gamma+1} - \alpha_\gamma)} \left(\dfrac{e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{n-2} \dfrac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!}}{(1-\alpha_\gamma)^{n-1}} \right. & \\ \left. - \dfrac{e^{(1-\alpha_{\gamma+1})\lambda t} - \sum_{i=0}^{n-2} \dfrac{\left((1-\alpha_{\gamma+1})\lambda t\right)^i}{i!}}{(1-\alpha_{\gamma+1})^{n-1}} \right) & \text{for } n = \gamma+1 \\ \dfrac{\alpha_\gamma \alpha_{\gamma+1} e^{-\lambda t}}{(\alpha_{\gamma+1} - \alpha_\gamma)} \left(\dfrac{e^{(1-\alpha_\gamma)\lambda t} - \sum_{i=0}^{n-2} \dfrac{\left((1-\alpha_\gamma)\lambda t\right)^i}{i!}}{(1-\alpha_\gamma)^{n-1}} \right. & \\ \left. - \dfrac{e^{(1-\alpha_{\gamma+1})\lambda t} - \sum_{i=0}^{n-2} \dfrac{\left((1-\alpha_{\gamma+1})\lambda t\right)^i}{i!}}{(1-\alpha_{\gamma+1})^{n-1}} \right) & \text{for } n > \gamma+1. \end{cases} \tag{B.3}$$

Inserting the Taylor series expansion of $e^{(1-\alpha_\gamma)\lambda t}$ and $e^{(1-\alpha_{\gamma+1})\lambda t}$, the DUE ($\boldsymbol{\gamma} = \{\gamma, \gamma+1\}$) probability distribution can be rewritten as

$$P_n(t) = \begin{cases} \dfrac{(\lambda t)^n e^{-\lambda t}}{n!} & \text{for } n < \gamma \\ (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{\left((1-\alpha_\gamma)\lambda t\right)^i}{(i+n)!} & \text{for } n = \gamma \\ \alpha_\gamma (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{c_i (\lambda t)^i}{(i+n)!} & \text{for } n = \gamma+1 \\ \alpha_\gamma \alpha_{\gamma+1} (\lambda t)^n e^{-\lambda t} \displaystyle\sum_{i=0}^{\infty} \frac{c_i (\lambda t)^i}{(i+n)!} & \text{for } n > \gamma+1, \end{cases} \tag{B.4}$$

where $c_i = \displaystyle\sum_{k=0}^{i} \left(1-\alpha_\gamma\right)^k \left(1-\alpha_{\gamma+1}\right)^{i-k}$.

References

Ball, F. (1995). A note on variation in birth processes. *The Mathematical Scientist* 20, 50–55.

Bartlett, M. S. (1978). *An Introduction to Stochastic Processes*. Cambridge University Press, Cambridge, UK.

Cameron, A. C. and Johansson, P. (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics* 12(3), 203–233.

Chanialidis, C., Evers, L., Neocleous, T., and Nobile, A. (2018). Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing* 28, 595–608.

Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering* 12, 132–136.

Cox, D. R. and Miller, H.D. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London, UK.

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *WIREs Computational Statistics* 10(2), 1–22.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Bates, D., and Chambers, J. (2021). Rcpp: Seamless R and C++ integration. R package version 1.0.6.

Faddy, M. J. (1994). On variation in Poisson processes. *The Mathematical Scientist* 19, 47–51.

Faddy, M. J. (1997). Extended Poisson process modelling and analysis of count data. *Biometrical Journal* 39(4), 431–440.

Faddy, M. J. and Smith, D. M. (2008). Extended Poisson process modelling of dilution series data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 57(4), 461–471.

Faddy, M. J. and Smith, D. M. (2011). Analysis of count data with covariate dependence in both mean and variance. *Journal of Applied Statistics* 38(12), 2683–2694.

Fung, T., Alwan, A., Wishart, J., and Huang, A. (2019). mpcmp: Mean-parametrized Conway-Maxwell Poisson (COM-Poisson) regression. R package version 0.1.3.

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (2020). expm: Matrix exponential, log, 'etc'. R package version 0.999-5.

Jaggia, S. and Thosar, S. (1993). Multiple bids as a consequence of target management resistance: a count data approach. *Review of Quantitative Finance & Accounting* 3(4), 447–457.

Jensen, A. (1953). Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal* sup1, 87–91.

Kharrat, T. and Boshnakov, G. N. (2018). Countr: Flexible univariate count models based on renewal processes. R package version 3.5.2.

Kharrat, T., Boshnakov, G. N., McHale, I., and Baker, R. (2019). Flexible regression models for count data based on renewal processes. *Journal of Statistical Software* 90(13), 1–35.

McShane, B., Adrian, M., Bradlow, E. T., and Fader, P. S. (2008). Count models based on Weibull interarrival times. *Journal of Business & Economic Statistics* 26(3), 369–378.

Parthasarathy, P. R. and Sudhesh, R. (2006). Exact transient solution of a state-dependent birth-death process. *Journal of Applied Mathematics and Stochastic Analysis* 82(6), 1–16.

Podlich, H. M., Faddy, M. J., and Smyth, G. K.(2004). Semi-parametric extended Poisson process models for count data. *Statistics and Computing* 14, 311–321.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ross, S. M. (2010). *Introduction to Probability Models*, 10th Ed. Academic Press, San Diego, CA, USA.

Saez-Castillo, A. J. and Conde-Sanchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data. *Journal of Computational Statistics & Data Analysis* 61, 148–157.

Sellers, K. F., Lotze, T., and Raim, A. M. (2018). COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) regression. R package version 0.6.1.

Sellers, K. F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics* 4(2), 943–961.

Smith, D. M. and Faddy, M. J. (2016). Mean and variance modelling of under- and overdispersed count data. *Journal of Statistical Software* 69(6), 1–23.

Smith, D. M. and Faddy, M. J. (2018). CountsEPPM: Mean and variance modeling of count data. R package version 3.0.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics* 13(4), 467–474.

SORT 46 (1) January-June 2022, 67-94 DOI: 10.2436/20.8080.02.118

Estimation of finite population distribution function with auxiliary information in a complex survey sampling

Mohsin Abbas[†] and Abdul Haq[†], *

Abstract

In this paper, we consider the problem of estimating the finite population cumulative distribution function (CDF) in a complex survey sampling, which includes two-stage and three-stage cluster sampling schemes with and without stratification. We propose two new families of CDF estimators using supplementary information on a single auxiliary variable. Explicit mathematical expressions of the biases and mean squared errors of the proposed CDF estimators are developed under the first order of the approximation. Real datasets are also considered to support the proposed theory.

MSC: *62D05, 62F10.*

Keywords: *Ratio estimator, exponential ratio estimator, auxiliary information, stratification, two-stage and three-stage cluster sampling, relative efficiencies, bias, mean-squared error.*

1. Introduction

An important problem in the inferential statistics is to estimate the cumulative distribution function (CDF) of a finite population. This problem frequently arises when the underlying interest is to determine the proportion of values of a study variable that are less than or equal to a certain value. For instance, for a nutritionist, it is important to know the proportion of a population that consumes 25% or less of the calories from a saturated fat. Likewise, the policy makers, in a developing country, are mostly interested in knowing the proportion of people living below the poverty line. In the context of survey sampling, it is common to develop CDF estimators with different sampling schemes,

* Corresponding author. E-mail address: aaabdulhaq@yahoo.com

† Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.

Received: July 2021
Accepted: April 2022

which include simple random sampling (SRS), stratified random sampling, cluster sampling (CS), ranked-set sampling, to name a few. For more details, see Francisco and Fuller (1986), Haq (2017a), Stokes and Sager (1988) and the references cited therein.

A common approach in survey sampling is to increase the precision of an estimator with suitable use of auxiliary information. The ratio, regression and product-type estimators are prime examples as these estimators require supplementary information on one or more auxiliary variables along with the information on a study variable to increase their relative efficiencies. For example, when estimating the total household income, the age and total expenditure may be used as two auxiliary variables. A significant amount of research work has been done in the literature of survey sampling to develop new improved estimators of the population parameters, which include the population mean, total, CDF, median, etc. Here, our focus is on the estimation of the finite population CDF with the auxiliary information. Chambers and Dunstan (1986) considered estimation of the population CDF and quantiles with the model-based approach. On similar lines, Rao, Kovar and Mantel (1990) proposed ratio and difference/regression estimators for estimating the CDF under a general sampling scheme. Singh, Singh and Kozak (2008) considered the problem of estimating the CDF and quantiles with the use of auxiliary information at the estimation stage of a survey. To our knowledge, recent works on the CDF estimation with auxiliary information may be seen in Tarima and Pavlov (2006), Martínez et al. (2010), Berger and Muñoz (2015), Mayor-Gallego, Moreno-Rebollo and Jiménez-Gamero. (2019), Hussain et al. (2020), Yaqub and Shabbir (2020) and Martínez, Rueda and Illescas (2022), to name a few.

In survey sampling, when the available population is in the form of clusters, that is, households in villages and their members, then it is useful to employ CS instead of SRS. In CS, clusters are randomly selected (with a sampling scheme) from a population, and the data pertaining to a study variable are then collected from all of the units of the selected cluster. However, CS is less efficient than SRS when estimating a population parameter and the former restricts the spread of sampling units across the population. One possible solution is to increase the number of clusters in the sample, and then select representative samples via a sampling scheme from the sampled clusters. This sampling scheme has two stages. It is thus called two-stage CS (2SCS), where the first-stage and second-stage units are called primary stage units (PSUs) and secondary stage units (SSUs), respectively. The 2SCS method is an improvement over CS when it may not be possible or difficult to enumerate all the units of the selected clusters, thereby reducing the cost of the survey. A natural extension of a 2SCS is a three-stage CS (3SCS), where third-stage units are called tertiary stage units (TSUs). This scheme is adopted for inpatients' care cost estimation, where hospitals are selected at the first stage, the selection of wards at the second stage, and the patients at the third stage. Moreover, in large-scale health and demographic surveys, where the population is not only heterogeneous but also more graphically spread, both 2SCS and 3SCS schemes may be combined with the stratified random sampling to get more representative samples, where the stratifying variable may be regions, rural and urban, plan and hilly regions, agro-climatic zones,

etc. For more details see, Cochran (1977), Deville and Särndal (1992), Hansen and Hurwitz (1943), Lee, Lee and Shin (2016), Murthy (1967), Nafiu, Oshungade and Adewara (2012), Rustagi (1978) and references cited therein.

In the survey sampling literature, several authors have considered estimation of the population parameters under 2S and 2SCS schemes. Sukhatme et al. (1984) and Sahoo (1987) considered the estimation of the finite population mean using regression-type estimators in 2S sampling. Smith (1969) studied the ratio estimator for estimation of the finite population mean under multi-stage sampling. Särndal, Swensson and Wretman (2003) considered a regression estimator using 2S sampling under a variety of options. In another study, Nematollahi, Salehi and Aliakbari (2008) developed a new estimator of the population mean using 2SCS, where ranked-set sampling (RSS) was considered in the secondary sampling frame. Srivastava and Garg (2009) used multi-auxiliary information for estimating the population mean in 2S sampling, and they proposed separate-type general class of estimators. Following Nematollahi et al. (2008), Haq (2017b) has considered a hybrid RSS scheme in the secondary sampling frame for developing an improved estimator of the population mean in 2SCS. Recently, Haq, Abbas and Khan (2021) have considered estimation of the finite population CDF under a complex survey sampling scheme, which includes 2SCS, 3SCS, stratified 2SCS (S2SCS) and stratified 3SCS (S3SCS). Under these sampling schemes, they have derived unbiased CDF estimators along with their variances, and the unbiased estimators of the variances of these CDF estimators.

In this study, on the lines of Haq et al. (2021), we consider estimation of the finite population CDF with auxiliary information under 2SCS/3SCS and S2SCS/S3SCS schemes. Following the works of Khoshnevisan et al. (2007) and Singh et al. (2009), we propose two families of classical ratio/product and exponential ratio/product-type estimators for estimating the population CDF under the aforementioned sampling schemes. Moreover, on the lines of Sukhatme et al. (1984) and Sahoo (1987), regression/difference estimators CDF are also developed. Explicit mathematical expressions are obtained for the biases and mean squared errors (MSEs) of the proposed estimators. Real datasets are also considered for the application of the proposed estimators.

The rest of the paper is as follows: In Section 2, CDF estimation is reviewed under 2SCS and 3SCS schemes. In Section 3, we develop explicit mathematical expressions for the covariances of the CDF estimators based on 2SCS/3SCS and S2SCS/S3SCS. In addition, the unbiased estimators of the covariances of the CDF estimators are also derived. In Section 4, two families of estimators, say ratio/product and exponential ratio/product, are proposed for estimating the population CDF. An empirical study is conducted in Section 5. Finally, Section 6 summarizes the main findings and concludes the paper.

2. Estimation of the population CDF

In this section, we briefly review the CDF estimators under 2SCS/S2SCS and 3SCS/S3SCS, which will be used in the subsequent sections.

2.1. Two-stage cluster sampling

The 2SCS uses two stages to select a sample. Assume that the target population, denoted by U, comprises N PSUs, where the ith PSU contains M_i SSUs for $i = 1,2,\ldots,N$. Let $Y_{i,j}$ denote the jth SSU that is present in the ith PSU, where $j = 1,2,\ldots,M_i$ with M_i being the total number of SSUs within the ith PSU. Under 2SCS, the population CDF, $F(y)$, may be written as

$$F(y) \;=\; \frac{1}{N\overline{M}}\sum_{i=1}^{N} M_i F_i(y), \tag{1}$$

where

$$\overline{M} \;=\; \frac{1}{N}\sum_{i=1}^{N} M_i \quad \text{and} \quad F_i(y) \;=\; \frac{1}{M_i}\sum_{j=1}^{M_i} I(Y_{i,j} \le y)$$

are the average cluster size and the CDF computed from the ith PSU, respectively.

In order to estimate $F(y)$ under 2SCS, let n denote the number of PSUs selected in the first stage, and let m_i be the number of SSUs selected from the ith PSU. It is to be noted that, with the 2SCS scheme, the samples under both stages are selected using SRS without replacement. An estimator of $F(y)$ under 2SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{2S}(y) = \frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i \hat{F}_i(y) \;=\; \frac{1}{n\overline{M}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} I(Y_{i,j} \le y), \tag{2}$$

where $I(\cdot)$ is an indicator variable. It can be shown that $\hat{F}_{2S}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{2S}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{2S}(y)) = \frac{\lambda\,\sigma^2_{Y,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\zeta_i M_i^2 \sigma^2_{Y,2i}}{m_i} \quad \text{and} \tag{3}$$

$$\widehat{V}(\hat{F}_{2S}(y)) = \frac{\lambda\,\hat{\sigma}^2_{Y,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\zeta_i M_i^2 \hat{\sigma}^2_{Y,2i}}{m_i}, \tag{4}$$

respectively, where

$$\sigma^2_{Y,2b} \;=\; \frac{1}{N-1}\sum_{i=1}^{N}\big(M_iF_i(y) - \overline{M}F(y)\big)^2,\;\; \sigma^2_{Y,2i} \;=\; F_i(y)(1-F_i(y)),$$

$$\hat{\sigma}^2_{Y,2b} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\big(M_i\hat{F}_i(y) - \overline{M}\hat{F}_{2S}(y)\big)^2, \hat{\sigma}^2_{Y,2i} \;=\; \frac{M_i(m_i-1)}{m_i(M_i-1)}\hat{F}_i(y)(1-\hat{F}_i(y)),$$

$$\lambda \;=\; \left(1-\frac{n}{N}\right), \text{and } \zeta_i \;=\; \frac{(M_i-m_i)}{(M_i-1)}.$$

In an 2SCS scheme, two types of variations may be considered. The first is the variation between the clusters, and the second is the variation within the clusters. In 2SCS, $\sigma^2_{Y,2b}$ denotes the variance between clusters and $\sigma^2_{Y,2i}$ denotes the variance within the ith cluster. Moreover, $\hat{\sigma}^2_{Y,2i}$ is an unbiased estimator of $\sigma^2_{Y,2i}$.

2.2. Three-stage cluster sampling

The 3SCS requires samples to be selected in three different stages. In the first stage, samples are selected from the PSUs; in the second stage, samples are selected from the SSUs of the selected PSUs; and, in the third stage, the tertiary units are selected from the selected SSUs. Similar to 2SCS, the SRS scheme may be used to select samples at three different stages of the 3SCS.

Suppose that the target population U consists of N PSUs, where each PSU contains M_i SSUs, and each SSU has T_{ij} TSUs. Let $Y_{ij,k}$ denote the kth TSU with the jth SSU of the ith PSU, where $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M_i$, and $k = 1,2,\ldots,T_{ij}$. Under 3SCS, the population CDF, $F(y)$, may be written as

$$F(y) \;=\; \frac{1}{N\overline{T}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} T_{ij} F_{ij}(y), \tag{5}$$

where

$$\overline{T} \;=\; \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} T_{ij} \quad \text{and} \quad F_{ij}(y) \;=\; \frac{1}{T_{ij}} \sum_{k=1}^{T_{ij}} I(Y_{ij,k} \le y).$$

Here, $\overline{T}$ denotes the average cluster size and $F_{ij}(y)$ be the CDF computed from the jth SSU of the ith PSU.

In order to estimate $F(y)$ under 3SCS, let n denote the number of PSUs selected in the first-stage, let m_i be the number of SSUs selected from the ith PSU, and let t_{ij} be the number of tertiary units selected from the jth SSU. An estimator of $F(y)$ under 3SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{3S}(y) = \frac{1}{n\overline{T}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{T_{ij}}{t_{ij}} \sum_{k=1}^{t_{ij}} I(Y_{ij,k} \le y) \;=\; \frac{1}{n\overline{T}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} T_{ij} \hat{F}_{ij}(y) \tag{6}$$

It can be shown that $\hat{F}_{3S}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{3S}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{3S}(y)) = \frac{\lambda \sigma^2_{Y,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{N} \frac{\lambda_i M_i^2 \sigma^2_{Y,3i}}{m_i} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{N} \frac{M_i}{m_i} \sum_{j=1}^{M_i} \frac{\zeta_{ij} T_{ij}^2 \sigma^2_{Y,3ij}}{t_{ij}} \quad \text{and} \tag{7}$$

$$\widehat{V}(\hat{F}_{3S}(y)) = \frac{\lambda \hat{\sigma}^2_{Y,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{n} \frac{\lambda_i M_i^2 \hat{\sigma}^2_{Y,3i}}{m_i} + \frac{1}{nN\overline{T}^2} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} \frac{\zeta_{ij} T_{ij}^2 \hat{\sigma}^2_{Y,3ij}}{t_{ij}}, \tag{8}$$

respectively, where

$$\sigma^2_{Y,3b} = \frac{1}{N-1}\sum_{i=1}^{N}\left(M_iF_i(y)-\overline{T}F(y)\right)^2, \hat{\sigma}^2_{Y,3b} = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)-\overline{T}\hat{F}_{3S}(y)\right)^2,$$

$$\sigma^2_{Y,3i} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(T_{ij}F_{ij}(y)-F_i(y)\right)^2, \hat{\sigma}^2_{Y,3i} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(T_{ij}\hat{F}_{ij}(y)-\hat{F}_i(y)\right)^2,$$

$$\sigma^2_{Y,3ij} = F_{ij}(y)(1-F_{ij}(y)), \qquad \hat{\sigma}^2_{Y,3ij} = \frac{t_{ij}(T_{ij}-1)}{T_{ij}(t_ij-1)}\hat{F}_{ij}(y)(1-\hat{F}_{ij}(y)),,$$

$$F_i(y) = \frac{1}{M_i}\sum_{j=1}^{M_i}T_{ij}F_{ij}(y), \qquad \hat{F}_i(y) = \frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}\hat{F}_{ij}(y),$$

$$\lambda = \left(1-\frac{n}{N}\right), \lambda_i = \left(1-\frac{m_i}{M_i}\right), \quad \zeta_{ij} = \frac{T_{ij}-t_{ij}}{T_{ij}-1},$$

where $\sigma^2_{Y,3b}$, $\sigma^2_{Y,3i}$ and $\sigma^2_{Y,3ij}$ have their usual meanings. Moreover, $\hat{\sigma}^2_{Y,3ij}$ is an unbiased estimator of $\sigma^2_{Y,3ij}$. But, $\hat{\sigma}^2_{Y,3b}$ and $\hat{\sigma}^2_{Y,3i}$ are biased estimators of $\sigma^2_{Y,3b}$ and $\sigma^2_{Y,3i}$, respectively. For more detail, see Haq et al. (2021).

2.3. Stratified two-stage cluster sampling

Suppose that the target population Y may be partitioned into L strata, where the hth stratum contains N_h units for $h = 1,2,\ldots,L$. In addition, there are N_h PSUs within the hth stratum, where the ith PSU contains $M_{i,h}$ SSUs for $i = 1,2,\ldots,N_h$. Let $Y_{i,j,h}$ denote the jth SSU that is present in the ith PSU of the hth stratum, where $j = 1,2,\ldots,M_{i,h}$ with $M_{i,h}$ be the total number of SSUs within the ith PSU. Then the population CDF, $F(y)$, under S2SCS, may be written as

$$F(y) = \sum_{h=1}^{L}W_hF_h(y) = \frac{1}{\sum_{h=1}^{L}N_h\overline{M}_h}\sum_{h=1}^{L}N_h\overline{M}_hF_h(y), \tag{9}$$

where

$$W_h = \frac{N_h\overline{M}_h}{\sum_{h=1}^{L}N_h\overline{M}_h}, \qquad F_h(y) = \frac{1}{N_h\overline{M}_h}\sum_{i=1}^{N_h}M_{i,h}F_{i,h}(y),$$

$$F_{i,h}(y) = \frac{1}{M_{i,h}}\sum_{j=1}^{M_{i,h}}I(Y_{i,j,h}\le y), \quad \overline{M}_h = \frac{1}{N_h}\sum_{i=1}^{N_h}M_{i,h}, \tag{10}$$

are computed for the hth stratum.

In order to estimate $F(y)$ under S2SCS, a two-stage cluster sample of size n_h is selected from the hth stratum, where the sample sizes n_h may be allocated using an allocation scheme, like proportional, equal or Neyman allocation. An estimator of $F(y)$ under S2SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{S2S}(y) = \sum_{h=1}^{L}W_h\hat{F}_{2S,h}(y), \tag{11}$$

where

$$\hat{F}_{2S,h}(y) = \frac{1}{n_h \overline{M}_h} \sum_{i=1}^{n_h} M_{i,h} \hat{F}_{i,h}(y) \quad \text{and} \tag{12}$$

$$\hat{F}_{i,h}(y) = \frac{1}{m_{i,h}} \sum_{j=1}^{m_{i,h}} I(Y_{i,j,h} \leq y).$$

It can be shown that $\hat{F}_{S2S}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{S2S}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{S2S}(y)) = \sum_{h=1}^{L} W_h^2 V(\hat{F}_{2S,h}(y)) \quad \text{and} \tag{13}$$

$$\widehat{V}(\hat{F}_{S2S}(y)) = \sum_{h=1}^{L} W_h^2 \widehat{V}(\hat{F}_{2S,h}(y)), \tag{14}$$

respectively. Note that the mathematical expressions of $V(\hat{F}_{2S,h}(y))$ and $\widehat{V}(\hat{F}_{2S,h}(y))$ (given in Eqs. (3) and (4)) are similar to $V(\hat{F}_{2S}(y))$ and $\widehat{V}(\hat{F}_{2S}(y))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1, 2, \ldots, L$.

2.4. Stratified three-stage cluster sampling

Suppose that the target population U is partitioned into L strata, where the hth stratum contains N_h units for $h = 1, 2, \ldots, L$. In addition, there are N_h PSUs in the hth stratum, where the ith PSU contains $M_{i,h}$ SSUs for $i = 1, 2, \ldots, N_h$. Moreover, each SSU contain $T_{ij,h}$ TSUs for $j = 1, 2, \ldots, M_{i,h}$. Let $Y_{ij,k,h}$ denote the kth TSU that is present in the jth SSU of the ith PSU within the hth stratum, where $k = 1, 2, \ldots, T_{ij,h}$, and $T_{ij,h}$ be the total number of TSUs within the jth SSU of the ith PSU. Then the population CDF, $F(y)$, under S3SCS, may be written as

$$F(y) \;=\; \sum_{h=1}^{L} W_h F_h(y) \;=\; \frac{1}{\sum_{h=1}^{L} N_h \overline{T}_h} \sum_{h=1}^{L} N_h \overline{T}_h F_h(y), \tag{15}$$

where

$$W_h \;= \frac{N_h \overline{T}_h}{\sum_{h=1}^{L} N_h \overline{T}_h}, \qquad F_h(y) \;=\; \frac{1}{N_h \overline{T}_h} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{i,h}} T_{ij,h} F_{ij,h}(y),$$

$$F_{ij,h}(y) \;= \frac{1}{T_{ij,h}} \sum_{k=1}^{T_{ij,h}} I(Y_{ij,k,h} \leq y), \quad \overline{T}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{i,h}} T_{ij,h}. \tag{16}$$

are computed for the hth stratum.

In order to estimate $F(y)$ with S3SCS, a stratified three-stage cluster sample of size n_h is selected from the hth stratum, where the sample size n_h may be allocated with an

allocation scheme, like equal, proportional or Neyman allocation. An estimator of $F(y)$ under S3SCS, developed by Haq et al. (2021), is given by

$$\hat{F}_{\mathrm{S3S}}(y) \;=\; \sum_{h=1}^{L} W_h \hat{F}_{\mathrm{3S},h}(y), \tag{17}$$

where

$$\begin{aligned}\hat{F}_{\mathrm{3S},h}(y) &= \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} \frac{M_{i,h}}{m_{i,h}} \sum_{j=1}^{m_{i,h}} \frac{T_{ij,h}}{t_{ij,h}} \sum_{k=1}^{t_{ij,h}} I(Y_{ij,k,h} \le y), \\ &= \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} \frac{M_{i,h}}{m_{i,h}} \sum_{j=1}^{m_{i,h}} T_{ij,h} \hat{F}_{ij,h}(y) \;=\; \frac{1}{n_h \overline{T}_h} \sum_{i=1}^{n_h} M_{i,h} \hat{F}_{i,h}(y), \end{aligned} \tag{18}$$

and

$$\hat{F}_{ij,h}(y) \;=\; \frac{1}{t_{ij,h}} \sum_{k=1}^{t_{ij,h}} I(Y_{ij,k,h} \le y), \quad \hat{F}_{i,h}(y) \;=\; \frac{1}{m_{i,h}} \sum_{j=1}^{m_{i,h}} T_{ij,h} \hat{F}_{ij,h}(y). \tag{19}$$

It can be shown that $\hat{F}_{\mathrm{S3S}}(y)$ is an unbiased estimator of $F(y)$. The variance of $\hat{F}_{\mathrm{S3S}}(y)$ along with its unbiased estimator are given by

$$V(\hat{F}_{\mathrm{S3S}}(y)) = \sum_{h=1}^{L} W_h^2 V(\hat{F}_{\mathrm{3S},h}(y)) \quad \text{and} \tag{20}$$

$$\widehat{V}(\hat{F}_{\mathrm{S3S}}(y)) = \sum_{h=1}^{L} W_h^2 \widehat{V}(\hat{F}_{\mathrm{3S},h}(y)), \tag{21}$$

respectively. Note that the mathematical expressions of $V(\hat{F}_{\mathrm{3S},h}(y))$ and $\widehat{V}(\hat{F}_{\mathrm{3S},h}(y))$ (given in Eqs. (3) and (4)) are similar to $V(\hat{F}_{\mathrm{3S}}(y))$ and $\widehat{V}(\hat{F}_{\mathrm{3S}}(y))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1,2,\ldots,L$, which can be found in Haq et al. (2021).

3. Covariance computation and estimation under a complex survey sampling

In this section, we develop explicit mathematical expressions for the covariances of the CDF estimators based on aforementioned complex survey sampling schemes. In addition, the unbiased estimators of these covariances of the CDF estimators are also derived, which may be used to develop regression-type estimators of the population CDF.

3.1. Two-stage and stratified two-stage cluster sampling

Let Y be the study variable and let X be an auxiliary variable in a finite population U. In order to estimate $(F(y),F(x))$ under 2SCS and S2SCS, let $(\hat{F}_{\mathrm{2S}}(y),\hat{F}_{\mathrm{2S}}(x))$ and $(\hat{F}_{\mathrm{S2S}}(y),\hat{F}_{\mathrm{S2S}}(x))$ be the respective CDF estimators that are based on (Y,X), respectively.

Lemma 1. *Under 2SCS scheme, the covariance between $\hat{F}_{2S}(y)$ and $\hat{F}_{2S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) = \frac{\lambda\sigma_{XY,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} \quad \textit{and} \tag{22}$$

$$\widehat{C}(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) = \frac{\lambda\hat{\sigma}_{XY,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,2i}}{m_i}, \tag{23}$$

respectively, where

$$\sigma_{XY,2b} = \frac{1}{N-1}\sum_{i=1}^{N}\Big((M_iF_i(y)-\overline{M}F(y))(M_iF_i(x)-\overline{M}F(x))\Big), \tag{24}$$

$$\hat{\sigma}_{XY,2b} = \frac{1}{n-1}\sum_{i=1}^{n}\left((M_i\hat{F}_i(y)-\overline{M}\hat{F}_{2S}(y))(M_i\hat{F}_i(x)-\overline{M}\hat{F}_{2S}(x))\right), \tag{25}$$

$$\sigma_{XY,2i} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}\Big((I(Y_{i,j}\le y)-F_i(y))(I(X_{i,j}\le x)-F_i(x))\Big), \tag{26}$$

$$\hat{\sigma}_{XY,2i} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left((I(Y_{i,j}\le y)-\hat{F}_i(y))(I(X_{i,j}\le x)-\hat{F}_i(x))\right). \tag{27}$$

Proof. Here, $\sigma_{XY,2b}$ and $\sigma_{XY,2i}$ have their usual meanings. The proof of this Lemma may be seen in the Appendix.

Lemma 2. *Under S2SCS scheme, the covariance between $\hat{F}_{S2S}(y)$ and $\hat{F}_{S2S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{S2S}(y),\hat{F}_{S2S}(x)) = \sum_{h=1}^{L}W_h^2\,C(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x)) \quad \textit{and} \tag{28}$$

$$\widehat{C}(\hat{F}_{S2S}(y),\hat{F}_{S2S}(x)) = \sum_{h=1}^{L}W_h^2\,\widehat{C}(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x)), \tag{29}$$

respectively, where W_h is given in Eq. (10).

Proof. The proof of Lemma 2 is similar to that of Lemma 1. Note that the mathematical expressions of $C(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x))$ and $\widehat{C}(\hat{F}_{2S,h}(y),\hat{F}_{2S,h}(x))$ are similar to those of $C(\hat{F}_{2S}(y),\hat{F}_{2S}(x))$ and $\widehat{C}(\hat{F}_{2S}(y),\hat{F}_{2S}(x))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1,2,\ldots,L$.

3.2. Three-stage and stratified three-stage cluster sampling

In order to estimate $(F(y),F(x))$ under 3SCS and S3SCS, let $(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$ and $(\hat{F}_{S3S}(y),\hat{F}_{S3S}(x))$ be the respective CDF estimators that are based on (Y,X), respectively.

Lemma 3. *Under 3SCS scheme, the covariance between $\hat{F}_{3S}(y)$ and $\hat{F}_{3S}(x)$, along with its unbiased estimators are given by*

$$C\big(\hat{F}_{3S}(y),\hat{F}_{3S}(x)\big) = \frac{\lambda\sigma_{XY,3b}}{n\overline{T}^2}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \tag{30}$$

and

$$\widehat{C}\big(\hat{F}_{3S}(y),\hat{F}_{3S}(x)\big) = \frac{\lambda\hat{\sigma}_{XY,3b}}{n\overline{T}^2}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,3i}}{m_i}+\frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\hat{\sigma}_{XY,3ij}}{t_{ij}}, \tag{31}$$

respectively, where

$$\sigma_{XY,3b}=\frac{1}{N-1}\sum_{i=1}^{N}\Big((M_iF_i(y)-\overline{T}F(y))(M_iF_i(x)-\overline{T}F(x))\Big), \tag{32}$$

$$\hat{\sigma}_{XY,3b}=\frac{1}{n-1}\sum_{i=1}^{n}\left((M_i\hat{F}_i(y)-\overline{T}\hat{F}_{3S}(y))(M_i\hat{F}_i(x)-\overline{T}\hat{F}_{3S}(x))\right), \tag{33}$$

$$\sigma_{XY,3i}=\frac{1}{M_i-1}\sum_{j=1}^{M_i}\Big((T_{ij}F_{ij}(y)-F_i(y))(T_{ij}F_{ij}(x)-F_i(x))\Big), \tag{34}$$

$$\hat{\sigma}_{XY,3i}=\frac{1}{m_i-1}\sum_{j=1}^{m_i}\left((T_{ij}\hat{F}_{ij}(y)-\hat{F}_i(y))(T_{ij}\hat{F}_{ij}(x)-\hat{F}_i(x))\right), \tag{35}$$

$$\sigma_{XY,3ij}=\frac{1}{T_{ij}-1}\sum_{k=1}^{T_{ij}}\Big((I(Y_{ij,k}\le y)-F_{ij}(y))(I(X_{ij,k}\le x)-F_{ij}(x))\Big), \tag{36}$$

$$\hat{\sigma}_{XY,3ij}=\frac{1}{t_{ij}-1}\sum_{k=1}^{t_{ij}}\left((I(Y_{ij,k}\le y)-\hat{F}_{ij}(y))(I(X_{ij,k}\le x)-\hat{F}_{ij}(x))\right), \tag{37}$$

and $\lambda_{ij}=(1-t_{ij}/T_{ij})$.

Proof. Here, $\sigma_{XY,3b}$ and $\sigma_{XY,3i}$ have their usual meanings. The proof of this Lemma may be seen in the Appendix.

Lemma 4. *Under S3SCS scheme, the covariance between $\hat{F}_{S3S}(y)$ and $\hat{F}_{S3S}(x)$, along with its unbiased estimator are given by*

$$C(\hat{F}_{S3S}(y),\hat{F}_{S3S}(x))=\sum_{h=1}^{L}W_h^2\,C(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x)) \quad \textit{and} \tag{38}$$

$$\widehat{C}(\hat{F}_{S3S}(y),\hat{F}_{S3S}(x))=\sum_{h=1}^{L}W_h^2\,\widehat{C}(\hat{F}_{3S,h}(y),\hat{F}_{3S,h}(x)), \tag{39}$$

respectively, where W_h is given in Eq. (16).

Proof. The proof of Lemma 4 is similar to that of Lemma 3. Note that the mathematical expressions of $C(\hat{F}_{3S,h}(y), \hat{F}_{3S,h}(x))$ and $\widehat{C}(\hat{F}_{3S,h}(y), \hat{F}_{3S,h}(x))$ are similar to those of $C(\hat{F}_{3S}(y), \hat{F}_{3S}(x))$ and $\widehat{C}(\hat{F}_{3S}(y), \hat{F}_{3S}(x))$, respectively, with the exception that the former are computed from the hth stratum for $h = 1, 2, \ldots, L$.

4. The CDF estimation with auxiliary information

In this section, we develop two auxiliary-information-based families of estimators, say ratio/product and exponential ratio/product, for estimating the population CDF $F(y)$ under the aforementioned complex survey sampling schemes.

In order to obtain the biases and MSEs of the proposed families of estimators of $F(y)$, we may consider the following relative error terms: Let

$$\xi_0 = \frac{\hat{F}_S(y) - F(y)}{F(y)} \quad \text{and} \quad \xi_1 = \frac{\hat{F}_S(x) - F(x)}{F(x)},$$

such that $E(\xi_0) = E(\xi_1) = 0$. Let us denote

$$V_{rs} = E(\xi_0^r \xi_1^s) = E\left[\left(\frac{\hat{F}_S(y) - F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_S(x) - F(x)}{F(x)}\right)^s\right], \tag{40}$$

which gives

$$\begin{aligned}
V_{20} &= E(\xi_0)^2 = E\left(\frac{\hat{F}_S(y) - F(y)}{F(y)}\right)^2 = \frac{V(\hat{F}_S(y))}{(F(y))^2}, \\
V_{02} &= E(\xi_1)^2 = E\left(\frac{\hat{F}_S(x) - F(x)}{F(x)}\right)^2 = \frac{V(\hat{F}_S(x))}{(F(x))^2}, \\
V_{11} &= E(\xi_0\xi_1) = E\left[\left(\frac{\hat{F}_S(y) - F(y)}{F(y)}\right)\left(\frac{\hat{F}_S(x) - F(x)}{F(x)}\right)\right] = \frac{C(\hat{F}_S(y), \hat{F}_S(x))}{F(y)F(x)},
\end{aligned}$$

where $\hat{F}_S$ denotes an CDF estimator based on an S sampling scheme, where S = 2S, S2S, 3S and S3S.

4.1. First proposed family of CDF estimators

On the lines of Khoshnevisan et al. (2007), we propose a family of ratio/product-type estimators for estimating the population CDF $F(y)$, given by

$$\hat{F}_R(y) = \hat{F}_S(y)\left(\frac{aF(x) + b}{\alpha(a\hat{F}_S(x) + b) + (1 - \alpha)(aF(x) + b)}\right)^g, \tag{41}$$

where $a \neq 0$ and b are either real numbers or functions of the known parameters of the auxiliary variable X such as coefficient of variation (C_X), correlation coefficient (ρ_{XY}), coefficient of skewness $(\beta_{1,X})$ and coefficient of kurtosis $(\beta_{2,X})$ etc. Here, $g \in \{-1, 1\}$

and $\alpha\,(0 \le \alpha \le 1)$ are suitably chosen scalars which make the MSE of $\hat{F}_R(y)$ minimum. It is possible to develop different estimators of $\hat{F}_R(y)$ with suitable choices of a, b, g and α. In Table 1, some members of $\hat{F}_R(y)$ are given for different values of a, b, α, and g.

In order to derive approximate mathematical expressions for the bias and MSE of $\hat{F}_R(y)$, we can write $\hat{F}_S(y) = F(y)(1+\xi_0)$ and $\hat{F}_S(x) = F(x)(1+\xi_1)$. Express the right-hand side (RHS) of (41) in terms of ξs to get:

$$\hat{F}_R(y) = F(y)(1+\xi_0)(1+\alpha\nu\xi_1)^{-g}, \tag{42}$$

where $\nu = aF(x)/(aF(x)+b)$. Expand the RHS of Eq. (42) and retain terms up to 2nd power of ξs, we have

$$\hat{F}_R(y) \;\approx\; F(y)\left(1+\xi_0-\alpha\nu g\xi_1+\frac{g(g+1)}{2}\alpha^2\nu^2\xi_1^2-\alpha\nu g\xi_0\xi_1\right) \tag{43}$$

Take expectation on both sides of Eq. (43) after subtracting $F(y)$ on both sides to get the bias of $\hat{F}_R(y)$ up to the first order of approximation, which is given by

$$\mathrm{Bias}(\hat{F}_R(y)) \;\approx\; F(y)\left(\frac{g(g+1)}{2}\,\alpha^2\nu^2V_{02}-\alpha\nu gV_{11}\right). \tag{44}$$

From Eq. (43), we can write

$$\hat{F}_R(y)-F(y) \approx F(y)(\xi_0-\alpha\nu g\xi_1) \tag{45}$$

Take square on both sides of Eq. (45) and then taking its expectation to get the MSE of $\hat{F}_R(y)$ under first order of approximation, which is given by

$$\mathrm{MSE}(\hat{F}_R(y)) \;\approx\; F^2(y)\left(V_{20}+\alpha^2\nu^2g^2V_{02}-2\alpha\nu gV_{11}\right), \tag{46}$$

The minimum MSE at the optimum value of $(\alpha\nu g)$, say $(\alpha\nu g)_{\mathrm{opt}} = V_{11}/V_{02}$, is given by

$$\mathrm{MSE}_{\min}(\hat{F}_R(y)) \approx F^2(y)\left(V_{20}-\frac{V_{11}^2}{V_{02}}\right) \tag{47}$$

$$\approx F^2(y)V_{20}\,(1-\rho^2), \tag{48}$$

where $\rho = V_{11}/\sqrt{V_{20}V_{02}}$ is the correlation coefficient between $\hat{F}_S(y)$ and $\hat{F}_S(x)$ with an S sampling scheme.

4.2. Second proposed family of CDF estimators

On the lines of Singh et al. (2009), we propose another family of exponential ratio/product-type estimators for estimating the population CDF $F(y)$, given by

$$\hat{F}_E(y) \;=\; \hat{F}_S(y)\exp\left(\frac{(agF(x)+b)-(ag\hat{F}_S(x)+b)}{(aF(x)+b)+(a\hat{F}_S(x)+b)}\right), \tag{49}$$

where $a = 0$ and b are either real numbers or functions of the known parameters of the auxiliary variable X, but $g \in \{-1, 1\}$. In Table 1, some members of $\hat{F}_E(y)$ are given for different values of a, b, α, and g.

In order to obtain the bias and MSE of $\hat{F}_E(y)$, express $\hat{F}_E(y)$ in terms of ξs to get

$$\begin{aligned}\hat{F}_E(y) &= F(y)(1+\xi_0)\exp\left(\frac{agF(x) - agF(x)(1+\xi_1)}{aF(x)+2b+aF(x)(1+\xi_1)}\right)\\ &= F(y)(1+\xi_0)\exp\left(-\theta g\xi_1(1+\theta\xi_1)^{-1}\right),\end{aligned} \tag{50}$$

where $\theta = aF(x)/(2aF(x)+2b)$. After expanding the RHS of Eq. (50) up to 2nd power of ξs, we have

$$\hat{F}_E(y) \;\approx\; F(y)\left(1+\xi_0-\theta g\xi_1+\frac{g(g+1)}{2}\theta^2\xi_1^2-\theta g\xi_0\xi_1\right). \tag{51}$$

Take expectation after subtracting $F(y)$ on both sides of Eq. (51) to get the bias of $\hat{F}_E(y)$, which under the first order of approximation is given by

$$\text{Bias}(\hat{F}_E(y)) \;\approx\; F(y)\left(\frac{g(g+1)}{2}\theta^2 V_{02}-\theta g V_{11}\right). \tag{52}$$

From Eq. (51), we can write

$$\hat{F}_E(y) - F(y) \approx F(y)(\xi_0 - \theta g\xi_1). \tag{53}$$

Take square on both sides of Eq. (53), and then take it expectation to get the MSE of $\hat{F}_E(y)$ under the first order of approximation, which is given by

$$\text{MSE}(\hat{F}_E(y)) \approx F(y)^2\left(V_{20}+\theta^2 g^2 V_{02}-2\theta g V_{11}\right). \tag{54}$$

The minimum MSE at the optimum value of (θg), say $(\theta g)_{\text{opt}} = V_{11}/V_{02}$, is given by

$$\text{MSE}_{\min}(\hat{F}_E(y)) \approx F^2(y)V_{20}(1-\rho^2), \tag{55}$$

which is equivalent to that of $\hat{F}_R(y)$.

In addition to these estimators a large number of estimators can also be generated from the proposed families of estimators $\hat{F}_R(y)$ and $\hat{F}_E(y)$ given in Eq. (41) and Eq. (49) respectively, just by putting values of a, b, α, and g.

It is observed that the expression of the first order approximation of bias and MSE/Variance of the given member of the families $\hat{F}_R(y)$ and $\hat{F}_E(y)$ can be obtained by mere substituting the values of α, g, a and b in (Eq. (44) and Eq. (46)) and (Eq. (52) and Eq. (54)), respectively. It is to be noted that, based on S scheme, the proposed families of estimators, $\hat{F}_R(y)$ and $\hat{F}_E(y)$, are more precise than $\hat{F}_R(y)$ when the following conditions hold in practice:

$$\begin{aligned}\text{MSE}(\hat{F}_R(y)) < V(\hat{F}_S(y)) &\implies \nu < \frac{2V_{11}}{(\alpha g V_{02})},\\ \text{MSE}(\hat{F}_E(y)) < V(\hat{F}_S(y)) &\implies \theta < \frac{2V_{11}}{(gV_{02})}.\end{aligned} \tag{56}$$

Table 1. *Some members of proposed families of CDF estimators.*

$\hat{F}_R(y)$	$\hat{F}_E(y)$	g	α	a	b
$\hat{F}_R^{(1)}(y)=\hat{F}_S(y)\left(\frac{F(x)}{\hat{F}_S(x)}\right)$	$\hat{F}_E^{(1)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)}\right)$	1	1	1	0
$\hat{F}_R^{(2)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\rho_{XY}}{\hat{F}_S(x)+\rho_{XY}}\right)$	$\hat{F}_E^{(2)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\rho_{XY}}\right)$	1	1	1	ρ_{XY}
$\hat{F}_R^{(3)}(y)=\hat{F}_S(y)\left(\frac{F(x)+C_X}{\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(3)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2C_X}\right)$	1	1	1	C_X
$\hat{F}_R^{(4)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\beta_{2,X}}{\hat{F}_S(x)+\beta_{2,X}}\right)$	$\hat{F}_E^{(4)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\beta_{2,X}}\right)$	1	1	1	$\beta_{2,X}$
$\hat{F}_R^{(5)}(y)=\hat{F}_S(y)\left(\frac{C_XF(x)+\beta_{2,X}}{C_X\hat{F}_S(x)+\beta_{2,X}}\right)$	$\hat{F}_E^{(5)}(y)=\hat{F}_S(y)\exp\left(\frac{C_X(F(x)-\hat{F}_S(x))}{C_X(F(x)+\hat{F}_S(x))+2\beta_{2,X}}\right)$	1	1	C_X	$\beta_{2,X}$
$\hat{F}_R^{(6)}(y)=\hat{F}_S(y)\left(\frac{\beta_{2,X}F(x)+C_X}{\beta_{2,X}\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(6)}(y)=\hat{F}_S(y)\exp\left(\frac{\beta_{2,X}(F(x)-\hat{F}_S(x))}{\beta_{2,X}(F(x)+\hat{F}_S(x))+2C_X}\right)$	1	1	$\beta_{2,X}$	C_X
$\hat{F}_R^{(7)}(y)=\hat{F}_S(y)\left(\frac{\rho_{XY}F(x)+C_X}{\rho_{XY}\hat{F}_S(x)+C_X}\right)$	$\hat{F}_E^{(7)}(y)=\hat{F}_S(y)\exp\left(\frac{\rho_{XY}(F(x)-\hat{F}_S(x))}{\rho_{XY}(F(x)+\hat{F}_S(x))+2C_X}\right)$	1	1	ρ_{XY}	C_X
$\hat{F}_R^{(8)}(y)=\hat{F}_S(y)\left(\frac{C_XF(x)+\rho_{XY}}{C_X\hat{F}_S(x)+\rho_{XY}}\right)$	$\hat{F}_E^{(8)}(y)=\hat{F}_S(y)\exp\left(\frac{C_X(F(x)-\hat{F}_S(x))}{C_X(F(x)+\hat{F}_S(x))+2\rho_{XY}}\right)$	1	1	C_X	ρ_{XY}
$\hat{F}_R^{(9)}(y)=\hat{F}_S(y)\left(\frac{F(x)+\beta_{1,X}}{\hat{F}_S(x)+\beta_{1,X}}\right)$	$\hat{F}_E^{(9)}(y)=\hat{F}_S(y)\exp\left(\frac{F(x)-\hat{F}_S(x)}{F(x)+\hat{F}_S(x)+2\beta_{1,X}}\right)$	1	1	1	$\beta_{1,X}$

ρ_{XY} is correlation coefficient between X and Y, C_X is coefficient of variation of X

$\beta_{1,X}$ is coefficient of skewness of X, $\beta_{2,X}$ is coefficient of kurtosis of X

4.3. Difference and regression CDF estimators

It is possible to further enhance the precision of the aforementioned families of estimators $(\hat{F}_S(y),\hat{F}_S(x))$ when the supplementary information in terms of the covariance between the CDF estimators based on Y and X, and on the variance of the CDF estimator of X are utilized.

Under a sampling scheme S, let β denote the ratio of the covariance of $\hat{F}_S(y)$ and $\hat{F}_S(x)$ to the variance of $\hat{F}_S(x)$, i.e.

$$\beta_S \;=\; \frac{C(\hat{F}_S(y),\hat{F}_S(x))}{V(\hat{F}_S(x))}. \tag{57}$$

In addition, it is also possible to have information available on the population CDF of X, say $F(x)$. The difference estimator of the population CDF $F(y)$, say $\hat{F}_D(y)$, that requires information on β_S, $\hat{F}_S(y)$ and $\hat{F}_S(x)$ is given by

$$\hat{F}_D(y) \;=\; \hat{F}_S(y)+\beta_S\left(F(x)-\hat{F}_S(x)\right), \tag{58}$$

where $\hat{F}_D(y)$ is a linear combination of $\hat{F}_{\mathrm{S}}(y)$ and $\hat{F}_{\mathrm{S}}(x)$. It can easily be shown that the $\hat{F}_D(y)$ is an unbiased estimator of $F(y)$.

In order to obtain the variance of $\hat{F}_D(y)$, we express $\hat{F}_D(y)$ in terms of ξs, i.e.

$$\begin{aligned} \hat{F}_D(y) &= F(y)(1+\xi_0) - \beta_{\mathrm{S}} F(x)\xi_1 \\ \hat{F}_D(y) - F(y) &= F(y)\xi_0 - \beta_{\mathrm{S}} F(x)\xi_1. \end{aligned} \tag{59}$$

Take square on both sides of Eq. (59) and then apply expectation to get the variance of $\hat{F}_D(y)$, which is given by

$$V(\hat{F}_D(y)) \;=\; F^2(y)V_{20} + \beta_{\mathrm{S}}^2 F^2(x)V_{02} - 2\beta_{\mathrm{S}} F(x)F(y)V_{11}. \tag{60}$$

The simplified expression for the variance of $\hat{F}_D(y)$, after replacing the value of β_{S} into $V(\hat{F}_D(y))$, is given by

$$V(\hat{F}_D(y)) \;=\; F^2(y)V_{20}(1-\rho^2), \tag{61}$$

which is equivalent to the minimum MSE of $\hat{F}_R(y)$ and $\hat{F}_E(y)$.

It is to be noted that the value of β_{S} may be taken from previous studies, surveys or census. In case the value of β_{S} is not known, then it is possible to estimate it with a large sample size. The estimated value of β_{S} may be obtained by replacing the covariance of $(\hat{F}_{\mathrm{S}}(y), \hat{F}_{\mathrm{S}}(x))$ and the variance of $\hat{F}_{\mathrm{S}}(x)$ by their respective unbiased estimators, which gives

$$\hat{\beta}_{\mathrm{S}} \;=\; \frac{\widehat{C}(\hat{F}_{\mathrm{S}}(y), \hat{F}_{\mathrm{S}}(x))}{\widehat{V}(\hat{F}_{\mathrm{S}}(x))}. \tag{62}$$

It is a well-known fact under SRS that the sample covariance $\widehat{C}(\hat{F}_{\mathrm{S}}(y), \hat{F}_{\mathrm{S}}(x))$ and sample variance $\widehat{V}(\hat{F}_{\mathrm{S}}(x))$ are weakly-consistent estimators of $C(\hat{F}_{\mathrm{S}}(y), \hat{F}_{\mathrm{S}}(x))$ and $V(\hat{F}_{\mathrm{S}}(x))$, respectively. Thus, for a large sample size, $\hat{\beta}_{\mathrm{S}}$ is also a weakly-consistent estimator of β_{S}.

In the survey sampling literature, the difference estimator $\hat{F}_D(y)$ with estimated value of β_{S} is called a regression estimator, given by

$$\hat{F}_{Reg}(y) \;=\; \hat{F}_{\mathrm{S}}(y) + \hat{\beta}_{\mathrm{S}}\left(F(x) - \hat{F}_{\mathrm{S}}(x)\right). \tag{63}$$

It can be shown that $\hat{F}_{Reg}(y)$ is a biased estimator of $F(y)$. Moreover, for a large sample size, we have

$$\mathrm{MSE}(\hat{F}_{Reg}(y)) \;\approx\; V(\hat{F}_D(y)) \;=\; F^2(y)V_{20}(1-\rho^2). \tag{64}$$

5. Empirical Study

In this section, real datasets are considered and the relative efficiencies (REs) of the proposed CDF estimators of $F(y)$ are computed with respect to $\hat{F}_{\mathrm{S}}(y)$ based on sampling scheme S.

5.1. Population I

This dataset is taken from Social & Household Integrated Economic Survey (HIES), conducted in Pakistan during the years 2011-12, which comprises 14722 households (after removing the missing observations). The entire dataset is partitioned into two strata, where Stratum-I and Stratum-II correspond to Urban and Rural (U-R) areas. These areas are further partitioned into four provinces of Pakistan, namely Punjab, Khyber Pakhtunkhwa (KPK), Sindh and Balochistan, where (Punjab - KPK) and (Sindh - Balochistan) belong to Stratum-I and Stratum-II, respectively. Moreover, where each province is further partitioned into different enumeration blocks (EBs). This dataset may be downloaded from the Pakistan Bureau of Statistics web-page via the link: https://www.pbs.gov.pk/content/microdata. The study variable Y and the auxiliary variable X are total income and total expenditure of a household (HH), respectively. Here, our objective is to estimate the proportion of HH whose yearly total income is less than or equal to $y = \$1.9 \times 365$, which is considered as the poverty line for Pakistan according to the World bank's website: https://data.worldbank.org/indicator/SI.POV.NAHC?locations=PK. The yearly total income is converted from USD to PKR by multiplying $1.9 \times 365 \times 86.3198$ PKR. For example, if the total income of a HH is less than or equal to 59862.7813 PKR, it is then considered on or below the poverty line using auxiliary variable X while $x = 226386.0582$ (yearly average expenditure of a HH). Note that (province and yearly total income of a HH) and (province, EB and yearly total income of a HH) are taken as (PSU and SSU) and (PSU, SSU and TSU) for 2SCS/S2SCS and 3SCS/S3SCS, respectively. The values of the population parameters are given below:

$$F(y) = 0.0474, F(x) = 0.6587, C_X = 0.8161,$$
$$\rho_{XY} = 0.7662, \beta_{1,X} = 4.5387 \text{ and } \beta_{2,X} = 43.4005.$$

The values of V_{rs} based on an S sampling scheme are computed and then reported in Table 2, where

$$V_{rs} = E\left(\xi_0^r \xi_1^s\right) = E\left[\left(\frac{\hat{F}_{\mathrm{S}}(y) - F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_{\mathrm{S}}(x) - F(x)}{F(x)}\right)^s\right],$$

where $r, s = 0, 1, 2$.

Table 2. *The V_{rs} values based on scheme S using Population-I.*

S	$\mathrm{S_{tr}}$ – Variable	PSU	SSU	TSU	n	m_i	t_{ij}	V_{20}	V_{02}	V_{11}
2SCS	––	Province	HH	––	3	40	––	0.31145	0.03899	0.04976
S2SCS	U/R	Province	HH	––	1	40	––	0.44729	0.12185	0.10904
3SCS	––	Province	EB	HH	3	15	4	0.27805	0.04275	0.05609
S3SCS	U/R	Province	EB	HH	1	15	4	0.39719	0.12749	0.11854

Note: Stratifying is abbreviated as $\mathrm{S_{tr}}$.

5.2. Population II

Another dataset is taken from Center of Disease Control (CDC), which is related to the Second National Health and Nutrition Examination Survey (NHANES-II). The NHANES sample (comprising 10351 units) represents the total non-institutionalized civilian (NIC) US population that resides in 50 states and the district of Columbia. This dataset is divided into four regions (REGs), namely southern, western, mid-western and north-eastern, where each REG is further divided into different locations (LOCs). The entire dataset is stratified into two strata, which are formed by generating random numbers from Bernoulli distribution with 0.50 as the probability of success, where 0 and 1 correspond to Stratum-I and Stratum-II, respectively. This dataset is available at the https://www.stata-press.com/data/r15/svy.html. Here, the body mass index (BMI) is taken as the study variable Y and weight is taken as the auxiliary variable X. Our objective is to estimate the proportion of people (in the NIC US population) that are under-weight, i.e., an individual is classified as under-weight if the BMI values are less than or equal to $y = 18.50$ using auxiliary variable X while $x = 71.8975$ (average weight of NIC US population) under sampling scheme S. Note that the (REG and BMI) and (REG, LOC and BMI) are taken as (PSU and SSU) and (PSU, SSU and TSU) for the 2SCS/S2SCS and 3SCS/S3SCS, respectively. The values of the population parameters are given below:

$$F(y) = 0.0318, F(x) = 0.5401, C_X = 0.2136,$$
$$\rho_{XY} = 0.8338, \beta_{1,X} = 0.7364 \text{ and } \beta_{2,X} = 4.0614.$$

The values of V_{rs} based on an sampling scheme S are computed and then reported in Table 3, where

$$V_{rs} = E\left(\xi_0^r \xi_1^s\right) = E\left[\left(\frac{\hat{F}_{\mathrm{S}}(y) - F(y)}{F(y)}\right)^r \left(\frac{\hat{F}_{\mathrm{S}}(x) - F(x)}{F(x)}\right)^s\right],$$

where $r, s = 0, 1, 2$.

Table 3. *The V_{rs} values based on scheme S using Population-II.*

S	$\mathrm{S_{tr}}-$Variable	PSU	SSU	TSU	n	m_i	t_{ij}	V_{20}	V_{02}	V_{11}
2SCS	$--$	REG	BMI	$--$	3	50	$--$	0.20634	0.00721	0.00766
S2SCS	0/1	REG	BMI	$--$	3	50	$--$	0.10207	0.00359	0.00386
3SCS	$--$	REG	LOC	BMI	3	3	50	0.07641	0.00892	0.00937
S3SCS	0/1	REG	LOC	BMI	3	3	50	0.03714	0.00476	0.00478

Using the aforementioned datasets, the REs of the CDF estimators based on a sampling scheme S are computed with different values of n, m_i and t_{ij}. The REs of the

proposed CDF estimators of $F(y)$ with auxiliary information with respect to usual unbiased CDF estimator of $\hat{F}_{\mathrm{S}}(y)$ without auxiliary information are given by

$$\mathrm{RE}_R = \frac{V(\hat{F}_{\mathrm{S}}(y))}{\mathrm{MSE}(\hat{F}_R^{(t)}(y))}, \quad \mathrm{RE}_E = \frac{V(\hat{F}_{\mathrm{S}}(y))}{\mathrm{MSE}(\hat{F}_E^{(t)}(y))}, \quad \mathrm{RE}_D = \frac{V(\hat{F}_{\mathrm{S}}(y))}{V(\hat{F}_D(y))}, \tag{65}$$

where $t = 1,2,\ldots,9$. The REs of these CDF estimators are reported in Tables 4 and 5.

Table 4. *REs of proposed CDF estimators with respect to $\hat{F}_S(y)$ using Population-I.*

$\hat{F}_R(y)$	2SCS	S2SCS	3SCS	S3SCS	$\hat{F}_E(y)$	2SCS	S2SCS	3SCS	S3SCS
$\hat{F}_R^{(1)}(y)$	1.2412	1.2741	1.3328	1.3810	$\hat{F}_E^{(1)}(y)$	1.1474	1.2131	1.1952	1.2791
$\hat{F}_R^{(2)}(y)$	1.1376	1.2007	1.1815	1.2616	$\hat{F}_E^{(2)}(y)$	1.0720	1.1088	1.0929	1.1374
$\hat{F}_R^{(3)}(y)$	1.1335	1.1953	1.1758	1.2540	$\hat{F}_E^{(3)}(y)$	1.0697	1.1053	1.0898	1.1329
$\hat{F}_R^{(4)}(y)$	1.0048	1.0073	1.0060	1.0089	$\hat{F}_E^{(4)}(y)$	1.0024	1.0036	1.0030	1.0045
$\hat{F}_R^{(5)}(y)$	1.0039	1.0060	1.0049	1.0073	$\hat{F}_E^{(5)}(y)$	1.0020	1.0030	1.0025	1.0037
$\hat{F}_R^{(6)}(y)$	1.2381	1.2764	1.3279	1.3829	$\hat{F}_E^{(6)}(y)$	1.1438	1.2087	1.1902	1.2728
$\hat{F}_R^{(7)}(y)$	1.1158	1.1717	1.1517	1.2213	$\hat{F}_E^{(7)}(y)$	1.0599	1.0908	1.0770	1.1140
$\hat{F}_R^{(8)}(y)$	1.1242	1.1830	1.1631	1.2369	$\hat{F}_E^{(8)}(y)$	1.0645	1.0976	1.0830	1.1228
$\hat{F}_R^{(9)}(y)$	1.0400	1.0609	1.0512	1.0758	$\hat{F}_E^{(9)}(y)$	1.0201	1.0307	1.0256	1.0379
$\hat{F}_D(y)$	1.2562	1.2790	1.3600	1.3841					

Table 5. *REs of proposed CDF estimators with respect to $\hat{F}_S(y)$ using Population-II.*

$\hat{F}_R(y)$	2SCS	S2SCS	3SCS	S3SCS	$\hat{F}_E(y)$	2SCS	S2SCS	3SCS	S3SCS
$\hat{F}_R^{(1)}(y)$	1.0409	1.0421	1.1473	1.1488	$\hat{F}_E^{(1)}(y)$	1.0292	1.0299	1.1030	1.1072
$\hat{F}_R^{(2)}(y)$	1.0244	1.0249	1.0850	1.0887	$\hat{F}_E^{(2)}(y)$	1.0134	1.0137	1.0457	1.0479
$\hat{F}_R^{(3)}(y)$	1.0365	1.0375	1.1309	1.1349	$\hat{F}_E^{(3)}(y)$	1.0226	1.0231	1.0786	1.0821
$\hat{F}_R^{(4)}(y)$	1.0083	1.0085	1.0279	1.0293	$\hat{F}_E^{(4)}(y)$	1.0043	1.0043	1.0142	1.0149
$\hat{F}_R^{(5)}(y)$	1.0020	1.0021	1.0067	1.0071	$\hat{F}_E^{(5)}(y)$	1.0010	1.0010	1.0034	1.0035
$\hat{F}_R^{(6)}(y)$	1.0402	1.0413	1.1448	1.1473	$\hat{F}_E^{(6)}(y)$	1.0273	1.0279	1.0958	1.0999
$\hat{F}_R^{(7)}(y)$	1.0355	1.0364	1.1269	1.1310	$\hat{F}_E^{(7)}(y)$	1.0216	1.0221	1.0749	1.0783
$\hat{F}_R^{(8)}(y)$	1.0086	1.0088	1.0289	1.0303	$\hat{F}_E^{(8)}(y)$	1.0044	1.0045	1.0147	1.0154
$\hat{F}_R^{(9)}(y)$	1.0258	1.0264	1.0903	1.0942	$\hat{F}_E^{(9)}(y)$	1.0143	1.0146	1.0489	1.0513
$\hat{F}_D(y)$	1.0410	1.0424	1.1477	1.1488					

It can be seen that the proposed CDF estimators under complex survey sampling with auxiliary information are slightly more efficient than those that are without the auxiliary information, that is, all values of the REs are greater than one. It can also be seen that the proposed CDF estimators under a sampling scheme S with stratification are slightly more efficient than those without stratification and the REs tend to increase with increasing the sampling stages. Generally, with an increase in the sample size at the primary, secondary or tertiary stage of sampling, the REs may tend to increase and vice versa. Among all estimators, as expected, the REs of $\hat{F}_D(y)$ are higher than those of other considered CDF estimators.

It is to be noted that the proposed families of estimators, $\hat{F}_R(y)$ and $\hat{F}_E(y)$, are conditionally better than $\hat{F}_S(y)$, i.e. when the conditions given in Eq. (56) hold. However, the difference and regression estimators, $\hat{F}_D(y)$ and $\hat{F}_{Reg}(y)$, respectively, are always more precise than $\hat{F}_S(y)$, $\hat{F}_R(y)$ and $\hat{F}_E(y)$. In usual practice, if no information is available to check these conditions, it is preferable to use $\hat{F}_{Reg}(y)$ when estimating the population CDF under scheme S.

6. Conclusion

In this paper, we have considered the problem of estimating the finite population CDF in 2SCS and 3SCS schemes with and without stratification. Two families of classical ratio/product-type and exponential ratio/product-type CDF estimators have been proposed that require supplementary information on a single auxiliary variable. In addition, difference and regression estimators of the CDF have also been proposed. Explicit mathematical expressions of the biases and MSEs of the proposed CDF estimators have been developed under first order of the approximation. Real datasets were also considered to support the proposed theory.

Along the lines of Nematollahi et al. (2008) and Haq et al. (2021), it is also possible to increase the precision of proposed families of the CDF estimators by employing RSS and double RSS schemes in the secondary and tertiary sampling frames. Moreover, the current work may be extended to develop new CDF estimators that require supplementary information on two or more auxiliary variables. In addition, it may be possible to develop the CDF estimators when using probability proportional to size sampling to select units at the first stage of sampling under the 2SCS/3SCS and S2SCS/S3SCS schemes.

Acknowledgments

The authors are thankful to the editor and two anonymous reviewers for providing many useful comments that led to an improved version of the article.

Appendix

In this Appendix, we present the proofs of the Lemmas in Section 3.

1: Proof of Lemma 1

Here, the indices 1 and 2 are used for the first-stage and second-stage of sampling under 2SCS, respectively.

1. The covariance between $\hat{F}_{2S}(y)$ and $\hat{F}_{2S}(x)$ can be written as:

$$C(\hat{F}_{2S}(y),\hat{F}_{2S}(x)) \;=\; C_1[E_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))] + E_1[C_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))]. \tag{66}$$

It can be shown that $E_2(\hat{F}_{2S}(y)) = \sum_{i=1}^{n} M_i F_i(y)/(n\overline{M})$. Based on this result, we have

$$C_1\left[E_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))\right] = C_1\left(\frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i F_i(y), \frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i F_i(x)\right) \;=\; \frac{\lambda\sigma_{XY,2b}}{n\overline{M}^2} \tag{67}$$

$$\begin{aligned} E_1\left[C_2(\hat{F}_{2S}(y),\hat{F}_{2S}(x))\right] &= E_1\left[\frac{1}{n^2\overline{M}^2}\sum_{i=1}^{n} M_i^2 C_2(\hat{F}_i(y),\hat{F}_i(x))\right], \\ &= \frac{1}{nN\overline{M}^2}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i}, \end{aligned} \tag{68}$$

which completes the proof.

2. An unbiased estimator of $C\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right)$ is given by

$$\widehat{C}\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right) = \frac{\lambda\hat{\sigma}_{XY,2b}}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,2i}}{m_i}, \tag{69}$$

From Eq. (25), we can write

$$\hat{\sigma}_{XY,2b} \;=\; \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(M_i\hat{F}_i(y)M_i\hat{F}_i(x)) - \overline{M}\hat{F}_{2S}(y)\overline{M}\hat{F}_{2S}(x)\right]. \tag{70}$$

Consider the mathematical expectation on the RHS of Eq. (70) to get:

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right] &= E_1\left[\frac{1}{n}\sum_{i=1}^{n} E_2\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right] \\ &= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(C_2(M_i\hat{F}_i(y),M_i\hat{F}_i(x))\right)\right. \\ &\qquad \left. + \frac{1}{n}\sum_{i=1}^{n}\left(E_2(M_i\hat{F}_i(y))E_2(M_i\hat{F}_i(x))\right)\right] \end{aligned}$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + M_iF_i(y)M_iF_i(x)\right)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + \frac{1}{N}\sum_{i=1}^{N}M_iF_i(y)M_iF_i(x), \tag{71}$$

and

$$E\left[\left(\overline{M}\hat{F}_{2S}(y)\overline{M}\hat{F}_{2S}(x)\right)\right] = C\left(\overline{M}\hat{F}_{2S}(y),\overline{M}\hat{F}_{2S}(x)\right) + E(\overline{M}\hat{F}_{2S}(y))E(\overline{M}\hat{F}_{2S}(x))$$
$$= \frac{\lambda\sigma_{XY,2b}}{n} + \frac{1}{nN}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i} + \overline{M}F(y)\overline{M}F(x). \tag{72}$$

Using Eqs. (71) and (72) and then take the mathematical expectation of Eq. (25) to get:

$$E(\hat{\sigma}_{XY,2b}) \;=\; \sigma_{XY,2b} + \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,2i}}{m_i}, \tag{73}$$

which shows that $\hat{\sigma}_{XY,2b}$ is a biased estimator of $\sigma_{XY,2b}$.

Similarly, from Eq. (27), we have

$$\hat{\sigma}_{XY,2i} \;=\; \frac{m_i}{m_i-1}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\Big(I(Y_{i,j}\le y)I(X_{i,j}\le x)\Big) - \hat{F}_i(y)\hat{F}_i(x)\right]. \tag{74}$$

Consider the mathematical expectation on the RHS of Eq. (74) to get:

$$E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(I(Y_{ij}\le y)I(X_{ij}\le x)\right)\right] = \frac{1}{M_i}\sum_{j=1}^{M_i}\left(I(Y_{ij}\le y)I(X_{ij}\le x)\right) \tag{75}$$
$$E_2\left[\hat{F}_i(y)\hat{F}_i(x)\right] = C_2\left(\hat{F}_i(y),\hat{F}_i(x)\right) - E_2(\hat{F}_i(y))E_2(\hat{F}_i(x))$$
$$= \frac{\lambda_i\sigma_{XY,2i}}{m_i} - F_i(y)F_i(x). \tag{76}$$

Using Eqs. (75)-(76) and then take the expectation of Eq. (27) to get

$$E_2(\hat{\sigma}_{XY,i}) \;=\; \sigma_{XY,2i}, \tag{77}$$

which shows that $\hat{\sigma}_{XY,2i}$ is an unbiased estimator of $\sigma_{XY,2i}$.

Now take the mathematical expectation of Eq. (69), and use the results given in Eqs. (73) and (77) to show that

$$E\left[\widehat{C}\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right)\right] \;=\; C\left(\hat{F}_{2S}(y),\hat{F}_{2S}(x)\right), \tag{78}$$

which completes the proof.

2: Proof of Lemma 3

Here, the indices 1, 2 and 3 are used for the first stage, second stage and third stage of sampling under 3SCS, respectively.

1. The covariance between $\hat{F}_{3S}(y)$ and $\hat{F}_{3S}(x)$ can be written as:

$$C\left(\hat{F}_{3S}(y),\hat{F}_{3S}(y)\right) = C_1E_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] + E_1C_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] + E_1E_2C_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right]. \quad (79)$$

It can be shown that $E_3(\hat{F}_{3S}(y)) = \sum_{i=1}^{n}(M_i/m_i)\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)/n\overline{T}$. Based on this result, we have

$$\begin{aligned} C_1E_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= C_1E_2\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\ &= C_1\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}M_iF_i(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}M_iF_i(x)\right] = \frac{\lambda\sigma_{XY,3b}}{n\overline{T}^2}. \quad (80) \\ E_1C_2E_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= E_1C_2\left[\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{n\overline{T}}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\ &= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}M_i^2\,C_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big)\right] \\ &= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_iM_i^2\sigma_{XY,3i}}{m_i}\right] \\ &= \frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{\lambda_iM_i^2\sigma_{XY,3i}}{m_i}. \quad (81) \\ E_1E_2C_3\left[\hat{F}_{3S}(y),\hat{F}_{3S}(x)\right] &= E_1E_2\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}T_{ij}^2\,C_3\big(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\big)\right] \\ &= E_1E_2\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right] \\ &= E_1\left[\frac{1}{n^2\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right] \\ &= \frac{1}{nN\overline{T}^2}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}. \quad (82) \end{aligned}$$

Add Eqs. (80)–(82), which completes the proof.

2. An unbiased estimator of $C(\hat{F}_{3S}(y),\hat{F}_{3S}(x))$ is given by

$$\widehat{C}(\hat{F}_{3S}(y),\hat{F}_{3S}(x)) = \frac{\lambda\hat{\sigma}_{XY,3b}}{n\overline{T}^2} + \frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{\lambda_i M_i^2\hat{\sigma}_{XY,3i}}{m_i} + \frac{1}{nN\overline{T}^2}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\hat{\sigma}_{XY,3ij}}{t_{ij}}, \tag{83}$$

From Eq. (33), we can write

$$\hat{\sigma}_{XY,3b} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right) - \left(\overline{T}\hat{F}_{3S}(y)\overline{T}\hat{F}_{3S}(x)\right)\right]. \tag{84}$$

Consider the mathematical expectation of the RHS of the above equation to get:

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n}(M_i\hat{F}_i(y)M_i\hat{F}_i(x))\right] &= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}E_3\left(M_i\hat{F}_i(y)M_i\hat{F}_i(x)\right)\right] \\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(C_3\left(M_i\hat{F}_i(y),M_i\hat{F}_i(x)\right)\right)\right. \\
&\quad \left.+\frac{1}{n}\sum_{i=1}^{n}\left\{E_3(M_i\hat{F}_i(y))E_3(M_i\hat{F}_i(x))\right\}\right] \\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}C_3\left(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\right)\right)\right. \\
&\quad \left.+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}E_3(T_{ij}\hat{F}_{ij}(y))\frac{M_i}{m_i}\sum_{j=1}^{m_i}E_3(T_{ij}\hat{F}_{ij}(x))\right\}\right] \\
&= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i^2}{m_i^2}\sum_{j=1}^{m_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right)\right. \\
&\quad \left.+\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right\}\right] \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}\right)\right. \\
&\quad \left.+\frac{1}{n}\sum_{i=1}^{n}\left\{E_2\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{M_i}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right)\right\}\right]
\end{aligned}$$

$$= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}\right)\right.$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n}\left(M_iF_i(y)M_iF_i(x)\right)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}$$
$$+\frac{1}{N}\sum_{i=1}^{N}M_iF_i(y)M_iF_i(x), \quad \text{and} \tag{85}$$

$$E\left[\left(\overline{T}\hat{F}_{3S}(y)\overline{T}\hat{F}_{3S}(x)\right)\right] = C\left(\overline{T}\hat{F}_{3S}(y),\overline{T}\hat{F}_{3S}(x)\right)+E(\overline{T}\hat{F}_{3S}(y))E(\overline{T}\hat{F}_{3S}(x)),$$
$$= \frac{\lambda\sigma_{XY,3b}}{n}+\frac{1}{nN}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{nN}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}$$
$$+\overline{T}F(y)\overline{T}F(x). \tag{86}$$

Using Eqs. (85) and (86) in Eq. (33), and then take expectation to show that

$$E(\hat{\sigma}_{XY,3b}) = \sigma_{XY,3b}+\frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i M_i^2\sigma_{XY,3i}}{m_i}+\frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \tag{87}$$

which shows that $\hat{\sigma}_{XY,3b}$ is a biased estimator of $\sigma_{XY,3b}$.

Similarly, we can write from Eq. (35):

$$\hat{\sigma}_{XY,3i} = \frac{m_i}{m_i-1}\left[\frac{1}{m_i}\sum_{j=1}^{m_i}(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x))-(\hat{F}_i(y)\hat{F}_i(x))\right]. \tag{88}$$

Consider the mathematical expectation on the RHS of the above equation to get:

$$E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x))\right] = E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}E_3\left(T_{ij}\hat{F}_{ij}(y)T_{ij}\hat{F}_{ij}(x)\right)\right]$$
$$= E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(T_{ij}^2\,C_3\left(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\right)\right)\right.$$
$$\left.+\frac{1}{m_i}\sum_{j=1}^{m_i}\left\{E_3\left(T_{ij}\hat{F}_{ij}(y)\right)E_3\left(T_{ij}\hat{F}_{ij}(x)\right)\right\}\right]$$
$$= E_2\left[\frac{1}{m_i}\sum_{j=1}^{m_i}\left(\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+T_{ij}F_{ij}(y)T_{ij}F_{ij}(x)\right)\right]$$
$$= \frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{1}{M_i}\sum_{j=1}^{M_i}T_{ij}F_{ij}(y)T_{ij}F_{ij}(x) \tag{89}$$

and

$$\begin{aligned}
E_2\left[(\hat{F}_i(y)\hat{F}_i(x))\right] &= E_2\left[E_3\left(\hat{F}_i(y)\hat{F}_i(x)\right)\right] \\
&= E_2\left[C_3\left(\hat{F}_i(y),\hat{F}_i(x)\right)+E_3\left(\hat{F}_i(y)\right)E_3\left(\hat{F}_i(x)\right)\right] \\
&= E_2\left[\frac{T_{ij}^2}{m_i^2}\sum_{j=1}^{m_i}C_3\left(\hat{F}_{ij}(y),\hat{F}_{ij}(x)\right)+\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\right] \\
&= \frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+E_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big) \\
&= \left[\frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+C_2\Big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y),\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\Big)\right. \\
&\quad \left.+\left\{E_2\big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(y)\big)E_2\big(\frac{1}{m_i}\sum_{j=1}^{m_i}T_{ij}F_{ij}(x)\big)\right\}\right] \\
&= \frac{1}{m_iM_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}+\frac{\lambda_i\sigma_{XY,3i}}{m_i}+F_i(y)F_i(x) \qquad (90)
\end{aligned}$$

Using Eqs. (89) and (90) in Eq. (35), and then take expectation to show that

$$E_2(\hat{\sigma}_{XY,3i}) \;=\; \sigma_{XY,3i}+\frac{1}{M_i}\sum_{j=1}^{M_i}\frac{\lambda_{ij}T_{ij}^2\sigma_{XY,3ij}}{t_{ij}}, \qquad (91)$$

which shows that $\hat{\sigma}_{XY,3i}$ is also a biased estimator of $\sigma_{XY,3i}$.

Similarly, we can write from Eq. (37):

$$\hat{\sigma}_{XY,3ij}=\frac{t_{ij}}{t_{ij}-1}\left[\frac{1}{t_{ij}}\sum_{k=1}^{t_{ij}}(I(Y_{ij,k}\le y)I(X_{ij,k}\le x))-\hat{F}_{ij}(x)\hat{F}_{ij}(y)\right]. \qquad (92)$$

Consider the RHS of Eq. (92):

$$E_3\left[\frac{1}{t_{ij}}\sum_{k=1}^{t_{ij}}\Big(I(Y_{ij,k}\le y)I(X_{ij,k}\le x)\Big)\right] \;=\; \frac{1}{T_{ij}}\sum_{k=1}^{T_{ij}}\big(I(Y_{ij,k}\le y)I(X_{ij,k}\le x)\big), \qquad (93)$$

and

$$\begin{aligned}
E_3\left[(\hat{F}_{ij}(x)\hat{F}_{ij}(y))\right] &= C_3\left(\hat{F}_{ij}(x),\hat{F}_{ij}(y)\right)+E_3\left(\hat{F}_{ij}(x)\right)E_3\left(\hat{F}_{ij}(y)\right) \\
&= \frac{\lambda_{ij}\sigma_{XY,3ij}}{t_{ij}}+F_{ij}(y)F_{ij}(x). \qquad (94)
\end{aligned}$$

Use Eqs. (93) and (94) in Eq. (37), and then take expectation to show that

$$E_3(\hat{\sigma}_{XY,3ij}) \;=\; \sigma_{XY,3ij} \qquad (95)$$

which show that $\hat{\sigma}_{XY,3ij}$ is an unbiased estimator of $\sigma_{XY,3ij}$. Now Eq. (83) follows from the results given in Eqs. (87), (91) and (95), which completes the proof.

References

Berger, Y. G. and Muñoz, J. F. (2015). On estimating quantiles using auxiliary information. *Journal of Official Statistics*, 31(1):101–119.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3):597–604.

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition.* John Wiley.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Francisco, C. A. and Fuller, W. A. (1986). Estimation of the distribution function with a complex survey. In *JSM Proceedings, Survey Research Methods Section, American Statistical Association*, pages 37–45.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

Haq, A. (2017a). Estimation of the distribution function under hybrid ranked set sampling. *Journal of Statistical Computation and Simulation*, 87(2):313–327.

Haq, A. (2017b). Two-stage cluster sampling with hybrid ranked set sampling in the secondary sampling frame. *Communications in Statistics-Theory and Methods*, 46(17):8450–8467.

Haq, A., Abbas, M., and Khan, M. (2021). Estimation of finite population distribution function in a complex survey sampling. *Communications in Statistics - Theory and Methods*, 0(0):1–23.

Hussain, S., Ahmad, S., Saleem, M., and Akhtar, S. (2020). Finite population distribution function estimation with dual use of auxiliary information under simple and stratified random sampling. *Plos one*, 15(9):e0239098.

Khoshnevisan, M., Singh, R., Chauhan, P., and Sawan, N. (2007). A general family of estimators for estimating population mean using known value of some population parameter (s). *Far East Journal of Theoretical Statistics*, 22(2):181–191.

Lee, S. E., Lee, P. R., and Shin, K.-I. (2016). A composite estimator for stratified two stage cluster sampling. *Communications for Statistical Applications and Methods*, 23(1):47–55.

Martínez, S., Rueda, M., Arcos, A., and Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics*, 233(9):2265–2277.

Martínez, S., Rueda, M., and Illescas, M. (2022). The optimization problem of quantile and poverty measures estimation based on calibration. *Journal of Computational and Applied Mathematics*, 405(15):113054.

Mayor-Gallego, J. A., Moreno-Rebollo, J. L., and Jiménez-Gamero, M. D. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, 103(1):1–35.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta-35: Statistical Publishing Society.

Nafiu, L., Oshungade, I., and Adewara, A. (2012). Alternative estimation method for a three-stage cluster sampling in finite population. *American Journal of Mathematics and Statistics*, 2(6):199–205.

Nematollahi, N., M, M. S., and Saba, R. A. (2008). Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. *Communications in Statistics - Theory and Methods*, 37(15):2404–2415.

Rao, J., Kovar, J., and Mantel, H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77:365–375.

Rustagi, R. K. (1978). *Some theory of the prediction approach to two stage and stratified two stage cluster sampling*. PhD thesis, The Ohio State University.

Sahoo, L. (1987). A regression-type estimator in two-stage sampling. *Calcutta Statistical Association Bulletin*, 36(1-2):97–100.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Singh, H. P., Singh, S., and Kozak, M. (2008). A family of estimators of finite-population distribution function using auxiliary information. *Acta Applicandae Mathematicae*, 104(2):115–130.

Singh, R., Chauhan, P., Sawan, N., and Smarandache, F. (2009). Improvement in estimating the population mean using exponential estimator in simple random sampling. *International Journal of Statistics & Economics*, 3(A09):13–18.

Smith, T. (1969). A note on ratio estimates in multistage sampling. *Journal of the Royal Statistical Society: Series A (General)*, 132(3):426–430.

Srivastava, M. and Garg, N. (2009). A general class of estimators of a finite population mean using multi-auxiliary information under two stage sampling scheme. *Journal of Reliability and Statistical Studies*, 2(1):103–118.

Stokes, S. L. and Sager, T. W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83(402):374–381.

Sukhatme, P. V., Sukhatme, B., Sukhatme, S., and Asok, C. (1984). Sampling theory with applications. *Indian Society of Agricultural Statistics, New Delhi & IOWA State University Press, Ames, USA*.

Tarima, S. and Pavlov, D. (2006). Using auxiliary information in statistical function estimation. *ESAIM: Probability and Statistics*, 10:11–23.

Yaqub, M. and Shabbir, J. (2020). Estimation of population distribution function involving measurement error in the presence of non response. *Communications in Statistics - Theory and Methods*, 49(10):2540–2559.

SORT 46 (1) January-June 2022, 95-114 DOI: 10.2436/20.8080.02.119

Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models

Jesus Orbe* and Jorge Virto*

Abstract

In this article we consider an extension of the penalized splines approach in the context of censored semiparametric modelling using Kaplan-Meier weights to take into account the effect of censorship. We proposed an estimation method and develop statistical inferences in the model. Using various simulation studies we show that the performance of the method is quite satisfactory. A real data set is used to illustrate that the proposed method is comparable to parametric approaches when assuming a probability distribution of the response variable and/or the functional form. However, our proposal does not need these assumptions since it avoids model specification problems.

MSC: *62N02.*

Keywords: *Censored data, Kaplan-Meier weights, P-splines, semiparametric models, survival analysis.*

1. Introduction

In this paper we present a proposal for estimating regression models where the variable to be explained is censored. That is, our research context is a scenario where the values of the explanatory variables are fully known but some observations of the variable to be explained are not known because there is censored data. This problem is very common in survival or duration analyses, where the sample individuals analysed are tracked over time until the specific event studied occurs (death, failure, breakdown, etc) or the study ends. In practice, there are various types of censoring, but the most common is right censoring. There is an a large body of literature on censored data, much of which can be grouped into two main approaches: one comprising models that directly specify

* Department of Quantitative Methods, University of the Basque Country UPV/EHU, Bilbao, Spain
Received: October 2021
Accepted: April 2022

the effect of the explanatory variables on the variable to be explained (the most widely used of which are those known as Accelerated Failure Times (AFT) see for example Kalbfleisch and Prentice, 2002) and the other comprising hazard models, the best known and most widely applied of which is Proportional Hazard (PH), proposed by Cox (1972). In the former a regression model is specified between the logarithmic transformation of the variable to be explained and the explanatory variables. The latter specifies a relationship between the hazard function of the variable to be explained and the explanatory variables.

PH models have the advantage that the effects of the explanatory variables can be estimated without having to assume a probability distribution for the variable to be explained which is usually unknown. However they also have the drawback that the assumption of proportional hazard functions must be imposed. Another drawback of the hazard functions approach is that the effect of the explanatory variables on the variable to be explained is hard to interpret: the results obtained from Cox model fits are harder to explain to non-statisticians and provide less information than AFT-type models, which are more attractive because they can be interpreted simply and straightforwardly (Wei, 1992; Reid, 1994; Stare, Heinzl and Harrel, 2000; Swindell, 2009). Therefore, in terms of interpretability of results the linear regression model is an attractive alternative to models for hazard functions or hazard ratios. However, its main disadvantage is that the usual estimation procedure for AFT-type models requires a probability distribution to be assumed.

The proposal presented here seeks to make the modelling of this type of data more flexible without imposing restrictions or assumptions that may prove restrictive or false in practice. We also propose an approach for making inferences in this flexible model. Our proposal can be classed as an AFT type model. Several papers using this particular approach can be found in the literature which enable the regression model to be estimated with no need to choose a specific probability distribution. They consider various least squares approaches, and include the papers by Koul, Susarla and Van-Ryzin (1981) and Leurgans (1987), who propose transforming the censored variable, and those by Miller (1976), Buckley and James (1979) and Stute (1993), which present proposals with a similar approach but without transforming the variable to be explained. There is also the rank-based estimation methods approach (see for example Tsiatis, 1990; Lai and Ying, 1992; Jin et al., 2003).

These proposals represent considerable progress in the specification of the model, avoiding the biases derived from wrong choices of probability distribution. But it is possible to go even further in making these methodologies more flexible, since all these proposals consider a known parametric relationship to specify the effect of the explanatory variables on the variable to be explained. In practice, it is quite common for the functional relationship between regressor variables and outcome not to be known. One way of avoiding errors likely to lead to biased conclusions in specifying these effects is not to impose a specific parametric functional relationship between the variable to be explained and the explanatory variable, but to assume only that that relationship is a

smooth function, *i.e.* to consider a nonparametric estimation of that specific effect. The estimation of nonparametric functional relationships involving non-censored data has been widely studied and various proposals have been presented in the literature. They can be grouped into two different approaches: methods based on kernel smoothers (Silverman, 1986; Härdle, 1990) and methods based on spline smoothers (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Eilers and Marx, 1996; Wood, 2017).

Applying these nonparametric estimation techniques is not straightforward in the case of censored data, so the earlier studies must be adapted to take into account the effect of censoring in the estimation process. Our proposal falls under the spline smoothers approach in the specific context of semiparametric regression models with censored data. This semiparametric regression model has already been studied and discussed in regard to samples without censored observations. It was initially analysed by Heckman (1986) and Rice (1986) using an approach with spline smoothers and by Speckman (1988) using an approach with kernel smoothers. Several authors have investigated inference in the semiparametric regression model when the response variable is subject to right censoring. Orbe, Ferreira and Núñez Antón (2003) use an approach based on smoothing splines while Zou, Zhang and Qin (2011) and Chen et al. (2015) use penalized splines and monotone B-splines, respectively. Aydin and Yilmaz (2018) apply the ideas proposed by Koul et al. (1981) in the context of a partial linear regression model and De Uña Álvarez and Roca Pardiñas (2009) consider the use of kernel smoothers in an additive censored regression model.

A previous paper by Orbe and Virto (2018) proposes an extension of the P-splines method of Eilers and Marx (1996), which has become very popular in applications and in theoretical work and is an active area of research (Eilers, Marx and Durbán, 2015), to handle censored responses using Kaplan-Meier weights (Kaplan and Meier, 1958). But the proposal by Orbe and Virto provides no tools to allow statistical inferences to be made, and considers the case of a unique covariate. It is therefore of limited use in practice, where the response variable usually depends on a large set of explanatory variables and it is of interest to draw inferences. Here we propose an extension of that previous paper that enables the technique to be applied to more general problems where the effect of other covariates is incorporated parametrically (parametric component) in addition to the nonparametric component for modelling effects where the functional relationship is not known, that is, a semiparametric regression model. Such extension is a well-studied problem for case of uncensored data (see, for example, Heckman, 1986; Schimek, 2000; Holland, 2017). We also develop variance estimators for both the parametric and nonparametric components and provide the tools needed to develop statistical inferences in this general framework and study performance by calculating coverage probabilities of the confidence intervals for the true values of interest in several simulation studies.

The rest of the paper is organized as follows. Section 2 shows how to extend the P-splines method when the sample has censored observations and proposes a censored data version of penalized splines. Section 3 examines the methodology proposed using simulation studies. Section 4 presents an application of the method to a real data set and Section 5 concludes.

2. Methodology

The existence of censored observations is very common in survival analysis or duration analysis, where the aim is to analyse a variable that measures the duration of an event or state or the time that elapses until a specific event occurs. In other words, we consider a model that allows us to analyse the effect of certain explanatory variables on a variable to be explained T, the duration variable or usually its logarithmic transformation, where some of its observations are censored. Furthermore, we separate the effects of the explanatory variables of the model into two components: a component captures the relationship between some explanatory variables (X) and the response variable assuming a specific parametric functional form (parametric component) and the other component captures the effects of other explanatory variables (Z) whose functional form is unknown (nonparametric component) and which we leave unspecified, without assuming a particular parametric relationship. Therefore, we are considering a semiparametric regression model but in a context where the variable to be explained in the model is right-censored:

$$T_i = X_i^{\intercal}\alpha + f(Z_i) + \varepsilon_i \qquad i = 1,\ldots,n \tag{1}$$

where we assume that the values of the variable T: $t_1,\ldots,t_n$ are independent and generated with an unknown probability distribution function F. That is, we are not assuming any probability distribution for the error term. In addition some observations of that variable T are not known due to the problem known as right censoring. Therefore, what we actually observe in the sample is the variable $y_i = \min(t_i, c_i)$, where the values $c_1,\ldots,c_n$ are the values of the censoring variable C. For the censoring mechanism it will be assume: a) the lifetimes and the censoring times are independent and, b) given the lifetime, the covariates do not provide any further information as to whether censoring will take place or not, *i.e.*, $P[T \leq C|X,Z,T] = P[T \leq C|T]$ (see Stute, 1993, 1999, for a discussion of these assumptions).

We use the indicator $\delta_i = I(t_i \leq c_i)$ to show whether in particular the value t_i is observed, *i.e.*, it is not censored. In addition, X_i is the $(p \times 1)$ vector that collects the values of the p explanatory variables of the parametric component for the i-th individual, α is the $(p \times 1)$ coefficients vector of the model associated with those regressor variables, $f(Z)$ represents the nonparametric component of the model, which captures the unknown functional form of the effect of the regressor variable Z and ε is the error term satisfying $E(\varepsilon|X,Z) = 0$ and $Var(\varepsilon|X,Z) = \sigma^2$.

2.1. Estimation method

Our proposal is based on the nonparametric estimation approach proposed by Eilers and Marx (1996) together with the idea of using Kaplan-Meier weights, proposed by Stute (1993), to control the effect of censoring in the estimation of the model. Thus following this particular approach, if we want to estimate the nonparametric component of the model without assuming a particular functional form $f(\cdot)$ to the unknown effect

of the regressor variable Z, we will use an approximation that rewrites or represents that effect by using a set of q B-splines type basis functions: $B_1(z),\ldots,B_q(z)$ (see, for example, Dierckx, 1993; De Boor, 2001). Thus we rewrite the unknown function as $f(z)=\sum_{j=1}^{q}\gamma_j B_j(z)$.

In order to solve the problem of choosing the number of the knots of the bases, we use the proposal of Eilers and Marx (1996) which introduces a penalty term in the estimation process of the model. This penalty term is based on the idea of previous works by O'Sullivan (1986, 1988) that propose to use a penalty term that measures the smoothness of the function through the integrated squared second derivative of the fitted function. Eilers and Marx (1996) in their proposal of the P-splines methodology suggest using, with the same idea, a different penalty term, which generalizes and simplifies the previous proposal, introducing a penalty but on the difference of the γ_j coefficients of the adjacent B-splines.

In order to account for the effect of censoring we follow the ideas of Orbe and Virto (2018) who extend the possibility of applying the P-spline methodology to the context of samples with censored observations in a simple model. Thus, to estimate the model (1) we propose to minimize the following expression:

$$\sum_{i=1}^{n} w_{[i]}\left[y_{(i)}-x_{[i]}^{\mathsf{T}}\alpha-\sum_{j=1}^{q}\gamma_j B_j(z_{[i]})\right]^2+\lambda\sum_{j=k+1}^{q}(\Delta^k\gamma_j)^2 \tag{2}$$

where $y_{(1)},\ldots,y_{(n)}$ are the ordered values of the observed variable $y_i=min(t_i,c_i)$, $x_{[i]}^{\mathsf{T}}$ is the $(1\times p)$ vector with the values of the regressors of the parametric component for the individual corresponding to the ordered observation $y_{(i)}$, $w_{[i]}$ is the Kaplan-Meier weight associated with that observation $y_{(i)}$ and this weight is calculated using the estimator $(\hat{F}_n)$ (Kaplan and Meier, 1958) of the probability distribution function F of the variable to be explained T:

$$w_{[i]}=\hat{F}_n(y_{(i)})-\hat{F}_n(y_{(i-1)})=\frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j]}} \tag{3}$$

without the need to assume a probability distribution for the error term, therefore a flexible methodology is used regarding to parametric assumption of the error. Furthermore $\Delta\gamma_j$ denotes the difference between the coefficients of adjacent B-splines $(\gamma_j-\gamma_{j-1})$ and $\Delta^k\gamma_j$ indicates that this difference is of order k. This difference measures the smoothness of the function $f(z)$, the larger the difference between the coefficients of adjacent B-splines the less smooth the function. Finally the parameter λ is the smoothing parameter that controls the degree of the smoothness of the estimated function in the estimation process.

The expression to minimize (2) can be rewritten in matrix form as follows:

$$(Y-X\alpha-B\gamma)^{\mathsf{T}}W(Y-X\alpha-B\gamma)+\lambda\gamma^{\mathsf{T}}D_k^{\mathsf{T}}D_k\gamma \tag{4}$$

where X is the $(n \times p)$ design matrix for the variables of the parametric component. Y is the vector of the observed variable to be explained. B is a $(n \times q)$ matrix where $B_{ij} = B_j(z_i)$. W is a $(n \times n)$ diagonal matrix with Kaplan-Meier weights. D_k is the matrix used to rewrite the Δ^k term in matrix form.

2.2. Algorithm

The optimization process of the expression (4) leads to the following equations:

$$\left(X^{\intercal}WX\right)\alpha = X^{\intercal}W(Y - B\gamma) \tag{5}$$

$$\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)\gamma = B^{\intercal}W(Y - X\alpha) \tag{6}$$

In practice, the estimations of α and γ can be obtained by means of an iterative process or backfitting algorithm that iteratively solves each set of equations (5) and (6) until the convergence of the estimators is reached. We describe the algorithm process as follows:

- Step 1. In equation (6) give initial value of $\widehat{\alpha}_{(0)} = \vec{0}$ and estimate γ by $\widehat{\gamma}_{(0)} = \left[B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}WY$.
- Step 2. Substitute γ by $\widehat{\gamma}_{(0)}$ in equation (5) and estimate α by $\widehat{\alpha}_{(1)} = [X^{\intercal}WX]^{-1} X^{\intercal}W(Y - B\widehat{\gamma}_{(0)}) = [X^{\intercal}WX]^{-1} X^{\intercal}W(I - H_c)Y$ where $H_c = B\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)^{-1} B^{\intercal}W$ is the smoothing matrix for the censored case obtained from equation (6).
- Step 3. Substitute α by $\widehat{\alpha}_{(1)}$ in equation (6) and estimate γ by $\widehat{\gamma}_{(1)} = \left[B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(Y - X\widehat{\alpha}_{(1)})$.
- Step 4. Iterate step 2 and step 3 until convergence is achieved.

The algorithm is considered to have converged when the difference between the GCV_c (see equation 8) of two successive iterations is less than a really small threshold: $|GCV_c(new) - GCV_c(old)| < 0.00001 \cdot GCV_c(new)$.

2.3. Choice of smoothing parameter and knots

It should be noted that in this iterative process we need to make a number of choices, such as the number of knots (K_c) and the choice of the smoothing parameter λ, in order to estimate the components of the model. The use of a penalty term in the optimization criterion makes the determination of the number of knots not a crucial decision as long as a sufficient number of knots is chosen. To choose this number of knots in samples with censored data we propose the following automatic choice criterion that takes into account the sample information available due to the existence of censored data by multiplying by one minus the proportion of censored observations:

$$K_c = round\left(\min\left(\frac{m}{4}, 40\right) \cdot (1 - PC)\right) \tag{7}$$

where m is the number of distinct values of the Z variable of the nonparametric component and PC represents the level of censoring, measured as a percentage, existing in the analysed sample. The expression (7) is a modification to the one proposed for the choice of the number of knots in Ruppert (2002) that we propose for application in contexts with censored data.

The choice of the smoothing parameter is a more relevant choice. To choose an optimal smoothing level we propose to use the following version of the generalized cross validation (GCV) criterion adapted for application in contexts with censored data:

$$GCV_c = \sum_{i=1}^{n} \frac{w_{[i]}\left(y_{(i)} - \hat{y}_{(i)}\right)^2}{(n - \phi tr(H_c))^2} \tag{8}$$

where ϕ is a parameter that tries to correct for the overfitting problem that occurs when using the ordinary GCV criterion. Wood (2017) proposes to use what he refers to as the double cross validation and suggests using a value of $\phi = 1.5$. This value is justified in different ways in the literature, see for example Kim and Gu (2004) for the uncensored case and Orbe and Virto (2021) for the censored case. The performance of proposal (8) has been analysed using a simulation study and, as in the uncensored case, the choice of $\phi = 1.5$ is better in almost all situations than $\phi = 1$, with the difference increasing as the censoring increases.

2.4. Variances estimation

In this section we develop the necessary tools to perform statistical inferences for the parametric and nonparametric components.

In order to determine the variance of the parametric component, we first solve equation (6) getting $\gamma = \left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)^{-1} B^{\intercal}W(Y - X\alpha)$. Therefore, substituting $B\gamma = H_c(Y - X\alpha)$ in equation (5) we get $(X^{\intercal}WX)\,\alpha = X^{\intercal}W\left[Y - H_c(Y - X\alpha)\right]$. Solving for α we obtain $\widehat{\alpha} = \left[X^{\intercal}W(I - H_c)X\right]^{-1} X^{\intercal}W(I - H_c)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\begin{aligned}\widehat{Var}(\widehat{\alpha}) = \widehat{\sigma}^2 \Big\{ &\left(X^{\intercal}W(I - H_c)X\right)^{-1} X^{\intercal}W(I - H_c)(I - H_c)^t WX \\ &\left(\left(X^{\intercal}W(I - H_c)X\right)^{-1}\right)^t \Big\}\end{aligned} \tag{9}$$

In a similar way, we solve equation (5) getting $\alpha = \left(X^{\intercal}WX\right)^{-1} X^{\intercal}W(Y - B\gamma)$. Plugging $X\alpha = X\left(X^{\intercal}WX\right)^{-1} X^{\intercal}W(Y - B\gamma) = H_p(Y - B\gamma)$, where $H_p = X\left(X^{\intercal}WX\right)^{-1} X^{\intercal}W$, in equation (6) we get $\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)\gamma = B^{\intercal}W\left[Y - H_p(Y - B\gamma)\right]$. Solving for γ we get $\widehat{\gamma} = \left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(I - H_p)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\begin{aligned}\widehat{Var}(\widehat{\gamma}) = \widehat{\sigma}^2 \Big\{ &\left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(I - H_p)(I - H_p)^t WB \\ &\left(\left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1}\right)^t \Big\}\end{aligned} \tag{10}$$

In order to calculate these estimated variances we need to estimate the σ^2 parameter. We propose the estimator given by the following expression:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} nw_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{n - tr(H_c) - p}$$

3. Simulation study

In this section the performance of the proposed methodology is studied using a simulation study. In order to do that we consider the next semiparametric model:

$$T_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + f(Z_i) + \varepsilon_i \tag{11}$$

where for the parametric component of the model: the variable X_1 is generated from a uniform distribution $U(0,2)$, X_2 from a uniform distribution $U(-1,3)$, being $\alpha_1 = -1$ and $\alpha_2 = 1$ the values of the coefficients. For the nonparametric component, we consider three different cases for the relationship $f(\cdot)$ between T and a relevant covariate Z, see Table 1 for the chosen functional forms and the probability distribution of the variable Z. For the distribution of the error term (ε) has been used the normal distribution $N(0,\sigma^2)$, where the value of σ^2 parameter has been chosen to obtain a similar signal/noise (SN) ratio in each example (see Table 1). In order to study the effect of censoring, we consider a censoring variable C generated independently from a uniform distribution $U(1,b)$. The value of parameter b changes to consider three different levels of censored data: 10%, 25% and 40%. Therefore, we observe $(y_1, x_{11}, x_{21}, z_1, \delta_1), \ldots, (y_n, x_{1n}, x_{2n}, z_n, \delta_n)$ a sample of size n, where $y_i = min(t_i, c_i)$ is the observed survival time, *i.e.*, the minimum between the survival time t_i and the censoring value c_i. In addition, it is known through the indicator variable $\delta_i = I(t_i \leq c_i)$ which observations are not censored. We use three sample sizes: $n = 200$, $n = 500$ and $n = 1000$. For each of the nine scenarios, three sample sizes for three levels of censorship, we consider 1000 Monte Carlo replications.

Table 1. *Three Case Studies.*

Name	z_i	$f(z_i)$	σ_ε^2	SN ratio
Case (i): Quadratic	$z_i \sim U[0,4]$	$2 + 4z_i - z_i^2$	0.40	3.5
Case (ii): Sinusoidal	$z_i \sim U[0,10]$	$2 + exp\{sin(z_i)\}$	0.20	3.3
Case (iii): Logit	$z_i \sim U[0,1]$	$2 + \frac{1}{1 + exp\{-20(z_i - 0.5)\}}$	0.06	3.3

For each of the 27 cases analysed in this simulation study we have estimated model (11) following the estimation proposal presented in the previous section, the censored P-

spline estimator (CPS), where the choice of the smoothing parameter λ and the number of knots of B-splines have been chosen using formulas (8) and (7), respectively.

Tables 2, 3 and 4 present a general summary of the results obtained for each combination of censoring level and sample size in each of the three cases of functional forms studied for model (11). That is, Table 2 summarizes the estimation of case (i), where $f(z)$ is a quadratic function. The first two rows of Table 2 present the estimated Mean Square Error (MSE) of each coefficient (α_1 and α_2) of the parametric component:

$$MSE(\hat{\alpha}_p) = \frac{1}{1000}\sum_{j=1}^{1000}(\alpha_p - \hat{\alpha}_{pj})^2 \qquad p = 1,2$$

and the third row the Averaged Mean Square Error (AMSE) of the nonparametric component:

$$AMSE = \frac{1}{1000}\sum_{j=1}^{1000}\left(\frac{\sum_{i=1}^{n}(f(z_i) - \hat{f}_j(z_i))^2}{n}\right)$$

Rows four to six of Table 2 present the empirical bias and rows seven to nine the coverage probabilities of the 95% confidence intervals based on the resampling.

Tables 3 and 4 present the same information for the estimates of case (ii) and (iii), where $f(z)$ is a sinusoidal function and a logit function, respectively. Tables 2 to 4 show the good performance of the proposed method in terms of MSE and AMSE, empirical bias and coverage probabilities.

Furthermore, if we focus on the estimation of each component of model (11), we have that for case (i), quadratic function: Figure 1(a) presents the MSE estimates for the nonparametric component using different censoring levels and sample sizes, where, as can be seen, the estimates of the nonparametric component improve as the sample size increases and the level of censoring in the sample decreases. Figures 1(b) and (c) show the estimates of the coefficients of the parametric component (α_1 and α_2), where it can be seen that the coefficient estimates are good and that their accuracy also improves as the sample size increases and the level of censoring in the sample decreases. In addition, Figure 1(d) presents the mean value of the estimates of the quadratic form function compared to the true functional form to be estimated. As can be seen, the proposal we made works very well reflecting the true functional form of $f(\cdot)$. In this Figure 1(d), we can also verify the good performance of the asymptotic confidence intervals generated with the estimates of the variances proposed in the previous section. As can be seen, for a confidence level of 95%, the proposed mean confidence interval (blue lines) is consistent with the corresponding 95th percentile interval of the simulations (green lines). Finally, the coverage probabilities of the confidence intervals presented in Table 2 show that the actual coverage probability is quite close to the nominal coverage probability.

Similar results, where the good performance of our proposals can be appreciated, are obtained for case (ii), sinusoidal function, see Figures 2(a)-(d), and for case (iii), logit function, see Figures 3(a)-(d).

As suggested by the referees, we conduct additional simulations considering a normal distribution for the censoring variable and also additional simulations considering

non-normal error distributions such as the Weibull distribution. The new results obtained (not shown) confirm the good performance of the proposed method and are consistent with those presented in this section.

Table 2. *Results of simulation study for the quadratic function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	3.090	3.741	5.965	1.324	1.440	2.370	0.521	0.656	0.992
$\widehat{\alpha}_2$	0.722	0.906	1.581	0.275	0.302	0.541	0.121	0.181	0.259
$\widehat{f}$	9.783	12.170	21.109	4.126	5.056	8.730	2.105	2.580	4.424
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00149	0.00099	0.01130	-0.00319	-0.00290	0.00042	0.00214	0.00354	0.00575
$\widehat{\alpha}_2$	0.00289	0.00206	-0.00257	0.00049	0.00039	-0.00335	0.00036	-0.00036	-0.00195
$\widehat{f}$	-0.00033	-0.00148	-0.01630	0.00239	0.00272	0.00067	-0.00232	-0.00253	-0.00575
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.938	0.955	0.947	0.928	0.950	0.947	0.946	0.946	0.948
$\widehat{\alpha}_2$	0.945	0.946	0.934	0.941	0.960	0.943	0.960	0.926	0.957
$\widehat{f}$	0.938	0.941	0.923	0.939	0.939	0.925	0.946	0.936	0.933

Table 3. *Results of simulation study for the sinusoidal function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.806	1.060	1.521	0.285	0.362	0.560	0.136	0.154	0.233
$\widehat{\alpha}_2$	0.189	0.266	0.376	0.062	0.087	0.132	0.035	0.044	0.064
$\widehat{f}$	4.088	5.205	7.970	1.702	2.023	3.072	0.870	1.047	1.545
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00543	-0.00202	0.00083	0.00085	0.00098	0.00209	0.00060	0.00016	0.00148
$\widehat{\alpha}_2$	-0.00060	-0.00073	-0.00145	0.00051	0.00058	-0.00093	-0.00016	-0.00017	-0.00136
$\widehat{f}$	0.00674	0.00311	-0.00152	-0.00116	-0.00167	-0.00324	-0.00021	0.00010	-0.00077
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.936	0.925	0.956	0.955	0.944	0.948	0.956	0.940
$\widehat{\alpha}_2$	0.949	0.930	0.938	0.952	0.938	0.944	0.944	0.943	0.962
$\widehat{f}$	0.930	0.927	0.918	0.932	0.941	0.932	0.942	0.938	0.941

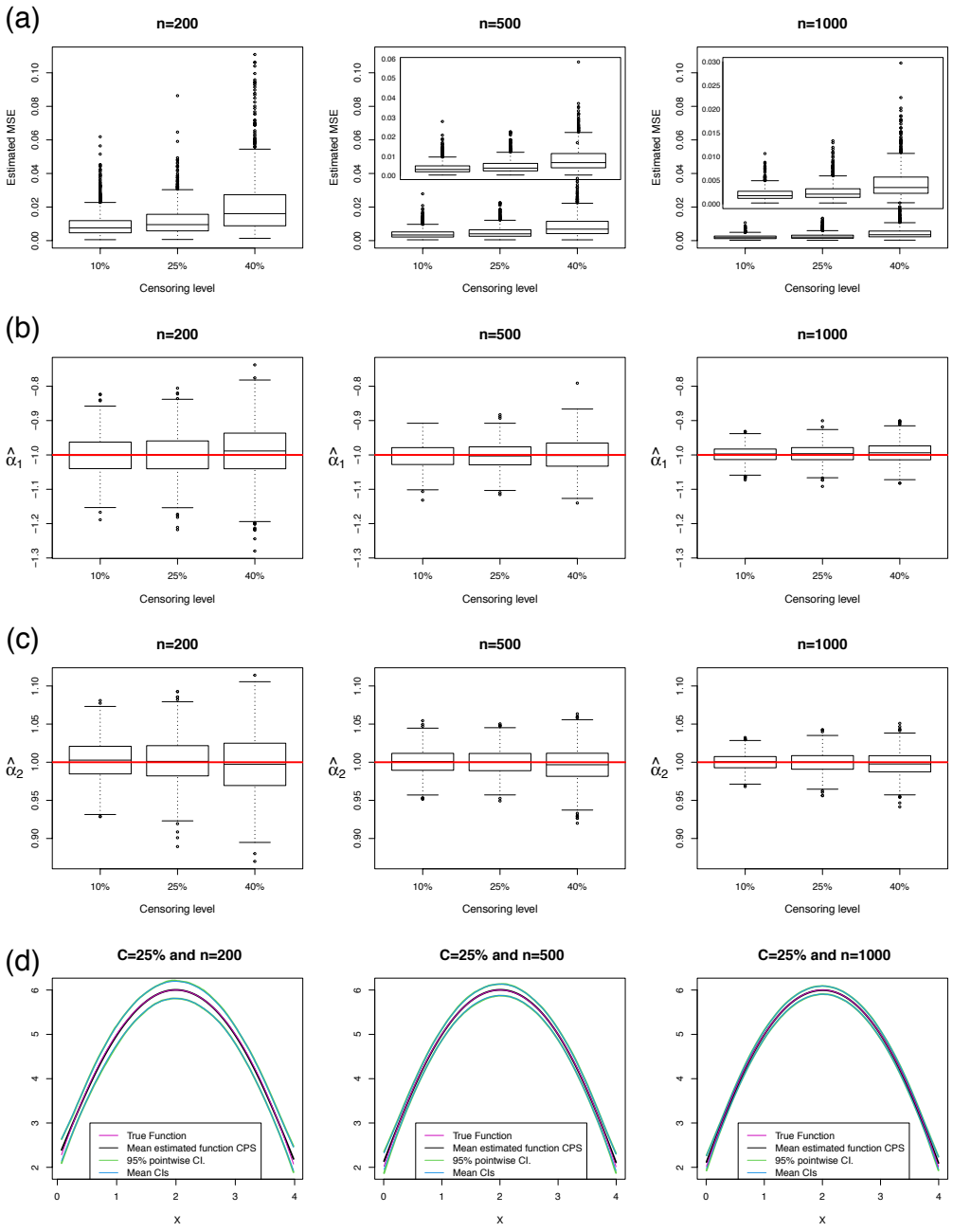


Figure 1. *Results of simulation study for the quadratic function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the quadratic form function compared to the true functional form to be estimated.*

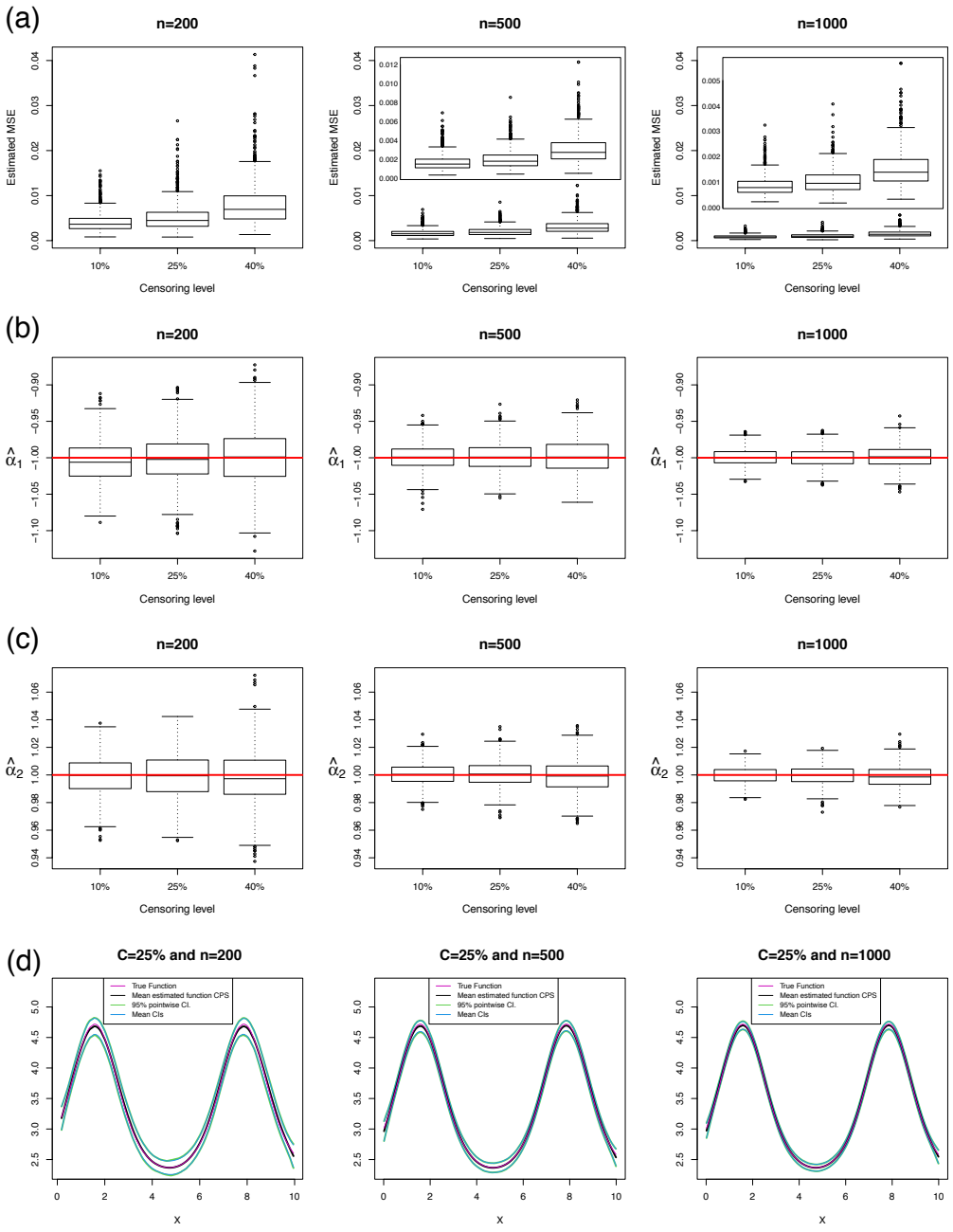


Figure 2. *Results of simulation study for the sinusoidal function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b)* $\hat{\alpha}_1$*. (c)* $\hat{\alpha}_2$*. (d) Mean value of the estimates of the sinusoidal form function compared to the true functional form to be estimated.*

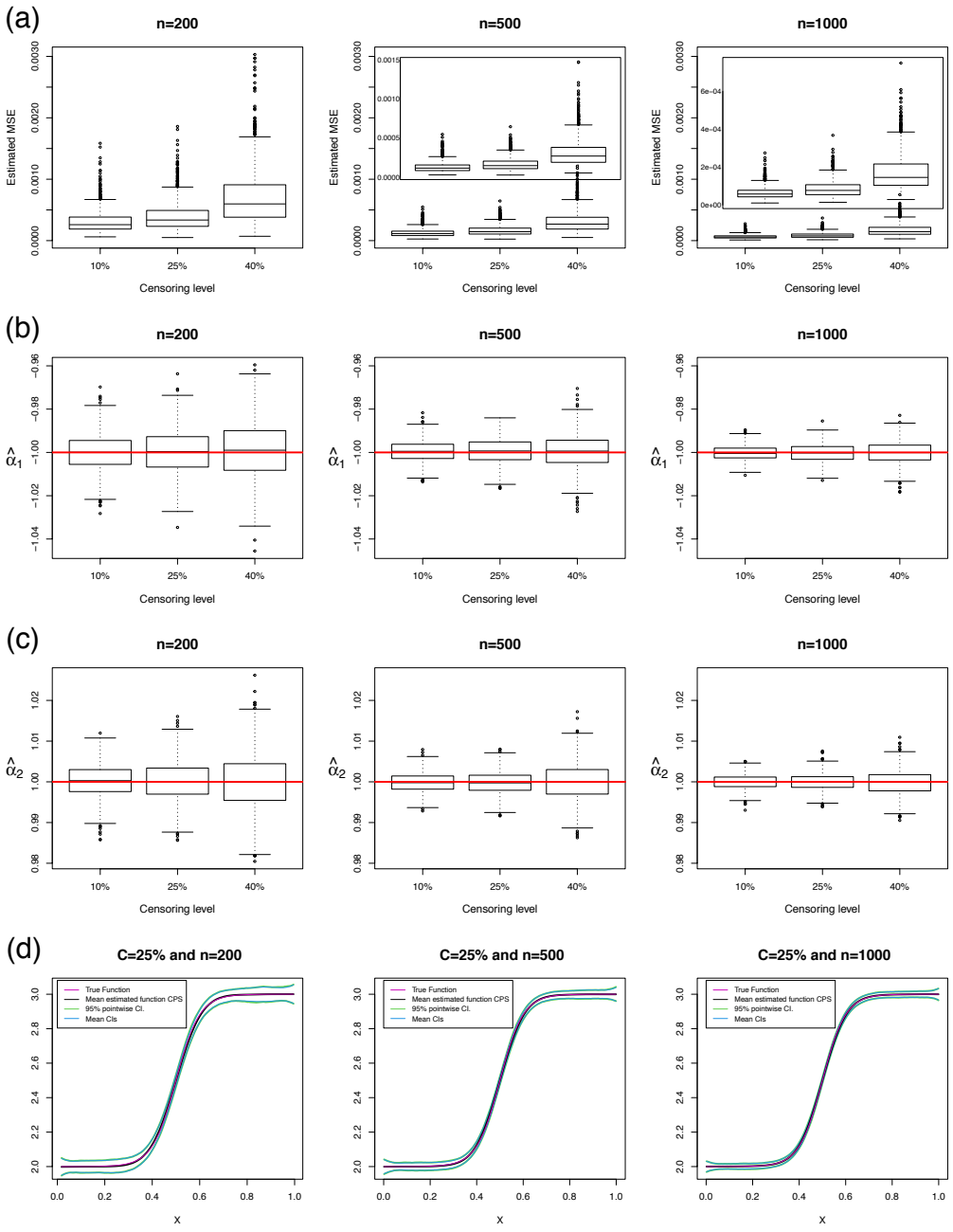


Figure 3. *Results of simulation study for the logit function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b)* $\hat{\alpha}_1$*. (c)* $\hat{\alpha}_2$*. (d) Mean value of the estimates of the logit form function compared to the true functional form to be estimated.*

Table 4. *Results of simulation study for the logit function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.072	0.098	0.172	0.024	0.036	0.064	0.011	0.016	0.027
$\widehat{\alpha}_2$	0.017	0.025	0.046	0.006	0.008	0.019	0.003	0.004	0.009
$\widehat{f}$	0.309	0.397	0.710	0.128	0.164	0.311	0.065	0.085	0.169
	Empirical Bias								
$\widehat{\alpha}_1$	0.00012	0.00029	0.00101	0.00042	0.00061	0.00035	-0.00029	-0.00022	-0.00011
$\widehat{\alpha}_2$	0.00026	0.00015	-0.00008	-0.00015	-0.00026	-0.00002	-0.00002	-0.00003	-0.00014
$\widehat{f}$	-0.00037	-0.00038	-0.00098	-0.00022	-0.00029	-0.00040	0.00035	0.00020	0.00014
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.955	0.933	0.953	0.944	0.941	0.946	0.941	0.948
$\widehat{\alpha}_2$	0.950	0.930	0.924	0.939	0.947	0.938	0.957	0.938	0.956
$\widehat{f}$	0.944	0.939	0.916	0.943	0.938	0.93 0	0.949	0.941	0.938

4. Empirical application: PBC data

The Mayo Clinic Primary Biliary Cirrhosis dataset contains information from 418 Mayo Clinic patients with primary biliary cholangitis (PBC), previously called primary biliary cirrhosis, an autoimmune disease of the liver. The first 312 cases in the dataset participated in a Mayo Clinic trial in PBC conducted between 1974 and 1984 comparing the drug D-penicillamine (treatment) with a placebo. The dataset provides information about the observed survival time from the date of registration in the trial and a large number of clinical, biochemical, serologic and histologic variables such as patient's age at first diagnosis, severity of edema (0 no edema, 0.5 moderate and 1 for severe edema), blood values related to liver function such as bilirubin, albumin, alkaline phosphotase and prothrombin time amid other explanatory variables, and an indicator of patient status (dead or alive) in July 1986. The dataset can be downloaded from the R package *survival* (Therneau, 2021; R Core Team, 2018). The additional cases are from an independent set of 106 Mayo Clinic primary biliary cholangitis patients who were elegible for the trial but declined to participate. This dataset has been previously used, for example, in Dickson et al. (1989), Therneau and Grambsch (2000) and Fleming and Harrington (2005), in censored regression models.

The studies by Therneau and Grambsch (2000) and Fleming and Harrington (2005) deal with the relationship between the covariates and the survival response variable. They conclude that age, edema score, bilirubin and albumin logarithms and prothrombin time are the variables that best explain patient survival. In addition, these studies analyse the need for transformations of the continuous variables in the proposed model

Table 5. *Estimate and standard deviation (SD) of estimated parameters for the Mayo Clinic Primary Biliary Cirrhosis dataset from AFT, Stute and CPS methods.*

	age	edema	trt	log(albumin)	log(bili)
AFT	-0.0246	-0.7692	-0.0627	1.4880	-0.5356
	(0.0065)	(0.2303)	(0.1273)	(0.5268)	(0.0694)
Stute	-0.0166	-0.9249	-0.0950	1.6161	-0.3028
	(0.0076)	(0.3489)	(0.1371)	(0.6015)	(0.0732)
CPS	-0.0168	-0.9163	-0.0991	1.6197	-0.3061
	(0.0064)	(0.1900)	(0.1291)	(0.4578)	(0.0633)

concluding that the relationship between prothrombin time (protime) and patient survival is likely to be non-linear.

In this application we incorporate the protime variable into the model in a flexible way only assuming that prothrombin time enters in the model via some unknown smooth function $f(\cdot)$:

$$log(T) = \alpha_1 + \alpha_2 age + \alpha_3 edema + \alpha_4 trt + \alpha_5 log(albumin) + \alpha_6 log(bili) + f(protime) + \varepsilon \tag{12}$$

We estimated model (12) using the censored P-spline method proposed in section 2. To evaluate the performance of the censored P-spline estimator, a quadratic relationship between the logarithm of survival and the protime variable has been proposed as an alternative, *i.e.*, $f(protime) = \alpha_7 protime + \alpha_8 protime^2$ in equation (12). Assuming that this parametric specification is correct, two methodologies known and proposed in the literature on survival analysis can be used to fit the model (12). These estimators can be used as a benchmark to evaluate the performance of the censored P-spline method proposed. The first and more restrictive approach is the parametric Accelerated Failure Time (AFT) methodology (Kalbfleisch and Prentice, 2002), based on the restricted assumptions of knowing the probability distribution of the response variable and the functional form relating the protime variable and patient survival, that estimates the α coefficients of the model using the maximum likelihood estimator. Thus, considering an AFT lognormal model, we estimate the α coefficients assuming a normal probability distribution. The second methodology, proposed by Stute (1993), is less restrictive in that it does not need the assumption of the probability distribution of the response variable, but it also trusts the quadratic functional form. That is, it needs to know the form of the relationship between the response variable and the covariate. This methodology estimates coefficients using weighted least squares via Kaplan-Meier weights (Stute, 1993).

Table 5 presents the estimates of the parametric components of the model (12) using these three methods. It can be seen that all three methods generate similar estimates and result in a biologically reasonable model estimate. As previously reported in the literature, all three methods agree that treatment with the drug D-penicillamine (treatment) has no significant effect on patient survival.

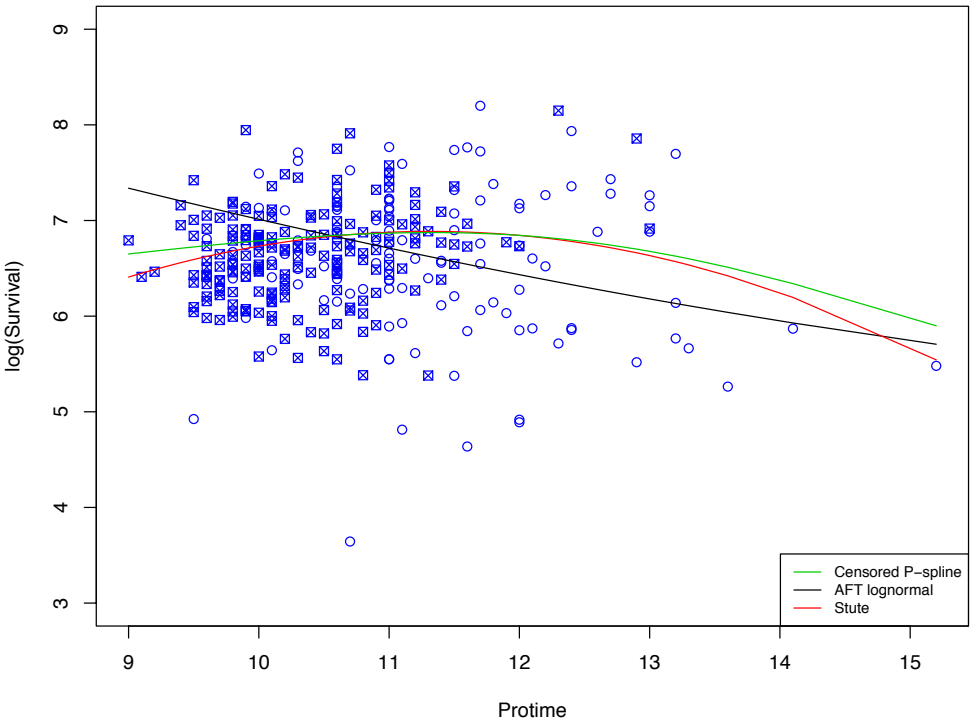


Figure 4. *Estimated relationship using three methodologies: AFT lognormal, Stute's approach and CPS estimator*

Figure 4 shows the estimates of the unknown function $f(protime)$ for the three approaches with the scatterplot of observed log survival time versus prothrombin time. Patients indicated by ◯ are dead and those indicated by ☒ are alive in July 1986; that is, the dead patients have uncensored survival times and the live patients have censored survival times.

In conclusion, the AFT methodology and Stute's proposals performance depends on the correct specification of the relationship between the duration and the protime variable. In this application it seems that the relationship between log survival and prothrombin time is quadratic, so both these methodologies perform reasonably well. Our proposal does not need to assume a specific parametric functional form and, however, it adequately estimates the relationship obtaining very similar results to the previous ones. However, if the functional form had been wrongly chosen these parametric methods would have led to a serious problem of incorrect specification and therefore to wrong conclusions. Therefore, we can see our approach as a robust solution to misspecification of the model.

5. Discussion and conclusion

In this paper, we have proposed an estimation method in the context of censored semiparametric models based on the P-spline approach of Eilers and Marx (1996) using Kaplan-Meier weights to take into account the effect of censorship. We present an extension of the estimation methodology proposed by Orbe and Virto (2018) to a context with more than one explanatory variable, which is very useful from a practical point of view. Furthermore, we develop the necessary tools to perform statistical inferences in this general framework, providing, for example, confidence intervals for both the nonparametric component and the coefficients associated with the regressors of the parametric component. The simulation studies conducted illustrate the good performance of the estimation method which satisfactorily estimates both the nonparametric component and the coefficients associated with the parametric part in the various examples studied. Furthermore, the accuracy of estimates improves as the censored level reduces and the sample size is increased. The coverage probabilities of the confidence intervals proposed have been calculated in several simulation studies and it has been found that the actual coverage probability is quite close to the nominal coverage probability in all the scenarios analysed.

The application to real data serves to illustrate the potential advantages of our proposal which is comparable with the parametric method AFT and Stute's approach when the functional form chosen is correct. Otherwise, it must be mentioned that if the functional form or the probability distribution are wrongly chosen this would lead to a serious problem of incorrect specification of the model and therefore to incorrect conclusions. The proposed method would be more flexible and robust as it does not need to impose a specific probability distribution for the response variable, nor assume a functional form for the relationship between the censored response variable and the covariate, which are usually unknown in practice. Therefore, its application in samples with censored data is particularly useful in contexts of survival or duration analysis where censored observations are common.

Funding

This study was supported by the Basque Government and the Spanish Research Agency of the Ministry of Science and Innovation under research grants UPV/EHU Econometrics Research Group IT1359-19 and PID2019-105183GB-I00.

References

Aydin, D. and Yilmaz, E. (2018). Modified estimators in semiparametric regression models with right-censored data. *Journal of Statistical Computation and Simulation*, 88:1470–1498.

Buckley, J. J. and James, I. R. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Chen, W., Li, X., Wang, D., and Shi, G. (2015). Parameter estimation of partial linear model under monotonicity constraints with censored data. *Journal of the Korean Statistical Society*, 44:410–418.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:187–202.

De Boor, C. (2001). *A Practical Guide to Splines, revised version*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

De Uña Álvarez, J. and Roca Pardiñas, J. (2009). Additive models in censored regression. *Computational Statistics and Data Analysis*, 53:3490–3501.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10:1–7.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11:89–121.

Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. John Wiley & Sons, Hoboken: New Jersey.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Härdle, W. (1990). *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48:244–248.

Holland, A. D. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis*, 153:211–235.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90:341–353.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kim, Y. J. and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:337–356.

Koul, H., Susarla, V., and Van-Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9:1276 – 1288.

Lai, T. L. and Ying, Z. (1992). Linear rank statistics in regression analysis with censored or truncated data. *Journal of Multivariate Analysis*, 40:13–45.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika*, 74:301–309.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, 63:449–464.

Orbe, J., Ferreira, E., and Núñez Antón, V. (2003). Censored partial regression. *Biostatistics*, 4:109–121.

Orbe, J. and Virto, J. (2018). Penalized spline smoothing using Kaplan-Meier weights with censored data. *Biometrical Journal*, 60:947–961.

Orbe, J. and Virto, J. (2021). Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. *Journal of Statistical Computation and Simulation*, 91:1–33.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, 1:502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9:363–379.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1994). A conversation with sir david cox. *Statistical Science*, 9:439–455.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letter*, 4:203–208.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, 91:525–540.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50:413–436.

Stare, J., Heinzl, H., and Harrel, F. (2000). On the use of buckley and james least squares regression for survival data. In Ferligoj, A. and Mrvar, A., editors, *New Approaches in Applied Statistics*, volume 16, pages 125–134. Metodološki zvezki, Ljubljana: Eslovenia.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45:89–103.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 9:1089–1102.

Swindell, W. R. (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology*, 44:190–200.

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-11.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18:354–372.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science Series. CRC press, Boca Raton: Florida.

Zou, Y., Zhang, J., and Qin, G. (2011). A semiparametric accelerated failure time partial linear model and its application to breast cancer. *Computational Statistics and Data Analysis*, 55:1479–1487.

SORT 46 (1) January-June 2022, 115-136 DOI: 10.2436/20.8080.02.120

Topological Data Analysis and its usefulness for precision medicine studies

Raquel Iniesta*,[1], Ewan Carr[1], Mathieu Carrière[2], Naya Yerolemou[3], Bertrand Michel[4] and Frédéric Chazal[5]

Abstract

Precision medicine allows the extraction of information from complex datasets to facilitate clinical decision-making at the individual level. Topological Data Analysis (TDA) offers promising tools that complement current analytical methods in precision medicine studies. We introduce the fundamental concepts of the TDA corpus (the simplicial complex, the Mapper graph, the persistence diagram and persistence landscape). We show how these can be used to enhance the prediction of clinical outcomes and to identify novel subpopulations of interest, particularly applied to understand remission of depression in data from the GENDEP clinical trial.

MSC: *Statistical aspects of big data and data science (62R07) and Topological data analysis (62R40)*

Keywords: *Precision medicine, data shape, topology, topological data analysis, persistence diagram, Mapper, persistence landscapes, machine learning.*

1. Precision medicine: what are the current needs?

The field of precision medicine is focused on the development of sophisticated algorithms that, by exploiting patient data – on clinical measurements, genomics, proteomics, medical imaging, etc. – can guide clinicians to make more accurate diagnoses, prognoses and treatment choices tailored to individual patients. The datasets used to develop these

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. *Corresponding author: raquel.iniesta@kcl.ac.uk

[2] Inria Sophia-Antipolis, DataShape Team, Biot, France.

[3] The University of Oxford and The Alan Turing Institute. UK.

[4] Ecole Centrale de Nantes,LMJL – UMR CNRS 6629, Nantes, France.

[5] Inria Saclay - Ile-de-France, Alan Turing Bldg, Palaiseau, France.

Received: September 2021.
Accepted: April 2022.

models present multiple complexities. They routinely include information for thousands of subjects, and the number of included variables can easily exceed millions (i.e., these datasets are high-dimensional), variables tend to be highly correlated, and may interact in complex ways that may not be immediately obvious. These factors combined often limit the utility of classical statistical procedures in the analysis of these data. In recent years, machine learning (ML) (Mitchell, 1997, 2006), a set of tools at the interface between computer sciences and statistics, has been used in precision medicine to overcome some of these limitations. The use of ML has led to the development of interesting predictive models built from complex data sets (Ekins et al., 2019; Ho et al., 2019; Rajkomar, Dean and Kohane, 2019; Iniesta, Stahl and McGuffin, 2017). However, the success of ML for these datasets has varied across medical areas – performing moderately well in some diseases but very poorly in others (Adamson and Welch, 2019; Iniesta et al., 2016, 2017, 2018), leaving considerable room for improvement. Recently, several works including studies on COVID-19 research, have emphasised the increasing demand of novel methods that can better deal with such complexity (Khan et al., 2019, 2021).

One of the key challenges in building models that can accurately predict outcomes for new patients is correctly identifying sources of heterogeneity among patients (i.e., sources that could contribute to observed differences in patient outcomes) and including these in the model in the form of predictor variables. When tailoring the choice of medical treatment to patients' pre-treatment characteristics, methods to identify subgroups in terms of treatment effectiveness – for example, where patients respond similarly to treatment within the group, and differently between groups – constitute one of the most prominent challenges currently for medical statisticians (Sies, Demyttenaere and Mechelen, 2019).

In addition to developing predictive models, methods for visualising data in high dimensions can facilitate decision-making for diagnosis and treatments targeting. Most classical tools, such as scatter plots or heat maps, are often restricted to two dimensions (Qu et al., 2019). Although new technologies have been used to create visualisation tools applicable to complex data, fields like genomics research are rapidly evolving and continuous advancement in visualisation techniques is needed (Nusrat, Harbig and Gehlenborg, 2019).

In recent years a growing literature has highlighted the benefits of applying topological techniques in precision medicine studies. For example, to identify genetic influences on patient survival in breast cancer (Nicolau, Levine and Carlsson, 2011), to improve treatment targeting for patients with spinal cord or traumatic brain injury by uncovering previously hidden data relationships in 20-year old data (Nielson et al., 2017), or to identify disease trajectories in type 2 diabetes data (Dagliati et al., 2020).

This paper aims to provide a first introduction to some of the basic topological concepts that form the field of Topological Data Analysis (TDA): the simplicial complex, the Mapper graph, the persistence diagram and the persistence landscape. We show how these techniques offer promising tools to reveal data structures not readily accessible using other statistical techniques, which may subsequently help machine learning models

in predicting clinical outcomes. We show an application of these methods to investigate remission of depression in data from the GENDEP clinical trial. We also summarise the software implementations of these techniques.

2. Introducing Topological Data Analysis

TDA is a promising field that has emerged from different works in applied algebraic topology (Edelsbrunner, Letscher and Zomorodian, 2000; Zomorodian and Carlsson, 2005; Ghrist, 2018). It aims to provide well-founded mathematical, statistical, and algorithmic methods to infer, analyse, visualise and exploit the complex topological and geometric structure of data (Chazal, 2016). The field is based on topology, the branch of mathematics born in response to Riemann's request in 1867 for "a good foundation of the concept of space" (Riemann and Clifford, 1998). In contrast with the more familiar field of geometry – the study of the shape of the space, that is, what the space *looks like* – topology can be broadly defined as the study of only those shape properties that are unaffected by continuous transformations such as stretching, shrinking, bending and twisting (examples of non-continuous transformations are cutting or gluing) (Kosniowski, 1980). For example, if a torus (a surface like a ring doughnut, as shown in Figure 1) is stretched horizontally, it does not change the fact that there is only one 'hole' on the inside; thus, this property is preserved despite transformation. Moreover, topological techniques assume *coordinate invariance*, the property that topological features are defined not in terms of their position on a coordinate system, but rather, in terms of their shape. Therefore, TDA can identify a torus regardless of whether the torus is compressed or stretched; the torus and its transformations are said to be *topologically equivalent*. Topological invariants like the number of holes and cavities are properties of a topological space that are shared by the space and all its topological equivalents (Henle, 1994). Properties such as these characterise the *invariant shape* of a space.

If we now move to the world of data, as Prof Gunnar Carlson reminds in his landmark paper (Carlsson, 2009) *data have shape* and this shape has a meaning. This idea is not new: linear regression, for example, is a well-established statistical technique based on the idea that the shape of data is linear – a line in two dimensions and a hyperplane in higher dimensions. Understanding the linear shape is key to understanding the relationship between dependent and independent variables. However, data may resemble many

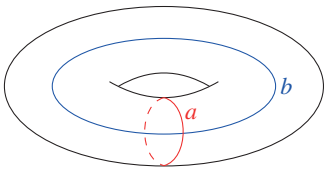


Figure 1. *The torus: has one connected component; two loops, since loops a and b are 'distinct' i.e. one cannot be transformed to the other along the torus' surface; and one void, since there is one void in the centre of the doughnut.*

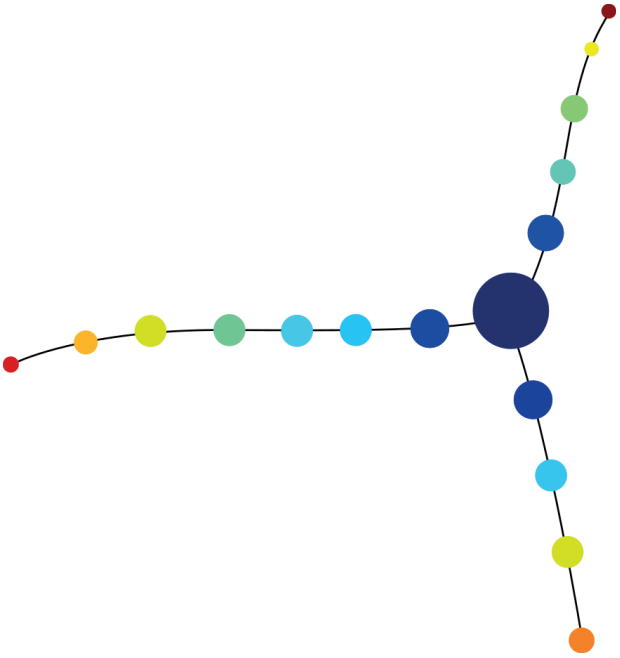


Figure 2. *Example of a flare, a single connected component consisting of three distinct groups of data*

other shapes which are harder to understand. Imagine, for example, data points that split into three distinct lines at a single point, forming a flare, i.e. a Y shape (Figure 2). Besides the flare, data may take on more complex shapes and unexpected behaviours, especially as the number of dimensions increases (e.g., the trefoil knot is knotted in three dimensions, but falls apart and becomes a trivial loop in four dimensions; see Figure 3).

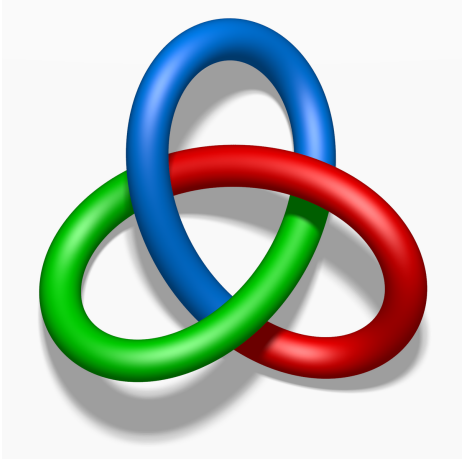


Figure 3. *A trefoil knot*

TDA provides techniques to describe these shapes by listing their topological invariants (such as holes or cavities), and to investigate the meaning of these topological features in terms of the specific data problem or clinical context.

2.1. Using TDA to help understand data structure

The question to answer is how we can 'build a bridge' from the collected patient data to a space in which topological invariants can be computed. This can be achieved in three steps.

Consider a dataset with m rows and n columns, where m is the number of observations in our sample (the number of patients, for example) and n is the number of measures collected for each patient.

1. Firstly, we need to define a measure to assess the proximity between any two data points – that is, to be able to measure how similar two patients are given the information we have for them. Interestingly, our data do not necessarily need to lie in the Euclidean space. As long as a distance can be computed between data points, we will be able to apply TDA tools. For ease of understanding, let us consider our data is made of numerical variables and let us represent them in a n-dimensional *point cloud* living in R^d, such that each patient becomes a point in the space, with each variable represented on a different coordinate axis. For the straightforward example of only two variables, values are drawn on the X and Y axes, but this can be extended to any number of dimensions, for three, four or more variables. In this way, we convert the patient data into a point cloud. For this particular case, we would assume that the point cloud is a finite sample drawn from an existing topological space. In case of a circle (see Figure 4a), we would assume that our data are a finite sample drawn from a 3D representation of a circle. In the Euclidean space R^d the natural choice of distance to assess similarity would be the Euclidean distance. There are also other distances that can be defined on numeric data as for example the Variance Normalised Euclidean or the Minkowski distance. When data are categorical rather than numeric, we can also define many different distances as for example the Gower distance.
2. Second, to highlight the underlying topology of the data we consider the construction of continuous shapes on top of the point cloud. These continuous shapes very commonly will be graphs. A graph is a finite, discrete representation of the set of points that encodes a one (or higher) dimensional skeleton of the data (Chartrand, 1985). Graphs are used in many data analysis applications and are much easier to visualise than the high-dimensional data used to construct them (see Figure 2).
3. Lastly, having built graphs based on the point cloud, we are able to compute the persistence diagram (and the extended-persistence diagram). These are topological signatures representing our data shape summary (Edelsbrunner et al., 2000; Zomorodian and Carlsson, 2005; Cohen-Steiner, Edelsbrunner and Harer, 2007; Carrière and Oudot, 2018).

Starting with a finite point cloud (Step 1), the following sections will introduce two approaches to constructing graphs on top of the point cloud (Step 2): the simplicial complex[1] and the Mapper graph. We will present the concept of persistence diagram, and will see (Step 3) how a persistence diagram can be derived from a family of simplicial complexes, and how a extended-persistence diagram can be derived from a single Mapper graph.

2.2. The simplicial complex

One way to construct a graph on top of point cloud data is by drawing a circle of radius ρ around each point in the cloud (Figure 4). If the corresponding circles for two points intersect, we connect the points with a line. If three circles intersect, we connect the three points to form a triangle, and so on. This particular graph is called a Čech complex and is a type of *simplicial complex*. A simplicial complex is a graph formed by a set of points, lines, triangles, etc. Simplicial complexes generalise the concept of one-dimensional graphs (formed only of edges and nodes) to allow other dimensional blocks like triangles (dimension 2), tetrahedrons (dimension 3) and so on. Besides the Čech complex, there are other types of simplicial complexes that can be constructed on top of point cloud data such as the Vietoris-Rips and the Alpha complex. In a Vietoris-Rips complex (Zomorodian, 2010) when 3 balls intersect (it can be a pairwise intersection, not all balls need to intersect), a triangle of dimension 2 is built. When 4 balls have a none empty intersection, a tetrahedron of dimension 3 is build, and so on. An Alpha complex is a simplicial complex constructed from the finite cells of a Delaunay Triangulation (Devillers, Hornus and Jamin, 2022). In terms of the topology of the Alpha complex (and its relationship with persistence theory) the Alpha complex is equivalent to the Čech complex and much smaller if one does not bound the radii.

2.3. The Mapper graph

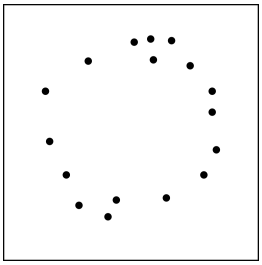
A second approach to constructing a graph on top of a point cloud is by using the Mapper algorithm (Singh, Mémoli and Carlsson, 2007). The Mapper algorithm reduces complex data to produce a one-dimensional graph – the Mapper graph. This consists of nodes (sets of clustered subjects) and edges connecting those nodes and edges connecting those nodes with non-empty intersections (that is, subjects can appear in more than one node).

The Mapper graph is built as follows. Suppose we have a finite point cloud and can compute all distances between pairs of points within the cloud. Suppose also, that we have a function called the *filter* that assigns a real value to each point in the data set. Then, the Mapper algorithm proceeds in the following steps (Figure 5):

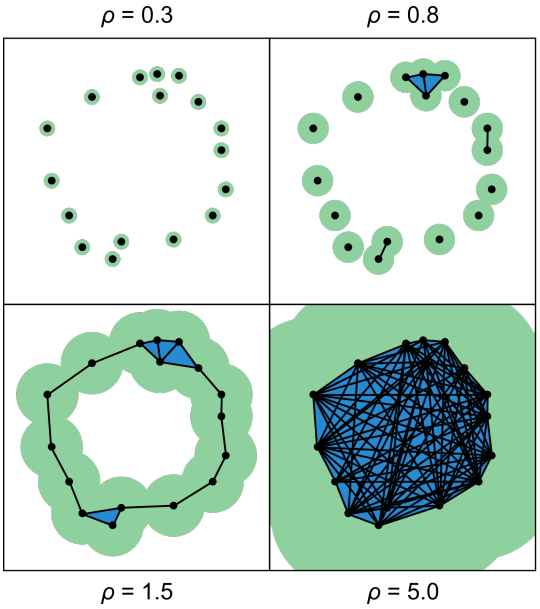
1. Find the range of the filter function (i.e., the interval of all values that the function takes);

[1] As simplicial complexes can be seen as higher dimensional generalizations of neighbouring graphs, we will make an abuse of notation and we will refer them as "graphs" throughout the paper

(a) *A point cloud sampled from a circle*



(b) *The sampled circle, now with smaller circles of an increasing radius ρ on top of each point*



(c) *The resulting persistence diagram, for a single topological feature in dimension 1.*

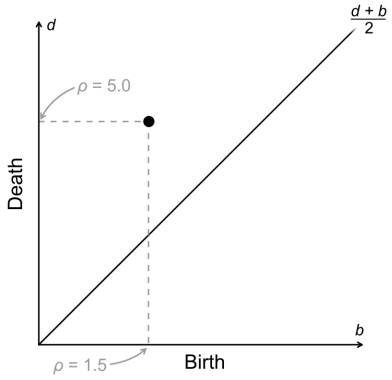


Figure 4. *Constructing the Čech complex of points sampled from a circle*

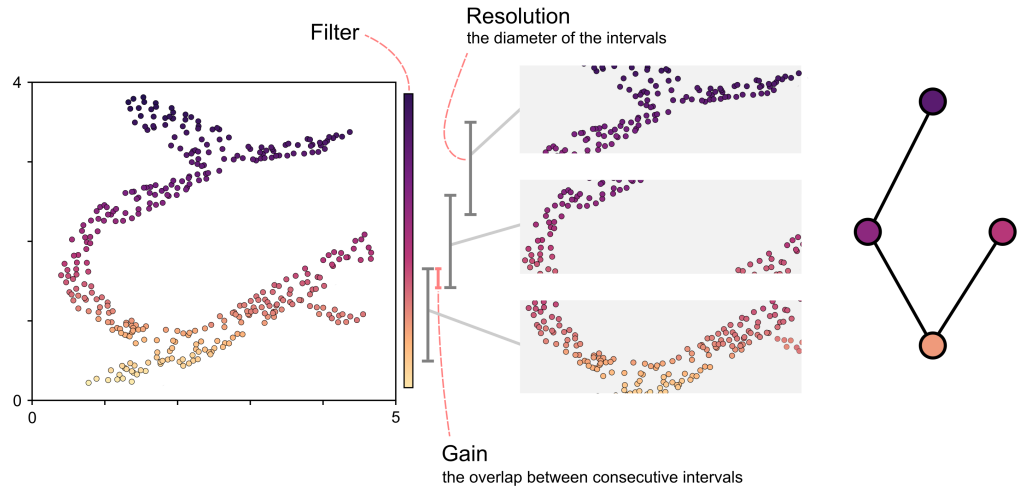


Figure 5. *The mapper graph: This shows the point cloud separated into intervals with diameter set by 'resolution' and overlap by 'gain'. Figure adapted from Munch (2017)*

2. Divide the range into smaller, overlapping intervals;

3. For each interval, find the set of data points whose values assigned by the filter function lie in the interval;

4. Decompose each of these sets into clusters based on a chosen clustering algorithm[2];

5. Represent each cluster by a node and connect nodes by an edge if the clusters intersect non-trivially, that is, they share data points.

The algorithm leaves various important choices to the user: the choice of the filter, the number of intervals and their percentage of overlap, and the clustering algorithm. See Chazal (2016) and Carrière, Michel and Oudot (2018) for a formal and complete discussion on parameter selection for Mapper. Several past studies have suggested the approach of selecting Mapper parameters based on exploration of a grid of possible values — selecting values that produce interesting or stable graphs (Carrière et al., 2018). However, as emphasised by Carrière et al. (2019), while useful for a data-driven exploratory phase, in many situations this approach may produce sub-optimal results, especially for non-trivial datasets. An alternative approach (Carrière, 2019) is to perform automatic tuning of Mapper parameters based on the rate of convergence of the Mapper graph to its continuous analogue, the Reeb graph. We return to this below.

[2]Any suitable clustering algorithm can be used; Refined analysis on the influence of the clustering method on the Mapper has been recently investigated (Belchí et al., 2019).

2.4. The persistence diagram

Recall our original goal is to obtain topological summaries that can describe complex structure in our data. Having constructed graphs based on the point cloud as described above, we can now use these graphs to compute topological invariants in our data (Chazal and Michel, 2021). One way to extract topological information is considering a family of simplicial complexes and coding the topological invariants in a two-dimensional diagram called the *persistence diagram* (Edelsbrunner et al., 2000). One can also identify the topological variants in a Mapper graph and represent them by means of the *extended-persistence diagram* (Carrière, 2019). Let us introduce both approaches below.

2.5. Persistence diagram for simplicial complexes

Let us consider a family of complexes constructed over an increasing range of values of the radius ρ (see Figure 4). This gives a *filtered complex*, a sequence of complexes such that each one is contained in the next. For each complex, we can deduce its topological invariants and trace them through the filtration as ρ increases, thus identifying their 'birth' (the radius at which they first appear) and 'death' (the radius at which they disappear). In Figure 4, the initial Čech complex for radius $\rho = 0.3$ and 0.8 shows no hole. A hole appears at $\rho = 1.5$ (the birth time of the hole) but disappears at $\rho = 5.0$ (the death time of the hole). So birth and death times represent radiuses at which the hole appears and disappears across the range of ρ values. The persistence diagram is a two dimensional plot where the X axis represents the birth time of a topological feature (a hole, in the example) and the Y axis represents its death time. The diagram includes a diagonal that represents the features that are born and die at equal time. The closer a point in the diagram is from the diagonal, the shorter was the life of that feature across the range of ρ values, i.e. the less persistent was the feature.

Intuitively, *persistent homology* captures how topological features of a space persist through the filtration, for some given time-span. The term homology refers to a mathematical (vector) space that represent the topological invariants in different dimensions. The homology group in dimension 0 represents the connectedness of the data space – a topological space is connected if it cannot be represented as the union of two or more disjoint non-empty open subsets. In dimension 1 the homology group represents the space of holes. In dimension 2 it represents the space of cavities, like the one we see in the torus or the 'bubble' inside the sphere, and so on (See Hatcher (2002) for a comprehensible introduction to homology). By identifying persistent features across a range of radiuses one avoids the need to choose a single radius ρ that would reveal the 'essential' topological features of the space. This ρ exists, and is mathematically proven, thanks to the combination of the nerve theorem and the reconstruction theorem (see Chazal and Michel (2021) for a formal formulation of both theorems). From a practical perspective, computing ρ rises many practical issues; a multiscale strategy has been introduced in (Chazal and Oudot, 2008).

Persistence diagrams of filtrations built on top of datasets are very stable with respect to some perturbations of the data. Thus, even for a dataset with some noise, the

persistence diagram obtained from this data is approximately correct because it is close to the diagram we would have obtained from the noise-free data (because the Gromov-Hausdorff distance between both datasets is assumed to be small, see Chazal and Michel, 2021).

2.6. Persistence diagram for the Mapper graph

The Mapper graph built under an optimal selection of the parameters involved in the algorithm (i.e. the filter function, intervals covering the range of the image of the filter function, and their overlap) is a discrete and computable optimal estimator of its continuous counterpart, the Reeb graph (the Mapper graph is said to 'converge' onto the Reeb graph) (Carrière and Oudot, 2018). A Reeb graph is a mathematical object reflecting the evolution of the level sets of a real-valued function on a topological space that locally resembles Euclidean space (see a Reeb graph in Figure 6 (iii)). From the Mapper graph, we can derive the extended persistence diagram (Figure 6) by tracing up and down the Reeb graph to identify pairs of critical points that mark the beginning ('birth' time) and the end ('death' time) of a topological feature in the associated Reeb graph. For example, Figure 6 shows the birth and death times for trunks, branches, and holes.

2.7. Statistical stability of points in the persistence diagram

As mentioned above, points on a persistence diagram with very short time spans, i.e. those points located close to the diagonal (the line representing points with equal birth and death), indicate features that appear and disappear quickly and which are more likely to be noise. We therefore may wish to discard 'non-significant' points that are close to the diagonal. One approach to assessing 'closeness' is to use the bootstrap to estimate and draw confidence bands on the persistence diagram, along the diagonal. Significant topological features will lie outside the confidence bands, whilst non-significant features will lie close to the diagonal, within the confidence bands, helping to distinguish between signal and noise (see Figure 7b) (Chazal, 2016). The bootstrap is a popular re-sampling method to quantify uncertainty around sample statistics (e.g. to estimate confidence intervals around a mean). To derive the confidence interval for a persistence diagram we:

1. Generate B bootstrap samples by re-sampling with replacement from the original source data, and construct a persistence diagram for each sample.

2. We then derive the 'distance' between the original persistence diagram (built using the source data) and each bootstrapped persistence diagram using the *Bottleneck distance*: two persistence diagrams are superimposed and each dot in the first diagram is assigned to its closest counterpart on the second. The Bottleneck distance is then defined as the maximum distance between any pair of matching dots. This way we get a distribution of distances for which a central 95% of values can be computed and a confidence interval (D) derived.

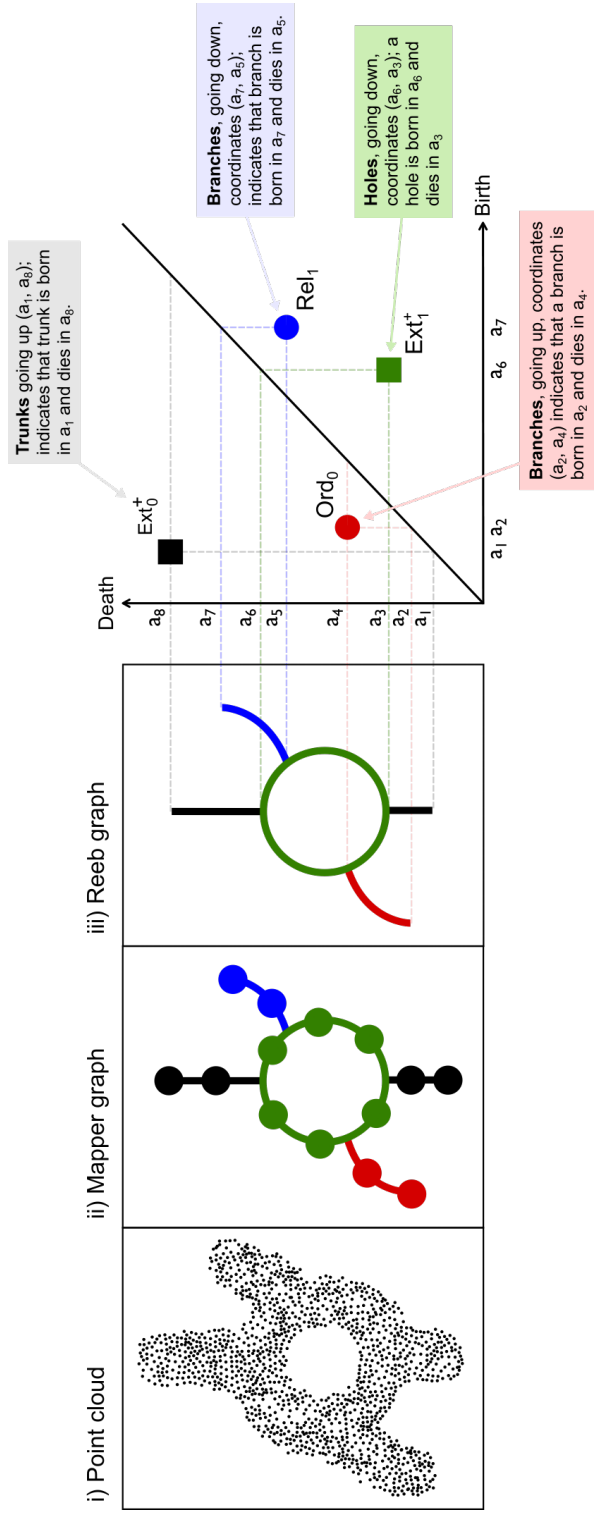


Figure 6. *Extended persistence diagram*

3. Finally, this confidence interval is drawn on the graph as a band spreading away from the diagonal (in both directions, each with width D), or as boxes around each point (of radius D).

The points outside the confidence band are considered as significant topological structures in the data, whereas those lying within the band's limits around the diagonal represent *insignificant* structures in the data set, and therefore are considered as noise and should not be interpreted nor processed for further analysis. This is a developing field, and while the validity of the use of the bottleneck bootstrap has been proven for the persistence diagram computed for a filtration of simplicial complexes, its use still remains as an open problem for the extended persistence diagrams computed for Mapper graphs (Carrière et al., 2018).

2.8. Use of Persistence Landscapes for outcomes prediction

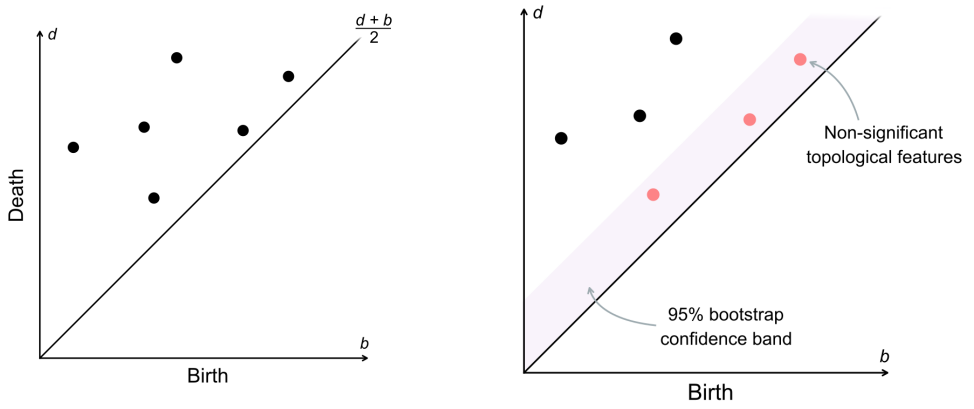
Persistence Landscapes (Bubenik, 2015) can be used to convert a persistence diagram (built from a filtration of simplicial complexes) into a vector space suitable for inclusion in ML models. Suppose we have a persistence diagram where each point represents the birth and death of a hole in our data. The corresponding persistence landscapes are constructed by 'tenting' each point in the diagram as shown in Figure 7c, to produce a collection of continuous *piecewise* linear functions, i.e. functions whose graph is composed of straight-line sections. Discretising the landscapes in a number of points produce a set of variables that encode the topological structure of data and can be included as predictors in a ML model. Interestingly, persistence landscapes share the same stability properties as persistence diagrams, described above.

2.9. Use of Mapper for subgroups detection, variable selection and data visualisation

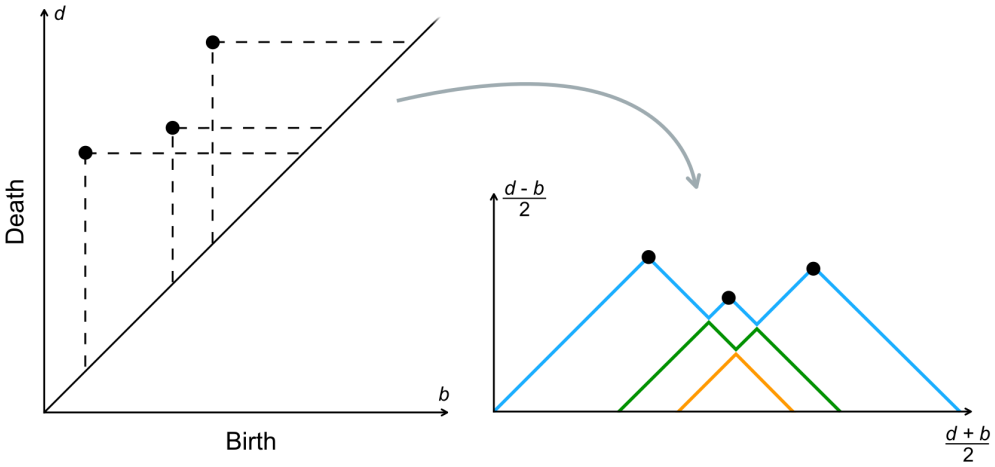
The Mapper graph can be useful to identify homogeneous subgroups of patients with regards of a characteristic of interest (Carr et al., 2021). The Mapper algorithm can highlight interesting clusters in data that might not be recoverable with traditional statistical clustering methods. Consider a data-based Mapper graph following the *flare* shape (the Y shape mentioned earlier; Figure 2). This could be interpreted as a single cluster of data. However, each arm could potentially represent a distinct data sub-population. The characterisation of topological features in a graph, like particular flares or loops, can help identify clinically relevant groups of nodes comprising subjects that experience particular prognostic outcomes or levels of treatment response.
Mapper can also be used to perform variable selection. One can build a Mapper graph from data, identify interesting structures as flares, loops or distinguished groups of coloured nodes, and then select the variables that best discriminate the data in these structures. Variables can then be assessed one-by-one for their ability to discriminate the potential sub-populations from the rest of the data using classical tests, as Kolmogorov-Smirnov. Interestingly, one can also consider a multivariate feature selection for which

(a) *The persistence diagram for a single point cloud*

(b) *Persistence diagram, showing the bootstrapped confidence interval. Points outside the interval are considered statistically significant.*



(c) *The computed persistence landscape, formed by 'tenting' the significant points on the persistence diagram. The first landscape is in blue, the second in green, and the last in orange.*



(d) *'Discretising' each landscape on a number of points, by selecting a discrete grid of values on the X-axis, and computing their corresponding Y-value on each persistence landscape.*

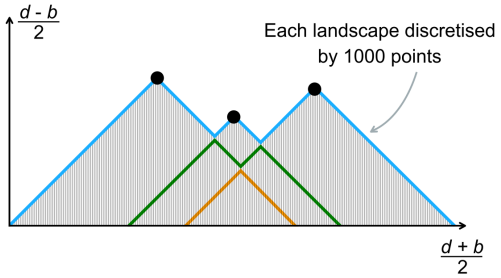


Figure 7. *Constructing the persistence landscape based on significant topological features*

Mapper can be used in conjunction with ML. Those detected flares and loops are given class labels, and a ML model including the desired set of predictors is tuned to solve the classification task of distinguishing one class from the other. This way, Mapper achieves two goals: identifying new sub-populations and selecting the combination of features that best differentiate them. We have implemented this ML based procedure to identify interesting subgroups and features in a pipeline that we will use below (https://github.com/kcl-bhi/mapper-pipeline).

The Mapper graph is also useful as a visualisation tool. If we select a set of intervals where no more than two intervals can intersect at once, Mapper becomes a visualisation tool that reflects the topology of the data. Mapper has a *multi-resolution structure*, i.e. by choosing the number of intervals and the percentage overlap between them, the user can adjust the level of the detail at which to view their data.

3. Application to a data case: using TDA to characterise depression remission in the GENDEP study

The Genome-based Therapeutic Drugs for Depression (GENDEP) is a pharmacogenetic study of antidepressant treatment response (Uher et al., 2010a). The GENDEP study aims to find a way to use clinical and genetic information about patients to help doctors decide which antidepressant treatment will work best for each patient, and with the least side-effects. A total of 220 patients were randomly allocated to be treated with escitalopram drug, a standard drug that is commonly prescribed to treat depressive symptoms. Over 12 weeks the study collected sociodemographic and clinical data including depressive symptoms. For each participant there were available sociodemographic variables (at baseline only) as age, age at onset, gender, smoking (yes/no), occupation (yes/no), partner (yes/no), years of education, number of children and body mass index. There were also available weekly repeated measures (from baseline to week 12) of depression severity by means of several standard scales: MADRS (Montgomery and Åsberg, 1979), Hamilton-17 (Hamilton, 1967), BDI (Beck et al., 1961), SCAN (Wing et al., 1990) and suicidal ideation (Perroud et al., 2012). Each scale assessed several individual items and was coded as a number (between 4 and 6, depending on the scale) of possible answers to a statement or question that allows respondents to indicate their positive-to-negative strength of agreement or strength of feeling regarding the question or statement. For example, the MADRS included 10 items assessing aspects such (1) apparent sadness; (2) reported sadness; (3) inner tension; (4) reduced sleep; (5) reduced appetite; (6) concentration difficulties; (7) lassitude; (8) inability to feel; (9) pessimistic thoughts; and (10) suicidal thoughts. Then each of these items was measured following a numerical codification ranging from 0 to 6 depending on the patient's strength of agreement. The ten resulting scores were then added to build a total numerical score. The rest of the scales were defined similarly. Data also included six symptoms dimensions (mood, anxiety, pessimism, interest-activity, sleep, appetite) from a published factor analysis (Uher et al., 2008, 2012). Remission was assessed for each patient at the last available mea-

surement after 4 – 12 weeks of treatment. Remission was defined as scoring $\leqslant 7$ on the Hamilton-17 scale (Hamilton, 1967), a commonly used definition for remission of depressive symptoms. A total of 94 patients remitted.

3.1. An analytical pipeline to predict remission depression

We aimed to use sociodemographic and clinical repeated measures in GENDEP to predict remission of depression. We implemented an analytical pipeline to compute persistence landscapes (and thus summaries of topological features) of our data, and including them in a ML model to predict remission. The pipeline requires Python 3.6 or higher (van Rossum, 1995) and R 4.1.2 or higher (R Core Team, 2020). It uses Scikit learn (Pedregosa et al., 2011) and Gudhi (The GUDHI Project, 2015) Python packages to derive the topological features based on the construction of persistence landscapes, and caret (Kuhn, 2008) and glmnet (Friedman, Hastie and Tibshirani, 2010) R packages to fit an elastic net logistic regression model that includes the topological features as predictors of a binary outcome. The pipeline can be freely downloaded at http://github.com/kcl-bhi/topological-review.

We used the pipeline to compute topological summaries on longitudinal measures of depression severity from baseline up to week 4 (a total of 5 time points). We included weekly total scores for MADRS, Hamilton-17, BDI and suicidal ideation, and a composite score for suicidal ideation[3]. We additionally included observed mood, cognitive and neurovegetative symptoms measured by means of the SCAN interview and the six symptoms dimensions from Uher et al. (2008, 2012) giving a total of 14 items measured on 5 occasions (a 14×5 matrix for each participant).

The detailed analytical pipeline we used was:

1. For each patient, compute the persistence diagram for a complex filtration based on the available data matrix (Figure 7a). For this case we computed an Alpha Complex filtration based on a 14×5 matrix (14 points in dimension 5). We considered the *connected components* and *holes* from the complex and created the associated persistence diagrams.

2. Compute the persistence landscape for each persistence diagram (Figure 7c). For this example, we computed the first three landscapes. The choice of how many landscapes to include can be guided by the predictive performance of the model (i.e. select the number of landscapes that maximises the predictive ability).

3. Discretise each landscape on a number of points (i.e. consider a discrete grid of values on the X-axis, and their corresponding Y-value on each persistence landscapes) (see Figure 7d). In our example, we considered the values of each persistence landscape on a grid of 1000 equidistant points, so that each patient was

[3]Composite scores are combinations of items that are highly related. They are computed from data in multiple variables in order to form reliable and valid measures of latent, theoretical constructs. These can be tested through factor analysis and reliability analysis (Ioannidis, Klavans and Boyack, 2016).

described by 6000 topological variables (3 landscapes $\times$ 1000 points $\times$ 2 dimensions). As one increases the number of discretisation points, the discretisation error will decrease but the resulting number of variables will increase. This decision should be based on the sample size available, although variable selection can help.

4. Include the topological variables together with the baseline variables as predictors in an elastic net logistic regression model to predict remission (yes or not). We chose a regularised regression model as this is efficient in preventing the risk of overfitting in complex data (i.e., when the model predicts well in known data, but generalises poorly to new cases) and performs variable selection, which helps to remove from the model topological variables that are not adding relevant information.

Parameter tuning for the elastic net regression model was performed with repeated (100 repetitions) 10-fold cross-validation. We compared (1) a model including sociodemographic and clinical variables only at baseline, and (2) a model including baseline sociodemographic and clinical variables and the topological variables derived from longitudinal measures on depression severity up to week 4, as described in the pipeline.

3.2. Results

In this preliminary analysis, the area under the ROC curve (AUC) for predicting remission was 0.746 for the model only including baseline measurements, which compared to an AUC of 0.799 when topological variables were added. This represented a promising improvement in predictive performance resulting from the inclusion of topological variables. Interestingly, the automated feature selection by the elastic net selected topological variables for both dimensions, that is, connected components and holes, as relevant variables for the prediction. The first landscapes tended to capture the most topological information, with subsequent landscapes bringing diminishing returns.

3.3. Using Mapper for subgroups detection in GENDEP

We used the Mapper algorithm to derive interesting clusters of patients in GENDEP based on their clinical and genetic baseline characteristics (full results are presented at Carr et al., 2021). We implemented a pipeline to tune the parameters of the Mapper graph seeking to maximise the purity of a given outcome variable within derived clusters of patients. A cluster of patients was defined as those patients with data belonging to a topological feature identified in the Mapper graph (i.e., a flare, a loop...). Mapper parameters were tuned to maximise the within clusters' level of purity with regards of depression remission (purity was computed by means of the Gini coefficient). Our pipeline allows predicting membership to a cluster using gradient boosted trees (XGBoost). This way it allows selecting the combination of variables that best differentiate a cluster of patients. The protocol also allows to consider both categorical and continuous variables (recent research in COVID-19 indicated high demand of such type of algorithms that are

suitable for mixed data types, see Khan et al., 2021). The pipeline can be freely downloaded under the GNU GPLv3 license at https://github.com/kcl-bhi/mapper-pipeline.

As it is shown in detail in Carr et al. (2021) when we applied our pipeline to the GENDEP dataset remission purity increased in the resulting clusters in comparison with the whole sample. We ranked the resulting clusters according to their remission purity, and, interestingly, the top five clusters from our pipeline outperformed the five-cluster solution from k-means clustering in terms of remission purity. Gini index in our clusters ranged from 0.30 to 0.38, whilst in clusters from k-means ranged from 0.33 to 0.50. A combination of clinical and genetic baseline measurements was able to discriminate patients in one of our top clusters with excellent discrimination.

4. TDA software

In practice, there are various algorithms implementing methods to produce simplicial complexes (the Čech complex and others) and compute topological invariants such as persistence diagrams and persistence landscapes. A good summary of software to implement persistent homology is given in Otter et al. (2017). There exist several general purpose libraries for topological data analysis including GUDHI (The GUDHI Project, 2020), Dionysus (Morozov, 2007), and PHAT (Bauer et al., 2017). All are written in `C++` and provide fast and efficient implementations of common topological invariants, with interfaces available for R and Python. Several packages have built upon these libraries to facilitate the application of common topological algorithms. The TDA package for `R` (Fasy et al., 2014) provides a user-friendly interface for `R` users. The `statmapper` (Carrière, 2020) Python package functions to derive extended persistence diagrams, to compute topological features in a Mapper graph and evaluate their statistical significance, using the bootstrap.

We have presented a pipeline that allows including summaries of topological features in a ML predictive model using persistence landscapes (http://github.com/kcl-bhi/topological-review) and a pipeline to identify sub-populations and perform multivariable selection using Mapper (https://github.com/kcl-bhi/mapper-pipeline).

5. Conclusion

TDA is a rapidly growing field that offers a unique set of tools with considerable potential for precision medicine. Topological summaries derived from persistence diagrams and landscapes have shown promising results in specific examples when included in machine learning predictive models, resulting in improved model performance, as we show in an application to a clinical trial on major depression. The Mapper algorithm makes it possible to identify homogeneous sub-populations of interest in complex data and deriving features that can be used to discriminate these groups. This paper provides a basis for the promising role that TDA can play in precision medicine using large biomedical datasets.

Acknowledgements

This work was supported by a 2017 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation granted to Dr Raquel Iniesta.
This work has been funded by the European Commission Framework 6 grant, EC Contract LSHB-CT-2003-503428 and an Innovative Medicine Initiative Joint Undertaking (IMI-JU) grant n-115008 of which resources are composed of European Union and the European Federation of Pharmaceutical Industries and Associations (EFPIA) in-kind contribution and financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013).
This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The authors further acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

Adamson, A. S. and Welch, H. G. (2019). Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard. *New England Journal of Medicine*, 381(24):2285–2287.

Bauer, U., Kerber, M., Reininghaus, J., and Wagner, H. (2017). Phat – Persistent Homology Algorithms Toolbox. *Journal of Symbolic Computation*, 78:76–90.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571.

Belchí, F., Brodzki, J., Burfitt, M., and Niranjan, M. (2019). A numerical measure of the instability of mapper-type algorithms. *Journal of Machine Learning Research*, 21:1–45.

Bubenik, P. (2015). Statistical Topological Data Analysis using Persistence Landscapes. page 26.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Carr, E., Carrière, M., Michel, B., Chazal, F., and Iniesta, R. (2021). Identifying homogeneous subgroups of patients and important features: a topological machine learning approach.

Carrière, M. (2019). MathieuCarriere/Sklearn-Tda.

Carrière, M. (2020). MathieuCarriere/Statmapper.

Carrière, M., Michel, B., and Oudot, S. (2018). Statistical Analysis and Parameter Selection for Mapper. *Journal of Machine Learning Research*, 19(1):1–39.

Carrière, M. and Oudot, S. (2017). Local Equivalence and Induced Metrics for Reeb Graphs. In *Proceedings of the 33rd Symposium on Computational Geometry*.

Carrière, M. and Oudot, S. (2018). Structure and Stability of the 1-Dimensional Mapper. *Foundations of Computational Mathematics*, 18(6):1333–1396.

Chartrand, G. (1985). *Introductory Graph Theory*. Dover Publications Inc., New York, abridged edition edition edition.

Chazal, F. (2016). High-Dimensional Topological Data Analysis. In *3rd Handbook of Discrete and Computational Geometry*. CRC Press.

Chazal, F., Massart, P., and Michel, B. (2016). Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10(2):2243–2286.

Chazal, F. and Michel, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4.

Chazal, F. and Oudot, S. Y. (2008). Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry - SCG '08*, page 232, College Park, MD, USA. ACM Press.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., and Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, 3(3):243–250.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of Persistence Diagrams. *Discrete & Computational Geometry*, 37(1):103–120.

Dagliati, A., Geifman, N., Peek, N., Holmes, J. H., Sacchi, L., Bellazzi, R., Sajjadi, S. E., and Tucker, A. (2020). Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine*, 108:101930.

Devillers, O., Hornus, S., and Jamin, C. (2022). dD triangulations. In *CGAL User and Reference Manual*. CGAL Editorial Board, 5.4 edition.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., and Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials*, 18(5):435–441.

Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014). Introduction to the R package TDA. *arXiv:1411.1830 [cs, stat]*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Ghrist, R. (2018). Homological algebra and data. In Mahoney, M., Duchi, J., and Gilbert, A., editors, *IAS/Park City Mathematics Series*, volume 25, pages 273–325. American Mathematical Society, Providence, Rhode Island.

Hamilton, M. (1967). Development of a Rating Scale for Primary Depressive Illness. *British Journal of Social and Clinical Psychology*, 6(4):278–296.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Number 1 in Springer Series in Statistics. Springer, New York.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press.

Henle, M. (1994). *A Combinatorial Introduction to Topology*. Dover Books, Inc, New York.

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine Learning SNP Based Prediction for Precision Medicine. *Frontiers in Genetics*, 10:267.

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Dobson, R., Aitchison, K. J., Farmer, A., McGuffin, P., Lewis, C. M., and Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, 8(1):1–9.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Stahl, D., Dobson, R., Aitchison, K. J., Farmer, A., Lewis, C. M., McGuffin, P., and Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, 78:94–102.

Iniesta, R., Stahl, D., and McGuffin, P. (2017). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46(12):2455–2465.

Ioannidis, J., Klavans, R., and Boyack, K. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLoS Biol*, 14(7).

Khan, W., Crockett, K., O'Shea, J., Hussain, A., and Khan, B. M. (2021). Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection. *Expert Systems with Applications*, 169.

Khan, W., Hussain, A., Ahmed Khan, S., Al-Jumailey, M., Raheel, N., and Liatsis, P. (2019). Analysing the impact of global demographic characteristics over the covid-19 spread using class rule mining and pattern matching. *Royal Society Open Science*, 8.

Kosniowski, C. (1980). *A First Course in Algebraic Topology*. CUP Archive.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York.

Mitchell, T. M. (2006). The Discipline of Machine Learning. page 9.

Montgomery, S. A. and Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134:382–389.

Morozov, D. (2007). Dionysus, a C++ library for computing persistent homology.

Müllner, D. and Babu, A. (2013). Python Mapper: An open-source toolchain for data exploration.

Munch, E. (2017). A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, 4(2):47–61.

Nature (2019). Ascent of machine learning in medicine. *Nature Materials*, 18(5):407–407.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

Nielson, J. L., Cooper, S. R., Yue, J. K., Sorani, M. D., Inoue, T., Yuh, E. L., Mukherjee, P., Petrossian, T. C., Paquette, J., Lum, P. Y., Carlsson, G. E., Vassar, M. J., Lingsma, H. F., Gordon, W. A., Valadka, A. B., Okonkwo, D. O., Manley, G. T., Ferguson, A. R., and TRACK-TBI Investigators (2017). Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis. *PLOS ONE*, 12(3):e0169490.

Nusrat, S., Harbig, T., and Gehlenborg, N. (2019). Tasks, Techniques, and Tools for Genomic Data Visualization. *Computer Graphics Forum*, 38(3):781–805.

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perroud, N., Uher, R., Ng, M. Y. M., Guipponi, M., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Gennarelli, M., Rietschel, M., Souery, D., Dernovsek, M. Z., Stamp, A. S., Lathrop, M., Farmer, A., Breen, G., Aitchison, K. J., Lewis, C. M., Craig, I. W., and McGuffin, P. (2012). Genome-wide association study of increasing suicidal ideation during antidepressant treatment in the GENDEP project. *The Pharmacogenomics Journal*, 12(1):68–77.

Qu, Z., Lau, C. W., Nguyen, Q. V., Zhou, Y., and Catchpoole, D. R. (2019). Visual Analytics of Genomic and Cancer Data: A Systematic Review. *Cancer Informatics*, 18:117693511983554.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Riemann, B. and Clifford, T. W. K. (1998). On the Hypotheses which lie at the Bases of Geometry. page 15.

Sies, A., Demyttenaere, K., and Mechelen, I. V. (2019). Studying treatment-effect heterogeneity in precision medicine through induced subgroups. *Journal of Biopharmaceutical Statistics*, 29(3):491–507.

Singh, G., Mémoli, F., and Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *PBG@Eurographics*.

The GUDHI Project (2015). *GUDHI User and Reference Manual*. GUDHI Editorial Board.

The GUDHI Project (2020). *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.1.1 edition.

Traylor, M., Markus, H., and Lewis, C. M. (2015). Homogeneous case subgroups increase power in genetic association studies. *European Journal of Human Genetics*, 23(6):863–869.

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., Mors, O., Elkin, A., Williamson, R. J., Schmael, C., Henigsberg, N., Perez, J., Mendlewicz, J., Janzing, J. G. E., Zobel, A., Skibinska, M., Kozel, D., Stamp, A. S., Bajs, M., Placentino, A., Barreto, M., McGuffin, P., and Aitchison, K. J. (2008). Measuring depression: Comparison and integration of three scales in the GENDEP study. *Psychological Medicine*, 38(2):289–300.

Uher, R., Muthén, B., Souery, D., Mors, O., Jaracz, J., Placentino, A., Petrovic, A., Zobel, A., Henigsberg, N., Rietschel, M., Aitchison, K. J., Farmer, A., and McGuffin, P. (2010a). Trajectories of change in depression severity during treatment with antidepressants. *Psychological Medicine*, 40(8):1367–1377.

Uher, R., Perlis, R. H., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., Hauser, J., Dernovsek, M. Z., Souery, D., Bajs, M., Maier, W., Aitchison, K. J., Farmer, A., and McGuffin, P. (2012). Depression symptom dimensions as predictors of antidepressant treatment outcome: Replicable evidence for interest-activity symptoms. *Psychological medicine*, 42(5):967–980.

Uher, R., Perroud, N., Ng, M. Y., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Placentino, A., Rietschel, M., Souery, D., Žagar, T., Czerski, P. M., Jerman, B., Larsen, E. R., Schulze, T. G., Zobel, A., Cohen-Woods, S., Pirlo, K., Butler, A. W., Muglia, P., Barnes, M. R., Lathrop<, M., Farmer, A., Breen, G., Aitchison, K. J., Craig, I., Lewis, C. M., and McGuffin, P. (2010b). Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project. *American Journal of Psychiatry*, 167(5):555–564.

van Rossum, G. (1995). Python reference manual.

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D., and Sartorius, N. (1990). SCAN: Schedules fonr Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry*, 47(6):589–593.

Zomorodian, A. (2010). Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271.

Zomorodian, A. and Carlsson, G. (2005). Computing Persistent Homology. *Discrete & Computational Geometry*, 33(2):249–274.

SORT 46 (1) January-June 2022, 137-158 DOI: 10.2436/20.8080.02.121

Estimation of cut-off points under complex-sampling design data

Amaia Iparragirre*,[1], Irantzu Barrio[1,3], Jorge Aramendi[2] and Inmaculada Arostegui[1,3]

Abstract

In the context of logistic regression models, a cut-off point is usually selected to dichotomize the estimated predicted probabilities based on the model. The techniques proposed to estimate optimal cut-off points in the literature, are commonly developed to be applied in simple random samples and their applicability to complex sampling designs could be limited. Therefore, in this work we propose a methodology to incorporate sampling weights in the estimation process of the optimal cut-off points, and we evaluate its performance using a real data-based simulation study. The results suggest the convenience of considering sampling weights for estimating optimal cut-off points.

MSC: *62J12, 62P25, 62D05*

Keywords: *Optimal cut-off points, complex survey data, sampling weights.*

1. Introduction

Survey data are gaining popularity in a number of fields, including but not limited to, social and health sciences. This type of data is data collected from a finite population, concerned to be studied, by some complex sampling design such as stratification or clustering, among others (Kalton, 1983). One of the differences between complex survey data and simple random samples is that, in the first, each sampled observation has assigned a sampling weight, which indicates the number of units that this observation represents in the finite population. Therefore, the straightforward application of the most

* *Corresponding author:* E-mail: amaia.iparragirre@ehu.eus, Address: Departamento de Matemáticas. Facultad de Ciencia y Tecnología. Universidad del País Vasco (UPV/EHU). Barrio Sarriena s/n. 48940 Leioa.

[1] Departamento de Matemáticas, Universidad del País Vasco (UPV/EHU).

[2] Eustat - Euskal Estatistika Erakundea - Instituto Vasco de Estadística.

[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain.

Received: September 2021.
Accepted: May 2022.

commonly applied statistical techniques, which are typically designed to be applied to simple random samples, is usually not suitable for complex survey data (Skinner, Holt and Smith, 1989).

In this paper, we focus on the particular case of a binary response variable Y and, specifically, on the logistic regression model to predict Y according to a collection of covariates whose distribution may be discrete or continuous. From a practical point of view, one of the most important characteristics of this kind of model is the support they provide for decision-making, since increasing knowledge about potential predictors helps the decision-making process (Steyerberg, 2008; Baker and Gerdin, 2017). In this context, decisions such as whether or not to recommend a patient to start treatment, or to give a diagnosis about a disease, are based on the individual risk (probability) of event given by the estimates of the logistic regression model. In order to make these decisions, first, for each individual, the predicted probability of event is classified based on a cut-off point. In this way, for example, if the individual's probability of suffering from extreme poverty is greater than the selected cut-off point, he or she is assigned a social benefit, while in contrast, if that is lower no social support is provided (Steyerberg, 2008; Pauker and Kassirer, 1980). Hence, cut-off point estimation is widely employed in practice, in the field of prediction models, especially, but not exclusively, in clinical prediction models (Steyerberg et al., 1999; Chen et al., 2015; Spence et al., 2018).

At this point, the main issue is usually to select a valid cut-off point that will provide the best classification of individuals in practice. Many strategies have been proposed in the literature in order to estimate optimal cut-off points. It should be noted that we can not talk about optimal cut-off points in general terms. In contrast, a cut-off point will or will not be the optimal depending on the objective of a particular study. Therefore, when we talk about selecting an optimal cut-off point, we are talking about selecting the one which satisfies a certain optimality criterion. Hence, as we have mentioned above, different techniques have been proposed to select optimal cut-off points, given a particular criterion. For instance, some of those methods select the optimal cut-off point with the aim of obtaining a certain value of sensitivity/specificity (i.e., probability of classifying correctly an individual with/without the event of interest) or to maximize a function of these two parameters as for example the Youden index (Youden, 1950). Some others select the cut-off point that maximizes some particular indexes, such as Kappa (Cohen, 1960; Greiner, Pfeiffer and Smith, 2000). Greiner (1995, 1996) proposed a method to select the optimal cut-off point that minimizes the error or either maximizes the accuracy of the classification rule. There are some other methods that select optimal cut-off points based on some other criteria related to several parameters such as predicted values (i.e., probability of event/non-event for an individual classified as event/non-event) (Vermont et al., 1991) or prevalence (i.e., the probability of event in the population) (Manel, Williams and Ormerod, 2001), among others. Besides, other methods are based on the analysis of the cost of incorrect and the benefit of correct diagnosis (Swets, 1992; Pauker and Kassirer, 1980; Wynants et al., 2019). An extensive review of those techniques can be found in López-Ratón et al. (2014).

However, those techniques have usually been designed and applied for simple random samples and, as far as we know, there is a lack of proposals to consider complex sampling designs, and in particular sampling weights, throughout the estimation process of optimal cut-off points. It is widely known that when the sampling designs are not considered for the analysis of data derived from complex surveys the variances tend to be underestimated, which can lead to biased estimates of test statistics (Yao, Li and Graubard, 2015; Skinner et al., 1989; Heeringa, West and Berglund, 2017; Binder and Roberts, 2009). In the same way, we believe that sampling weights should not be ignored when estimating optimal cut-off points when working with complex survey data. Therefore, in this work, we propose a methodology to modify some of the methods to select optimal cut-off points of the probability of event in the logistic regression framework that have been previously proposed in the literature, so that they take into account sampling weights in the estimation process. In addition, the performance of the proposed methods is compared to the performance of those which ignore the sampling weights, by means of a simulation study. In particular, we focus on surveys which are based on one-step stratified samples.

The rest of the paper is organized as follows. Section 2 describes the real survey that has motivated this work. Section 3 defines some basic notation that will be used along the rest of the paper. Furthermore, we describe some of the methods that are usually applied in practice to estimate optimal cut-off points of the probability of event in the logistic regression framework and finally we propose a new methodology which takes into account the effect of the sampling weights in the cut-off point estimation process. In Section 4, we describe the simulation process that has been carried out so as to study the performance and effectiveness of the proposed method to incorporate sampling weights into the estimation process of optimal cut-off points and we show the results we have obtained in the mentioned simulation study. The methodology proposed in this work has been applied to real survey data and this application is described in Section 5. Finally, we conclude with a discussion in Section 6.

2. Motivating data set

This work has been motivated by the Survey on the Information Society in Companies[1], which has been designed, conducted and collected by the Official Statistics Basque Office (Eustat). This survey, which is usually denoted as ESIE survey due to its Spanish acronym, is carried out annually among the companies in the Basque Country (BC) in order to collect information about the implementation of New Information and Communication Technology in the companies of the BC. In particular, the information considered in this study is related to the survey carried out in 2010.

The finite population is defined by a total of 14 200 companies, all of which have at least 10 employees. From this population a sample of 2 852 was obtained by means

[1] https://en.eustat.eus/estadisticas/tema_150/opt_1/tipo_7/temas.html

of one-step stratified sampling technique with simple random sampling in each stratum. Strata are defined by means of the combination of three categorical variables: the province where the company is located (3 categories), activity of the company (65 categories) and the number of employees (2 categories). In this way, a total of 390 different strata have been defined. However, it should be noted that in some of these strata there are no units in the population, so in fact we have 325 strata in total ($h = 1, \ldots, H$, where $H = 325$). Once the sample is obtained, a sampling weight is assigned to the companies sampled in each stratum. The sampling weight (w_i, $\forall i \in S$) is calculated per stratum as the total number of companies in the finite population of the stratum (let us denote it as N_h, $\forall h \in \{1, \ldots, H\}$) divided by the number of companies sampled in that stratum (denoted as n_h, $\forall h \in \{1, \ldots, H\}$). In other words, for a unit i sampled from stratum h its sampling weight is computed as follows:

$$w_i = \frac{N_h}{n_h}, \quad \forall i \in S. \tag{1}$$

Each sampling weight indicates the number of companies that this sampled company represents in the finite population.

In the survey data considered for this paper, strata sizes in the finite population (i.e., N_h, $\forall h \in \{1, \ldots, H\}$) ranges from 1 to 860, where the median is 12 and the interquartile range $4 - 44$. An unequal probability sampling design has been applied in the sampling process, in which the probabilities of being sampled from each stratum (i.e., n_h/N_h, $\forall h \in \{1, \ldots, H\}$) range from 0.0391 to 1 (with a median of 0.6667 and an interquartile range of $0.2604 - 1$). The dichotomous response variable considered for this work indicates whether a company has its own website (1) or not (0). The probability of event in the sample (without considering the sampling weights) is 0.8222, while the weighted estimate of the probability of event (computed by taking into account the number of units that each element represents in the finite population by means of the sampling weights w_i, $\forall i \in S$) is 0.7544.

3. Methods

In this section, first of all, we introduce the basic notation that we will use throughout this document. In addition, we describe some of the methods that are usually applied for estimating optimal cut-off points in this context based on different optimality criteria for simple random samples. Finally, we develop a new estimation method, in which we propose to introduce the sampling weights in these methods so that they are valid in complex design samples.

3.1. Basic notation and preliminaries

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ be a vector of p random predictor variables denoting the covariates and Y a random variable denoting the dichotomous response variable. Without loss of generality, and in order to ease the notation, suppose that the covariates $\boldsymbol{X}$ are continuous

and the response variable Y takes the value 1 to represent the event or the presence of the characteristic of interest, and 0 otherwise. Let $P(Y=1|\boldsymbol{X})$ represent the conditional probability of event given the vector of covariates $\boldsymbol{X}$. Then, the linear form of the logistic regression model for Y is written as follows:

$$\text{logit}\,(P(Y=1|\boldsymbol{X})) = \ln\left[\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right] = \boldsymbol{\beta}^{\top}\boldsymbol{X}, \tag{2}$$

being $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)^{\top}$ the vector of regression coefficients.

Consider $U = \{1,\ldots,N\}$ a finite population of N units. In the context of complex survey data, let S be a sample of n units drawn from the finite population by some complex sampling design. To each sampled observation $i \in S$, a set of values $(y_i,\boldsymbol{x}_i,w_i)$ is associated where each sampling weight w_i indicates the number of units that $i \in S$ represents in the finite population (note that $\sum_{i\in S} w_i = N$) and y_i and $\boldsymbol{x}_i$ indicate the realizations of the variables Y and $\boldsymbol{X}$ for the sampled units, respectively. For each $i \in S$ let us define its probability of event as $p(\boldsymbol{x}_i) = P(Y=1|\boldsymbol{X}=\boldsymbol{x}_i)$, which can be estimated as follows:

$$\hat{p}(\boldsymbol{x}_i) = \frac{e^{\hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}_i}}{1+e^{\hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}_i}} \quad (i \in S), \tag{3}$$

where the estimated regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\ldots,\hat{\beta}_p)^{\top}$, are usually obtained by maximizing the weighted pseudo-likelihood function, defined as (Binder, 1981, 1983):

$$PL(\boldsymbol{\beta}) = \prod_{i\in S} p(\boldsymbol{x}_i)^{y_i w_i}(1-p(\boldsymbol{x}_i))^{(1-y_i)w_i}. \tag{4}$$

3.2. Optimal cut-off point estimation methods

It is usually very useful in practice to select a cut-off point in order to distinguish between units with and without the event of interest. In our particular case, we are interested in discriminating between units with and without the event of interest based on their estimated probability of event. In this context, one observation $i \in S$ is usually classified as event if its estimated probability of event exceeds a determined threshold c which has been previously selected (Magder and Fix, 2003; Pepe, 2003). The correct classification of an observation with the event of interest is usually denoted as *true positive* (TP), while the correct classification of an observation without the event of interest is commonly denoted as *true negative* (TN). But usually, those classifications are not entirely accurate. Therefore, some of the observations are commonly classified incorrectly: an observation with the event of interest may be classified as non-event (*false negative* (FN)) or an observation without the event of interest may be classified as event (*false positive* (FP)).

Methods of estimation of the optimal cut-off point have been developed in the literature, with the aim of optimizing diverse measures. In particular, many methods consist on the optimization of an objective function of the Receiver Operating Characteristic (ROC) curve, which is a curve that describes the global accuracy of a model (Bamber,

1975; Pepe, 2003). Coming back to our particular case, taking into account that the predicted probabilities range from 0 to 1, the ROC curve of a logistic regression model can be defined as follows (Hosmer and Lemeshow, 2000; Pepe, 2003):

$$ROC(\cdot) = \{(1 - Sp(c), Se(c)),\ c \in (0,1)\}, \tag{5}$$

where $Se(c)$ and $Sp(c)$ are defined as follows and denoted as sensitivity and specificity, respectively:

$$\begin{array}{l} Se(c) = P\left[P(Y=1|\boldsymbol{X}) \geq c | Y=1\right], \\ Sp(c) = P\left[P(Y=0|\boldsymbol{X}) < c | Y=0\right]. \end{array} \tag{6}$$

In practice, following the notation defined so far, assume that to each sampled observation $i \in S$ a set of values $(y_i, \boldsymbol{x}_i, w_i)$ is associated. Suppose that the vector $\hat{\boldsymbol{\beta}}$ is obtained by means of the pseudo-likelihood function in (4) and $\hat{p}(\boldsymbol{x}_i)$ are estimated for $i \in S$ following (3). Let us define the following groups of correctly or incorrectly classified observations, for a specific cut-off point c:

$$\begin{array}{ll} TP_c = \{i \in S : y_i = 1 \text{ and } \hat{p}(\boldsymbol{x}_i) \geq c\}, & TN_c = \{i \in S : y_i = 0 \text{ and } \hat{p}(\boldsymbol{x}_i) < c\}, \\ FP_c = \{i \in S : y_i = 0 \text{ and } \hat{p}(\boldsymbol{x}_i) \geq c\}, & FN_c = \{i \in S : y_i = 1 \text{ and } \hat{p}(\boldsymbol{x}_i) < c\}. \end{array} \tag{7}$$

In addition, let us define an indicator function associated to each of the sets defined in (7) as follows. For example, for the set TP_c:

$$1_{TP_c}(i) = \left\{ \begin{array}{lll} 1 & \text{if} & i \in TP_c, \\ 0 & \text{if} & i \notin TP_c. \end{array} \right. \tag{8}$$

In the same way, indicator functions can be defined as in (8) for the rest of the sets described in (7), which will be denoted as $1_{TN_c}(i)$, $1_{FP_c}(i)$ and $1_{FN_c}(i)$, hereinafter. Then, for a specific cut-off point c, sensitivity and specificity parameters can be estimated based on sample S as follows:

$$\widehat{Se}(c) = \frac{\sum_{i \in S} 1_{TP_c}(i)}{\sum_{i \in S}\left[1_{FN_c}(i) + 1_{TP_c}(i)\right]}, \quad \widehat{Sp}(c) = \frac{\sum_{i \in S} 1_{TN_c}(i)}{\sum_{i \in S}\left[1_{TN_c}(i) + 1_{FP_c}(i)\right]}. \tag{9}$$

For this study, we have selected some of those methods which are based on several optimality criteria related to sensitivity and specificty parameters:

- *Youden* (Youden, 1950; Greiner et al., 2000): This method selects the cut-off point (c^{Youden}) that maximizes the Youden Index, which is defined as the sum of sensitivity and specificity parameters minus one, i.e.,

$$c^{\text{Youden}} = \underset{c \in (0,1)}{\text{argmax}} \left\{ \widehat{Se}(c) + \widehat{Sp}(c) - 1 \right\}. \tag{10}$$

- *MaxProdSpSe* (Lewis et al., 2008): This method selects the cut-off point c that maximizes the product between sensitivity and specificity parameters, i.e.,

$$c^{\text{MaxProdSpSe}} = \underset{c \in (0,1)}{\text{argmax}} \left\{ \widehat{Se}(c) \cdot \widehat{Sp}(c) \right\}. \tag{11}$$

- *ROC01* (Metz, 1978; Vermont et al., 1991): This method selects the cut-off point c that minimizes the distance between the ROC curve and the point (0,1), i.e.,

$$c^{\text{ROC01}} = \underset{c\in(0,1)}{\text{argmin}} \left\{ (\widehat{Se}(c)-1)^2 + (\widehat{Sp}(c)-1)^2 \right\}. \tag{12}$$

- *MaxEfficiency* (Greiner, 1995, 1996): This method selects the cut-off point c that maximizes the efficiency or, in other words, minimizes the error, i.e.,

$$c^{\text{MaxEff}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{p}_Y \widehat{Se}(c) + (1-\widehat{p}_Y)\widehat{Sp}(c) \right\}, \tag{13}$$

where $\widehat{p}_Y$ is the estimated prevalence which is calculated as follows:

$$\widehat{p}_Y = \frac{1}{n}\sum_{i\in S} [1_{FN_c}(i) + 1_{TP_c}(i)]. \tag{14}$$

3.3. Cut-off point estimation proposal with sampling weights

Although sensitivity and specificity parameters, as well as the prevalence, can be estimated by expressions (9) and (14) in any kind of data, including complex survey data, these expressions have been defined in a simple random sampling scenario. However, in complex survey data each of the sampled units has a sampling weight associated, which indicates the importance of each of them within the sample. Thus, the influence of all sampled units is not uniform. Therefore, we believe that the estimates obtained by means of the above-mentioned formulas may be misleading for complex survey data and they should be pondered, so that they incorporate the sampling weights. In this way, instead of the number of correct or incorrect classifications in sample S, it should be considered the number of units that these correctly or incorrectly classified observations represent in the finite population. For this reason, we propose to consider the sampling weights w_i to estimate sensitivity ($\widehat{Se}_w(c)$) and specificity ($\widehat{Sp}_w(c)$) parameters as follows:

$$\widehat{Se}_w(c) = \frac{\sum_{i\in S} w_i \cdot 1_{TP_c}(i)}{\sum_{i\in S} w_i \cdot [1_{FN_c}(i) + 1_{TP_c}(i)]}, \quad \widehat{Sp}_w(c) = \frac{\sum_{i\in S} w_i \cdot 1_{TN_c}(i)}{\sum_{i\in S} w_i \cdot [1_{TN_c}(i) + 1_{FP_c}(i)]}. \tag{15}$$

where the indicator functions are the ones described in (8).

In addition, note that sampling weights should also be considered to estimate the prevalence ($\widehat{p}_{Y,w}$):

$$\widehat{p}_{Y,w} = \frac{1}{N}\sum_{i\in S} w_i \cdot [1_{FN_c}(i) + 1_{TP_c}(i)]. \tag{16}$$

Therefore, we propose to estimate the optimal cut-off points based on the modified parameters of sensitivity ($\widehat{Se}_w(c)$) and specificity ($\widehat{Sp}_w(c)$) when working with complex survey data, i.e.:

$$c_w^{\text{Youden}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{Se}_w(c) + \widehat{Sp}_w(c) - 1 \right\}, \tag{17}$$

$$c_w^{\text{MaxProdSpSe}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{Se}_w(c) \cdot \widehat{Sp}_w(c) \right\}, \tag{18}$$

$$c_w^{\text{ROC01}} = \underset{c\in(0,1)}{\text{argmin}} \left\{ (\widehat{Se}_w(c) - 1)^2 + (\widehat{Sp}_w(c) - 1)^2 \right\}, \tag{19}$$

$$c_w^{\text{MaxEff}} = \underset{c\in(0,1)}{\text{argmax}} \left\{ \widehat{p}_{Y,w}\widehat{Se}_w(c) + (1 - \widehat{p}_{Y,w})\widehat{Sp}_w(c) \right\}. \tag{20}$$

4. Simulation study

This section describes the simulation process developed in this work and the scenarios that have been drawn. The results obtained in this simulation study are also presented in this section.

As stated above, the aim of this work is to study the influence of sampling weights in the estimation process of optimal cut-off points for the methods described in Section 3.2. Since the decision of which optimal cut-off point estimation method to use in practice depends on the research of interest, the objective of this work is not to compare the behaviour of the methods among them, but to compare the estimates that we obtain for each of these methods when sampling weights are considered or not in the estimation of sensitivity and specificity parameters.

In addition, we study the impact that the proposed estimators have in the estimation of the probability of event in the finite population. Therefore, a theoretical finite population is required, in which the response variable is known for all the units in the finite population. Thus, a pseudo-population has been generated based on real survey data. The real survey on which this pseudo-population is based is described in Section 2 and the process followed to generate it is explained in detail in Appendix A. The pseudo-population sampling process, which is replicated several times in the simulation study, is also based on the same real-life survey. This sampling process is described in Appendix B.

4.1. Scenarios and set up

Let $U = \{1,\dots,N\}$ be the pseudo-population generated by following the steps described in Appendix A to which $\{(y_i,\boldsymbol{x}_i)\}_{i=1}^N$ are assigned. From this pseudo-population, a total of $R = 500$ samples have been obtained and the sampling weights have been assigned to the sampled units by the sampling process described in Appendix B. The optimal cut-off points estimation methods that have been applied in this study are the ones described in Section 3.2, i.e., $m \in \{\text{Youden, MaxProdSpSe, ROC01, MaxEfficiency}\}$.

The steps that have been followed in the simulation study are described below. For $r = 1,\dots,500$:

Step 1. Draw a sample $S^r \subset U$ by one-step stratification with simple random sampling without replacement in each stratum (Appendix B, mimicking the sampling process carried out for the real-life dataset described in Section 2).

Step 2. Fit the logistic regression model to S^r and estimate $\hat{\boldsymbol{\beta}}^r$ by (4).

Step 3. For $i \in S^r$, estimate $\hat{p}^r(\boldsymbol{x}_i)$ by means of $\hat{\boldsymbol{\beta}}^r$ following (3).

Step 4. Estimate the optimal cut-off points, $c^{m,r}$ (see (10), (11), (12), (13)) and $c_w^{m,r}$ (see (17), (18), (19), (20)) for each method m.

As mentioned above, the selection of the optimality criteria for selecting the cut-off points is based on the particular goal of each study. Therefore, our goal is not to compare the performance of the described methods between them. That is, the aim is not to compare the performance of a method $m \in \{$Youden, MaxProdSpSe, ROC01, MaxEfficiency$\}$, to the rest of the methods, but to compare the cut-off points selected by means of the method m when sampling weights are considered or not in the estimation process. Thus, we define the difference and absolute difference between weighted and unweighted cut-off points as follows:

$$\text{Diff}^{\,m,r} = c^{m,r} - c_w^{m,r} \quad \text{and} \quad \text{AbsDiff}^{\,m,r} = |c^{m,r} - c_w^{m,r}|. \tag{21}$$

In addition, we would also like to regard the impact that may have the decision to select weighted or unweighted optimal cut-off points in the classification of all the units in the finite population. Thus, we continue with the simulation study as follows:

Step 5. For $i = 1, \ldots, N$ calculate $\hat{p}^r(\boldsymbol{x}_i)$ by means of $\hat{\boldsymbol{\beta}}^r$ (**Step 3.**) following (3).

Step 6. For $i = 1, \ldots, N$ classify each unit as event or non-event based on $\hat{p}^r(\boldsymbol{x}_i)$. Define two estimated responses ($\hat{y}_i^{m,r}$ and $\hat{y}_{w,i}^{m,r}$) for each unit based on the cut-off points $c^{m,r}$ and $c_w^{m,r}$ (selected in **Step 4.**) as follows. For each method m and $i = 1, \ldots, N$:

$$\hat{y}_i^{m,r} = \begin{cases} 1 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) \geq c^{m,r}, \\ 0 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) < c^{m,r}, \end{cases} \quad \text{and} \quad \hat{y}_{w,i}^{m,r} = \begin{cases} 1 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) \geq c_w^{m,r}, \\ 0 & \text{if} \quad \hat{p}^r(\boldsymbol{x}_i) < c_w^{m,r}. \end{cases}$$

Finally, in order to account for the error that may be introduced in the classification of the units in the finite population by the selected optimal cut-off points, one more parameter is defined. The error is estimated by comparing the prevalence estimated by means of the estimated responses (**Step 6**) to the true prevalence in the finite population. We split the finite population U in K disjointed subsets of the same size where $U = U_1 \cup \ldots \cup U_K$. We repeat this process $L = 10$ times, where for each $l = 1, \ldots, L$, $U = U_1^l \cup \ldots \cup U_K^l$. In this way, we get $L \times K$ subsets from U and the prevalence will be estimated in each one of these subsets. Let us define the following indicator functions:

$$1_{U_k^l}(i) = \begin{cases} 1 & \text{if} \quad i \in U_k^l, \\ 0 & \text{if} \quad i \notin U_k^l, \end{cases} \quad \text{for} \quad l = 1, \ldots, L \quad \text{and} \quad k = 1, \ldots K. \tag{22}$$

We denote as global mean squared error (GMSE) of the prevalence with $L = 10$ replicates the following parameters:

$$\begin{aligned} \text{GMSE}^{m,r} &= \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{N} \hat{y}_i^{m,r} \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} \right)^2, \\ \text{GMSE}_w^{m,r} &= \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{N} \hat{y}_{w,i}^{m,r} \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_k^l}(i)}{\sum_{i=1}^{N} 1_{U_k^l}(i)} \right)^2. \end{aligned} \tag{23}$$

Different number of subsets have been selected in order to evaluate the impact the sample size of each subset may have: $K \in \{1, 10, 100, 500\}$. In addition, we considered the GMSE evaluated considering the H strata as the subsets where U_h, $\forall h = 1, \ldots, H$ indicates the subset corresponding to stratum h and $U = \bigcup_{h=1}^{H} U_h$:

$$\begin{aligned} \text{GMSE}_h^{m,r} &= \frac{1}{H} \sum_{h=1}^{H} \left(\frac{\sum_{i=1}^{N} \hat{y}_i^{m,r} \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} \right)^2, \\ \text{GMSE}_{w,h}^{m,r} &= \frac{1}{H} \sum_{h=1}^{H} \left(\frac{\sum_{i=1}^{N} \hat{y}_{w,i}^{m,r} \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} - \frac{\sum_{i=1}^{N} y_i \cdot 1_{U_h}(i)}{\sum_{i=1}^{N} 1_{U_h}(i)} \right)^2, \end{aligned} \tag{24}$$

where,

$$1_{U_h}(i) = \begin{cases} 1 & \text{if} \quad i \in U_h, \\ 0 & \text{if} \quad i \notin U_h, \end{cases} \qquad \text{for} \quad h = 1, \ldots, H. \tag{25}$$

This simulation study has been carried out by means of the statistical software R. In particular, some functions of the R package `OptimalCutpoints` (López-Ratón et al., 2014) have been modified in order to incorporate an argument that provides us with the option to consider sampling weights in the estimation process of the optimal cut-off points for the described methods.

4.2. Results

In this Section we show the results obtained in the simulation study described in Section 4.1. Figures 1, 2, 3 and 4 depict the box-plots of unweighted and weighted estimates of the optimal cut-off points and the results of the parameters Diff and GMSE (see (21) and (23)) for Youden, MaxProdSpSe, ROC01 and MaxEfficiency methods, respectively. Numerical results of the simulation study are summarized in Table 1.

In general, except for the MaxEfficiency method, the results suggest that the optimal cut-off point estimates differ when sampling weights are ignored or considered in the estimation process. The difference has always been positive (i.e. the unweighted estimates have been greater than the weighted ones), except in the MaxEfficiency method where both positive and negative differences have been observed. For this reason, the mean and standard deviation of the difference and absolute difference parameters are equal for all the methods except for MaxEfficiency (see Table 1). The error generated and accounted in terms of GMSE described in (23) decreases considerably when sampling

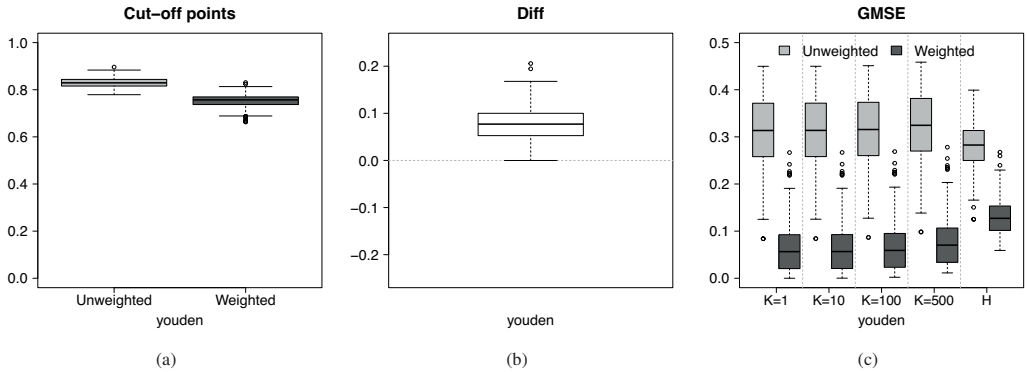
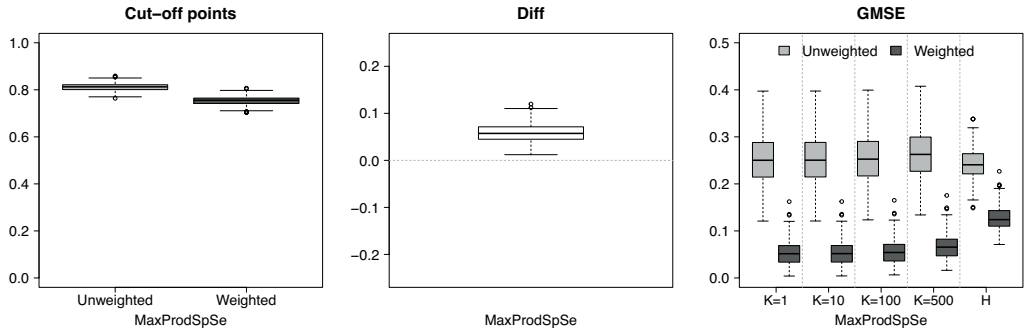


Figure 1. *Box-plots of the results obtained for the Youden method across $R = 500$ samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for $K \in \{1, 10, 100, 500\}$ and H.*

weights are taken into account. In addition, similar results have been obtained for different $K \in \{1, 10, 100, 500\}$ values, which indicates that the difference between estimated and true prevalence is similar in smaller homogeneous subsets and in the total population. However, it could be observed that the average of GMSE becomes slightly greater as the number of subsets K increases (for both, weighted and unweighted estimates), indicating that the differences between the estimated and true prevalence tend to be a little bit greater in smaller subsets. When considering the strata as non-homogeneous subsets defined by the H strata of the population, the GMSE obtained as described in (24) with the weighted estimates is still smaller than with the unweighted ones. However, the difference between weighted and unweighted GMSE is slightly smaller for the non-homogenous partition than for homogeneous partitions. We believe that the reason is that the difference obtained between estimated and true prevalence differs depending on the number of individuals sampled in each strata, being increased in very small strata. Note that if the population size of a particular stratum is 1 then the error in this stratum is 0 (if the unit is classified correctly) or 1 (otherwise). This is not common when working with homogeneous strata where in all the randomly selected subsets the difference between estimated and true prevalence seem to be similar (results not shown). In addition, note that even though strata are of different sizes, the stratum size is not taken into account when computing the GMSE parameter. Below, the behaviour of each of the methods that have been studied throughout this work will be analysed one by one.

The optimal cut-off point estimated by the Youden method in this simulation study, is 0.8304 on average when sampling weights are not taken into account while the weighted estimates are smaller on average (0.7524), with standard deviations of 0.0208 and 0.0277, respectively. The difference among the unweighted and weighted estimates is on average 0.0780 with a standard deviation of 0.0343 (see Figure 1). The smallest difference observed among the unweighted and weighted estimates is 0 while the largest difference is 0.2057, with a median of 0.0771. The impact of the differences between these estimates



(a) (b) (c)

Figure 2. *Box-plots of the results obtained for the MaxProdSpSe method across $R = 500$ samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for $K \in \{1, 10, 100, 500\}$ and H.*

in the total population was measured by the GMSE parameter. In terms of GMSE, the error produced by means of the weighted estimates in the finite population is more or less 5 times smaller than the error produced by means of the unweighted estimates on average. The standard deviation is also smaller for the weighted estimates. For example, for $K = 1$ the GMSE of the unweighted estimates is 0.3110 on average with a standard deviation of 0.0747, while the GMSE of the weighted estimates is 0.0630 on average with a standard deviation of 0.0503. When the GMSE is computed over the $H = 325$ strata, the GMSE turns out to be 0.1298 and 0.2809, for weighted and unweighted estimates, respectively.

The unweighted estimates obtained by the MaxProdSpSe method are again greater than the weighted ones, being on average 0.8117 and 0.7534, respectively (see Figure 2). The difference between those estimates is 0.0584 on average with a standard deviation of 0.0190. The smallest difference observed among the unweighted and weighted estimates is 0.0121 while the largest difference is 0.1198, with a median of 0.0573. GMSE becomes again 5 times smaller when sampling weights are considered in the estimation process and the standard deviation of the weighted estimates is half of that of the unweighted ones. For example, for $K = 100$ the GMSE is reduced from 0.2532 to 0.0556 on average when considering sampling weights, being the standard deviations of 0.0708 and 0.0342, respectively. The GMSE measured over the different strata for weighted and unweighted estimates is 0.1261 and 0.2425, respectively.

For the ROC01 method weighted estimates are also lower than the unweighted ones (0.7526 and 0.8078 on average, respectively) and the standard deviations are slightly greater (0.0174 and 0.0151, respectively) (see Figure 3). The smallest difference observed among the unweighted and weighted estimates is 0.0121 while the largest difference is 0.1088, being the median of 0.0540 and the average of 0.0552 with a standard deviation of 0.0166. The error generated by the weighted estimates in the finite population is again lower than the error produced by the unweighted estimates in terms of GMSE.

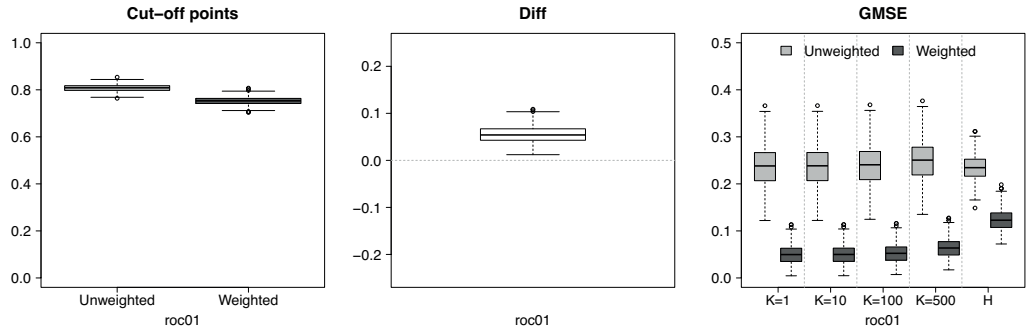


Figure 3. *Box-plots of the results obtained for the ROC01 method across $R = 500$ samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for $K \in \{1, 10, 100, 500\}$ and H.*

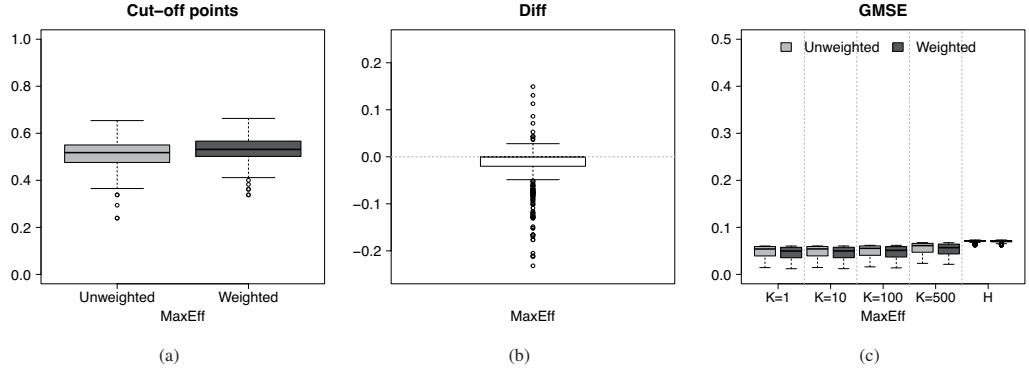


Figure 4. *Box-plots of the results obtained for the MaxEfficiency method across $R = 500$ samples: (a) unweighted and weighted estimates of the optimal cut-off points, (b) differences between unweighted and weighted estimates (Diff), and (c) GMSE produced by the unweighted and weighted estimates for $K \in \{1, 10, 100, 500\}$ and H.*

For example, for $K = 10$, the error obtained by the weighted estimates is 0.0507 on average with a standard deviation of 0.0210, while for the unweighted estimates the error is 0.2368 on average with a standard deviation of 0.0462. The GMSE computed over the different strata takes the value of 0.1245 and 0.2340 for weighted and unweighted estimates, respectively.

Finally, in contrast to the results obtained by the rest of the methods, for the MaxEfficiency method no significant differences are observed among the unweighted and weighted estimates. Optimal cut-off point estimates throughout the $R = 500$ samples are quite similar in terms of mean and standard deviation. The average of the unweighted estimates is of 0.5106 while for the weighted estimates the average is of 0.5297. The standard deviation of the weighted estimates (0.0522) is slightly lower than the standard deviation of the unweighted estimates (0.0579). The smallest absolute difference

Table 1. *Average (mean) and standard deviation (sd) of the a) unweighted and weighted optimal cut-off points, b) difference (Diff) and absolute difference (AbsDiff) among them and, c) GMSE produced by the unweighted and weighted optimal cut-off points when classifying units in the finite population for* $K \in \{1, 10, 100, 500\}$ *and* H *across* $R = 500$ *samples for all the methods considered.*

		Youden Mean (sd)	MaxProdSpSe Mean (sd)	ROC01 Mean (sd)	MaxEff Mean (sd)
Cut-off points	**Unweighted**	0.8304 (0.0208)	0.8117 (0.0157)	0.8078 (0.0151)	0.5106 (0.0579)
	Weighted	0.7524 (0.0277)	0.7534 (0.0183)	0.7526 (0.0174)	0.5297 (0.0522)
Diff		0.0780 (0.0343)	0.0584 (0.0190)	0.0552 (0.0166)	-0.0191 (0.0456)
AbsDiff		0.0780 (0.0343)	0.0584 (0.0190)	0.0552 (0.0166)	0.0232 (0.0436)
GMSE (K=1)	**Unweighted**	0.3110 (0.0747)	0.2509 (0.0525)	0.2366 (0.0440)	0.0482 (0.0132)
	Weighted	0.0630 (0.0503)	0.0530 (0.0243)	0.0505 (0.0198)	0.0454 (0.0136)
GMSE (K=10)	**Unweighted**	0.3112 (0.0762)	0.2511 (0.0544)	0.2368 (0.0462)	0.0483 (0.0140)
	Weighted	0.0632 (0.0509)	0.0532 (0.0253)	0.0507 (0.0210)	0.0456 (0.0144)
GMSE (K=100)	**Unweighted**	0.3131 (0.0899)	0.2532 (0.0708)	0.2390 (0.0642)	0.0496 (0.0211)
	Weighted	0.0656 (0.0566)	0.0556 (0.0342)	0.0531 (0.0307)	0.0469 (0.0211)
GMSE (K=500)	**Unweighted**	0.3219 (0.1361)	0.2628 (0.1203)	0.2488 (0.1153)	0.0556 (0.0419)
	Weighted	0.0764 (0.0791)	0.0667 (0.0621)	0.0642 (0.0594)	0.0530 (0.0413)
GMSE (H)	**Unweighted**	0.2809 (0.0470)	0.2425 (0.0325)	0.2340 (0.0270)	0.0706 (0.0022)
	Weighted	0.1298 (0.0377)	0.1261 (0.0250)	0.1245 (0.0235)	0.0701 (0.0025)

observed among the unweighted and weighted estimates is 0 while the largest absolute difference is 0.2318. In particular, in more than 50% of the cases the difference between weighted and weighted estimates is 0. The difference of the error produced by those estimates in the finite population is also negligible. For $K = 1$ for example, the GMSE produced by the unweighted estimates is on average of 0.0482 with a standard deviation of 0.0132, while the average of GMSE of the weighted estimates is 0.0454 with a standard deviation of 0.0136. The GMSE calculated over the $H = 325$ strata, is 0.0701 for weighted estimates and 0.0706 for unweighted estimates.

5. Application to a real survey data

The methodology proposed in Section 3 could be applied to real-world surveys. In particular, for illustration purposes, we have applied this methodology to the ESIE survey data described in Section 2.

In this case, the response variable Y in which we are interested in indicates the availability of the website for each company: it takes the value $y_i = 1$ if a company has its own website and $y_i = 0$ otherwise. Assume that the goal is to estimate the probability

Table 2. *Optimal cut-off point estimates obtained by means of Youden, MaxProdSpSe, ROC01 and MaxEfficiency methods, considering or not the sampling weights.*

	Youden	**MaxProdSpSe**	**ROC01**	**MaxEff**
Unweighted	0.7998	0.7998	0.7998	0.3882
Weighted	0.7518	0.7518	0.7470	0.3882

of event for Y of the companies in the finite population. Thus, we want to fit a logistic regression model to our sample. Four categorical variables that are also available in the finite population will be used as predictors: X_1 (which indicates the province where the company is located, in 3 categories), X_2 (indicates the activity of the company, in 9 categories), X_3 (indicates the ownership of the company, in 7 categories) and X_4 (indicates the number of employees of the company, in 4 categories). In this way, a logistic regression model was fitted to the sample considering these four covariates, the regression coefficients where estimated and $\hat{p}(\boldsymbol{x}_i)$ where calculated for each sampled unit.

We have applied the methods described in Section 3 for the selection of optimal cut-off points, which have been estimated by both, ignoring and considering sampling weights. The results are shown in Table 2. It can be observed that the unweighted and weighted estimates differ when Youden, MaxProdSpSe and ROC01 methods are applied, which is in line with the results obtained in the simulation study. In particular, the unweighted estimates are greater than the weighted estimates, which are similar to the ones observed in Section 4.2 (see Table 1). The unweighted and weighted estimates obtained by means of the MaxEfficiency method are equal, which is also in line with the results observed in the simulation study. Those estimates obtained by the MaxEfficiency method are lower than the average of the estimates obtained in the simulation study. However, it should be noted that this may be justified by the large standard deviation observed previously for the cut-off points estimated by means of the MaxEfficiency method (see Figure 4 and Table 1).

6. Discussion

In this work, a methodology has been proposed for estimating optimal cut-off points of the probability of event in the logistic regression framework cons

idering sampling weights in the estimation process. In particular, we have focused on data derived from complex sampling designs. For this purpose, four optimal cut-off point estimation methods (which are denoted as Youden, MaxProdSpSe, ROC01 and MaxEfficiency (López-Ratón et al., 2014)) have been selected and modified in order to incorporate sampling weights in the estimation process. These four methods have been selected for being the ones most commonly applied in the literature. In particular, the so widely used `pROC` package in R (Robin et al., 2011) has incorporated the Youden and ROC01 methods for the estimation of optimal cut-off points. All these methods are based on different optimality criteria that are related to sensitivity and specificity param-

eters. Therefore, we propose a methodology for considering sampling weights in the estimation process of sensitivity and specificity parameters, as well as in the estimation of prevalence, in order to estimate optimal cut-off points based on these parameters by taking into account the sampling weights. A simulation study has been carried out in order to analyse the behaviour of both methodologies by comparing the optimal cut-off point estimates obtained by means of the above-mentioned methods when sampling weights are considered or ignored in the estimation process. The error that those estimates generate in the estimation of the probability of event of interest in the finite population has also been analysed in this simulation study. In particular, we considered the GMSE in order to evaluate the behaviour of the prevalence once the cut-off point was estimated, by comparing it with the true prevalence. We also considered it interesting to study the differences in estimating sensitivity and specificity based on the cut-off points estimated with and without sampling weights. However, in this case, the theoretical value of these parameters in the population are unknown and therefore the comparison is not so direct. Even so, we have observed (results not shown) that the differences are in line with those observed when studying the GMSE.

In general, the results suggest the convenience of incorporating sampling weights into the estimation process of optimal cut-off points. For three out of the four methods studied, estimates obtained differ depending on whether the sampling weights were considered or not. Furthermore, it can be observed that the error in the estimates of the response variable obtained by taking into account sampling weights is much smaller than that generated by the estimates obtained by ignoring them for the units in the finite population. Although the cut-off point estimates may not seem very different from each other in some cases, it is observed that the effect of applying one or the other estimate for the classification of units in the population is considerable. In our opinion, the reason for this is that a large amount of individuals of the finite population (specifically, more than 20% of all the units on average) has estimated probabilities which range in the interval defined by the unweighted and weighted estimates and thus, choosing the unweighted cut-off point leads to misclassify a larger number of units in the finite population.

Nevertheless, the results related to the MaxEfficiency method appear to be different compared to Youden, MaxProdSpSe and ROC01. In general, in the results obtained using this method, there are no great differences between the estimates obtained by ignoring or considering the sampling weights, and furthermore, in most cases, the two estimates coincide. Therefore, the errors generated in the population by these estimates are also similar and there are no significant differences among them. Hence, we can say that, at least under the scenario we have worked on, there is no difference among the unweighted and weighted estimates obtained by the MaxEfficiency method. However, we believe that this could be due to a particular characteristic of the scenario in which we have worked and not a specific property of the method itself. Specifically, we believe that differences among those estimates obtained by using or not sampling weights could occur when there are also significant differences between unweighted and weighted estimates of the prevalence, which is not the case in the scenario that has been

studied. In particular, the unweighted estimate of the prevalence is 0.8330 on average in the simulated samples, while the weighted estimate is 0.7552. Due to the properties of the efficiency function, we believe that different cut-off point estimates may be obtained for this method when one of the prevalence estimates (either weighted or unweighted) is greater than 0.5, while the other is smaller (results not shown). Nevertheless, studying the mathematical properties of this behaviour is part of a further research, which is out of the scope of this paper.

Finally, we would like to comment on the limitations of this study. First of all, it should be noted that we have conducted this simulation study based on a real survey data. Therefore, the effect that the sampling technique chosen may have on the differences between weighted and unweighted optimal cut-off point estimates remains to be studied as further work. For example, it should be mentioned that in this study we have only analysed the effect of the sampling weights obtained by means of one-stage stratification. Data derived from other sampling techniques such as clustering or two-stage sampling have not been considered. It would also be interesting to study the behaviour of the studied methods under non-informative complex sampling designs. Secondly, it would be interesting to analyse and compare the behaviour of the methods that have been studied throughout this document in different scenarios, for instance, with different prevalence values. Nevertheless, it should be noted that as the simulation study we have used is based on a real survey, the prevalence of the scenario we have analysed was also described by the observed data.

In conclusion, in this work we have implemented four of the most commonly used optimal cut-off point estimation methods, which are implemented in diverse software. Out of these four methods, in three of them the use of sampling weights highly improve the results, while in the fourth, the results do not differ whether you use the sampling weights or not. Therefore, our recommendation is to incorporate the sampling weights in the estimation process of optimal cut-off points when working with data derived from complex sampling designs. However, it should be noted that if one is interested in applying other methods, different from those studied throughout this paper, it should be considered whether it is appropriate or not the use of sampling weights in each particular case.

Acknowledgment

This work was financially supported in part by grants from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco [IT1294-19] and by the Spanish State Research Agency through BCAM Severo Ochoa excellence accreditation [SEV-2017-0718] and through project [PID2020-115882RB-I00 / AEI / 10.13039/5011 00011033] funded by Agencia Estatal de Investigación and acronym “S3M1P4R”; Basque Government through the [BERC 2018-2021]; the SPRI-Basque Government through the Elkartek project 3KIA [KK-2020/00049]. The work of AI was supported by grant [PIF18/213] and [PES18/26] from the Universidad del País Vasco UPV/EHU.

We would like to acknowledge the Official Statistics Basque Office (Eustat) for providing us with the ESIE survey data. We also gratefully acknowledge María Xosé Rodríguez Álvarez for helping us incorporate the sampling weights into the `Optimal Cutpoints` R package functions.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Baker, T. and Gerdin, M. (2017). The clinical usefulness of prognostic prediction models in critical illness. *European Journal of Internal Medicine*, 45:37–40.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Binder, D. A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7(2):157–170.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292.

Binder, D. A. and Roberts, G. (2009). Design- and model-based inference for model parameters. *Handbook of Statistics*, 29:33–54.

Chen, J.-Y., Feng, J., Wang, X.-Q., Cai, S.-W., Dong, J.-H., and Chen, Y.-L. (2015). Risk scoring system and predictor for clinically relevant pancreatic fistula after pancreaticoduodenectomy. *World Journal of Gastroenterology*, 21(19):5926–5933.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Filella, X., Alcover, J., Molina, R., Giménez, N., Rodríguez, A., Jo, J., Carretero, P., and Ballesta, A. M. (1995). Clinical usefulness of free PSA fraction as an indicator of prostate cancer. *International Journal of Cancer*, 63(6):780–784.

Greiner, M. (1995). Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. *Journal of Immunological Methods*, 185(1):145–146.

Greiner, M. (1996). Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. *Journal of Immunological Methods*, 191(1):93–94.

Greiner, M., Pfeiffer, D., and Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1-2):23–41.

Hanley, J. A. and Mcneil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017). *Applied Survey Data Analysis (2nd ed.)*. Chapman and Hall/CRC.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley New York.

Kalton, G. (1983). *Introduction to Survey Sampling*. Thousand Oaks, CA: Sage.

Lewis, J. D., Chuai, S., Nessel, L., Lichtenstein, G. R., Aberra, F. N., and Ellenberg, J. H. (2008). Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. *Inflammatory Bowel Diseases*, 14(12):1660–1666.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., Gude-Sampedro, F., et al. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61(8):1–36.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.

Magder, L. S. and Fix, A. D. (2003). Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *Journal of Clinical Epidemiology*, 56(10):956–962.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyse and compare ROC curves. *BMC Bioinformatics*, 12(77).

Rutter, C. M. and Miglioretti, D. L. (2003). Estimating the accuracy of psychological scales using longitudinal data. *Biostatistics*, 4(1):97–107.

Skinner, C. J., Holt, D., and Smith, T. F. (1989). *Analysis of Complex Surveys*. John Wiley & Sons.

Spence, R. T., Chang, D. C., Kaafarani, H. M., Panieri, E., Anderson, G. A., and Hutter, M. M. (2018). Derivation, validation and application of a pragmatic risk prediction index for benchmarking of surgical outcomes. *World Journal of Surgery*, 42(2):533–540.

Steyerberg, E. W. (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media.

Steyerberg, E. W., Marshall, P. B., Jan Keizer, H., and Habbema, J. D. F. (1999). Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. *Cancer*, 85(6):1331–1341.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47(4):522–532.

Vermont, J., Bosson, J., Francois, P., Robert, C., Rueff, A., and Demongeot, J. (1991). Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35(2):141–150.

Wynants, L., van Smeden, M., McLernon, D. J., Timmerman, D., Steyerberg, E. W., Van Calster, B., et al. (2019). Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(192).

Yao, W., Li, Z., and Graubard, B. I. (2015). Estimation of ROC curve with complex survey data. *Statistics in Medicine*, 34(8):1293–1303.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

A. Generation of the pseudo-population

This section describes the process of generating the pseudo-population that has been implemented in the simulation study described in Section 4. The pseudo-population has been generated based on a real survey data, which is described in Section 2. Let us denote as S_{ESIE} the original survey sample and U_{ESIE} the real finite population of size N (note that $S_{\text{ESIE}} \subset U_{\text{ESIE}}$). It should be noted that some information of the finite population U_{ESIE} and the real sample S_{ESIE} is known for us. In particular, for the N units in the finite population the values for the vector of covariates $X_1,\ldots,X_p$ are known, i.e. $\left\{(x_{1j},\ldots,x_{pj})\right\}_{j\in U_{\text{ESIE}}}$. In addition to the values for the covariates, the values of the response variables $Y_1,\ldots,Y_q$ are also known for the units in the sample, i.e. $\left\{(y_{1j},\ldots,y_{qj},x_{1j},\ldots,x_{pj})\right\}_{j\in S_{\text{ESIE}}}$. In the ESIE survey, a total of H strata have been defined (i.e., $\{1,\ldots,H\}$) combining information of three categorical design variables, which will be denoted as X_1, X_2 and X_3. Therefore, the finite population can be partitioned in subsets defined by means of these strata, i.e., $U_{\text{ESIE}} = \bigcup_{h=1}^{H} U_{\text{ESIE},h}$. $\forall h \in \{1,\ldots,H\}$ let us indicate as N_h the size of stratum h in the finite population U_{ESIE} ($U_{\text{ESIE},h}$) and as n_h the size of this stratum in the sample S_{ESIE}. Then, the sampling weight associated to a unit $j \in S_{\text{ESIE}}$ from stratum h is the following:

$$w_j = \frac{N_h}{n_h}. \tag{26}$$

Our goal is to generate a pseudo-population (U) based on the known real ESIE survey data, for which all the information of the covariates $X_1,\ldots,X_p$ and the response variables $Y_1,\ldots,Y_q$ will be available. This new pseudo-population U will be the same size as the true ESIE population (N). In order to ease the notation, the variable names of the pseudo-population are the same as in the real finite population and the units of the real ESIE population will be denoted as $j \in U_{\text{ESIE}}$ while the units that are artificially generated for the pseudo-population will be denoted as $i \in U$.

Several dichotomous response variables are available in the original survey (being the response variable Y, which we have applied in the simulation study, one of them). All

possible combinations of these response variables have been examined. For instance, assuming that $Y_1,\ldots,Y_q$ are all the response variables that are available in the survey (where $Y \in \{Y_1,\ldots,Y_q\}$), for some unit $j \in S_{\text{ESIE}}$: $\boldsymbol{y}_j = (y_{1j},\ldots,y_{qj}) = \alpha$, $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$, where $\{\alpha_1,\ldots,\alpha_A\}$ is the set of all of possible combinations of the responses. For each stratum (i.e., $\forall h \in \{1,\ldots,H\}$) and for each possible combination of the responses (i.e., $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$) we generate $N_{h,\alpha}$ units in the pseudo-population (U) as:

$$N_{h,\alpha} = \sum_{j \in S_{\text{ESIE}}} w_j \cdot 1_{U_{\text{ESIE},h}}(j) \cdot [\boldsymbol{y}_j = \alpha], \tag{27}$$

where,

$$1_{U_{\text{ESIE},h}}(j) = \begin{cases} 1, & \text{if } j \in U_{\text{ESIE},h}, \\ 0, & \text{if } j \notin U_{\text{ESIE},h}, \end{cases} \tag{28}$$

and

$$[\boldsymbol{y}_j = \alpha] = \begin{cases} 1, & \text{if } \ (y_{1j},\ldots,y_{qj}) = \alpha, \\ 0, & \text{if } \ (y_{1j},\ldots,y_{qj}) \neq \alpha. \end{cases} \tag{29}$$

In this way, $N_{h,\alpha}$ is the number of units of the pseudo-population U in stratum h, which take the values of responses $(y_{1j},\ldots,y_{qj}) = \alpha$. Once we repeat the process for $\forall h \in \{1,\ldots,H\}$ and $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$ a pseudo-population of $N = \sum_{h\in\{1,\ldots,H\}} \sum_{\alpha\in\{\alpha_1,\ldots,\alpha_A\}} N_{h,\alpha} = \sum_{j\in S_{\text{ESIE}}} w_j$ units will be generated with the information of response variables (Y, among others) and strata (hence, information of the design variables X_1, X_2 and X_3 will also be generated). Note that the pseudo-population U has been created in such a way that has the same number of individuals N as the ESIE finite population U_{ESIE}.

Finally, we generate the rest of the covariates as follows. $\forall s \in \{4,\ldots,p\}$ assume that X_s is a categorical variable with a total of D categories: $\{1,\ldots,D\}$. Then, for each unit generated in the pseudo-population ($\forall i \in U$) from stratum h, we generate $x_{si} \in \{1,\ldots,D\}$ following a categorical distribution (i.e., $x_{si} \sim Cat(\pi_{s1},\ldots,\pi_{sD})$) where the probability corresponding to each category $d \in \{1,\ldots,D\}$ is calculated as follows based on the known ESIE finite population U_{ESIE}.

$$\pi_{sd} = \frac{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},h}}(j) \cdot [x_{sj} = d]}{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},h}}(j)}, \quad \forall d \in \{1,\ldots,D\}, \tag{30}$$

where $1_{U_{\text{ESIE},h}}(j)$ is defined in (28) and,

$$[x_{sj} = d] = \begin{cases} 1 & \text{if } \ x_{sj} = d, \\ 0 & \text{if } \ x_{sj} \neq d, \end{cases} \quad \forall j \in U_{\text{ESIE}} \text{ and } \forall d \in \{1,\ldots,D\}. \tag{31}$$

In this way, the pseudo-population has been generated with the response variable of interest Y, the vector of covariates $\boldsymbol{X}$ and the strata.

B. Pseudo-population sampling process

The pseudo-population generated following the steps described in Appendix A, has been sampled by one-step stratified sampling with simple random sampling without replacement in each stratum, in the same way as the real survey data described in Section 2.

In the sampling process, first, we identify how many units have been sampled from a stratum h ($\forall h \in \{1,\ldots,H\}$) in the real survey sample S_{ESIE} (let us denote this amount as n_h). Then, we sample randomly n_h units from stratum h of size N_h from the pseudo-population U. In this way, repeating the process for $\forall h \in \{1,\ldots,H\}$ we sample a total of n units (where $n < N$) to the sample $S \subset U$.

Finally, sampling weights are assigned to each sampled unit as follows. For $\forall i^* \in S$ (assume that $i^* \in h$ ($\forall h \in \{1,\ldots,H\}$)), then:

$$w_{i^*} = \frac{N_h}{n_h}. \tag{32}$$

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process. The article must also use inclusive language, that is, language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as “man”, “mankind”, and masculine pronouns, the use of which might be considered to exclude women. The article should also inform about whether the original data of the research takes gender into account, in order to allow the identification of possible differences.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

- Chapter in book: Engelmann, B. (2006). Measures of a rating’s discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT *(Statistics and Operations Research Transactions)*

Name _____

Organisation _____

Street Address _____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)***) from now on

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:
sort@idescat.cat