SORT

Statistics and Operations Research Transactions

Volume
49

Number 1, January-June 2025

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 49, Number 1, January-June 2025

eISSN: 2013-8830

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) —formerly Qüestiió— is an international journal launched in 2003 and distributed in printed form as well as in digital form online. From 2024 it will be published in digital form only. It is 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, the Catalan Statistical Society and the Departament de Recerca i Universitats, of the Generalitat de Catalunya. 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

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

Publisher: Institut d'Estadística de Catalunya (Idescat)

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

FECYT-073/2024
Fecha de certificación: 20 de mayo de 2011 (2ª convocatoria)
Válido hasta: 24 de julio de 2025

eISSN: 2013-8830
SORT 49 (1) January-June (2025)

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
Departament de Recerca i Universitats

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 49
Number 1
January-June 2025
eISSN: 2013-8830

Invited article

Articles

SORT 49 (1) January-June 2025, 3-42 DOI: 10.57645/20.8080.02.21

Recent advances in copula-based methods for dependent censoring

Gilles Crommen*[1], Negera Wakgari Deresa*[1], Myrthe D'Haen*[1,2], Jie Ding*[1,3], Ilias Willems*[1] and Ingrid Van Keilegom[1]

Abstract

When modeling time-to-event data that are subject to right censoring, it is commonly assumed that the survival time T and the censoring time C are independent. However, this assumption frequently fails in practice, leading to biased estimators and testing procedures having invalid type 1 error rates. To overcome this issue, several models relaxing the independent censoring assumption have been proposed in the literature. Among these, copula-based approaches have become popular due to their ability to separately model the marginal distributions of T and C and their dependence structure. This review paper gives a comprehensive overview of recent advances in copula-based methods for dependent censoring, along with a discussion of the most important historical papers on this topic. As it is well known that the distribution of (T,C) (and hence of T) is not identified in a fully nonparametric way, we examine different strategies to achieve model identifiability. These strategies consist of imposing assumptions on either the copula or the marginal distributions of T and C. Both of these approaches will be discussed, with and without covariates. We also consider the case where a dependent censoring time is accompanied by an additional latent independent censoring time. Lastly, we briefly explain alternative approaches that are not based on copulas.

MSC: *62Nxx, 62-02.*

Keywords: *Copula, Dependent censoring, Identifiability, Survival analysis.*

[1] Research Centre for Operations Research and Statistics, KU Leuven, Leuven, Belgium.

[2] Data Science Institute, UHasselt, Diepenbeek, Belgium.

[3] School of Mathematical Sciences, Dalian University of Technology, Dalian, China.

* These authors contributed equally to this work.

Received: February 2025.

Accepted: April 2025.

1. Introduction

Survival analysis is concerned with the time until some event of interest occurs. This covers an endless amount of applications, ranging from failure of machine components to relapse of a certain disease, unemployment duration and many more. Depending on the application domain, common synonyms are *reliability theory* (for machine components) and *duration analysis* (for unemployment); the biomedical term *survival analysis* is usually used as an umbrella term. The time T until occurrence of the event of interest, usually called the *survival time*, *failure time* or even *time-to-event*, is the main quantity of interest. However, the very nature of survival data also presents us with the challenge of censoring. Indeed, in many cases, the event is not observed for all subjects within the time window of the study. This can be due to subjects surviving beyond the study, but also due to premature leaving caused by another, intervening event. For instance, machine components may be replaced before their complete breakdown, and in a medical setting, humans could drop out because they move to another region, die in a car crash, die from another disease than the one under study, because they simply no longer want to participate, because they experience undesired side effects of a treatment, etc. In all these cases, one observes the censoring time C of the other event that prohibits the observation of the survival time T.

To avoid some statistical difficulties that will be introduced in Section 1.1, one often treats censoring times as if they were 'unrelated' to the survival time (in a sense to be made more precise below). This assumption is not harmless and may result in biased estimates when incorrectly taken on (Moeschberger and Klein, 1984; Emura and Chen, 2016). In the above examples, people moving or dying in traffic is very likely unrelated to their health condition regarding the specific disease or treatment under study. On the other hand, in the machine components example, parts could be replaced in routine, yearly maintenance, but they could also be replaced precisely *because* they show early signs of failure. Similarly, in the medical setting, treating patient dropout as unrelated to their health condition (and hence their survival) is often not justifiable. Patients may, for instance, be withdrawn from a clinical study due to the occurrence of some side effects or treatment toxicity. The expected survival of the remaining patients after removing those with undesired side effects could be higher than that of the full (initial) population under study. Another example arises in transplant studies, where patients are selected for transplantation based on their medical condition. Since the most severely ill patients are prioritized, their expected survival on the waiting list may not be representative of those who were not selected for transplantation.

Many more such examples exist in which the assumption of censoring unrelated to survival is at least dubious. Increasing attention has therefore been given to approaches for dependent censoring in survival data. One popular and highly active such line of research is the one using copulas to accommodate for the dependence. The aim of this review is to provide an overview of both seminal articles and recent advances in the transition from independence to copula-based dependence modeling. First, however, we introduce some relevant basic notions and motivations.

1.1. Dependent censoring notions

Let us start by translating all of this into a more mathematical framework. In a typical survival study with censoring, one has access to the random variables Y and Δ for each subject, where $Y = \min(T, C)$ and $\Delta = \mathbb{1}(T \leq C) = \mathbb{1}(Y = T)$. In the latter formula, $\mathbb{1}(\cdot)$ denotes the indicator function. Thus, Y, usually called the *follow-up time*, corresponds to the time of the first occurring event, while the *censoring indicator* Δ is an indicator of the event type with $\Delta = 1$ for an observation of T, and $\Delta = 0$ for a censored observation. The observed data $\mathcal{D}_n$ of size n consist of n i.i.d. replicates of (Y, Δ) or, in a regression setting, of (Y, Δ, X), which we denote by $\{(Y_i, \Delta_i, X_i)\}_{i=1,\ldots,n}$. For ease of notation, we will assume throughout that the covariate vector $X = (1, \tilde{X}^\top)^\top$, including an intercept entry, contains the relevant components for both T and C (although in some articles, one works with separate covariate vectors for the individual margins). Throughout, T and C are continuous and positive random variables with densities f_T and f_C, respectively.

Uninformative vs. independent censoring. There are two concepts concerning the 'unrelatedness' of the survival and censoring time. For both concepts, multiple definitions exist in the literature; we first present those versions that are widely used and seem the most intuitive to understand, and some alternative references will be provided in another paragraph below. We say censoring is *uninformative* (or *noninformative*) if the distribution of C does not depend on the parameters of interest related to the survival time distribution (see Emura and Chen (2018), Chapter 2; Kalbfleisch and Prentice (2002), Chapter 3). In particular, in maximum likelihood methods, terms containing only the censoring distribution or density function can be removed from the likelihood since they yield no information for the modeling of T. Next, censoring is called *(unconditionally) independent* if the random variables T and C are stochastically independent. When both these 'unrelatedness' concepts hold, the likelihood for $\{(Y_i, \Delta_i)\}_{i=1,\ldots,n}$ can be written as

$$\prod_{i=1}^{n} f_T(Y_i)^{\Delta_i} \left(1 - F_T(Y_i)\right)^{1-\Delta_i} = \prod_{i=1}^{n} f_T(Y_i)^{\Delta_i} S_T(Y_i)^{1-\Delta_i}, \tag{1}$$

where we used the standard notation $F_T(y)$ and $S_T(y) = 1 - F_T(y) = \mathbb{P}(T > y)$ for the distribution and survival function, respectively, of T; the 'survival' function of C is analogously defined as $S_C(y) = 1 - F_C(y)$ in terms of its cumulative distribution $F_C(y)$. (We will sometimes explicitly indicate dependence on any marginal parameters by writing, for example, F_{T,θ_T}, including the parameter vector θ_T of T, but, when clear from the context, this may be omitted to avoid too heavy notation.)

The regression setting gives rise to an additional notion of *conditional independence*, requiring the conditional margins $T|X$ and $C|X$ to be stochastically independent. This notion is different from 'absolute' independence of T and C. For instance, patient drop-out (i.e., censoring) due to tumor progression could be dependent, whereas a patient's survival time and drop-out may be independent *conditionally* on the patient's cancer stage (Emura and Chen, 2018, p. 2).

Finally, we point out that one sometimes also includes a second, *administrative* censoring variable A, which usually corresponds to the end of the study. The simpler form of administrative censoring is called *type I censoring*. This occurs when all individuals enter the study simultaneously and are observed until a predetermined date, a setup commonly seen in industrial life testing (Andersen et al., 1993). In contrast, clinical trials often involve staggered entry, where patients are recruited at different times, but the study ends on a fixed date. In this case, the administrative censoring time is the length of time between the entrance into the study and the predetermined study end date. This type of censoring is known in the literature as *progressive Type I censoring* (see, e.g., Geskus, 2016; Andersen et al., 1993). For any type of administrative censoring, $Y = \min(T,C,A)$, and two indicators $\Delta_T = \mathbb{1}(Y = T)$ and $\Delta_C = \mathbb{1}(Y = C)$ together identify the observation type: $(\Delta_T = 1, \Delta_C = 0)$ corresponds to an event of interest, $(\Delta_T = 0, \Delta_C = 1)$ to a dependent censoring event and $(\Delta_T = 0, \Delta_C = 0)$ to an administrative censoring event. An equivalent convention is to denote observations as (Y,k), where $k = 0$ corresponds to observations that were administratively censored, $k = 1$ to observations that experienced the event of interest and $k = 2$ to observations that were dependently censored. Since one can generally assume A to be independent of (T,C) (even unconditionally so), including it in the model often only leads to limited complications.

Alternative definitions. Alternative definitions often rely on the probabilistic concepts of filtrations, counting processes and martingales. A detailed explanation is considered outside the scope of this review, but they tend to be slightly more generally satisfied. For instance, the notion of uninformative censoring can be extended in terms of filtrations (Kalbfleisch and Prentice, 2002, Chapter 6), leading to censoring probabilities that may actually depend on the parameters relevant to survival, as long as the censoring mechanism is somehow 'ancillary', that is, not contributing information about these parameters (Andersen et al., 1988, Section 1); such censoring schemes are still considered uninformative and allow reduction of the likelihood as above. Alternative notions of *independent* censoring, on the other hand, can be found in Fleming and Harrington (1991) and Andersen et al. (1988), for instance. The former defines independence in terms of the hazard rate

$$\lambda(y) = \lim_{h \to 0} \frac{\mathbb{P}(y \leq T \leq y+h | T \geq y)}{h} = -\frac{d}{dy} \log S_T(y),$$

by requiring that this hazard rate for T in the absence of censoring be the same as the one when censoring is present, that is, with additional conditioning on $C \geq y$ in the above formula (Fleming and Harrington, 1991, Section 1.3). As they point out, one could construct examples in which T and C are not stochastically independent as random variables, yet satisfy their notion of independence. Subsequently, they generalize this definition slightly more still, also allowing discontinuities in the distribution of T, by using martingales. The notion of independent censoring (as well as uninformative censoring) in Section 3.3 of Andersen et al. (1988) heavily relies on filtrations and counting processes and is considered too technical for this exposition, but they summarize it as a

probabilistic concept *"[...] very heuristically stating that the extra randomness, and the reduced information, caused by the censoring mechanism, should be 'orthogonal' to the (conceptual) situation without censoring"* (Andersen et al., 1988, Section 1).

Independence assumption: rationale and risks. Starting from the expression in (1), Kaplan and Meier (1958) construct their nonparametric estimator of the survival function S_T. While still highly popular in practice, all too often one takes the underlying independence assumption too lightly. However, Kaplan and Meier (1958) raise the following important caveat in their introduction: *"[...] it is usually assumed in this paper that the lifetime (age at death) is independent of the potential loss* [i.e., censoring] *time; in practice this assumption deserves careful scrutiny."* And yet, naively assuming independence remains tempting, especially in view of the identifiability issue addressed in Cox (1959) and Tsiatis (1975). Indeed, they showed that the joint model cannot be fully identifiable unless at least some formal assumptions are made. That is, solely based on observations (Y, Δ, X), it is impossible to determine the true underlying joint distribution of (T, C), as there exist infinitely many joint distributions of (T, C) that would all lead to the observed data generating process. Worse yet, even when (a part of) this joint distribution is parametrized, the problem might still persist in the sense that distinct parameters can lead to models which both imply the observed data distribution. We refer to van der Vaart (1998, p. 62) for a formal definition of model identification.

It is intuitively clear that this non-identifiability in a censoring context arises since one always observes either T or C, but never both simultaneously, such that too little information is available on their interdependence. Assuming independence therefore seems a natural remedy, but caution is advised. As Tsiatis (1975) puts it for his times Y_i to death of cause i, *"[...] the results [...] based on the assumption that $Y_1, \ldots, Y_k$ are independent, may have no resemblance to reality."* This resulted in the advent of dependence modeling within survival analysis, an early review paper of which is Moeschberger and Klein (1995).

Independence assumption: validity and counterexamples. While extending the applicability of independence models, the mere notion of *conditional* independence does not suffice to cover the full range of data encountered in practice. We will now give a few examples – and many more exist – where this conditional independence fails to hold, thus motivating the work on (conditionally) dependent censoring reviewed here.

First, consider unemployment studies. Denote T the time to re-employment and C the time to moving to a different city or region. One can think of cases where someone decides to continue their job search elsewhere, precisely because their unemployment duration gets too long. Correcting for individual information such as the individual's age, level of education, etc., will not remove this dependence. Similarly, Crommen, Beyhum and Van Keilegom (2024) analyzed data stemming from the so-called JTPA study (Job Training Partnership Act) to estimate the effect of job training on time until employment, in a dependent censoring context. Indeed, individuals may not show up for the follow-up interviews anymore (inducing censoring) precisely because they found

a job. Other examples of dependent censoring in unemployment data may be found in Han and Hausman (1990), Lo and Wilke (2010) and Blanco et al. (2020), among others.

Next, medical survival data are also often prone to dependent censoring. For example, a data set on liver transplants has been analyzed several times in this context (Collett, 2015; Deresa and Van Keilegom, 2020a; D'Haen et al., 2025). It is no surprise that the time until death while waiting for a transplant (T) is dependent on the time until transplant receipt (C): patients with an overall bad condition have a higher probability of both receiving a transplant sooner (as priority is given to more ill patients) and dying sooner. Conditioning on patient information does not remove this dependence as a patient's health condition cannot be fully captured in covariates.

Similarly, in many studies evaluating treatments for a disease, censoring by patient withdrawal from the study may be related to the patient's health condition. Drop-out might be caused by an improved (or deteriorated) health condition, causing them to feel like they no longer need (or benefit from) the treatment. A well-known example of this is the ACTG 175 (AIDS Clinical Trials Group) discussed in Huang and Zhang (2008) and Chen (2010) under the assumption of a known copula, and later in Deresa and Van Keilegom (2021) using an approach not based on copulas.

In this review article, *dependent censoring* will always refer to the absence of *conditional* independence as in the examples that were just mentioned, that is, after conditioning on any covariates present, the (conditional) margins are still stochastically dependent. As such, the methodology discussed always requires nontrivial modeling of the dependence.

1.2. Introduction to copulas

Copula models are a popular approach to incorporate dependence modeling in survival analysis, and are the core topic of this review paper. Despite its relatively recent publication, the book of Emura and Chen (2018), providing an overview of copula-based survival approaches, precedes numerous novel articles. This illustrates the considerable and ongoing research interest in this topic; we aim to update the overview of Emura and Chen (2018) by also including the latest advances. Before diving into both this recent work as well as some of the seminal papers, we will introduce the necessary preliminary material related to copulas. Useful references for copula introductions are Nelsen (2006) and Joe (2014).

Copula definition. Two-dimensional copulas are bivariate distribution functions defined on the unit square $[0,1] \times [0,1]$ that have uniform margins. Crucial to their usefulness in dependence modeling is the result by Sklar (1959) that whenever A and B are two random variables with continuous distribution functions F_A and F_B, respectively, there exists a unique copula $\mathcal{C}$ such that their joint distribution $F_{A,B}$ can be written as

$$F_{A,B}(a,b) = \mathbb{P}(A \leq a, B \leq b) = \mathcal{C}(F_A(a), F_B(b))$$

for any a, b in the respective ranges of A and B. In other words, the joint distribution can be written as a 'coupling' (hence the term 'copula') of its two marginal distributions. This result can also be extended to $d > 2$ margins $A_1, \ldots, A_d$ and d-dimensional copula functions, but in most censored survival contexts, $d = 2$ and the margins under consideration are T and C, whether or not after conditioning on a covariate. Note that copulas for discrete random variables are not identifiable, as discussed by Geenens (2020). A typical model in this review paper will be of the following form (or its covariate-dependent counterpart):

$$F_{T,C}(t,c) = \mathcal{C}\left(F_T(t), F_C(c)\right), \quad \text{or} \quad S_{T,C}(t,c) = \tilde{\mathcal{C}}\left(S_T(t), S_C(c)\right), \tag{2}$$

where the precise assumptions on the margins and on the copulas naturally depend on the specific model under consideration. The copula function $\tilde{\mathcal{C}}(u,v) = u + v - 1 + \mathcal{C}(1-u, 1-v)$ is referred to as the *survival copula* (Nelsen, 2006); it allows the translation of the joint distribution modeling framework on the left to the (equivalent) survival function setting on the right. In general, (2) allows separating the marginal distributions from the dependence structure between T and C. Clearly, the independence case is covered by $\mathcal{C}(u,v) = uv = \tilde{\mathcal{C}}(u,v)$, the independence copula.

Copula families. Two important types of copulas are the Archimedean copulas (see, for instance, Genest and MacKay (1986) and Nelsen (2006)) and the elliptical copulas (Frahm, Junker and Szimayer, 2003). In the former type, the copula $\mathcal{C}$ allows a formulation in terms of a so-called *(Archimedean) generator function* ψ. This is a continuous, convex and strictly decreasing function on $[0,1]$ such that $\psi(1) = 0$, to which the copula is related through

$$\mathcal{C}(u,v) = \psi^{[-1]}\left(\psi(u) + \psi(v)\right),$$

for any $0 \leq u, v \leq 1$. The pseudo-inverse $\psi^{[-1]}(\cdot)$ extends the regular inverse $\psi^{-1}(\cdot)$ by $\psi^{[-1]}(u) = 0$ whenever u exceeds the range of ψ; when $\psi(0) = \infty$, the pseudo-inverse coincides with the usual inverse and the generator ψ is called *strict*. Note that Archimedean copulas are always symmetric.

Next, elliptical copulas are those for which $\mathcal{C}(u,v) = F_{1,2}(F_1^{-1}(u), F_2^{-1}(v))$ for an elliptical (joint) distribution $F_{1,2}$ with marginal distributions F_1 and F_2 (Frahm et al., 2003). By far the most well-known and applied elliptical copula is the Gaussian copula

$$\mathcal{C}(u,v) = \Phi_R(\Phi^{-1}(u), \Phi^{-1}(v)), \tag{3}$$

where Φ_R denotes the two-dimensional Gaussian distribution with mean zero and correlation matrix R (equal to the covariance matrix), and Φ^{-1} is the inverse of the cumulative distribution function of a standard normal.

Parametric copula families. Many copulas belong to some parametric family of copula distributions, where the parameter value governs the strength of the association

between both margins. Thus, for instance, the family of all two-dimensional Gaussian copulas is given by $\{\mathcal{C}_\rho \mid \rho \in [-1,1]\}$, where each member $\mathcal{C}_\rho$ is a Gaussian copula (3) with correlation ρ (equal to the covariance between both variables) determining the nontrivial entries in the correlation matrix R. In the general, non-Gaussian case, we will write $\{\mathcal{C}_\xi \mid \xi \in \Xi\}$ for families of parametric copulas with parameter space Ξ for the corresponding *dependence* (or *association*) *parameter* ξ. When these families are Archimedean, their associated generator function will be denoted by ψ_ξ to indicate its dependence on the parameter ξ. Common parametric families of Archimedean copulas are the Frank family, for which

$$\mathcal{C}_\xi(u,v) = -\frac{1}{\xi}\log\left(1+\frac{(\exp(-\xi u)-1)(\exp(-\xi v)-1)}{\exp(-\xi)-1}\right), \qquad \xi \in \mathbb{R}\setminus\{0\},$$

the Gumbel family of copulas, for which

$$\mathcal{C}_\xi(u,v) = \exp\left(-\left[(-\log u)^\xi + (-\log v)^\xi\right]^{1/\xi}\right), \qquad \xi \in [1,\infty),$$

and the Clayton family, with

$$\mathcal{C}_\xi(u,v) = \left[\max\left(u^{-\xi}+v^{-\xi}-1,0\right)\right]^{-1/\xi}, \qquad \xi \in [-1,\infty)\setminus\{0\},$$

respectively (Nelsen, 2006, Table 4.1). Closed-form expressions for their corresponding generators ψ_ξ can also be found there.

Link to Kendall's tau and Spearman's rho. Finally, if A and B are two random variables with corresponding copula $\mathcal{C}$ as in Sklar's theorem, then Kendall's τ value related to these random variables can be computed by

$$\tau = 4\iint_{[0,1]^2}\mathcal{C}(u,v)c(u,v)dudv - 1 = 4\int_0^1 \frac{\psi(t)}{\psi'(t)}dt + 1,$$

where $c(u,v) = \partial^2\mathcal{C}(u,v)/\partial u\partial v$ denotes the copula density function, and the second equality holds in the case of an Archimedean copula $\mathcal{C}$ with generator ψ only (Genest and MacKay, 1986; Nelsen, 2006). Similarly, the following formula expresses the value of Spearman's ρ associated to A and B, in terms of their copula $\mathcal{C}$ (Nelsen, 2006):

$$\rho = 12\iint_{[0,1]^2}\mathcal{C}(u,v)dudv - 3.$$

For some families of parametric copulas $\mathcal{C}_\xi$, evaluating these integrals yields an easy, explicit expression for Kendall's τ or Spearman's ρ as a function of ξ or vice versa. Conversion formulas between these measures and ξ per parametric copula family may be found in Chapter 4 of Joe (2014).

1.3. Competing risks

The notion of dependent censoring is related to, yet different from, the notion of *competing risks* as in, for example, Fine and Gray (1999). Both are characterized by the presence of one or more other random variables whose occurrence precludes the observation of a survival time. However, in dependent censoring, there is a specific interest in some random variable T, whereas other possibly observation-prohibiting events are considered nuisance. By contrast, in competing risks, all these random variables are considered to be of equal importance. One may for example think of several types of cancer to which a patient may succumb, without there being one specific type of primary interest to the modelers.

Importantly, this difference in point of view also induces a different modeling approach, aligning with the *crude* vs. *net* dichotomy in Tsiatis (1975); the former corresponds to competing risks, the latter to the dependent censoring setting. Indeed, Tsiatis (1975) distinguishes between *net* survival functions $S_k(t) = \mathbb{P}(T_k > t)$ for $k = 1, \ldots, K$, and their *crude* counterparts $Q_k(t) = \mathbb{P}(T_k > t, T_k = \min_{j=1,\ldots,K}(T_j))$. The latter represent the probability of surviving beyond a certain time t *with cause of death* k; the corresponding subdistribution functions $\mathbb{P}(T_k \leq t, T_k = \min_{j=1,\ldots,K}(T_j))$ are called the *(cause-specific) cumulative incidence functions*. Note that, as opposed to proper distribution functions, these *sub*distributions tend to a value strictly smaller than one when $t \to \infty$: while eventual death to *some* cause $j = 1, \ldots, K$ is guaranteed, it is not for *a specific* death type k that may, indeed, be precluded by the occurrence of a competing ($j \neq k$) earlier death. Moreover, the quantities are *crude* in the sense that they directly correspond to the observations, whereas inference for the *net* quantities requires the hypothetical removal of all other potential risks T_j ($j \neq k$) to evaluate the behavior of the latent quantity T_k. In particular, working only on the level of the crude, observational quantities (i.e. in the competing risks setting), one does not suffer from the same identifiability issue raised in Tsiatis (1975), and hence this approach does not require imposing a dependence structure.

Unfortunately, the dependent censoring framework in which one is interested in the *net* survival quantities, but in which there are more than two random variables, is also – confusingly – referred to as *multiple competing risks*, even though in this case a dependence structure is required. All copula models in this review paper fall within this dependent censoring category; throughout, any usage of the term 'competing risks' is to be understood in the net sense. Note that this is in line with e.g. Zheng and Klein (1995) and Chen (2010), but Kalbfleisch and Prentice (2002) and Collett (2015) use 'competing risks' with its crude meaning instead.

1.4. Paper outline

The material discussed in this review paper is organized largely chronologically; most sections correspond to key steps in the development of copula-based dependence modeling in censored survival analysis. Section 2 discusses the pioneering work on the

so-called copula-graphic estimator (Zheng and Klein, 1995; Rivest and Wells, 2001) and its extensions, where the copula is assumed to be completely known. The recent work of Czado and Van Keilegom (2023), that is the core of Section 3, relaxed this heavy assumption at the cost of a fully parametric model specification. Extensions including a cure fraction and one nonparametric margin are also discussed. On the other hand, their work was also extended to a setting including covariates (Section 4), moreover gradually relaxing the parametric assumptions to semiparametric ones. This section also covers a separate line of research on parametric copulas with a nonparametric margin. Section 5 considers copula methods in machine learning, while Section 6 discusses copula models in dependent censoring with a focus on quantile regression and frailty modeling. In Section 7 we briefly touch upon some other approaches not based on copulas. For all methods considered throughout Sections 2–7, we provide references to corresponding software packages whenever available. Some concluding remarks on the limitations of the discussed literature and interesting future research directions, finally, are provided in Section 8.

2. Known copula

This section provides an overview of dependent censoring models under the assumption of a known copula for the dependence structure between the survival time T and censoring time C. In this context, a known copula implies that both the functional form and the parameters controlling the association between T and C are correctly specified. The overview begins with an examination of the copula-graphic estimator, as it was introduced by Zheng and Klein (1995). Following this, the work of Rivest and Wells (2001) is explored. Using a martingale approach, they investigated the copula-graphic estimator when the joint survival function is modeled with an Archimedean copula. Finally, various other extensions of the copula-graphic estimator are discussed.

2.1. Copula-graphic estimator

In their pioneering work, Zheng and Klein (1995) were among the first to propose modeling the dependence structure between the survival and censoring time through the use of a known copula function. Specifically, they proposed a nonparametric estimator for the marginal distribution functions of T and C, known as the copula-graphic estimator. Their approach begins by noticing that the observable data (Y,Δ) can be used to directly estimate the probabilities $\pi(y) = \mathbb{P}(Y > y)$, $I_1(y) = \mathbb{P}(Y \leq y, \Delta = 1)$ and $I_2(y) = \mathbb{P}(Y \leq y, \Delta = 0)$ with $0 \leq y < \infty$ and assuming that $\mathbb{P}(T = C) = 0$. Under the assumption of independence between T and C, these probabilities uniquely determine the marginal distribution of T. This result was generalized by Theorem 3.1 in Zheng and Klein (1995), which states that:

Theorem 3.1, Zheng and Klein (1995). *Suppose the marginal distribution functions of* (T,C) *are continuous and strictly increasing in* $(0,\infty)$. *Suppose the copula* $\mathcal{C}$, *of* (T,C),

is known, and the corresponding probability measure $\mu_{\mathcal{C}}(E) > 0$ *for any open set* E *in* $[0,1] \times [0,1]$. *Then* F_T *and* F_C*, the marginal distribution functions of* T *and* C*, are uniquely determined by* $\{\pi(y), I_1(y), I_2(y), y > 0\}$.

This theorem implies that, under the assumption of a known copula, the marginal distribution functions F_T and F_C are uniquely determined by the set of probabilities $\{\pi(y), I_1(y), I_2(y) \mid y \in \mathbb{R}_{>0}\}$, which can be estimated nonparametrically from the observable data. Using this insight, they showed that for any y, we have that

$$\mu_{\mathcal{C}}(A_y) = \mathbb{P}(T > y, C > y) = \pi(y), \quad \mu_{\mathcal{C}}(B_y) = \mathbb{P}(T \leq y, T < C) = I_1(y),$$

where

$$A_y = \{(u,v) \mid F_T(y) < u \leq 1, F_C(y) < v \leq 1\},$$
$$B_y = \{(u,v) \mid 0 \leq u \leq F_T(y), F_C(F_T^{-1}(u)) < v \leq 1\},$$

and $\mu_{\mathcal{C}}$ denotes the probability measure of the known copula $\mathcal{C}$. These relationships uniquely determine F_T and F_C, which allows for the construction of estimators $\hat{F}_T$ and $\hat{F}_C$ that preserve these equalities over a selected grid of points. Moreover, their Theorem 4.1 established that $\hat{F}_T$ and $\hat{F}_C$ are strongly consistent for F_T and F_C, respectively, under mild regularity conditions on the marginal distribution functions and copula density.

A natural choice for the grid on which the estimator is calculated is the set of distinct times $\{y_1, \ldots, y_m\}$ at which individuals experience the event of interest or are censored, with $\{\delta_1, \ldots, \delta_m\}$ the respective censoring indicators. This grid leads to $\hat{F}_T$ being a step function with jumps at the observed event times and is referred to as the *copula-graphic estimator* of F_T. Similarly, the copula-graphic estimator of F_C is a step function with jumps at the observed censoring times. Therefore, when $\delta_i = 0$ we have that $\hat{F}_T(y_i) = \hat{F}_T(y_{i-1})$ and when $\delta_i = 1$ it is clear that $\hat{F}_C(y_i) = \hat{F}_C(y_{i-1})$. Further, let $y_0 = 0$ with $\hat{F}_T(y_0) = \hat{F}_C(y_0) = 0$ and $\hat{\pi}(y) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(Y_i > y)$. Noting that $\mathbb{P}(T > y, C > y) = 1 - \mathbb{P}(T \leq y) - \mathbb{P}(C \leq y) + \mathbb{P}(T \leq y, C \leq y)$, it follows that when $\delta_i = 0$:

$$\mu_{\mathcal{C}}(\hat{A}_{y_i}) = 1 - \hat{F}_T(y_{i-1}) - \hat{F}_C(y_i) + \mathcal{C}\left(\hat{F}_T(y_{i-1}), \hat{F}_C(y_i)\right) = \hat{\pi}(y_i),$$

and similarly for $\delta_i = 1$. Since $\hat{\pi}(y_i)$ can be estimated directly from the observable data, $\hat{F}_T$ and $\hat{F}_C$ can be computed iteratively by solving these equations. Furthermore, Zheng and Klein (1995) established that the copula-graphic estimator is a maximum likelihood estimator, and that under the assumption of the independence copula, the copula-graphic estimates of marginal survival functions coincide with the Kaplan and Meier (1958) estimates.

It is to be noted that the assumption of a known copula may be overly restrictive, as the exact strength of the dependence between T and C is usually unknown in practice. However, this is exactly the assumption made when using the Kaplan-Meier estimator. For illustration, we simulated 1000 observations of $Y = \min(T, C)$ and $\Delta = \mathbb{1}(Y = T)$, where T and C are generated from a Frank copula with Kendall's tau equal to 0.5 and

standard log-normal margins. Figure 1 displays the copula-graphic estimator for various specified values of Kendall's tau and the true survival function of T. Note that the Kaplan-Meier estimator overlaps with the copula-graphic estimator when $\tau = 0$. This example demonstrates that the copula-graphic estimator is sensitive to misspecification of the association parameter, suggesting that its most evident use is to provide bounds on the survival function for varying strengths of dependence.

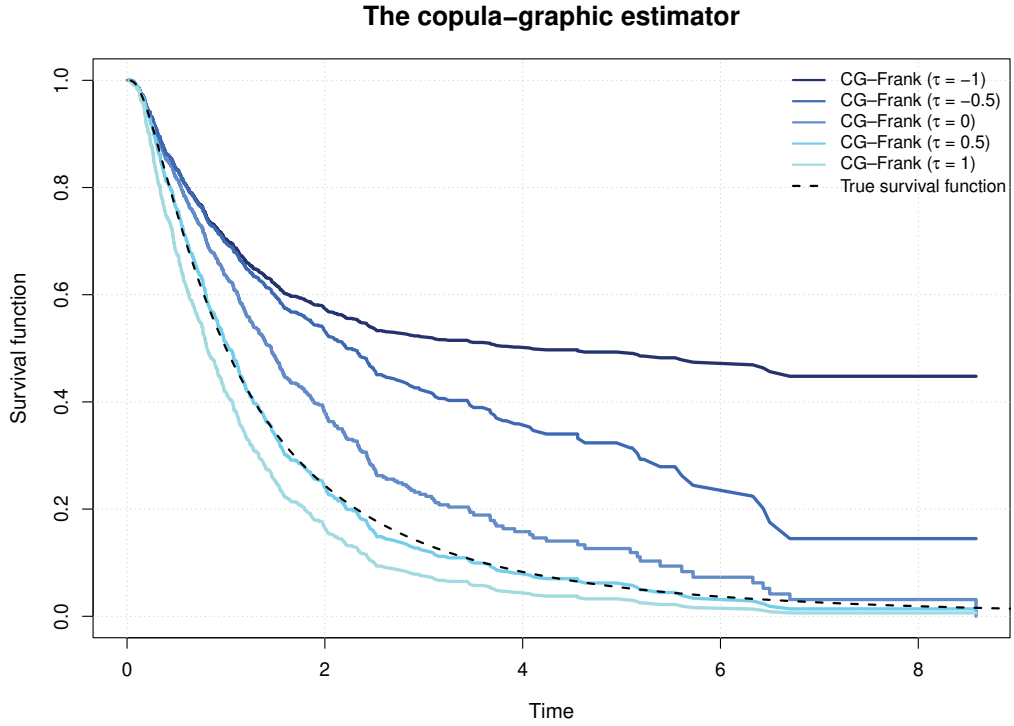


Figure 1. *The copula-graphic estimator for different values of Kendall's tau and the true survival function.*

2.2. Archimedean copulas

Subsequently, Rivest and Wells (2001) investigated the copula-graphic estimator where the dependence between T and C can be modeled by any known Archimedean copula. Note that many of the Archimedean copulas do not meet the regularity assumptions of the identifiability and consistency results given by Zheng and Klein (1995). As an example, the condition that the copula has a strictly positive density on $[0,1] \times [0,1]$ is not met by the Clayton copula. Under the assumption of a known Archimedean copula, Rivest and Wells (2001) derived the following closed-form expression for the copula-graphic estimator:

$$\hat{S}_T(t) = \psi^{-1}\left[- \sum_{Y_i \leq t, \Delta_i = 1} \left\{ \psi\left(\frac{\bar{Z}(Y_i)}{n} \right) - \psi\left(\frac{\bar{Z}(Y_i) - 1}{n} \right) \right\} \right],$$

where $\psi(\cdot)$ is a known strict generator function and $\bar{Z}(t) = \sum_{i=1}^{n} \mathbb{1}(Y_i \geq t)$. Note that when $\psi(y) = -\log(y)$, the survival and censoring time are independent and the copula-graphic estimator for Archimedean copulas reduces to the Kaplan and Meier (1958) estimator.

Using counting process notations, the copula-graphic estimator for Archimedean copulas can also be written as

$$\hat{S}_T(t) = \psi^{-1}\left[\int_0^t \mathbb{1}(\bar{Z}(u) > 0)\left\{\psi\left(\frac{\bar{Z}(u)-1}{n}\right) - \psi\left(\frac{\bar{Z}(u)}{n}\right)\right\} d\bar{N}(u)\right],$$

with $\bar{N}(t) = \sum_{i=1}^{n} \mathbb{1}(Y_i \leq t, \Delta_i = 1)$. From this representation, it can be deduced that the copula-graphic estimator for Archimedean copulas and the original copula-graphic estimator have the same asymptotic behavior. Moreover, Rivest and Wells (2001) showed in their Theorem 1 that $\hat{S}_T(t)$ is a uniformly consistent estimate of $S_T(t)$ under different regularity conditions from those given by Theorem 4.1 in Zheng and Klein (1995).

2.3. Extensions

The work of Rivest and Wells (2001) was extended to the regression setting by Braekers and Veraverbeke (2005), who applied the martingale approach to a fixed design regression model with one continuous covariate. Assuming the independence copula, they found that their estimator reduces to the Beran (1981) estimator as it was studied by Van Keilegom and Veraverbeke (1996). Building on this work, Veraverbeke (2006) considered the nonparametric estimation of conditional quantiles of T given X under the assumed copula model, while Sujica and Van Keilegom (2015) proposed estimators for location and scale functionals of T given X under a random design setting and studied their asymptotic properties. Subsequently, Sujica and Van Keilegom (2018) developed an estimator of the conditional distribution of T given X, where (T,X) satisfies a nonparametric location-scale model.

Further extensions of the copula-graphic estimator include a sensitivity analysis for the Cox (1972) proportional hazards model with respect to the association between T and C by Huang and Zhang (2008). Furthermore, Chen (2010) introduced a semiparametric transformation model using a known copula dependence structure to perform marginal regressions. Note that this approach accommodates proportional hazards and proportional odds models as special cases.

In a separate line of work, the copula-graphic estimator was generalized to multiple competing risks by Carrière (1995), who showed that the marginal survival functions remain identifiable. However, this approach, which relies on solving a system of nonlinear differential equations, can be computationally intensive as the number of competing risks increases. Under the assumption of an Archimedean copula, Lo and Wilke (2010) proposed a more practical method that pools the irrelevant risks to reduce a multiple competing risks model to a model with only two competing risks. In this way, the copula-graphic estimator can be applied directly. This work was later extended to the regression setting by Lo and Wilke (2014), allowing for an arbitrary number of risks and

covariates. By assuming an Archimedean copula, they derived closed-form solutions for the latent marginal distributions and the covariate effects on these distributions.

Additionally, Li, Tiwari and Guha (2007) expanded upon the work of Rivest and Wells (2001) by allowing for a cure fraction. Furthermore, de Uña-Álvarez and Veraverbeke (2013) extended the copula-graphic estimator for Archimedean copulas by allowing for an administrative censoring time A that censors the full process independently of T and C. This framework was further expanded by de Uña-Álvarez and Veraverbeke (2017) to accommodate left truncation.

2.4. Software

The functions `CG.Clayton`, `CG.Frank` and `CG.Gumbel` in the `compound.Cox` package (Emura et al., 2024) in R implement the known copula approach discussed in Section 2.1.

3. Unknown copula without covariates

In Section 2, the copula was assumed to be fully known. However, as previously mentioned, this assumption is quite restrictive in practice since the strength of the dependence is rarely known. As mentioned in Section 2, this limitation leads to estimators that are useful for sensitivity analyses, but are seldom effective for point estimation of the marginal distribution of T. Recently, new approaches have been proposed to address dependent censoring by using a copula to characterize dependence without needing to explicitly specify the association parameter. In this section, we will review the relevant literature, highlighting models that adopt the form given in (2) as per Sklar's theorem.

3.1. Fully parametric model

To relax the conditions on the copula, Czado and Van Keilegom (2023) investigated a scenario for dependent censoring where the copula is parametric, but the parameter describing the association between T and C is unknown. Specifically, they assumed that the copula $\mathcal{C}$ can be modeled as:

$$\mathcal{C} \in \{\mathcal{C}_\xi \mid \xi \in \Xi\}, \tag{4}$$

for some parameter space Ξ. Additionally, the marginal distributions of T and C belong to parametric families:

$$F_T \in \{F_{T,\theta_T} \mid \theta_T \in \Theta_T\} \quad \text{and} \quad F_C \in \{F_{C,\theta_C} \mid \theta_C \in \Theta_C\}, \tag{5}$$

for certain parameter spaces Θ_T and Θ_C. Under mild conditions given below, which can be verified for most common parametric families of copulas and marginal distributions, identifiability can be demonstrated. The work by Czado and Van Keilegom (2023) complements other studies focusing on unknown copula parameters (Schwarz, Jongbloed

and Van Keilegom, 2013; Emura and Chen, 2016; Fan, Wang and Ju, 2019; Shih et al., 2019) and shows that a completely known copula is not always necessary for identifying the joint distribution of the survival and censoring times. There is, however, no free lunch, and the trade-off for this flexibility is that both margins must be parametric.

To discuss the assumptions ensuring identifiability in more detail, let us define the partial derivatives with respect to the copula function $\mathcal{C}$ as follows:

$$h_{1,\xi}(u,v) = \frac{\partial}{\partial u}\mathcal{C}_\xi(u,v) \quad \text{and} \quad h_{2,\xi}(u,v) = \frac{\partial}{\partial v}\mathcal{C}_\xi(u,v). \tag{6}$$

Furthermore, denote $f_{T,\theta_T}(t) = (d/dt)F_{T,\theta_T}(t)$ and $f_{C,\theta_C}(t) = (d/dt)F_{C,\theta_C}(t)$. Identifiability means that two distinct parameter vectors $\theta_j = (\xi_j, \theta_{Tj}, \theta_{Cj}) \in \Xi \times \Theta_T \times \Theta_C$ $(j = 1,2)$ lead to two distinct distributions of (Y,Δ). Czado and Van Keilegom (2023) imposed the following two general conditions:

Condition 1. Suppose that we have the following four equivalences:

$$\lim_{t\to 0}\frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2}, \quad \lim_{t\to\infty}\frac{f_{T,\theta_{T1}}(t)}{f_{T,\theta_{T2}}(t)} = 1 \iff \theta_{T1} = \theta_{T2},$$
$$\lim_{t\to 0}\frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2}, \quad \lim_{t\to\infty}\frac{f_{C,\theta_{C1}}(t)}{f_{C,\theta_{C2}}(t)} = 1 \iff \theta_{C1} = \theta_{C2},$$

for $\theta_{T1}, \theta_{T2} \in \Theta_T$ and $\theta_{C1}, \theta_{C2} \in \Theta_C$.

Condition 2. One of the following two properties holds:

$$\lim_{t\to 0} h_{1,\xi}(F_{T,\theta_T}(t), F_{C,\theta_C}(t)) = 0 \quad \text{for all} \quad (\xi, \theta_T, \theta_C) \in \Xi \times \Theta_T \times \Theta_C,$$
$$\lim_{t\to\infty} h_{1,\xi}(F_{T,\theta_T}(t), F_{C,\theta_C}(t)) = 0 \quad \text{for all} \quad (\xi, \theta_T, \theta_C) \in \Xi \times \Theta_T \times \Theta_C.$$

Similarly, for $h_{2,\xi}(F_{T,\theta_T}(t), F_{C,\theta_C}(t))$, at least one of these limit statements holds.

Condition 1 refers only to the margins and is satisfied for a wide range of parametric families of densities, such as Weibull, log-normal, log-logistic and log-Student-t densities. It is also worth noting that in Condition 2, only one of the two limits needs to be zero. This condition can be demonstrated for popular classes of copulas under mild additional conditions, including most Archimedean copulas and Gaussian copulas. However, the Clayton copula requires a different set of assumptions, as identifiability cannot be proved the same way using Conditions 1 and 2 as for the other common copulas. For more details, we refer to the paper by Czado and Van Keilegom (2023).

Czado and Van Keilegom (2023) also considered the parameter estimation of the above joint parametric model. Recall that $\mathcal{D}_n$ is a random sample from the population (Y,Δ) of size n. The joint likelihood for the parameter vector $\boldsymbol{\theta} = (\xi, \theta_T, \theta_C)^\top$ is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\Big[\left\{f_{T,\theta_T}(Y_i)[1 - h_{1,\xi}(F_{T,\theta_T}(Y_i), F_{C,\theta_C}(Y_i))]\right\}^{\Delta_i} \times \left\{f_{C,\theta_C}(Y_i)[1 - h_{2,\xi}(F_{T,\theta_T}(Y_i), F_{C,\theta_C}(Y_i))]\right\}^{1-\Delta_i}\Big]. \tag{7}$$

A maximum likelihood approach is adopted based on the likelihood specified in (7), resulting in parameter estimators formally defined by:

$$\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\theta}_T, \hat{\theta}_C)^\top = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ \log L(\boldsymbol{\theta}),$$

where $\Theta = \Xi \times \Theta_T \times \Theta_C$. Based on the results of White (1982), the estimator $\hat{\boldsymbol{\theta}}$ can be shown to be consistent and asymptotically normal under certain sufficient conditions.

3.2. Inclusion of a cure fraction

In the previous modeling framework, it was assumed that all individuals in the population are susceptible to the event of interest, meaning that each individual has either already experienced the event or will experience it in the future. However, in practical applications, some individuals may never experience the event and can be considered cured (Amico and Van Keilegom, 2018), meaning that these subjects have infinite survival times ($T = \infty$).

To account for this feature, the mixture cure model can be adopted (Peng and Yu, 2021; Parsa and Van Keilegom, 2023), which expresses $S_T(t)$ or $F_T(t)$ as:

$$S_T(t) = 1 - p + pS_U(t) \quad \Longleftrightarrow \quad F_T(t) = pF_U(t), \tag{8}$$

where $S_U(t) = 1 - F_U(t) = \mathbb{P}(T > t | T < \infty)$ is the proper survival function of the susceptibles, called the *latency*, and $p = \mathbb{P}(T < \infty)$ is the *incidence*. The issue of dependent censoring in the presence of a cure fraction has been discussed by Delhelle and Van Keilegom (2025). They considered a fully parametric model for the bivariate distribution of (T,C) as in equations (2), (4) and (5), except that F_T is now modeled based on (8). They further assumed that F_U belongs to a parametric family of distributions:

$$F_U \in \{F_{U,\theta_U} \mid \theta_U \in \Theta_U\},$$

for a certain parameter space Θ_U. We remark that the support of U is assumed to be bounded, with details provided in the next paragraph.

Proving the identifiability of a model with both dependent censoring and a cure fraction is not an easy task. Delhelle and Van Keilegom (2025) presented sufficient conditions under which such a model is identifiable. Adaptations of both Conditions 1 and 2 are needed. Specifically, they imposed a condition similar to Condition 1 on the density of U instead of T, considering only the limit as t tends to zero since a right-truncated distribution is used for U. They also developed a more specialized key assumption:

Condition 3. If $h_{2,\xi}(p, F_{C,\theta_C}(y)) = h_{2,\tilde{\xi}}(\tilde{p}, F_{C,\theta_C}(y))$ for all $y > \tau_U$, then $p = \tilde{p}$ and $\xi = \tilde{\xi}$, where $\tau_U = \inf\{y \mid F_{U,\theta_U}(y) = 1\}$ is supposed to be finite.

This condition is essential for identifying p and can be shown to be satisfied for many combinations of copula families and censoring margins. For example, as demonstrated

in Delhelle and Van Keilegom (2025), this condition is satisfied by the Frank and Joe copulas, as well as the Clayton copula with 90, 180 or 270 degrees of rotations, regardless of the marginal distributions. For both Gumbel and Gaussian copulas, this condition also holds provided that the marginal distributions satisfy some additional mild regularity conditions listed in Delhelle and Van Keilegom (2025).

Following Czado and Van Keilegom (2023) as introduced in Section 3.1, a maximum likelihood-based estimation procedure for $\boldsymbol{\theta} = (p, \xi, \theta_U, \theta_C)^\top$ was proposed by Delhelle and Van Keilegom (2025) as well. Instead of (7), the joint likelihood becomes:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \Big[\left\{ p f_{U,\theta_U}(Y_i)[1 - h_{1,\xi}(pF_{U,\theta_U}(Y_i), F_{C,\theta_C}(Y_i))] \right\}^{\Delta_i} \times \left\{ f_{C,\theta_C}(Y_i)[1 - h_{2,\xi}(pF_{U,\theta_U}(Y_i), F_{C,\theta_C}(Y_i))] \right\}^{1-\Delta_i} \Big], \tag{9}$$

where $f_{U,\theta_U}(t) = (d/dt)F_{U,\theta_U}(t)$. Then, an estimator of $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood specified in (9) as follows:

$$\hat{\boldsymbol{\theta}} = (\hat{p}, \hat{\xi}, \hat{\theta}_U, \hat{\theta}_C)^\top = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max}\ \log L(\boldsymbol{\theta}),$$

where $\Theta = (0,1) \times \Xi \times \Theta_U \times \Theta_C$.

3.3. One nonparametric margin

Rather than imposing full parametric assumptions on both marginal distributions of T and C, it could be preferable to work under more relaxed constraints. However, maintaining identifiable copula structures generally requires parametric specifications for at least one margin, particularly with Archimedean copulas where non-identifiability arises when both margins are unspecified alongside unknown association parameters (Wang, 2012, 2014). Wang (2023) resolved the identifiability challenge by focusing exclusively on Archimedean copulas and imposing an exponential distribution on S_C, while the other margin is allowed to remain either parametric or fully nonparametric.

Using a different approach, Ding and Van Keilegom (2025) considered a semiparametric approach, modeling the marginal distribution of the survival time T nonparametrically while keeping the censoring time C parametric, with its parametric specification extending beyond the exponential family. In other words, while equations (2), (4) and (5) are imposed, F_T in (5) is now nonparametric. To facilitate the estimation procedure that will be described below, they modeled the joint survival function of T and C, instead of the joint distribution function.

Under certain conditions, they demonstrated that the joint model is identifiable, representing a significant advancement compared to Czado and Van Keilegom (2023) by relaxing the parametric assumption on one of the margins. Specifically, the underlying idea is to first identify the censoring parameter θ_C by imposing conditions similar to Conditions 1 and 2 above. Then, two additional conditions are introduced to identify the

association parameter of the copula ξ. Finally, the identifiability of S_T follows automatically. Although the marginal distribution of the censoring time remains parametric, this trade-off is often acceptable in practice because of the difficulty in determining the association parameter and the marginal distribution of the survival time, which are essential quantities in survival analysis.

We proceed to discuss the methodology for estimating the unknown parameters (ξ, θ_C, S_T) based on the pseudo-likelihood technique and the copula-graphic estimator from Rivest and Wells (2001) under the assumption that the copula is Archimedean. Specifically, the likelihood function for a single observation (y, δ) of (Y, Δ) is given by:

$$L(y,\delta|\xi,\theta_C,S_T) = \left\{\tilde{h}_{1,\xi}(S_T(y), S_{C,\theta_C}(y)) f_T(y)\right\}^{\delta} \times \left\{\tilde{h}_{2,\xi}(S_T(y), S_{C,\theta_C}(y)) f_{C,\theta_C}(y)\right\}^{1-\delta}, \tag{10}$$

in which both $\tilde{h}_1$ and $\tilde{h}_2$ are defined similarly to h_1 and h_2 with $\mathcal{C}$ replaced by $\tilde{\mathcal{C}}$, as defined in (2). Maximizing $\prod_{i=1}^n L(Y_i, \Delta_i|\xi, \theta_C, S_T)$ directly is infeasible and, therefore, a two-step solution is proposed. The key idea is to replace S_T with the copula-graphic estimator $\hat{S}_{T,\xi}$ (Rivest and Wells, 2001) for a fixed value of the parameter ξ, leading to a function that relies only on the parameters ξ and θ_C.

More specifically, the estimators of the copula's association parameter and censoring time parameter are calculated via:

$$(\hat{\xi}, \hat{\theta}_C) = \underset{(\xi,\theta_C)\in\Xi\times\Theta_C}{\arg\min} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\alpha}} \log L(Y_i, \Delta_i|\alpha, \hat{S}_{T,\xi}) \right\|,$$

where $\boldsymbol{\alpha} = (\xi, \theta_C)^\top$ and $\|\cdot\|$ is the Euclidean norm. The estimator of the marginal survival function S_T can then be formally defined by $\hat{S}_T(t) = \hat{S}_{T,\hat{\xi}}(t)$. The explicit expression of $\hat{S}_{T,\xi}(t)$ under Archimedean copulas ensures computationally efficient estimation of ξ, $\boldsymbol{\theta}_C$ and $S_T(t)$. Notably, when using the independence copula, the estimator $\hat{S}_T(t)$ coincides with the classical Kaplan-Meier estimator (Kaplan and Meier, 1958). Although this estimation procedure can be extended to other copulas like Gaussian copulas, the absence of closed-form solutions for copula-graphic estimators in such cases introduces significant computational overhead, thereby constraining practical applicability relative to Archimedean copulas. While the methodology for parameter estimation capitalizes on the computational tractability of Archimedean copulas, this choice entails certain limitations. Among them, what is most worth mentioning is that the restriction to this copula family precludes exploration of additional dependence patterns achievable with other copula families, such as elliptical copulas or vine copulas. Consequently, developing computationally tractable estimation frameworks for non-Archimedean copula structures represents a crucial avenue for future research. Also note that the above estimation procedure can be applied in the reverse scenario as well, where F_C is non-parametric and F_T is parametric.

As discussed in Section 3.2, it is possible that a subgroup of individuals is not susceptible at all and can be considered cured in practical applications. To account for this

characteristic, we can consider the mixture cure model and express the survival function $S_T(t)$ as shown in (8). Instead of modeling F_U parametrically, Ding and Van Keilegom (2025) considered modeling it nonparametrically. An assumption can be imposed that ensures sufficient information in the right tail (Maller et al., 2024) to guarantee the identifiability of the model. By first treating S_T as a whole, the estimator $\hat{S}_T$ can be obtained. Next, define:

$$\hat{p} = 1 - \hat{S}_T(Y^1_{\max}) \quad \text{and} \quad \hat{S}_U(t) = \frac{\hat{S}_T(t) - 1 + \hat{p}}{\hat{p}},$$

where $Y^1_{\max} = \max\{Y_i \mid \Delta_i = 1, i = 1, \ldots, n\}$. The estimators $\hat{p}$ and $\hat{S}_U(t)$ can then be used to estimate p and $S_U(t)$, respectively.

3.4. Software

The R package `depCensoring` implements the model introduced in Section 3.1 through the `ParamCop` function, while the cure model discussed in Section 3.2 can be implemented using the `SurvDC` function. The approach with one nonparametric margin, as presented in Section 3.3, is also available via the `SurvDC` function in the same package.

4. Unknown copula with covariates

This section reviews existing unknown copula approaches for analyzing survival data in the presence of covariates, particularly in scenarios where the data are subject to both dependent censoring and administrative censoring. In this setting, we have less information about the dependence between T and C compared to the case where there is no administrative censoring as in the previous section. Recall that we observe $Y = \min(T, C, A)$ and the censoring indicators (Δ_T, Δ_C) given by $\Delta_T = \mathbb{1}(Y = T)$ and $\Delta_C = \mathbb{1}(Y = C)$. Let X (of dimension p) denote the vector of covariates. It is assumed that (T, C) and A are independent given X, and that A and X are independent.

4.1. Fully parametric model

By building further on the results in Czado and Van Keilegom (2023), Deresa, Van Keilegom and Antonio (2022) suggested a fully parametric copula model for survival data with dependent censoring and left truncation. If Z denotes the left truncation time, we observe individual information only when $Z \leq Y$. They used a bivariate copula to model the relation between T and C for given X:

$$\mathbb{P}(T \leq t, C \leq c | X = x) = \mathcal{C}_\xi \left(F_{T|X}(t|x), F_{C|X}(c|x) \right), \tag{11}$$

where $F_{T|X}(t|x) = \mathbb{P}(T \leq t | X = x)$ and $F_{C|X}(c|x) = \mathbb{P}(C \leq c | X = x)$ are the conditional distributions of T and C, respectively. Note that the copula $\mathcal{C}_\xi$ is specified up to a

finite-dimensional parameter ξ as in (4) (and hence does not depend on covariates), the distributions $F_{T|X}$ and $F_{C|X}$ belong to parametric families depending on the covariate X:

$$F_{T|X} \in \left\{F_{T|X,\theta_T} \mid \theta_T \in \Theta_T\right\} \quad \text{and} \quad F_{C|X} \in \left\{F_{C|X,\theta_C} \mid \theta_C \in \Theta_C\right\}, \tag{12}$$

whereas the distributions of Z and A are left completely unspecified. Assuming that Z is independent of all other variables in the model, Deresa et al. (2022) demonstrated the identifiability of the dependence parameter and all other model parameters using the distribution of the observed data $(Y, \Delta_T, \Delta_C, X, Z)$ with $Y \geq Z$. This was achieved by adopting general high-level conditions on the marginal distributions and the copula structure similar to Conditions 1 and 2 of Section 3.1. Deresa et al. (2022) then developed the likelihood under the above model, which is similar to the one in (7), except that it needs to be adjusted for the presence of covariates and left truncation. This likelihood can be maximized numerically over the parameter space $\Theta = \Xi \times \Theta_T \times \Theta_C$, yielding the maximum likelihood estimator of θ: $\hat{\theta} = (\hat{\xi}, \hat{\theta}_T, \hat{\theta}_C)^\top = \arg\max_{\theta \in \Theta} L(\theta)$.

Finally, it is worth highlighting that the model proposed in Deresa et al. (2022) has significant applications in actuarial science, particularly in the modeling of lifetime data collected on portfolios of joint life annuities and joint life insurance policies issued to coupled lives. For joint life annuities, benefits are paid until the death of the first of two annuitants, while joint life insurance pays a death benefit upon the first death. Since couples are subject to similar risks, their lifetimes are related, which induces dependent censoring. Additionally, policyholders who died before the start of the study are excluded from the data, resulting in left truncation.

4.2. Semiparametric model

In a subsequent paper, Deresa and Van Keilegom (2024a) relaxed the parametric specification regarding the marginal distribution of T and proposed a semiparametric copula approach for dependent censoring. Let X_1 represent the covariates associated with T, and let X_2 represent the covariates that influence C. It is assumed that the vector X_1 contains at least one continuous variable that is not contained in X_2, and similarly, the vector X_2 contains at least one continuous variable that is not included in X_1. We consider the joint model given in (11) depending on a parametric model for the distribution of C, except that now the marginal model for T is given by the Cox proportional hazards model:

$$F_{T|X_1}(t|x_1) = \mathbb{P}(T \leq t | X_1 = x_1) = 1 - \exp\{-\Lambda(t) e^{x_1^\top \beta}\}, \tag{13}$$

where β is a vector of regression coefficients and Λ is an unknown increasing and differentiable cumulative baseline hazard function of unspecified form with $\Lambda(0) = 0$. The corresponding baseline hazard function is denoted by λ, defined as $\lambda(y) = d\Lambda(y)/dy$. In addition to dependent censoring, this model allows for administrative censoring.

Compared to the identifiability proof under a fully parametric copula model in Czado and Van Keilegom (2023) and Deresa et al. (2022), the nonparametric function Λ in (13)

introduces an additional challenge in establishing the identifiability of the semiparametric copula model. To circumvent this issue, Deresa and Van Keilegom (2024a) imposed the following additional assumption on the dependence structure:

Condition 4. For all θ_C, $\xi_k, \zeta_k = (\beta, \Lambda_k)$ $(k = 1,2)$ for which $\lim_{t\to 0} \lambda_1(t)/\lambda_2(t) = 1$,

$$\lim_{t\to 0} \frac{c_{\xi_1}(F_{T|X_1,\zeta_1}(t|x_1), F_{C|X_2,\theta_C}(t|x_2))}{c_{\xi_2}(F_{T|X_1,\zeta_2}(t|x_1), F_{C|X_2,\theta_C}(t|x_2))} = 1 \text{ for all } (x_1,x_2) \iff \xi_1 = \xi_2,$$

where c_ξ denotes the copula density.

This condition is satisfied for the families of Frank, Gumbel and Gaussian copulas (see Lemma 3.1 in Deresa and Van Keilegom (2024a) for a formal proof). Under Condition 4 and under conditions somewhat similar to Conditions 1 and 2 (but considering only limits when t tends to zero), the latter paper proved the identifiability of the semiparametric copula model. More recently, Crommen, Beyhum and Van Keilegom (2025) showed that the semiparametric copula model remains identified after removing the condition that the vector X_1 contains at least one continuous variable that is not contained in X_2 and the vector X_2 contains at least one continuous variable that is not included in X_1. This is done by replacing c_ξ with $h_{2,\xi}$ in *Condition 4* and adding an extra condition on the copula density. It is shown that these new conditions are also satisfied for the families of Frank, Gumbel and Gaussian copulas, thereby relaxing the required conditions for identifiability.

Assume that the data consist of n i.i.d. replications $\{(Y_i, \Delta_{T,i}, \Delta_{C,i}, X_{1i}, X_{2i})\}_{i=1,\dots,n}$ of $(Y, \Delta_T, \Delta_C, X_1, X_2)$, and let $\theta = (\xi, \beta, \theta_C)^\top$. Then, the joint likelihood function under dependent censoring is given by

$$\begin{aligned} L(\theta,\Lambda) = \prod_{i=1}^{n} &\Big[\lambda(Y_i) e^{X_{1i}^\top \beta} \exp\{-\Lambda(Y_i) e^{X_{1i}^\top \beta}\} \big\{1 - h_{1,\xi}(F_{T|X_1,\zeta}(Y_i|X_{1i}), F_{C|X_2,\theta_C}(Y_i|X_{2i}))\big\}\Big]^{\Delta_{T,i}} \\ &\times \Big[f_{C|X_2,\theta_C}(Y_i|X_{2i}) \big\{1 - h_{2,\xi}(F_{T|X_1,\zeta}(Y_i|X_{1i}), F_{C|X_2,\theta_C}(Y_i|X_{2i}))\big\}\Big]^{\Delta_{C,i}} \\ &\times \Big[\bar{\mathcal{C}}_\xi(F_{T|X_1,\zeta}(Y_i|X_{1i}), F_{C|X_2,\theta_C}(Y_i|X_{2i}))\Big]^{(1-\Delta_{T,i})(1-\Delta_{C,i})}, \end{aligned} \tag{14}$$

where $\bar{\mathcal{C}}_\xi(u,v) = 1 - u - v + \mathcal{C}_\xi(u,v)$. The presence of the nonparametric function in (13) not only complicates the identification of the joint model but also complicates its estimation. To address this issue, Deresa and Van Keilegom (2024a) derived the nonparametric estimator of Λ based on martingale ideas, assuming a fixed value for θ. Details can be found in the aforementioned paper, where it is shown that the estimator, denoted by $\hat{\Lambda}(\cdot,\theta)$, has the following explicit formula when computed sequentially over the ordered observed survival times $y_1, \dots, y_m$:

$$\Delta\hat{\Lambda}(y_k,\theta) = \frac{\sum_{i=1}^{n} \mathbb{1}(Y_i \le y_k, \Delta_{T,i} = 1)}{\sum_{i=1}^{n} \mathbb{1}(Y_i \ge y_k) \exp\{\omega_i(y_{k-1}, \theta, \hat{\Lambda})\}}, \tag{15}$$

with $\hat{\Lambda}(y_1-,\theta) = 0$, $\Delta\hat{\Lambda}(y_k,\theta) = \hat{\Lambda}(y_k,\theta) - \hat{\Lambda}(y_{k-1},\theta)$, and $\omega_i(y,\theta,\Lambda) = X_{1i}^\top\beta - \log(\bar{\mathcal{C}}_\xi(F_{T|X_1,\zeta}(y|X_{1i}),F_{C|X_2,\theta_C}(y|X_{2i}))) + \log\big(1-h_{1,\xi}(F_{T|X_1,\zeta}(y|X_{1i}),F_{C|X_2,\theta_C}(y|X_{2i}))\big) - \Lambda(y)\exp(X_{1i}^\top\beta)$, where $\zeta = (\beta,\Lambda)$. Deresa and Van Keilegom (2024a) suggested the following estimation algorithm that allows iterative estimation of θ and Λ.

- *Step 0:* Choose an initial value of θ.
- *Step 1*: Given that $\hat{\Lambda}(y_1-,\theta)=0$ and using (15), we obtain $\hat{\Lambda}(y_k,\theta)$, $k=1,\ldots,m$, one-by-one by solving the equation:

$$\hat{\Lambda}(y_k) = \hat{\Lambda}(y_{k-1}) + \frac{\sum_{i=1}^n \mathbb{1}(Y_i \le y_k, \Delta_{T,i}=1)}{\sum_{i=1}^n \mathbb{1}(Y_i \ge y_k)\exp\{\omega_i(y_{k-1},\theta,\hat{\Lambda})\}}.$$

- *Step 2*: Next, estimate θ by solving the score equation derived from the pseudo-likelihood function $L(\theta,\Lambda)$ in (14) with Λ replaced by $\hat{\Lambda}$ computed in Step 1.
- *Step 3*: Repeat Steps 1 and 2 until predefined convergence criteria are met.

Note that when the independence copula $\mathcal{C}_\xi(u,v) = uv$ is assumed, the nonparametric estimator of Λ in (15) reduces to the standard Breslow estimator of the cumulative hazard function in the Cox model (Breslow, 1974). Hence, the proposed estimation method extends the usual partial likelihood estimator to the case of dependent censoring. Deresa and Van Keilegom (2024a) established the uniform consistency of $\hat{\Lambda}(\cdot,\theta)$ (using similar arguments as in Zucker (2005)), and they also showed the consistency and asymptotic normality of $\hat{\theta}$ using the results in Chen, Linton and Van Keilegom (2003).

Deresa and Van Keilegom (2024b) extended the above copula-based Cox model to the case where both T and C follow a semiparametric model. Their method is based on a semiparametric transformation model (Zeng and Lin, 2007) for the marginal distributions of both T and C, which includes the Cox proportional hazards model used in Deresa and Van Keilegom (2024a) as a special case, while the dependence structure is modeled using a parametric Archimedean copula family. In addition, they addressed the technical challenges posed by the presence of two nonparametric functions by utilizing the bivariate survival function instead of the bivariate distribution. Specifically, the joint survival function of T and C given (X_1,X_2) is expressed as follows:

$$\mathbb{P}(T>t,C>c|X_1=x_1,X_2=x_2) = \psi_\xi^{-1}\Big[\psi_\xi\big\{S_{T|X_1}(t|x_1)\big\} + \psi_\xi\big\{S_{C|X_2}(c|x_2)\big\}\Big], \quad (16)$$

where $S_{T|X_1}(t|x_1) = \mathbb{P}(T>t|X_1=x_1)$ and $S_{C|X_2}(c|x_2) = \mathbb{P}(C>c|X_2=x_2)$, and ψ_ξ is a strict generator function. Compared to the Cox model presented above, the presence of two nonparametric functions (in the models for T and C) leads to additional challenges in establishing identifiability. We refer to Deresa and Van Keilegom (2024b) for more details about the conditions under which identifiability of the model can be shown, as well as how the model can be estimated. Finally, the latter paper also proved the asymptotic properties of the proposed estimators, which are based on a system of nonhomogeneous linear differential equations, given that the two nonparametric functions are entangled in a complex way.

4.3. Archimedean copulas with one nonparametric margin

A similar line of research focuses on parametric Archimedean copulas to model the dependence between T and C, and studies models of the form (16) in which at least one of the marginal distributions is specified fully nonparametrically.

In an earlier article, Lo and Wilke (2017) considered a more general model of the form (11) and were able to identify the sign of the covariate effect on the marginal distributions of T and C, both of which are unspecified. The restriction to Archimedean copulas, leading to model (16), is only needed in the case of multiple competing risks. The key idea used in their approach originated from the Hicksian decomposition in econometrics, which analyzes the change in demand for a good as a result of a change in its price. Lo and Wilke (2017) applied this idea to their setting by decomposing the covariate effect, and were able to link the elements in this decomposition to estimable quantities. Due to the upheld generality of the model, the identification of the sign only holds for time points belonging to a defined subset $\mathbb{G} \subsetneq \mathbb{R}_{\geq 0}$. Three different approaches to increase the size of $\mathbb{G}$ are discussed.

In subsequent work, the same authors proposed a parametric model for T alongside the Archimedean copula assumption (16) while leaving the distribution of C completely unspecified (Lo and Wilke, 2023). Their work extends the recent result in Wang (2023) regarding the identification of Archimedean copula models. Identification of the model for T proceeds as follows. First, model (16) can be solved for $S_{T|X}(t|x)$ in closed form as

$$S_\xi(t|x) = \psi_\xi \left[- \int_0^t (\psi_\xi^{-1})'[\pi(u|x)] f_t(u|x) du \right], \tag{17}$$

where $\pi(u|x)$ denotes the overall conditional survival function, i.e. $\pi(u|x) = \mathbb{P}(T \geq u, C \geq u | X = x)$ and $f_t(u|x)$ denotes the subdensity function, defined as the derivative of the cumulative incidence function of T. As already noted in Section 2.1, both π and f_t are identified and estimable based on the observed data, implying identifiability of $S_\xi(t|x)$ up to the association parameter ξ. Next, another model for T is imposed, namely $S(t|x) = \exp(-\Lambda_\beta(t|x))$, where the cumulative hazard function Λ_β is specified up to the finite-dimensional parameter vector β. By imposing equality of $S_\xi(t|x)$ and $S(t|x)$, the authors were able to identify and estimate the dependence parameter ξ as well as the regression coefficients β. In a follow-up paper, Wilke and Lo (2025) developed several useful extensions of this model, including a generalization to the setting with additional administrative censoring, and a goodness-of-fit test which relies on the Cramér–Von-Mises distance between $S_\xi(t|x)$ and $S(t|x)$.

Lastly, Hiabu, Lo, and Wilke (2025) studied ξ while leaving the marginals of both T and C unspecified. Similar to Deresa and Van Keilegom (2024b), their approach uses the assumed existence of two special covariates which only influence the distribution of T (C), while not affecting the distribution of C (T). The key insight in this setting is that variation in Y due to variation in one of these special covariates is entirely due to variation in the latent time corresponding to the special covariate under consideration. This property is then exploited to identify ξ.

4.4. Software

The approach of Deresa and Van Keilegom (2024a) is implemented in the `depCensoring` package in R, in the function `fitDepCens`. Siegfried et al. (2024) describe an alternative estimation framework for this approach, but additionally use Bernstein polynomials to flexibly model the log-cumulative baseline hazard instead of leaving it non-parametric. An implementation is available in the R package `tram` (Hothorn et al., 2024).

5. Machine learning methods

In recent years, methods such as neural networks, support vector machines and random forests have been adapted to accommodate time-to-event data. For a comprehensive review of these topics, see Bou-Hamad, Larocque and Ben-Ameur (2011) and Wang, Li and Reddy (2019). However, most of these approaches operate under the (conditional) independent censoring assumption.

One of the first machine learning approaches addressing dependent censoring was proposed by Moradian, Larocque and Bellavance (2019) who adapted existing methods for constructing survival forests to allow for dependent censoring. Firstly, the authors propose to modify how the trees are aggregated. In this approach, the forest can be constructed using established methods that rely on the independent censoring assumption. Subsequently, a weighted version of the copula-graphic estimator is used to estimate the survival function, with the weights inferred from the random forest. Therefore, a correction for dependent censoring is made only when the information in the trees is combined. The second approach involves modifying the tree-splitting rule. Most survival forests employ a splitting rule that is based on the log-rank statistic, which exhibits a significant loss of power when the proportional hazards assumption is violated. To overcome this limitation, Moradian, Larocque and Bellavance (2017) introduced a splitting rule that is based on the Kaplan-Meier estimator. This rule can easily be adapted to account for dependent censoring by substituting the Kaplan-Meier estimator with the copula-graphic estimator. However, adjusting the splitting rule requires specifying an appropriate copula function and association parameter during tree construction. As a result, conducting a sensitivity analysis by varying the association parameter will require generating distinct forests, which is computationally demanding. To resolve this issue, Moradian et al. (2019) also introduced a novel survival forest approach, called *p-forest*. Their method differs from standard survival forests because it constructs a series of classification forests at multiple time points. For each predetermined time t, the method redefines the outcome by classifying each subject as alive, dead or censored. This transformation enables the use of standard classification techniques without relying on the independent censoring assumption and eliminates the need to specify the association parameter during tree construction.

More recently, Midtfjord, De Bin and Huseby (2022) introduced a gradient-boosting approach called *Clayton-boost*. They assumed a generalized Accelerated Failure Time

(AFT) model for both the survival and censoring times, that is $\log(T) = h_T(X) + \sigma_T \varepsilon_T$, with a similar formulation for the censoring time. A Clayton copula accounts for the dependence between the survival and censoring distributions. The primary objective in this supervised learning framework is to estimate the function $h_T(x)$, for a given vector of covariates x, such that it provides a good estimate, in terms of a specified loss function, for the response $\log(T)$. An appropriate loss function that works for dependently censored data is derived by looking at the log-likelihood. To implement the approach, the authors propose to use the eXtreme Gradient Boosting algorithm (Chen and Guestrin, 2016), with statistical trees as the base learner. However, it is to be noted that their approach can be integrated into any statistical and machine learning frameworks that support custom loss functions. A downside of the method is that it requires specifying the association parameter and baseline distributions (i.e. the distributions of ε_T and ε_C) along with their standard deviations, which are rarely known in practice. Alternatively, they propose to use cross-validation to select the distributions and the value of the association parameter. In a similar line of work, Gharari et al. (2023) introduced a deep learning-based approach that assumes a generalized Weibull Cox proportional hazards model for the survival and censoring time. The dependence is modeled using a known Clayton or Frank copula.

So far, all the machine learning methods discussed rely on the assumption of a known copula to model the dependence between the survival and censoring distributions. However, correctly specifying the association parameter is difficult in practice and misspecification can lead to biased estimates. To address these limitations, Zhang, Ling and Zhang (2023) proposed a more flexible deep learning-based approach that allows for dependent censoring without having to specify even the parametric form of the copula. Identification of the model (including the association parameter) is shown under two high-level conditions similar to Condition 1 and 2 (cf. Section 3). These conditions are satisfied when the marginal distributions belong to the Weibull, log-normal, log-logistic or log-Student families and the dependence structure is modeled by a subset of Archimedean copulas. This subset consists of Archimedean copulas whose generator function can be represented by a neural network as described by Ling, Fang and Kolter (2020). To represent the conditional marginal distributions, their approach leverages a monotonic *neural density estimator* (NDE). While the NDE offers greater flexibility in capturing complex distributional shapes, it does not guarantee that the conditions for identifiability are met. Lastly, Strömer et al. (2025) introduced a boosting methodology to estimate the association and distribution parameters as functions of distinct covariates. To ensure identifiability, the method combines parametric margins and single-parameter copulas. Each distribution parameter is linked to possibly different covariates through structured additive predictors with parameter-specific link functions. The boosting process iteratively selects and updates only the best-fitting base learner until a maximum amount of iterations has been reached. This produces models similar to lasso regression in simpler settings, with unselected base learners excluded from the final model.

5.1. Software

The work by Strömer et al. (2025) is implemented as an add-on to the function `glmboost LSS` in the R package `gamboostLSS`. We refer to https://github.com/AnnikaStr/CopBoost DepCens for documentation of this add-on. Python implementations are available for the methodologies of Zhang et al. (2023) (https://github.com/WeijiaZhang24/DCSurvival) and Midtfjord et al. (2022) (https://github.com/alimid/clayton_boost).

6. More specialized topics

This section elaborates on an assortment of other copula-based models for dependently censored data. The works discussed below form by no means a complete overview. On the contrary: they should illustrate the vast corpus of literature that hasn't received attention in this review paper thus far. To keep this section brief, we mainly aspire to provide useful starting points for a more in-depth search into these topics.

6.1. Quantile regression

Regression models are widespread in both theoretical and applied statistics. Classical regression allows one to investigate the effects of covariates on a certain quantity of interest, often the mean or variance of a random variable. While certainly very useful, each of these moments only quantifies one aspect of the distribution and as such, does not fully characterize it. In quantile regression, one performs a regression on the quantiles of a distribution, which do characterize the distribution. In this sense, quantile regression can lead to richer analyses.

For survival data under dependent censoring, the literature surrounding quantile regression is sparse, especially when the dependence parameter between T and C is not assumed to be known. To our knowledge, only two papers exist in this setting. A first paper is by Fan and Liu (2018), who were able to partially identify (cf. Section 7.2) the parameters in the linear quantile regression model

$$Q_{\log(T)|X}(\lambda|x) = x^\top \beta, \tag{18}$$

where $Q_{\log(T)|X}(\cdot|x)$ represents the quantile function of $\log(T)$ conditional on the covariates and λ denotes the quantile level of interest. Furthermore, they did not assume a model on C and coupled $\log(T)$ and $\log(C)$ using an Archimedean copula with dependence parameter ξ. By exploiting the formulation for the marginal of $\log(T)$ that is implied by the copula (cf. equation (17)), Fan and Liu (2018) were able to derive a test for the hypothesis that for a given parameter vector β, there exists a dependence parameter ξ such that the implied model is consistent with the data. Applying this test over the entire parameter space and collecting all values of β that are not rejected, the set of parameters that are consistent with the model is obtained.

The second paper that can be used in this context is by D'Haen, Van Keilegom and Verhasselt (2025). They imposed the same linear quantile regression model (18) for

$\log(T)$, required a parametric marginal for $\log(C)$, and modeled the joint distribution of these variables using a (not necessarily Archimedean) parametric copula. Moreover, they assumed a location-scale-like specification for $\log(T)$, namely

$$\log(T) = X^\top \beta + \sigma(X;\gamma)\varepsilon_T,$$

with ε_T independent of X. Here, σ is assumed to be a known, parametric function of X that can be used to model heterogeneity, and ε_T belongs to a carefully constructed class of distributions, namely the *enriched asymmetric Laplace* (EAL) distributions.

The EAL distribution for ε_T could be regarded as the backbone of this methodology. It derives from the asymmetric Laplace (AL) distribution, which is already well-established in the quantile regression literature as it possesses the sought-after property of having zero as its λ-th quantile. It is specified as

$$f_{\text{AL}}(y|\lambda) = \lambda(1-\lambda)\begin{cases} e^{-\lambda y} & y > 0 \\ e^{-(\lambda-1)y} & y \leq 0. \end{cases} \tag{19}$$

Motivated by both theoretical and model flexibility arguments, this distribution is extended to the EAL class of distributions by multiplying with a Laguerre polynomial of order m and $\tilde{m}$, respectively, in both case distinctions,

$$f_{\text{EAL}}(y|\lambda) = \lambda(1-\lambda)\begin{cases} e^{-\lambda y}\|\varphi\|^{-2}\left(\sum_{k=0}^{m}\varphi_k L_k(\lambda y)\right)^2 & y > 0 \\ e^{-(\lambda-1)y}\|\tilde{\varphi}\|^{-2}\left(\sum_{k=0}^{\tilde{m}}\tilde{\varphi}_k L_k((\lambda-1)y)\right)^2 & y \leq 0. \end{cases} \tag{20}$$

In this notation, L_k represents the Laguerre basis function of order k, which enters the polynomial with coefficient φ_k. It can be shown (Kreiss and Van Keilegom, 2025) that (20) can approximate any continuous density in Hellinger sense, provided sufficiently large maximum degrees m and $\tilde{m}$. As such, despite being specified fully parametrically, the methodology of D'Haen et al. (2025) enjoys great flexibility.

6.2. Frailty modeling

Throughout this review paper, we have restricted to discussing research that models the dependence between T and C using a copula. In a seemingly other line of research, this dependence is modeled through an unobserved frailty term. Generally, one imposes the model

$$\begin{cases} S_{T|X,v}(t|x,v) = \exp(-v\Lambda_{T|X}(t|x)) \\ S_{C|X,v}(c|x,v) = \exp(-v\Lambda_{C|X}(c|x)), \end{cases}$$

where v represents the unobserved frailty with distribution F_v. Expressions for the marginal distributions are obtained by integrating out this frailty. Specifically for T, one obtains

$$S_{T|X}(t|x) = \int_v \exp(-v\Lambda_{T|X}(t|x))dF_v(v).$$

Moreover, writing $S_{T|X}(t|x) = \int_v \exp(-vu) f_v(v) dv$ it can be seen that $S_{T|X}$ is the Laplace transformation of f_v evaluated at $u = \Lambda_{T|X}(t|x)$, which we will denote as $\mathcal{L}_v(u)$. Hence, $S_{T|X}(t|x) = \mathcal{L}_v(\Lambda_{T|X}(t|x))$ and $S_{C|X}(c|x) = \mathcal{L}_v(\Lambda_{C|X}(c|x))$.

Though it is not immediately clear at first sight, the frailty model formulation is equivalent to the Archimedean copula formulation (Lo, Stephan and Wilke, 2017) discussed in equation (16). Indeed, by noting that in the frailty model formulation, T and C are independent given v, we have that

$$S_{T,C|X,v}(t,c|x,v) = \mathbb{P}(T > t, C > c|X = x, v) = \exp(-v\Lambda_{T|X}(t|x) - v\Lambda_{C|X}(c|x)).$$

Integrating out the unobserved frailty, we obtain

$$\begin{aligned} S_{T,C|X}(t,c|x) &= \int_v \exp(-v\Lambda_{T|X}(t|x) - v\Lambda_{C|X}(c|x)) f_v(v) dv \\ &= \mathcal{L}_v(\Lambda_{T|X}(t|x) + \Lambda_{C|X}(c|x)) \\ &= \mathcal{L}_v(\mathcal{L}_v^{-1}(S_{T|X}(t|x)) + \mathcal{L}_v^{-1}(S_{C|X}(c|x))), \end{aligned}$$

which is precisely the same as (16) with $\mathcal{L}_v^{-1} \equiv \psi_\xi$, exposing the relation between the assumed distribution F_v for the frailty term and the assumed Archimedean copula.

This link opens the door to the body of literature surrounding frailty models for survival analysis. This type of models is commonly used in the context of clustered data, and there exist some works that additionally consider the problem of dependent censoring. Among the first to explore this intersection were Huang and Wolfe (2002), who proposed a frailty modeling framework that can also take dependent censoring *between* clusters into account. In subsequent work, Huang, Wolfe and Hu (2004) proposed a test for dependent censoring between clusters. In the current state-of-the-art, Schneider et al. (2020) further studied the model of Huang and Wolfe (2002) in the special case of Weibull or piecewise exponential baseline hazards. Notice, however, that in all these works, censoring *within* each cluster is assumed to be independent. To the best of our knowledge, survival analysis methods for clustered data under dependent censoring both *between* and *within* clusters remains a gap in the literature.

Some general introductory sources to frailty models include Wienke (2011), Hanagal (2019) and Balan and Putter (2020). For an in-depth review on more recent advances, we refer to Gorfine and Zucker (2023).

6.3. Software

The `DepCens` package (Schneider and Grandemagne, 2023) in R fits the frailty approach of Schneider et al. (2020).

7. Other approaches not based on copulas

So far, the models included in this review primarily focused on copula-based approaches to dependent censoring. However, dependent censoring has also been explored in other

contexts or using other models. In this section we will illustrate this using two examples, that are by no means intended to be exhaustive or representative for the full literature on this topic.

7.1. Transformation models

Transformation models aim to make the data follow certain modeling assumptions, such as normality or homoscedasticity, by transforming the survival and censoring times. In this regard, Deresa and Van Keilegom (2020a) proposed a parametric normal transformation of T and C, assuming that the transformed variables follow a bivariate normal distribution after accounting for covariate effects. More specifically, their model has the following form:

$$\begin{cases} \Lambda_\kappa(\log(T)) = X^\top \beta + \varepsilon_T \\ \Lambda_\kappa(\log(C)) = X^\top \eta + \varepsilon_C, \end{cases} \tag{21}$$

where $\{\Lambda_\kappa \mid \kappa \in K\}$ is a parametric class of monotone increasing transformations defined on $(-\infty, +\infty)$, and β and η are the vectors of regression coefficients. The vector of error terms $(\varepsilon_T, \varepsilon_C)$ has a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_T \\ \varepsilon_C \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_T^2 & \rho\sigma_T\sigma_C \\ \rho\sigma_T\sigma_C & \sigma_C^2 \end{pmatrix}\right),$$

where Σ is assumed to be a positive definite matrix, i.e. $\sigma_T > 0, \sigma_C > 0$ and $|\rho| < 1$.

The parametric transformation model involving the parameters $\beta, \eta, \kappa, \sigma_T, \sigma_C$ and ρ is shown to be identifiable. One of the basic assumptions for establishing this identifiability involves the transformation family $\{\Lambda_\kappa \mid \kappa \in K\}$, which is required to map the whole real line to the whole real line, meaning $\lim_{t\to\pm\infty} \Lambda_\kappa(t) = \pm\infty$ for all κ in K. While this condition does not hold for the widely used Box-Cox transformation, it is satisfied by a family of Yeo and Johnson (2000) transformations, which is an extension of the Box-Cox family to the whole real line. In addition, the bivariate normality assumption for $(\varepsilon_T, \varepsilon_C)$ plays a crucial role in the proof of identifiability. Some earlier works in this area include those by Nádas (1971) and Basu and Ghosh (1978), though neither of these papers considers transformations nor incorporates covariates. Deresa and Van Keilegom (2020a) suggested estimating the transformation model parameters using a maximum likelihood estimation approach and showed the asymptotic properties of these parameters based on results in White (1982).

Deresa and Van Keilegom (2020b) later extended model (21) in several directions by incorporating features such as multivariate competing risks and administrative censoring. Their model accommodates these censoring mechanisms within a single model and can also account for the effect of covariates on event times. A multivariate normal distribution is assumed to model the correlations between competing event times. In particular, the extended model identifies and estimates all parameters without imposing additional assumptions on the covariance matrix. Furthermore, Deresa and Van Keilegom

(2021) relaxed model (21) by proposing a nonparametric specification of the transformation function, instead of relying on a parametric restriction. Their model is based on an arbitrary nonparametric transformation for both T and C, while assuming that the errors in the transformation model follow a standard bivariate normal distribution, which allows for nonzero correlation. This model is shown to be identifiable and estimable based on the observed data.

More recently, Crommen et al. (2024) proposed a fully parametric model that identifies the causal effect of an endogenous treatment variable on a possibly dependently censored event time. We refer the interested reader to Wooldridge (2010) for a more in-depth introduction to endogeneity (or causal inference). A bivariate normality assumption is placed on the error terms of a joint regression model for the logarithm of T and C, without specifying the correlation between them. Using a control function approach, the authors proposed a two-step estimation method that consistently estimates the causal effect of interest and the correlation parameter. In subsequent work, Willems et al. (2025c) increased the flexibility of this model by extending it to a multiple competing risks framework. The authors proposed applying a unique power transformation to each risk to make the multivariate normality assumption on the error terms more plausible.

This bivariate normality assumption has also been omitted altogether in the recent work of Yu and Liu (2024), which is an extension of the transformation model in (21). They apply the generalized estimating equation framework, including a working covariance matrix to improve model efficiency, to a dependent censoring context. While their work allows no transformation function and always uses the identity function $\Lambda_K(y) = y$, it is in turn less restrictive in terms of the error terms $(\varepsilon_T, \varepsilon_C)$. More specifically, Yu and Liu (2024) no longer require homoscedasticity, and moreover leave the distribution of $(\varepsilon_T, \varepsilon_C)$ unspecified. Their semiparametric approach leverages the fact that the estimating equations are equivalent with those under some bivariate normal distribution. Replacing unobserved values of T and C by their conditional expectation under normality, they subsequently show that these score functions remain asymptotically valid even under violation of the assumed normality.

7.2. Partial identification

Throughout the review, we have spent substantial attention on the (point) identifiability of most of the models discussed. And rightfully so: Tsiatis (1975) shows that the joint distribution of T and C cannot be identified nonparametrically, in the sense that for any observed data generating process (Y, Δ), there exist infinitely many joint distributions (T, C) that would imply it. Even after restricting to Archimedean dependence structures, Wang (2012) shows that the problem persists. As such, the first-order goal of many models in the literature is to conceive a set of assumptions under which the implied, initially infinite set of possible marginals for T shrinks to a singleton.

The assumptions required for this are often stringent and/or non-testable. This has motivated a small branch of survival analysts to study *partially identified* models. In essence, these models allow the set of possible marginals for T to remain infinite while

still doing inference on some quantities of interest. Typically, this leads to the estimation of bounds on said quantities instead of point estimates.

In the partial identification paradigm, Szydlowski (2019) studied the effect of the independence assumption on the parameters in a parametric model. When there is unmeasured confounding present in the data, Blanco et al. (2020) were able to derive bounds on the causal effect. In the most recent literature, Sakaguchi (2024) and Willems et al. (2025a) developed methodologies to obtain bounds on the parameters in different regression models for T, while leaving both the censoring distribution as well as the dependence structure completely unspecified.

7.3. Software

The `depCensoring` package in R implements the method of Deresa and Van Keilegom (2021) in the function `NonParTrans`, the method of Willems et al. (2025c) (which includes Crommen et al. (2024) as a special case) in the function `estimate.cmprsk`, and the method of Willems, Beyhum and Van Keilegom (2025a) in the function `pi.surv`.

8. Conclusion

While we have tried to give an extensive overview of the literature on copula-based dependent censoring, inevitably a number of topics have not been discussed. For instance, while we assumed throughout that for every subject either T or C is observed, but never both, there are cases where both variables are sometimes observed, which is referred to as *semicompeting risks* in the literature. The use of copulas for semicompeting risks is very natural and has been well documented in the literature (see e.g. Peng et al. (2007); Chen (2012)). Semicompeting risks can be considered as an example of a situation where additional information is available (in this case in the form of observing the full pair (T,C) for certain subjects), which helps to identify the dependent censoring model. Other situations where additional information is available (in the form of auxiliary data, transfer learning, data integration, etc.) would be worth exploring in the future. The work in Jo et al. (2023) can be considered as a first step in this context. However, they simply use the extra information (*in casu* of an observable variable influencing both T and C) to rather heuristically find a reasonable range for the dependence parameter in performing copula-graphic based sensitivity analysis. An extension to more rigorous models remains an interesting challenge.

Another topic that has not been mentioned in this review, is that of goodness-of-fit tests for the proposed models. A natural and common way to test whether a dependent censoring model is valid would be to compare its fit with that of a fully nonparametric model. However, this is not a feasible approach, as it is well known that the fully nonparametric dependent censoring model is not identified. Deresa and Van Keilegom (2020a, 2024a) tried to circumvent this problem by calculating the distribution of $Y = \min(T,C)$ under the fitted model and comparing it with its fully nonparametric distribution, which is always identified. While this is a correct testing procedure, it has

the drawback that it actually tests whether the distribution of Y is correct, instead of that of T. More research on this important topic would be needed, as well as on the related topic regarding the consequences of model misspecification on the sign and magnitude of parameters in a dependent censoring model.

Lastly, there are cases where the censoring time is related to the survival time in a specific way that differs from copulas. This results in particular forms of dependent censoring, such as in the case of medical cost data (Lin, 2003; Willan, Lin and Manca, 2004; Li et al., 2016), survival data with measurement errors (Van Keilegom and Kekeç, 2024), gap times in multistate models (de Uña-Álvarez and Meira-Machado, 2008), dynamic covariates (Beyhum et al., 2024), simultaneous shocks (Escobar-Bach and Helali, 2024) and many more.

Acknowledgements

G. Crommen is funded by a PhD fellowship from the Research Foundation - Flanders (grant number 11PKA24N). M. D'Haen is funded by a BOF PhD fellowship at Hasselt University (number R-12215). J. Ding is supported by the China Scholarship Council (grant number 202306060103). I. Van Keilegom acknowledges funding from the FWO and F.R.S. - FNRS (Excellence of Science programme, project ASTeRISK, grant no. 40007517), and from the FWO (senior research projects fundamental research, grant no. G047524N).

References

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5:311–342.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1988). Censoring, truncation and filtering in statistical models based on counting processes. In Prabhu, N. (Ed.), *Statistical Inference from Stochastic Processes*, volume 80 of *Contemporary Mathematics*. American Mathematical Society, Providence, Rhode Island.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer Verlag, New York.

Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11):3424–3454.

Basu, A. P. and Ghosh, J. K. (1978). Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8:413–429.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. (Technical report, University of California, Berkeley).

Beyhum, J., Centorrino, S., Florens, J.-P., and Van Keilegom, I. (2024). Instrumental variable estimation of dynamic treatment effects on a duration outcome. *Journal of Business and Economic Statistics*, 42:732–742.

Blanco, G., Chen, X., Flores, C. A., and Flores-Lagunes, A. (2020). Bounds on average and quantile treatment effects on duration outcomes under censoring, selection, and noncompliance. *Journal of Business & Economic Statistics*, 38(4):901–920.

Bou-Hamad, I., Larocque, D., and Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys*, 5:44 – 71.

Braekers, R. and Veraverbeke, N. (2005). A copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33(3):429–447.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Carrière, J. F. (1995). Removing cancer when it is correlated with other causes of death. *Biometrical Journal*, 37(3):339–350.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA. Association for Computing Machinery.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society – Series B*, 72(2):235–251.

Chen, Y.-H. (2012). Maximum likelihood analysis of semicompeting risks data with semiparametric regression models. *Lifetime Data Analysis*, 18:36–57.

Collett, D. (2015). *Modelling Survival Data in Medical Research*, volume 2. Chapman & Hall/CRC Press, Boca Raton, FL, third edition.

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society – Series B*, 21(2):411–421.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society – Series B*, 34(2):187–202.

Crommen, G., Beyhum, J., and Van Keilegom, I. (2024). An instrumental variable approach under dependent censoring. *TEST*, 33(2):473–495.

Crommen, G., Beyhum, J., and Van Keilegom, I. (2025). Estimation of the complier causal hazard ratio under dependent censoring. *arXiv:2504.02096*. (Under review).

Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738.

de Uña-Álvarez, J. and Meira-Machado, L. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics & Probability Letters*, 78(15):2440–2445.

de Uña-Álvarez, J. and Veraverbeke, N. (2013). Generalized copula-graphic estimator. *TEST*, 22:343–360.

de Uña-Álvarez, J. and Veraverbeke, N. (2017). Copula-graphic estimation with left-truncated and right-censored data. *Statistics*, 51(2):387–403.

Delhelle, M. and Van Keilegom, I. (2025). Copula based dependent censoring in cure models. *TEST*.

Deresa, N. W. and Van Keilegom, I. (2020a). Flexible parametric model for survival data subject to dependent censoring. *Biometrical Journal*, 62(1):136–156.

Deresa, N. W. and Van Keilegom, I. (2020b). A multivariate normal regression model for survival data subject to different types of dependent censoring. *Computational Statistics & Data Analysis*, 144:106879.

Deresa, N. W. and Van Keilegom, I. (2021). On semiparametric modelling, estimation and inference for survival data subject to dependent censoring. *Biometrika*, 108(4):965–979.

Deresa, N. W. and Van Keilegom, I. (2024a). Copula based Cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, 119(546):1044–1054.

Deresa, N. W. and Van Keilegom, I. (2024b). Semiparametric transformation models for survival data with dependent censoring. *Annals of the Institute of Statistical Mathematics*.

Deresa, N. W., Van Keilegom, I., and Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring. *Insurance: Mathematics and Economics*, 107:1–21.

D'Haen, M., Van Keilegom, I., and Verhasselt, A. (2025). Quantile regression under dependent censoring with unknown association. *Lifetime Data Analysis*, 31:253–299.

Ding, J. and Van Keilegom, I. (2025). A copula based extension of the Kaplan-Meier estimator under dependent censoring with unknown association. (Under review).

Emura, T. and Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring: a copula-based approach. *Statistical Methods in Medical Research*, 25(6):2840–2857.

Emura, T. and Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*, volume 450. Springer, Singapore.

Emura, T., Chen, H.-Y., Matsui, S., and Chen, Y.-H. (2024). *Compound.Cox: Univariate Feature Selection and Compound Covariate for Predicting Survival, Including Copula-Based Analyses for Dependent Censoring*. R package (version 3.31).

Escobar-Bach, M., and Helali, S. (2024). Dependent censoring with simultaneous death times based on the generalized Marshall-Olkin model. *Journal of Multivariate Analysis*, 204:105347.

Fan, T.-H., Wang, Y.-F., and Ju, S.-K. (2019). A competing risks model with multiply censored reliability data under multivariate Weibull distributions. *IEEE Transactions on Reliability*, 68(2):462–475.

Fan, Y. and Liu, R. (2018). Partial identification and inference in censored quantile regression. *Journal of Econometrics*, 206(1):1–38.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Frahm, G., Junker, M., and Szimayer, A. (2003). Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63(3):275–286.

Geenens, G. (2020). Copula modeling for discrete random vectors. *Dependence Modeling*, 8:417–440.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283.

Geskus, R. B. (2016). *Data Analysis with Competing Risks and Intermediate States*. Taylor and Francis.

Gharari, A. H. F., Cooper, M., Greiner, R., and Krishnan, R. G. (2023). Copula-based deep survival models for dependent censoring. In *Uncertainty in Artificial Intelligence*, pages 669–680.

Gorfine, M. and Zucker, D. M. (2023). Shared frailty methods for complex survival data: A review of recent advances. *Annual Review of Statistics and Its Application*, 10(1):51–73.

Han, A. and Hausman, J. A. (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics*, 5(1):1–28.

Hanagal, D. D. (2019). *Modeling Survival Data Using Frailty Models*. Industrial and Applied Mathematics. Springer, Singapore.

Hiabu, M., Lo, S. M. S., and Wilke, R. A. (2025). Identifiability and estimation of the competing risks model under exclusion restrictions. *Statistica Neerlandica*.

Hothorn, T., Barbanti, L., Siegfried, S., Ripley, B., Venables, B., Bates, D. M. and Klein, N. (2024). *tram: Transformation models*. R package (version 1.2-0)

Huang, X. and Wolfe, R. A. (2002). Frailty model for informative censoring. *Biometrics*, 58(3):510–520.

Huang, X., Wolfe, R. A., and Hu, C. (2004). A test for informative censoring in clustered survival data. *Statistics in Medicine*, 23(13):2089–2107.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4):1090–1099.

Jo, J.H., Gao, Z., Jung, I., Song, S.Y., Ridder, G. and Moon, H.R. (2023). Copula graphic estimation of the survival function with dependent censoring and its application to analysis of pancreatic cancer clinical trial. *Statistical Methods in Medical Research*, 32(5):944–962.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Press, New York.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, Hoboken, N.J, second edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kreiss, A. and Van Keilegom, I. (2025). Efficient quantile regression under censoring using Laguerre polynomials. *Bernoulli*. (To appear).

Li, J., Handorf, E., Bekelman, J., and Mitra, N. (2016). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in Medicine*, 35(12): 1985–1999.

Li, Y., Tiwari, R. C., and Guha, S. (2007). Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society – Series B*, 69(3):285–306.

Lin, D. (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine*, 22:1181–1200.

Ling, C. K., Fang, F., and Kolter, J. Z. (2020). Deep archimedean copulas. *Advances in Neural Information Processing Systems*, 33:1535–1545.

Lo, S. M. S., Stephan, G., and Wilke, R. A. (2017). Competing risks copula models for unemployment duration: An application to a German Hartz reform. *Journal of Econometric Methods*, 6(1):20150005.

Lo, S. M. S. and Wilke, R. A. (2010). A copula model for dependent competing risks. *Journal of the Royal Statistical Society – Series C*, 59(2):359–376.

Lo, S. M. S. and Wilke, R. A. (2014). A regression model for the copula-graphic estimator. *Journal of Econometric Methods*, 3(1):21–46.

Lo, S. M. S. and Wilke, R. A. (2017). Identifiability of the sign of covariate effects in the competing risks model. *Econometric Theory*, 33(5):1186–1217.

Lo, S. M. S. and Wilke, R. A. (2023). A parametric competing risks regression model with unknown dependent censoring. *Journal of the Royal Statistical Society – Series C*, 72(4):1079–1093.

Maller, R., Resnick, S., Shemehsavar, S., and Zhao, M. (2024). Mixture cure model methodology in survival analysis: some recent results for the one-sample case. *Statistic Surveys*, 18:82–138.

Midtfjord, A. D., De Bin, R., and Huseby, A. B. (2022). A copula-based boosting model for time-to-event prediction with dependent censoring. *arXiv:2210.04869*.

Moeschberger, M. L. and Klein, J. P. (1984). Consequences of departures from independence in exponential series systems. *Technometrics*, 26(3):277–284.

Moeschberger, M. L. and Klein, J. P. (1995). Statistical methods for dependent competing risks. *Lifetime Data Analysis*, 1(2):195–204.

Moradian, H., Larocque, D., and Bellavance, F. (2017). L_1 splitting rules in survival forests. *Lifetime data analysis*, 23:671–691.

Moradian, H., Larocque, D., and Bellavance, F. (2019). Survival forests for data with dependent censoring. *Statistical Methods in Medical Research*, 28(2):445–461.

Nádas, A. (1971). The distribution of the identified minimum of normal pair determines the distribution of the pair. *Technometrics*, 13:201–202.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York.

Parsa, M. and Van Keilegom, I. (2023). Accelerated failure time vs Cox proportional hazards mixture cure models: David vs Goliath? *Statistical Papers*, 64(3):835–855.

Peng, L., Jiang, H., Chappell, R., and Fine, J. (2007) An overview of the semi-competing risks problem. In *Statistical Advances in the Biomedical Sciences*, pages 177–192. John Wiley & Sons, Ltd.

Peng, Y. and Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation*. Chapman & Hall/CRC Press, New York.

Rivest, L.-P. and Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155.

Sakaguchi, S. (2024). Partial identification and inference in duration models with endogenous censoring. *Journal of Applied Econometrics*, 39(2):308-326.

Schneider, S., Demarqui, F. N., Colosimo, E. A., and Mayrink, V. D. (2020). An approach to model clustered survival data with dependent censoring. *Biometrical Journal*, 62(1):157–174.

Schneider, S. and Grandemagne, G. (2023). *Depcens: Dependent censoring regression models*. R package (version 0.2.3).

Schwarz, M., Jongbloed, G., and Van Keilegom, I. (2013). On the identifiability of copulas in bivariate competing risks models. *Canadian Journal of Statistics*, 41(2):291–303.

Shih, J.-H., Lee, W., Sun, L.-H., and Emura, T. (2019). Fitting competing risks data to bivariate Pareto models. *Communications in Statistics-Theory and Methods*, 48(5):1193–1220.

Siegfried, S., Tamási, B. and Hothorn, T. (2024). Smooth Transformation Models for Survival Analysis: A Tutorial Using R. *arXiv:2402.06428*.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231.

Strömer, A., Klein, N., Van Keilegom, I., and Mayr, A. (2025). Modelling dependent censoring in time-to-event data by boosting copula regression. (Under review).

Sujica, A. and Van Keilegom, I. (2015). Estimation of location and scale functionals in nonparametric regression under copula dependent censoring. *Canadian Journal of Statistics*, 43(2):306–335.

Sujica, A. and Van Keilegom, I. (2018). The copula-graphic estimator in censored nonparametric location-scale regression models. *Econometrics and Statistics*, 7:89–114.

Szydlowski, A. (2019). Endogenous censoring in the mixed proportional hazard model with an application to optimal unemployment insurance. *Journal of Applied Econometrics*, 34(7):1086–1101.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Van Keilegom, I. and Kekeç, E. (2024). Estimation of the density for censored and contaminated data. *STAT*, 13:e651.

Van Keilegom, I. and Veraverbeke, N. (1996). Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles. *Communications in Statistics–Theory and Methods*, 25(10):2251–2265.

Veraverbeke, N. (2006). Regression quantiles under dependent censoring. *Statistics*, 40(02):117–128.

Wang, A. (2012). On the nonidentifiability property of Archimedean copula models under dependent censoring. *Statistics & Probability Letters*, 82(3):621–625.

Wang, A. (2014). Properties of the marginal survival functions for dependent censored data under an assumed Archimedean copula. *Journal of Multivariate Analysis*, 129(3):57–68.

Wang, A. (2023). The identifiability of copula models for dependent competing risks data with exponentially distributed margins. *Statistica Sinica*, 33(2):983–1001.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC Press, Boca Raton, FL.

Wilke, R. A. and Lo, S. M. S. (2025). Accelerated failure time analysis for industrial life modeling in presence of unknown dependent and independent censoring. *Quality Engineering*.

Willan, A. R., Lin, D. Y., and Manca, A. (2004). Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine*, 24(1):131–145.

Willems, I., Beyhum, J., and Van Keilegom, I. (2025a). Bounds for the regression parameters in dependently censored survival models. *arXiv:2503.11210*. (Under review).

Willems, I., Crommen, G., Deresa, N. W., Ding, J., Czado, C., and Van Keilegom, I. (2025b). *depCensoring: Statistical Methods for Survival Data with Dependent Censoring*. R package (version 0.1.5).

Willems, I., Rutten, S., Crommen, G., and Van Keilegom, I. (2025c). A flexible control function approach for survival data subject to different types of censoring. *arXiv:2403.11860*. (Under review).

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, second edition.

Yeo, I. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87:954–959.

Yu, L. and Liu, L. (2024). Generalized estimating equations for survival data with dependent censoring. *Statistics in Medicine*, 43:5983–5995.

Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of Royal Statistical Society – Series B*, 69(4):507–564.

Zhang, W., Ling, C. K., and Zhang, X. (2023). Deep copula-based survival analysis for dependent censoring with identifiability guarantees. *arXiv:2312.15566*.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138.

Zucker, D. M. (2005). A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *Journal of the American Statistical Association*, 100:1264–1277.

SORT 49 (1) January-June 2025, 43-72 DOI: 10.57645/20.8080.02.22

On statistical model extensions based on randomly stopped extremes

Jordi Valero and Josep Ginebra[1]

Abstract

The maxima and the minima of a randomly stopped sample of a random variable, X, together with two newly defined random variables that make X into the maxima or minima of a randomly stopped sample of them, can be used to define statistical model transformation mechanisms. These transformations can be used to define models for extreme-value data that are not grounded on large sample theory. The relationship between the stopping model and characteristics of the corresponding model transformations obtained is investigated. In particular, one looks into which stopping models make these model transformations into model extensions, and which stopping models lead to statistically stable extensions in the sense that using the model extension a second time leaves the extended model unchanged. The stopping models under which the extensions based on randomly stopped maxima and their inverses coincide with the extensions based on randomly stopped minima and their inverses are also characterized. The advantages of using models obtained through these model extension mechanisms instead of resorting to extreme-value models grounded on asymptotic arguments is illustrated by way of examples.

MSC: *62E10, 62A99, 60E10.*

Keywords: *Marshall-Olkin extension, extreme value, randomly stopped maximum, randomly stopped minimum, statistical stability, stopping model.*

1. Introduction

In disciplines such as hydrology, meteorology, ecology, seismology, actuarial sciences, civil engineering or finance, there is a need for statistical models to analyze extreme-valued data, like the largest single-event rainfall or the magnitude of the strongest earth-

[1] Address for correspondence: Department of Statistics and O.R., Polytechnic University of Catalonia, Avgda. Diagonal 647, 6ª Planta, 08028 Barcelona, Spain (E-mail: josep.ginebra@upc.es)
Received: September 2023.
Accepted: November 2024.

quake in a year. In these settings researchers most often resort to the use of the generalized extreme-value model, which is grounded on large sample theory that only applies as an approximation when sample sizes are large enough.

Hence, there is a need for statistical models for extreme-valued data that can be grounded on finite-sample theory. One framework that provides that ground, models the number of events in a year, like the number of rainfalls or of earthquakes, through a random variable, N, with a given stopping model, it models the magnitude of the events in that year as a sample of i.i.d. observations, $(X_1,\ldots,X_N)$, with a given stopped model, and it assumes that one observes the maxima or the minima, Y, of that randomly stopped sample. Models defined like the one for Y are also useful in reliability, where the minimum (or maximum) of a randomly stopped sample from a lifetime distribution serves as a model for the lifetime of a series (or parallel) system.

Marshall and Olkin (1997) obtained statistical models of this kind by extending an initial statistical model through the distribution of the minimum and of the maximum of a geometrically stopped sample of independent observations with a distribution in the initial family. This statistical model transformation mechanism has proved extremely fruitful in practice, as the more than seventeen hundred citations of that paper indicate.

One nice feature of model transformations based on geometrically stopped extremes is that they always work as model extensions, because the initial family of distributions is always included in the new family. A second interesting feature of these geometrically stopped extreme extensions is that they are statistically stable in the sense that the extended model can not be further extended by using that same extension mechanism a second time. These two features are not in place in general, when transforming statistical models through randomly stopped maxima and minima with a stopping model different from geometric. In fact, Marshall and Olkin (1997) conjectures that this kind of stability can only be obtained through geometrically stopped extremes.

Here these issues are investigated in full generality, by looking into all model transformations defined through the maxima or the minima of N-stopped random samples of X, for any given stopping model for N and any given stopped model for X.

On top of looking into randomly stopped extreme model extensions beyond geometric stopping, we also propose two new model transformation mechanisms based on two new random variables defined to be the ones that make X into the randomly stopped maxima and the randomly stopped minima of them, which we label as the N-maxprecursor and the N-minprecursor of X. These transformations can be viewed as the inverse transformations of N-stopped maxima and of N-stopped minima of X, and the statistical models obtained through them can be used to learn about the magnitude of events, X, based on their frequency N and their extreme values Y.

Finally, on top of these four basic model transformation mechanisms based on randomly stopped maxima and minima and on their inverses, we also propose another two new pair of transformation mechanisms that combine N-stopped maxima of X with their inverses, and combine N-stopped minima of X with their inverses. Under geometric stopping, these combined model transformation mechanisms coincide with the

Marshall-Olkin extension mechanism, and they work as model extensions under any stopping model, which is why we consider them to be the natural way to generalize Marshall-Olkin when the stopping model is not geometric.

The relationship between characteristics of the stopping model and characteristics of all the corresponding model transformations considered is studied. The first objective is to identify which stopping models lead to transformations that always work as model extensions, and the second objective is to identify which stopping models lead to model extensions that are statistically stable, in the sense that they do not further extend the initial model beyond the first use.

The second objective leads to the investigation of the class of stopping models that are closed under probability generating function (pgf) composition, because that is a necessary condition for the corresponding randomly stopped extreme extensions to be stable. This investigation helps us to disprove by way of examples the conjecture that only geometric stopping models lead to statistically stable extensions.

The paper also looks into the reversibility conditions required of stopping models so that the model extensions built based on N-stopped maxima and their inverses coincide with the model extensions built based on N-stopped minima and their inverses, which is a property satisfied in particular by the extensions based on geometric stopping.

The paper illustrates through examples the advantages of modeling extreme-valued data with models obtained through randomly stopped extreme extensions instead of resorting to the usual generalized extreme-value model backed through large sample arguments. We also use examples to help understand the rationale behind the use of the models obtained through the new model extension mechanisms that use the inverse of N-stopped maxima or minima.

The paper is organized as follows. Section 2 defines randomly stopped extreme and extreme-precursor random variables, and presents the four basic and four combined model transformation mechanisms that will be investigated, and Section 3 illustrates the use of models obtained with these transformations to deal with extreme-value data. Section 4 introduces the definition of statistically stable model transformation. Section 5 defines extreme-reversible and auto-reversible stopping models, and Section 6 looks into stopping models that are closed under pgf composition, which are the ones that yield statistically stable transformations. Section 7 relates these and other properties of the stopping model with features of the corresponding model transformations, and Section 8 presents examples of statistically stable randomly stopped extreme extensions.

2. Statistical models based on randomly stopped extremes

2.1. Randomly stopped extremes and extreme precursors

Let X be a real valued random variable defined through its cumulative distribution function, F_X, and let N be a positive integer valued random variable, with $\Pr(N = 0) = 0$, defined through its probability generating function (pgf), h_N. Assume that one observes

n independent copies of X, X_i, where n is a realization of the stopping variable N independent of the X_i.

The *N-stopped maximum of X*, which we denote by $\max_N(X)$, is the random variable $Y = \max(X_1, \cdots X_N)$ with cumulative distribution function:

$$F_{\max_N(X)} = h_N(F_X),$$

and the *N-stopped minimum of X*, which we denote by $\min_N(X)$, is the random variable $Y = \min(X_1, \cdots, X_N)$ with survival function $h_N(S_X)$, where $S_X = 1 - F_X$, and therefore with cdf:

$$F_{\min_N(X)} = 1 - h_N(1 - F_X) = \overline{h}_N(F_X),$$

where $\overline{h}_N(t) = 1 - h_N(1-t)$, which will be denoted as the *conjugate function* of $h_N(t)$.

These two random variables are studied for example in Raghundanan and Patil (1972), Shaked (1975), Consul (1984), Gupta and Gupta (1984), Rohatgi (1987), Shaked and Wong (1997), in pp.155-157 of Arnold, Balakrishnan and Nagaraja (1992) and in Louzada, Beret and Franco (2012).

Next, two new random variables that play a central role in what follows are introduced. They arise from the fact that given any N and any X, one can always interpret X to be the N-stopped maximum and the N-stopped minimum of the two random variables defined next.

Definition 1. *Given any stopping variable N and any real valued random variable X as defined above, let the N-maxprecursor of X, denoted as* $\max_N^{-1}(X)$*, be the random variable Y with cdf:*

$$F_{\max_N^{-1}(X)} = h_N^{-1}(F_X),$$

and let the N-minprecursor of X, denoted as $\min_N^{-1}(X)$*, be the random variable Y with cdf:*

$$F_{\min_N^{-1}(X)} = \overline{h}_N^{-1}(F_X).$$

The properties of h_N and of $\overline{h}_N$ presented in Section 5.1 guarantee that they are always invertible and therefore that $F_{\max_N^{-1}(X)}$ and $F_{\min_N^{-1}(X)}$ are always properly defined cdf's. As a consequence, the random variables $\max_N^{-1}(X)$ and $\min_N^{-1}(X)$ will exist for any N and any X.

By definition, X is always the N-stopped maximum of $\max_N^{-1}(X)$, the N-stopped minimum of $\min_N^{-1}(X)$, the N-maxprecursor of $\max_N(X)$, and the N-minprecursor of $\min_N(X)$,

$$X = \max_N(\max_N^{-1}(X)) = \max_N^{-1}(\max_N(X)) = \min_N(\min_N^{-1}(X)) = \min_N^{-1}(\min_N(X)),$$

which is why we denote N-maxprecursors and minprecursors as the inverses of the N-stopped maxima and minima.

2.2. Statistical model transformations based on randomly stopped extremes

Let the family of distributions $\mathcal{X} = \{X_\theta : F_{X_\theta},\ \theta \in \Theta\}$ be a statistical model defined on $x \in S \subseteq \mathbb{R}$, with parameter space Θ, where F_{X_θ} is the cdf of X_θ.

Let $\mathcal{N} = \{N_\delta : h_{N_\delta} = \sum_{n=1}^{\infty} p_n(\delta)t^n,\ \delta \in \mathcal{D}\}$ be a statistical model defined on the positive integers, $n \in \mathbb{N}^+$, with parameter space $\mathcal{D}$, where $p_n(\delta) = \Pr(N_\delta = n)$ and where h_{N_δ} is the pgf of N_δ. We denote $\mathcal{N}$ as the stopping model.

Note that by definition in this paper it will always be assumed that stopping models, $\mathcal{N}$, are always such that $\Pr(N_\delta = 0) = 0$ for any $\delta \in \mathcal{D}$.

We next define four basic mechanisms,$\mathcal{T}(\cdot)$, that transform the initial statistical model, $\mathcal{X}$, into a new statistical model, $\mathcal{Y} = \mathcal{T}(\mathcal{X})$, through the N-stopped maximum (minimum) of $X \in \mathcal{X}$, and through the N-maxprecursors (N-minprecursors) of $X \in \mathcal{X}$, with $N \in \mathcal{N}$.

Definition 2. *Given any statistical model $\mathcal{X}$ and any stopping model $\mathcal{N}$ as defined above, let $\max_{\mathcal{N}}(\mathcal{X})$ and $\max_{\mathcal{N}}^{-1}(\mathcal{X})$ denote the statistical models defined as:*

$$\max_{\mathcal{N}}(\mathcal{X}) = \{Y_{\theta,\delta} : F_{Y_{\theta,\delta}} = h_{N_\delta}(F_{X_\theta}),\ \theta \in \Theta, \delta \in \mathcal{D}\},$$

$$\max_{\mathcal{N}}^{-1}(\mathcal{X}) = \{Y_{\theta,\delta} : F_{Y_{\theta,\delta}} = h_{N_\delta}^{-1}(F_{X_\theta}),\ \theta \in \Theta, \delta \in \mathcal{D}\}.$$

Likewise, let $\min_{\mathcal{N}}(\mathcal{X})$ and $\min_{\mathcal{N}}^{-1}(\mathcal{X})$ denote the statistical models defined as:

$$\min_{\mathcal{N}}(\mathcal{X}) = \{Y_{\theta,\delta} : F_{Y_{\theta,\delta}} = \overline{h}_{N_\delta}(F_{X_\theta}),\ \theta \in \Theta, \delta \in \mathcal{D}\},$$

$$\min_{\mathcal{N}}^{-1}(\mathcal{X}) = \{Y_{\theta,\delta} : F_{Y_{\theta,\delta}} = \overline{h}_{N_\delta}^{-1}(F_{X_\theta}),\ \theta \in \Theta, \delta \in \mathcal{D}\}.$$

These two pairs of basic transformations do not always work as model extensions. Instead, the two pairs of combined transformations defined next work as model extensions for any $\mathcal{X}$, even when one of the two new parameters is fixed. They are the family of all N-stopped maximum (minimum) of all N-maxprecursors (N-minprecursors) of X, and viceversa.

Definition 3. *Given any statistical model $\mathcal{X}$ and any stopping model $\mathcal{N}$ as defined above, let $\max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X}))$ and $\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X}))$ denote the statistical models defined as:*

$$\max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\delta_1,\delta_2} : F_{Y_{\theta,\delta_1,\delta_2}} = h_{N_{\delta_2}} \circ h_{N_{\delta_1}}^{-1}(F_{X_\theta}),\ \theta \in \Theta, \delta_1, \delta_2 \in \mathcal{D}\},$$

$$\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X})) = \{Y_{\theta,\delta_1,\delta_2} : F_{Y_{\theta,\delta_1,\delta_2}} = h_{N_{\delta_2}}^{-1} \circ h_{N_{\delta_1}}(F_{X_\theta}),\ \theta \in \Theta, \delta_1, \delta_2 \in \mathcal{D}\}.$$

Likewise, let $\min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X}))$ and $\min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X}))$ denote the statistical models:

$$\min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\delta_1,\delta_2} : F_{Y_{\theta,\delta_1,\delta_2}} = \overline{h}_{N_{\delta_2}} \circ \overline{h}_{N_{\delta_1}}^{-1}(F_{X_\theta}),\ \theta \in \Theta, \delta_1, \delta_2 \in \mathcal{D}\},$$

$$\min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X})) = \{Y_{\theta,\delta_1,\delta_2} : F_{Y_{\theta,\delta_1,\delta_2}} = \overline{h}_{N_{\delta_2}}^{-1} \circ \overline{h}_{N_{\delta_1}}(F_{X_\theta}),\ \theta \in \Theta, \delta_1, \delta_2 \in \mathcal{D}\}.$$

Note that these statistical model transformation mechanisms can also be used to generate statistical models, $\mathcal{Y}$, starting from a single initial random variable, $\mathcal{Y} = \mathcal{T}(X)$.

Using our notation, the Marshall-Olkin extension of $\mathcal{X}$ is defined to be $\max_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}(\mathcal{X})$ when $\mathcal{N}$ is the geometric stopping model. In Sections 7.2 and 7.4 it will be argued that for stopping models other than geometric the transformation $\max_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}(\mathcal{X})$ does not always work as an extension, but that under geometric stopping this transformation coincides with the four model transformations in Definition 3, which do work as extensions under any stopping model. As a consequence, we will propose Definition 3 and not $\max_{\mathcal{N}}(\cdot) \cup \min_{\mathcal{N}}(\cdot)$ to be the natural way to generalize Marshall-Olkin when using stopping models different from geometric.

3. Examples of the use of randomly stopped extreme models

The examples presented here illustrate the advantage in using models defined through the randomly stopped extreme transformations in Definition 2 instead of using the generalized extreme-value model, and they help understand the practical relevance of the randomly stopped extreme-precursor models also considered in that definition. The examples also touch on the rationale behind the use of the model extensions proposed in Definition 3.

3.1. On the usefulness of randomly stopped extreme models

Let's assume for example that one has data on the rainfall in the largest rain event of a year, Y_i, for a set of m years, $(y_1, \ldots, y_m)$. This kind of data is usually modeled through the three parameter generalized extreme-value model, because it is the limiting model for properly normalized extreme values when the rainfall in an event is i.i.d., and the number of rainfall events in a year grows.

As an alternative way to model this kind of data one can assume that the number of rain events in the i-th year, N_i, is random and can be modeled through a specific stopping model, $\mathcal{N}$, and that the rainfall in the set of N_i events is a sample of i.i.d. observations, $(X_1, \ldots, X_{N_i})$, from a specific model, $\mathcal{X}$. In this framework the statistical model for the largest rainfall in the i-th year, $Y_i = \max(X_1, \ldots, X_{N_i})$, is the $\mathcal{Y} = \max_{\mathcal{N}}(\mathcal{X})$ considered in Definition 2 for that $\mathcal{N}$ and that $\mathcal{X}$.

In particular, for simplicity here it will be assumed that the stopping model for the number of rain events, N_p, is the Logarithmic(p) model covered in Example 5.2 and in Appendix 1, and that the model for the rainfall in an event, X_λ, is the Exponential(λ) with cdf $F_{X_\lambda}(x) = 1 - e^{-\lambda x}$. In that case, the model for the largest rainfall of the year, $(Y_1, \ldots, Y_m)$, is the logarithmic stopped maximum of an exponential,

$$\mathcal{Y}_{Lg-Exp} = \{Y_{p,\lambda} : F_{Y_{p,\lambda}} = h_{N_p}(F_{X_\lambda}) = \frac{\log(1 - p + pe^{-\lambda y})}{\log(1-p)}, \lambda \in (0,\infty), p \in (0,1)\}.$$

To compare the use of the randomly stopped extreme models with the use of the generalized extreme-value model, we have simulated a sample for $m = 150$ years assuming

that N_i is Logarithmic$(p = .95)$ and X_i is Exponential$(\lambda = 0.01)$. We have fitted the true two-parameter $\mathcal{Y}_{Lg-Exp}$ model and the three parameter generalized extreme-value model,

$$\mathcal{Y}_{GEV} = \{Y_{\eta,\theta,\kappa} : F_{Y_{\eta,\theta,\kappa}} = e^{-(1-\kappa(x-\eta)/\theta)^{1/\kappa}}, \eta \in (-\infty,\infty), \theta \in (0,\infty), \kappa \in (-\infty,\infty)\},$$

on this data set by maximum likelihood. We have also fitted the $\mathcal{Y}_{TB2-Exp}$ and $\mathcal{Y}_{ETNB-Exp}$ models, which are the randomly stopped maxima of an exponential sample when the stopping model is the truncated binomial$(2,p)$ and the extended truncated negative binomial (ETNB) with pgf $h_N = \frac{\log(1-pt)^{-r}-1}{\log(1-p)^{-r}-1}$ where r is in $(-1,\infty)$. We also fit $\mathcal{Y}_{PC-LgNor}$, which is the randomly stopped maximum with $\mathcal{N}$ being the potential conjugate model considered in Example 6.1 and $\mathcal{X}$ being the lognormal model.

Table 1 presents the maximum likelihood estimates of the parameters of these five models together with the value of the log-likelihood at its maximum, and their AIC and BIC. Note that the $\mathcal{Y}_{ETNB-Exp}$ model fits the data slightly better than the actual $\mathcal{Y}_{Lg-Exp}$ model, but when $r = 0$ the $\mathcal{Y}_{ETNB-Exp}$ becomes the $\mathcal{Y}_{Lg-Exp}$ model and the likelihood ratio test between these two nested models does not reject the simpler actual model with a $p-val$ of 0.758.

Table 1. *Maximum likelihood parameter estimates, logarithm of the likelihood at the mle, and AIC and BIC for the five models considered for the data on the largest annual rainfall event.*

Model	N.par	MLE	loglikel	AIC	BIC
$\mathcal{Y}_{Lg-Exp}$	2	$\hat{p} = .9574$ $\hat{\lambda} = .0098$	-942.326	1888.65	1894.67
$\mathcal{Y}_{ETNB-Exp}$	3	$\hat{p} = .9844$ $\hat{r} = -.1177$ $\hat{\lambda} = .0108$	-942.172	1890.34	1899.38
$\mathcal{Y}_{TB2-Exp}$	2	$\hat{p} = .3949$ $\hat{\lambda} = .0063$	-945.196	1894.39	1900.41
$\mathcal{Y}_{PC-LgNor}$	3	$\hat{p} = .9352$ $\hat{\mu} = 4.9109$ $\hat{\sigma} = 1.1475$	-952.578	1911.15	1920.19
$\mathcal{Y}_{GEV}$	3	$\hat{\eta} = 129.01$ $\hat{\theta} = 111.25$ $\hat{\kappa} = -.1207$	-954.256	1914.51	1923.54

Even though the $\mathcal{Y}_{GEV}$ model has one more parameter than the actual $\mathcal{Y}_{Lg-Exp}$ model, it fits the simulated data significantly worse than this model, and worse than the other three stopped extreme models tried, even though two of these models assume a wrong stopping model and one of them assumes a wrong stopping and a wrong stopped model. Of course that will not always be the case, and the $\mathcal{Y}_{GEV}$ model will do better than other stopped extreme models, but when one has a good guess on what the stopping and the stopped models could be, the corresponding randomly stopped extreme model will tend to do better than $\mathcal{Y}_{GEV}$.

Note also that an important advantage of using randomly stopped extreme models is that through them one can interpret the estimated parameter values in terms of the parameters of the model for the stopping variable and the parameters of the model for the stopped variable. That provides useful information about the frequency of rain and

about the distribution of the amounts of rain in them, which is lacking when the analysis is based on the GEV model.

Remark: When $\mathcal{Y} = \max_{\mathcal{N}}(\mathcal{X})$ one has that $F_Y = h_N(F_X)$ and the pdf of Y is $f_Y = h_N^{'}(F_X)f_X$, where f_X is the pdf of X. Therefore f_Y is a weighted version of f_X and maximizing the likelihood function using data on Y is not any more complicated than doing it with data on X.

3.2. On the usefulness of randomly stopped extreme precursors

Lets assume here that one has data on the magnitude of the strongest earthquake on a given year for m_1 years, $(y_1,\ldots,y_{m_1})$, and data on the number of earthquakes in a year for m_2 years, $(n_1,\ldots,n_{m_2})$, where the set of years with available data might not coincide. Let's also assume that one has a good model $\mathcal{Y}$ for Y_i and a good model $\mathcal{N}$ for N_i.

Like in the previous example one can pose $Y_i = \max(X_1,\ldots,X_{N_i})$ where the magnitudes of the earthquakes, X_j, are i.i.d. realizations of a random variable, X, and hence one can assume that $\mathcal{Y} = \max_{\mathcal{N}}(\mathcal{X})$. In such a setting one might lack data about the X_j and yet the interest in the analysis might be to learn about the distribution of these X_j, and therefore about their cdf, F_X.

In particular, the stopping model for the number of earthquakes, N_i, could for example again be Logarithmic(p), and a good model for the magnitude of the strongest earthquake, Y_i, could be the GEV(η,θ,κ) model that was discarded in the previous example for the largest rainfall. If that was the case the magnitude of earthquakes, X_j, would be a sample from the random variable X that is the N_p-maxprecursor of the GEV r.v., $Y_{\eta,\theta,\kappa}$, and the cdf of X would be:

$$F_{X_{p,\eta,\theta,\kappa}} = F_{\max_{N_p}^{-1}(Y_{\eta,\theta,\kappa})} = h_{N_p}^{-1}(F_{Y_{\eta,\theta,\kappa}}).$$

Hence, by obtaining maximum likelihood estimates of p and of (η,θ,κ) and estimates of their standard deviations using the data on N_i and the data on Y_i one would obtain estimates and confidence intervals for the cdf of X, $\hat{F}_{X_{p,\eta,\theta,\kappa}} = h_{N_{\hat{p}}}^{-1}(F_{Y_{\hat{\eta},\hat{\theta},\hat{\kappa}}})$.

3.3. On the rationale behind using the extensions in Definition 3

Finally, lets assume that in either the hydrology or the seismology settings considered above one guesses that $\mathcal{N}_0$ is the stopping model for the number of events, N_i, and $\mathcal{X}_0$ is the model for the magnitude of the events X_j, but it turns that the statistical model $\mathcal{Y}_0 = \max_{\mathcal{N}_0}(\mathcal{X}_0)$ for $Y_i = \max(X_1,\ldots,X_{N_i})$ fails to fit properly the sample of extreme values available, $(y_1,\ldots,y_m)$.

In a case like this, if one is confident that $\mathcal{N}_0$ is the right stopping model one will want to extend $\mathcal{Y}_0$ by extending $\mathcal{X}_0$ while still using $\mathcal{N}_0$ as the stopping model. The first model extension in Definition 3 does that by replacing $\mathcal{Y}_0 = \max_{\mathcal{N}_0}(\mathcal{X}_0)$ by:

$$\mathcal{Y}_1 = \max_{\mathcal{N}_0}(\max_{\mathcal{N}_0}^{-1}(\mathcal{Y}_0)) = \{Y_{\xi,\delta_1,\delta_2} : F_{Y_{\xi,\delta_1,\delta_2}} = h_{N_{\delta_2}} \circ h_{N_{\delta_1}}^{-1}(F_{Y_\xi}),\ \xi \in \Xi, \delta_1, \delta_2 \in \mathcal{D}\},$$

where Ξ is the parameter space of $\mathcal{Y}_0$. In this way, the extended model can be posed as $\mathcal{Y}_1 = \max_{\mathcal{N}_0}(\mathcal{X}_1)$ where $\mathcal{X}_0$ has been replaced by its extension, $\mathcal{X}_1 = \max_{\mathcal{N}_0}^{-1}(\max_{\mathcal{N}_0}(\mathcal{X}_0))$. Note that this extension also applies when $\mathcal{Y}_0$ is chosen without making any $\mathcal{X}_0$ explicit, in which case the extended model is $\mathcal{Y}_1 = \max_{\mathcal{N}_0}(\mathcal{X}_1)$ with $\mathcal{X}_1 = \max_{\mathcal{N}_0}^{-1}(\mathcal{Y}_0)$.

By construction, the dimension of the parameter space of $\mathcal{Y}_1 = \max_{\mathcal{N}_0}(\max_{\mathcal{N}_0}^{-1}(\mathcal{Y}_0))$ is never smaller than the one of $\mathcal{Y}_1' = \max_{\mathcal{N}_0}^{-1}(\mathcal{Y}_0)$, which is never smaller than the one of $\mathcal{Y}_0$. This paper investigates when is the initial model always included in the transformed model, and when does repeated use of these extensions fail to keep extending the model.

4. Statistical stability of statistical model transformations

Transformations of a statistical model, $\mathcal{X}$, into a new model, $\mathcal{Y} = \mathcal{T}(\mathcal{X})$, can be classified depending on how initial and final models relate. Most often neither $\mathcal{X}$ nor $\mathcal{Y}$ are included into each other. The next definition distinguishes three possible relationships when they do.

Definition 4. *Let $\mathcal{T}(\cdot)$ transform a statistical model, $\mathcal{X}$, into $\mathcal{Y} = \mathcal{T}(\mathcal{X})$. Then*

1. *if $\mathcal{X} \subset \mathcal{T}(\mathcal{X})$, one says that $\mathcal{X}$ is extended by $\mathcal{T}(\cdot)$, and that $\mathcal{T}(\cdot)$ extends $\mathcal{X}$,*
2. *if $\mathcal{T}(\mathcal{X}) \subset \mathcal{X}$, one says that $\mathcal{X}$ is contracted by $\mathcal{T}(\cdot)$, and that $\mathcal{T}(\cdot)$ contracts $\mathcal{X}$,*
3. *if $\mathcal{T}(\mathcal{X}) = \mathcal{X}$, one says that $\mathcal{X}$ is invariant under $\mathcal{T}(\cdot)$.*

When $\mathcal{X}$ is extended by $\mathcal{T}(\cdot)$ for all $\mathcal{X}$, one says that $\mathcal{T}(\cdot)$ is a model extension. Most often, using a model extension repeatedly will keep extending the model, but some model extensions do not further extend models beyond their first use. These special model extensions are examples of the statistically stable transformations defined next.

Definition 5. *A statistical model transformation, $\mathcal{T}(\cdot)$, is said to be statistically stable if for any model $\mathcal{X}$ one has that $\mathcal{T}(\mathcal{X})$ is invariant under $\mathcal{T}(\cdot)$, and so if $\mathcal{T}(\mathcal{T}(\mathcal{X})) = \mathcal{T}(\mathcal{X})$ for any $\mathcal{X}$.*

When a model transformation is statistically stable, using that transformation twice in a row on any statistical model, $\mathcal{X}$, has the same effect as using it just once.

Definition 5 generalizes to any statistical model transformation the concept of geometric-extreme stability proposed in Marshall and Olkin (1997) in the special case of geometric stopped extreme transformations. Note that the statistical notion of stability presented here is different from probabilistic notions of stability, like the ones used in Rachev and Resnick (1991) or in Fama and Roll (1968), which apply to individual random variables and not to families of them.

The main purpose of the paper is to investigate the properties of the model transformations in Definitions 2 and 3, and to determine when do they work as model extensions,

and when are these model extensions statistically stable in the sense of Definition 5. This depends only on the characteristics of the stopping model, and in particular on whether they are extreme auto-reversible and/or closed under pgf composition, the way defined in the next two sections.

5. Stopping models that are extreme reversible or auto-reversible

5.1. Properties of h_N, $\overline{h}_N$, h_N^{-1}, and $\overline{h}_N^{-1}$ for positive count variables

A function, h_N, is the probability generating function of a positive integer-valued random variable N, if and only if it is real valued and such that $h_N(0) = 0$, that $h_N(1) = 1$, and that it is analytic at least on $[0,1)$, with all derivatives in that set being non-negative.

As a consequence, $\overline{h}_N(t) = 1 - h_N(1-t)$ is always such that $\overline{h}_N(0) = 0$, $\overline{h}_N(1) = 1$, and that it is analytic at least on $(0,1]$, with all of its odd derivatives in that set being non-negative, and all of its even derivatives non-positive. If all the moments of N are finite, analyticity and the declared signs of the derivatives of h_N and $\overline{h}_N$ hold at least on $[0,1]$.

From the characterization of h_N it also follows that h_N^{-1} and $\overline{h}_N^{-1}$ are always such that $h_N^{-1}(0) = \overline{h}_N^{-1}(0) = 0$ and $h_N^{-1}(1) = \overline{h}_N^{-1}(1) = 1$, and they are analytic at least on $(0,1)$ with a first derivative that is non-negative in that set. The second derivative of h_N^{-1} is non-positive, while the second derivative of $\overline{h}_N^{-1}$ is non-negative.

In particular, h_N, $\overline{h}_N$, h_N^{-1} and $\overline{h}_N^{-1}$ are always continuous and increasing on $[0,1]$, with h_N and $\overline{h}_N^{-1}$ being convex, and $\overline{h}_N$ and h_N^{-1} being concave.

For the limiting stopping random variable N_I with $\Pr(N_I = 1) = 1$, these four functions coincide, $h_{N_I}(t) = t = \overline{h}_{N_I}(t) = h_{N_I}^{-1}(t) = \overline{h}_{N_I}^{-1}(t)$. The next result will be used later on.

Proposition 1. *If N, N_1 and N_2 are positive integer valued random variables with pgfs h_N, h_{N_1} and h_{N_2}, then 1) $\overline{\overline{h}}_N = h_N$, 2) $\overline{h_N^{-1}} = \overline{h}_N^{-1}$, and 3) $\overline{h}_{N_1} \circ \overline{h}_{N_2} = \overline{(h_{N_1} \circ h_{N_2})}$.*

5.2. Extreme-reversible stopping models

As a consequence of the properties listed above, $\overline{h}_N$ and h_N^{-1} can only be the pgf of a positive integer valued random variable if $N = N_I$.

On the other hand, $\overline{h}_N^{-1}$ sometimes is the pgf of a non-degenerate positive integered random variable, N^*. That leads to the following definition.

Definition 6. *The pair of positive integer valued random variables, (N, N^*), is said to be extreme reversible if $\overline{h}_N^{-1} = h_{N^*}$, and therefore if $h_N^{-1} = \overline{h}_{N^*}$.*

When (N, N^*) are extreme reversible, their pgf's need to be such that:

$$h_{N^*} \circ \overline{h}_N(t) = \overline{h}_{N^*} \circ h_N(t) = t = \overline{h}_N \circ h_{N^*}(t) = h_N \circ \overline{h}_{N^*}(t), \quad \text{for } t \in [0,1],$$

and in that case, $\max_N^{-1}(X) = \min_{N^*}(X)$, $\min_N^{-1}(X) = \max_{N^*}(X)$, and therefore

$$X = \max_N(\min_{N^*}(X)) = \min_N(\max_{N^*}(X)) = \max_{N^*}(\min_N(X)) = \min_{N^*}(\max_N(X)).$$

It is important to emphasize that extreme reversibility is a property of (N, N^*), and that when it holds, this property applies for any real valued random variable, X.

Example 5.1: For the "potential conjugate" random variable N_b, with $h_{N_b} = 1-(1-t)^b$ for $b \in (0,1]$, one has that $\overline{h}_{N_b}^{-1} = t^{1/b}$, which is a pgf when $b = 1/m$ and m is a positive integer. Hence, for any positive integer m, the N_m with pgf $h_{N_m} = 1-(1-t)^{1/m}$, and N_m^* with pgf $h_{N_m^*} = t^m$, are extreme reversible.

Example 5.2: If N_α is zero-truncated Poisson(α), with:

$$h_{N_\alpha} = \frac{e^{\alpha t}-1}{e^{\alpha}-1}$$

for a given given $\alpha > 0$, then:

$$\overline{h}_{N_\alpha}^{-1} = -\frac{1}{\alpha}\ln(1-(1-e^{-\alpha})t) = h_{N_\alpha^*},$$

which is the pgf of a r.v. N_α^* with a Logarithmic(α) distribution, most often parametrized through $p = 1-e^{-\alpha}$. This means that each zero-truncated Poisson random variable is extreme reversible with one logarithmic random variable.

If a statistical model, $\mathcal{N}^*$, is the set of all random variables N^* that are extreme reversible with a random variable in $\mathcal{N}$, one says that $\mathcal{N}^*$ and $\mathcal{N}$ are a pair of extreme-reversible models.

Note that when $\mathcal{N}$ and $\mathcal{N}^*$ are extreme reversible one has that $\max_{\mathcal{N}}^{-1}(\cdot) = \min_{\mathcal{N}^*}(\cdot)$, and that $\min_{\mathcal{N}}^{-1}(\cdot) = \max_{\mathcal{N}^*}(\cdot)$, and one also has that:

$$\max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\cdot)) = \min_{\mathcal{N}^*}^{-1}(\min_{\mathcal{N}^*}(\cdot)),$$

$$\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\cdot)) = \min_{\mathcal{N}^*}(\min_{\mathcal{N}^*}^{-1}(\cdot)),$$

and viceversa. As an immediate consequence, when $\mathcal{N}$ and $\mathcal{N}^*$ are extreme-reversible models the set of transformations in Definitions 2 and 3 obtained with $\mathcal{N}$ and the set of transformations in these definitions obtained with $\mathcal{N}^*$ coincide.

5.3. Extreme auto-reversible stopping models

There are instances when N and N^* are the same, hence the next definition.

Definition 7. *The positive integer random variable N is extreme auto-reversible if* $\overline{h}_N^{-1} = h_N$, *and therefore if* $h_N^{-1} = \overline{h}_N$.

When N is extreme auto-reversible,

$$h_N \circ \overline{h}_N(t) = \overline{h}_N \circ h_N(t) = t, \text{ for } t \in [0,1],$$

which is a condition used in stochastic comparison theorems of Shaked (1975) and Shaked and Wong (1997). When it holds, $\max_N^{-1}(X) = \min_N(X)$, $\min_N^{-1}(X) = \max_N(X)$, and:

$$X = \max_N(\min_N(X)) = \min_N(\max_N(X)).$$

A necessary condition for a r.v. N to be auto-reversible is that $\Pr(N=1) = 1/E[N]$. The next result, providing a way to generate two auto-reversible random variables starting from any pair of reversible ones, will be used to find examples of auto-reversible variables.

Proposition 2. *If (N, N^*) are a pair of extreme-reversible random variables, with pgf's h_N and $h_{N^*} = \overline{h}_N^{-1}$, then the random variables N_1 and N_2, with pgfs $h_{N_1} = h_N \circ h_{N^*}$ and $h_{N_2} = h_{N^*} \circ h_N$, are both extreme auto-reversible.*

Proof: Given that $\overline{h}_N(t) = 1 - h_N(1-t)$, one has that:

$$h_{N_1} \circ \overline{h}_{N_1}(t) = h_N \circ h_{N^*} \circ \overline{h}_N \circ \overline{h}_{N^*}(t) = h_N \circ \overline{h}_{N^*}(t) = t,$$

where the last two steps use the fact that N and N^* are extreme reversible. ■

Corollary 1. *If N is extreme auto-reversible with pgf h_N, then the random variable N_3 with pgf $h_{N_3} = h_N \circ h_N$ is also extreme auto-reversible.*

Example 5.3: Using Proposition 2 with the random variables of Example 5.1 leads to $h_{N_1} = 1 - (1-t^m)^{1/m}$ and to $h_{N_2} = (1-(1-t)^{1/m})^m$, which whenever m is a positive integer are the pgf's of two auto-reversible random variables.

Example 5.4: Using Proposition 2 with the random variables of Example 5,2 yields:

$$h_{N_1} = \frac{pt}{1-(1-p)t},$$

for $0 < p = e^{-\alpha} \leq 1$, which is the pgf of the geometric distribution, and

$$h_{N_2} = 1 - (1/\alpha)\log(1 + e^{\alpha} - e^{\alpha t}),$$

for $\alpha > 0$, where N_1 and N_2 are extreme auto-reversible random variables.

When all random variables N in $\mathcal{N}$ are extreme auto-reversible, one says that the stopping model $\mathcal{N}$ is extreme auto-reversible.

When $\mathcal{N}$ is an extreme auto-reversible model one has that $\max_{\mathcal{N}}^{-1}(\cdot) = \min_{\mathcal{N}}(\cdot)$ and that $\min_{\mathcal{N}}^{-1}(\cdot) = \max_{\mathcal{N}}(\cdot)$, and therefore that:

$$\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\cdot)) = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\cdot)) = \min_{\mathcal{N}}(\max_{\mathcal{N}}(\cdot)),$$

$$\min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\cdot)) = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\cdot)) = \max_{\mathcal{N}}(\min_{\mathcal{N}}(\cdot)).$$

Therefore, when $\mathcal{N}$ is an extreme auto-reversible model, the four basic and four combined transformations in Definitions 2 and 3 collapse down into two basic and two combined transformations.

6. Stopping models closed under pgf composition

A necessary condition for the transformations in Definitions 2 and 3 to be statistically stable is that the corresponding stopping model be closed under pgf composition as defined next.

Definition 8. *The stopping model* $\mathcal{N} = \{N_\delta : h_{N_\delta},\ \delta \in \mathcal{D}\}$ *is said to be closed under pgf composition, if having* N_{δ_1} *and* N_{δ_2} *with pgfs* $h_{N_{\delta_1}}$ *and* $h_{N_{\delta_2}}$ *belonging to* $\mathcal{N}$*, implies that* N_{δ_3} *with pgf* $h_{N_{\delta_3}} = h_{N_{\delta_1}} \circ h_{N_{\delta_2}}$ *also belongs to* $\mathcal{N}$.

Requiring that $\mathcal{N}$ be closed under pgf composition is equivalent to requiring that if N_{δ_1} and N_{δ_2} belong to $\mathcal{N}$, then the N_{δ_1}-stopped sum of N_{δ_2} also belongs to $\mathcal{N}$, and it is thus equivalent to being closed under model compounding.

6.1. Uniparametric stopping models closed under pgf composition

Here we restrict consideration to stopping models, $\mathcal{N} = \{N_\delta : h_{N_\delta} = \sum_{i=1}^{\infty} p_i(\delta)t^i,\ \delta \in \mathcal{D}\}$, that i) are closed under pgf composition, ii) have a parametrization δ such that the $p_i(\delta) = \Pr(N_\delta = i)$ are continuously differentiable in δ for any i, and iii) have a parameter space, $\mathcal{D}$, that is a connected subset of $\mathbb{R}$ with a non-empty interior. From now on, this class of stopping models is denoted in a shorthanded way just as "*models uniparametric and closed under pgf composition*."

By focusing on stopping models continuously differentiable and with this kind of parameter space, we restrict consideration to the kind of stopping models useful in statistical practice. In particular, we essentially require that the parameter space be a non-empty interval, thus avoiding stopping models closed under pgf composition like $\mathcal{N} = \{N_k : h_{N_k} = t^k,\ k \in \mathbb{N}^+\}$, which lead to trivially stable transformations, and we also avoid parameter spaces with isolated points.

The following result, crucial in all that follows, is proved in Appendix 2 (Supplementary material).

Theorem 1. *If a stopping model,* $\mathcal{N} = \{N_\delta :\ \delta \in \mathcal{D}\}$*, is "*uniparametric and closed under pgf composition*" as defined above, then:*

1. $p_1(\delta) = \Pr(N_\delta = 1) > 0$ *for all* $N_\delta \in \mathcal{N}$,
2. $\mathcal{N}$ *can be parametrized in an identifiable way through* $\theta = \Pr(N_\delta = 1)$*, or through* $\eta = -\log \Pr(N_\delta = 1)$*, and*

3. *the parameter space is of the form* $(0,\theta_0]$ *for a given* $\theta_0 \leq 1$ *when using* θ*, and it is of the form* $\mathcal{H} = [\eta_0,\infty)$ *for a given* $\eta_0 \geq 0$ *when using* η.

From now on, we will always use η as the parametrization for models *uniparametric and closed under pgf composition*. Note that N_I, with $h_{N_I}(t) = t$, belongs to one of these models if, and only if, the lower limit of the parameter space, η_0, is equal to 0.

Next consequence of Theorem 1 relates to repeated use of the transformations in Definition 2.

Theorem 2. *If the stopping model,* $\mathcal{N} = \{N_\eta : h_{N_\eta},\ \eta \in [\eta_0,\infty)\}$*, is "*uniparametric and closed under pgf composition*" as defined above, then:*

1. $h_{N_{\eta_1}} \circ h_{N_{\eta_2}} = h_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h_{N_{\eta_1+\eta_2}}$,
2. $\overline{h}_{N_{\eta_1}} \circ \overline{h}_{N_{\eta_2}} = \overline{h}_{N_{\eta_2}} \circ \overline{h}_{N_{\eta_1}} = \overline{h}_{N_{\eta_1+\eta_2}}$,
3. $h^{-1}_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}} = h^{-1}_{N_{\eta_2}} \circ h^{-1}_{N_{\eta_1}} = h^{-1}_{N_{\eta_1+\eta_2}}$,
4. $\overline{h^{-1}}_{N_{\eta_1}} \circ \overline{h^{-1}}_{N_{\eta_2}} = \overline{h^{-1}}_{N_{\eta_2}} \circ \overline{h^{-1}}_{N_{\eta_1}} = \overline{h^{-1}}_{N_{\eta_1+\eta_2}}$,

for all $\eta_1,\eta_2 \in [\eta_0,\infty)$.

Proof: Given that $\mathcal{N}$ is closed under pgf composition, $h_{N_{\eta_1}} \circ h_{N_{\eta_2}} = h_{N_\eta}$, with:

$$\eta = -\log((h_{N_{\eta_1}}(h_{N_{\eta_2}}(t)))'_{|t=0}) = -\log((h'_{N_{\eta_1}}(h_{N_{\eta_2}}(t))h'_{N_{\eta_2}}(t))_{|t=0}) = \eta_1+\eta_2,$$

and commutativity follows from the commutativity of addition. The other three assertions follow from the fact that, because of Proposition 1,

$$\overline{h}_{N_{\eta_1}} \circ \overline{h}_{N_{\eta_2}} = \overline{(h_{N_{\eta_1}} \circ h_{N_{\eta_2}})} = \overline{h}_{N_{\eta_1+\eta_2}},$$

$$h^{-1}_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}} = (h_{N_{\eta_2}} \circ h_{N_{\eta_1}})^{-1} = h^{-1}_{N_{\eta_1+\eta_2}},$$

and

$$\overline{h^{-1}}_{N_{\eta_1}} \circ \overline{h^{-1}}_{N_{\eta_2}} = \overline{(h_{N_{\eta_2}} \circ h_{N_{\eta_1}})^{-1}} = \overline{h^{-1}}_{N_{\eta_1+\eta_2}}.$$

■

The second result that follows from Theorem 1 will imply that under stopping models closed under pgf composition, Definition 3 yields only two distinct extensions, and that the basic transformations in Definition 2 are restricted versions of them.

Theorem 3. *If the stopping model,* $\mathcal{N} = \{N_\eta : h_{N_\eta},\ \eta \in [\eta_0,\infty)\}$*, is "*uniparametric and closed under pgf composition*" as defined above, then:*

$$h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}}$$

for all $\eta_1, \eta_2 \in [\eta_0, \infty)$. *Furthermore,* $H_{\mathcal{N}}(t;\eta_1,\eta_2) = h_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}}(t)$ *can be parametrized in an identifiable way through* $\eta = -\log(H'_{\mathcal{N}}(0;\eta_1,\eta_2)) = \eta_1 - \eta_2$, *and if one denotes* $H_{\mathcal{N},\eta}(t) = H_{\mathcal{N}}(t;\eta_1,\eta_2)$ *with* $\eta \in \mathbb{R}$, *then*

1. $H_{\mathcal{N},\eta} \circ H_{\mathcal{N},\eta'} = H_{\mathcal{N},\eta+\eta'}$ *for all* $\eta, \eta' \in \mathbb{R}$,
2. *when* $\eta \geq \eta_0$, *then* $H_{\mathcal{N},\eta} = h_{N_\eta}$,
3. *when* $\eta \geq \eta_0$, *then* $H_{\mathcal{N},-\eta} = h^{-1}_{N_\eta}$,
4. $H_{\mathcal{N},\eta=0}(t) = t$.

Likewise, $\overline{h}_{N_{\eta_1}} \circ \overline{h}^{-1}_{N_{\eta_2}} = \overline{h}^{-1}_{N_{\eta_2}} \circ \overline{h}_{N_{\eta_1}}$, *and the properties listed above also apply for* $\overline{H}_{\mathcal{N},\eta}(t) = \overline{H}_{\mathcal{N}}(t;\eta_1,\eta_2) = \overline{h}_{N_{\eta_1}} \circ \overline{h}^{-1}_{N_{\eta_2}}(t) = 1 - H_{\mathcal{N},\eta}(1-t)$.

Proof: The commutativity for $\eta_1, \eta_2 \in [\eta_0, \infty)$ follows from:

$$h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} \circ h_{N_{\eta_2}} \circ h^{-1}_{N_{\eta_2}} = h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_2}} \circ h_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}} = h_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}}.$$

$H_{\mathcal{N}}(t;\eta_1,\eta_2)$ can be parametrized through $\eta = -\log(H'_{\mathcal{N}}(0;\eta_1,\eta_2)) = \eta_1 - \eta_2$ because

$$H'_{\mathcal{N}}(0;\eta_1,\eta_2) = \left(h^{-1}_{N_{\eta_2}}\right)'(0) \cdot h'_{N_{\eta_1}}(0) = \frac{1}{h'_{N_{\eta_2}}(0)} h'_{N_{\eta_1}}(0) = e^{-(\eta_1-\eta_2)},$$

and if $\eta = \eta_1 - \eta_2 = \eta'_1 - \eta'_2$ with $\eta_1, \eta_2, \eta'_1, \eta'_2 \geq \eta_0$, then:

$$h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_2}} \circ h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_2}} \circ h_{N_{\eta_1}} \circ h^{-1}_{N_{\eta_2}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta_1}+N_{\eta'_2}} \circ h^{-1}_{N_{\eta_2}} =$$

$$h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_1}+N_{\eta_2}} \circ h^{-1}_{N_{\eta_2}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_1}} \circ h_{N_{\eta_2}} \circ h^{-1}_{N_{\eta_2}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_1}},$$

and because if $h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}} = h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_1}}$ with $\eta_1, \eta_2, \eta'_1, \eta'_2 \geq \eta_0$, then:

$$\left(h^{-1}_{N_{\eta_2}} \circ h_{N_{\eta_1}}\right)'_{|t=0} = \left(h^{-1}_{N_{\eta'_2}} \circ h_{N_{\eta'_1}}\right)'_{|t=0},$$

and so $e^{-(\eta_1-\eta_2)} = e^{-(\eta'_1-\eta'_2)}$ and $\eta_1 - \eta_2 = \eta'_1 - \eta'_2$. To prove the additivity of $H_{\mathcal{N},\eta} \circ H_{\mathcal{N},\eta'}$, let $\beta \geq \eta_0 + \max(|\eta|, |\eta'|)$ and note that:

$$H_{\mathcal{N},\eta} \circ H_{\mathcal{N},\eta'} = (h^{-1}_{N_\beta} \circ h_{N_{\beta+\eta}}) \circ (h^{-1}_{N_\beta} \circ h_{N_{\beta+\eta'}}) =$$

$$h^{-1}_{N_\beta} \circ h^{-1}_{N_\beta} \circ h_{N_{\beta+\eta}} \circ h_{N_{\beta+\eta'}} = h^{-1}_{N_{2\beta}} \circ h_{N_{2\beta+\eta+\eta'}} = H_{\mathcal{N},\eta+\eta'}.$$

Furthermore, letting $\beta \geq \eta_0$ one has that for any $\eta \geq \eta_0$:

$$H_{\mathcal{N},\eta} = h_{N_\beta}^{-1} \circ h_{N_{\beta+\eta}} = h_{N_\beta}^{-1} \circ h_{N_\beta} \circ h_{N_\eta} = h_{N_\eta},$$

$$H_{\mathcal{N},-\eta} = h_{N_{\beta+\eta}}^{-1} \circ h_{N_\beta} = h_{N_\eta}^{-1} \circ h_{N_\beta}^{-1} \circ h_{N_\beta} = h_{N_\eta}^{-1},$$

and that, $H_{\mathcal{N},\eta=0}(t) = h_{N_\beta}^{-1} \circ h_{N_\beta}(t) = t$. ■

By using $H_{\mathcal{N},\eta}$ or $\overline{H}_{\mathcal{N},\eta}$ with $\eta \in \mathbb{R}$ in a model extension of Definition 3, one extends the parameter space through values of η in the whole real line and not just in $\mathcal{H} = [\eta_0, \infty)$.

Many stopping models satisfy the consequences of Theorem 1 without being closed under pgf composition. Next, an extra necessary condition for being a stopping model closed under pgf composition is obtained by imposing that the t^2 coefficients of the series expansion of $h_{N_{\eta_2}}^{-1} \circ h_{N_{\eta_1}}$ and of $h_{N_{\eta_1}} \circ h_{N_{\eta_2}}^{-1}$ have to be equal for any $\eta_1, \eta_2 \in [\eta_0, \infty)$. Imposing that higher order term coefficients of these expansions are equal leads to other necessary conditions.

Corollary 2. *If a stopping model* $\mathcal{N} = \{N_\eta : h_{N_\eta},\ \eta \in [\eta_0, \infty)\}$ *is closed under pgf composition,*

$$\frac{\Pr(N_\eta = 2)}{\Pr(N_\eta = 1)(1 - \Pr(N_\eta = 1))} = C, \ \textit{for all}\ \ \eta \in [\eta_0, \infty).$$

6.2. Examples of stopping models closed under pgf composition

Example 6.1: The *potential conjugate* model,

$$\mathcal{N} = \{N_p : h_{N_p} = 1 - (1-t)^p,\ \ p \in (0,1]\},$$

is closed under pgf composition with $E[N_p] = \infty$ and $\Pr(N_p = 1) = p$, and therefore with $\eta = -\log p \in [0, \infty)$. It includes N_I but it is not auto-reversible as described in Section 5.3.

Example 6.2: The zero-truncated geometric model,

$$\mathcal{N} = \{N_p : h_{N_p} = \frac{pt}{1-(1-p)t},\ \ p \in (0,1]\},$$

is closed under pgf composition, with $\Pr(N_p = 1) = p$ and $\eta = -\log p \in [0, \infty)$. It includes N_I and, as indicated in Example 5.4, it is auto-reversible.

The next result provides a way of generating a new model closed under pgf composition, starting from an initial model closed under pgf composition and two random variables that do not belong to the initial model but whose pgf composition does.

Proposition 3. *Let the stopping model* $\mathcal{N} = \{N_\eta : h_{N_\eta},\ \eta \in [\eta_0, \infty)\}$ *be closed under pgf composition, and let* N_1 *and* N_2 *be two random variables that do not belong to* $\mathcal{N}$ *but such that* $h_{N_1} \circ h_{N_2} = h_{N_\alpha}$ *with* $N_\alpha \in \mathcal{N}$*. Then, for any given* $\alpha > 0$ *the statistical model*

$$\mathcal{N}_\alpha = \{\tilde{N}_\eta : h_{\tilde{N}_\eta} = h_{N_2} \circ h_{N_{\eta-\alpha}} \circ h_{N_1},\ \ \eta \in [\alpha + \eta_0, \infty)\}$$

is also closed under pgf composition.

Proof: If $\tilde{N}_{\eta_1}$ and $\tilde{N}_{\eta_2}$ are random variables that belong to $\mathcal{N}_\alpha$, then

$$h_{\tilde{N}_{\eta_1}} \circ h_{\tilde{N}_{\eta_2}} = h_{N_2} \circ h_{N_{\eta_1-\alpha}} \circ h_{N_1} \circ h_{N_2} \circ h_{N_{\eta_2-\alpha}} \circ h_{N_1} =$$

$$h_{N_2} \circ h_{N_{\eta_1-\alpha}} \circ h_{N_\alpha} \circ h_{N_{\eta_2-\alpha}} \circ h_{N_1} = h_{N_2} \circ h_{N_{\eta_1+\eta_2-\alpha}} \circ h_{N_1},$$

which is the pgf of a random variable $\tilde{N}_{\eta_1+\eta_2}$ that also belongs to $\mathcal{N}_\alpha$. ■

Using Proposition 3 twice in a row does not generate any new family of models.

Next, this result is used to generate three families of stopping models closed under pgf composition starting from the geometric model.

Example 6.3: If N_1 is zero-truncated Poisson(α) and N_2 is Logarithmic(α), as in Example 5.2, then $h_{N_1} \circ h_{N_2}$ is the pgf of a Geometric($p = e^{-\alpha}$) and by Proposition 3 one has that for any given value of $\alpha > 0$ the statistical model:

$$\mathcal{N}_\alpha = \{N_\eta : h_{N_\eta} = \frac{1}{\alpha} \ln\left(1 + \frac{(e^{\alpha t} - 1)(e^\alpha - 1)}{(e^\eta - 1)(e^\alpha - e^{\alpha t}) + e^\alpha - 1}\right),\ \ \eta \in [\alpha, \infty)\},$$

is closed under pgf composition. In the limit, when α tends to 0 this model becomes the geometric model, and when α tends to ∞ it becomes N_I. The model $\mathcal{N}_\alpha$ is extreme auto-reversible for every α, but it only includes N_I in the limiting cases mentioned.

Example 6.4: Let N_1 be zero-truncated Negative-Binomial($\alpha\beta, \beta$), with:

$$h_{N_1} = \frac{(1 - (1 - e^{-\frac{\alpha}{\beta}})t)^{-\beta} - 1}{e^\alpha - 1},$$

and let N_2 be extended truncated Negative-Binomial($\alpha, -1/\beta$) in Engen (1974), with:

$$h_{N_2} = \frac{(1 - (1 - e^{-\alpha})t)^{\frac{1}{\beta}} - 1}{e^{-\frac{\alpha}{\beta}} - 1},$$

where $\alpha \geq 0$ and $\beta \geq 1$. Then $h_{N_1} \circ h_{N_2}$ is the pgf of a Geometric($p = e^{-\alpha}$), and by Proposition 3 one has that given any $\alpha \geq 0$ and $\beta \geq 1$ the statistical model:

$$\mathcal{N}_{\alpha,\beta} = \{N_\eta : h_{N_\eta} = \frac{1 - \left(\frac{(-e^{\alpha+\eta}+1)(1-t+te^{-\frac{\alpha}{\beta}})^\beta + e^\eta - 1}{(e^\alpha - e^{\alpha+\eta})(1-t+te^{-\frac{\alpha}{\beta}})^\beta + e^\eta - e^\alpha}\right)^{\frac{1}{\beta}}}{1 - e^{-\frac{\alpha}{\beta}}},\ \ \eta \in [\alpha, \infty)\},$$

is closed under pgf composition. When β tends to ∞ one obtains the models in Example 6.3, and when α tends to 0 or β converges to 1 one obtains the geometric model in Example 6.2. Other than in these limiting cases, $\mathcal{N}_{\alpha,\beta}$ is neither extreme auto-reversible, nor includes N_I.

Example 6.5: Let N_1 be zero-truncated Binomial$(n, p = 1 - e^{-\alpha/n})$, with:

$$h_{N_1} = \frac{(1 + (e^{\frac{\alpha}{n}} - 1)t)^n - 1}{e^{\alpha} - 1},$$

and let N_2 be zero-truncated Negative-Binomial$(\alpha, 1/n)$, with:

$$h_{N_2} = \frac{(1 - (1 - e^{-\alpha})t)^{-\frac{1}{n}} - 1}{e^{\frac{\alpha}{n}} - 1},$$

where $\alpha \geq 0$ and $n \in \mathbb{N}^+$. Then, $h_{N_1} \circ h_{N_2}$ is the pgf of a Geometric$(p = e^{-\alpha})$, and by Proposition 3 one has that for any given $\alpha \geq 0$ and $n \in \mathbb{N}^+$ the statistical model

$$\mathcal{N}_{\alpha,n} = \{N_\eta : h_{N_\eta} = \frac{\left(\frac{(\mathrm{e}^{\eta}-1)(-t+1+t\mathrm{e}^{\frac{\alpha}{n}})^n - \mathrm{e}^{\alpha+\eta}+1}{(-\mathrm{e}^{\alpha}+\mathrm{e}^{\eta})(-t+1+t\mathrm{e}^{\frac{\alpha}{n}})^n + \mathrm{e}^{\alpha} - \mathrm{e}^{\alpha+\eta}}\right)^{-\frac{1}{n}} - 1}{\mathrm{e}^{\frac{\alpha}{n}} - 1},\ \ \eta \in [\alpha, \infty)\},$$

is closed under pgf composition. In the limit, when n converges to ∞ one obtains the models in Example 6.3, and when α converges to 0, or when n is 1, one obtains the geometric model in Example 6.2. Other than in these limiting cases, $\mathcal{N}_{\alpha,n}$ is neither extreme auto-reversible, nor includes N_I, but it is extreme reversible with the $\mathcal{N}_{\alpha,\beta=n}$ in Example 6.4.

The next result provides a way of generating a family of statistical models closed under pgf composition starting from any model that is like that.

Proposition 4. *If the stopping model* $\mathcal{N} = \{N_\eta : h_{N_\eta}(t),\ \eta \in [\eta_0, \infty)\}$ *is closed under pgf composition then, for every given* $k \in \mathbb{N}^+$, *the statistical model*

$$\mathcal{N}_k = \{\tilde{N}_\eta : h_{\tilde{N}_\eta}(t) = \left(h_{N_\eta}(t^k)\right)^{1/k},\ \ \eta \in [\eta_0, \infty)\}$$

is also closed under pgf composition.

Using this result with Examples 6.1 and 6.2 one has that for every $k \in \mathbb{N}^+$ the models

$$\mathcal{N}_k = \{N_p : h_{N_p} = \left(1 - \left(1 - t^k\right)^p\right)^{1/k},\ \ p \in (0, 1]\},$$

and

$$\mathcal{N}_k = \{N_p : h_{N_p} = \left(\frac{pt^k}{1 - (1 - p)t^k}\right)^{1/k},\ \ p \in (0, 1]\},$$

are closed under pgf composition with $\eta = -(1/k)\log p$ and support $n = 1, k+1, 2k+1, \ldots$.

Finally we present a family of statistical models closed under pgf composition that embed Examples 6.1 and 6.2 as limiting cases and all include N_I.

Example 6.6: Given any value $\alpha \in (0,1)$, the statistical model

$$\mathcal{N}_\alpha = \{N_p : h_{N_p} = 1 - \frac{1-t}{(p+(1-p)(1-t)^\alpha)^{1/\alpha}},\ \ p \in (0,1]\},$$

is closed under pgf composition with $E[N_p] = p^{-1/\alpha}$, with $Var[N_p] = \infty$ and with $\eta = -\log p \in [0,\infty)$, and it contains N_I. In the limit, when α tends to 0 one obtains the model in Example 6.1, and when α tends to 1 one obtains the model in Example 6.2.

7. Randomly stopped extreme-based model transformations

7.1. Model transformations in Definitions 2 and 3

Given the properties of h_N and of $\overline{h}_N^{-1}$ it follows that $F_{\max_N(X)}(y) \leq F_X(y)$ and $F_{\min_N^{-1}(X)}(y) \leq F_X(y)$ for all y in their domain, and therefore that $\max_N(X)$ and $\min_N^{-1}(X)$ are random variables always larger than X in the usual stochastic order. Furthermore, given the properties of $\overline{h}_N$ and of h_N^{-1} it follows that $F_{\min_N(X)}(y) \geq F_X(y)$ and $F_{\max_N^{-1}(X)}(y) \geq F_X(y)$, and therefore that $\min_N(X)$ and $\max_N^{-1}(X)$ are always smaller than X in that stochastic order.

Hence, two of the basic transformations of Definition 2 transform any model $\mathcal{X} = \{X_\theta,\ \theta \in \Theta\}$ into a model $\mathcal{Y} = \{Y_{\theta,\delta},\ \theta \in \Theta, \delta \in \mathcal{D}\}$ with random variables $Y_{\theta,\delta}$ stochastically larger than X_θ, while the other two transform $\mathcal{X}$ into a model with $Y_{\theta,\delta}$ stochastically smaller than X_θ.

The four combined transformations of Definition 3 transform $\mathcal{X}$ into a model $\mathcal{Y}$ with random variables $Y_{\theta,\delta_1,\delta_2}$ that can be stochastically larger and smaller than X_θ.

By construction, the dimension of the parameter space of models obtained through transformations in Definition 3 is never smaller than the dimension of the parameter space of models obtained through transformations in Definition 2, which in turn is never smaller than the dimension of the parameter space of the initial model. We next investigate when is the initial model always included in the transformed model, and when does repeated use of these extensions leave the extended model unchanged.

7.2. When do transformations work as extensions?

A sufficient condition for basic transformations in Definition 2 to work as extensions for any model, $\mathcal{X}$, is that the identity belongs to the stopping model.

Proposition 5. *If $N_I \in \mathcal{N}$, with $\Pr(N_I = 1) = 1$, then the four basic model transformations in Definition 2 work as a model extension of $\mathcal{X}$, for any $\mathcal{X}$.*

Proof: If N_I, with $h_{N_I}(t) = t$, belongs to $\mathcal{N}$, then $X \in \mathcal{X}$ implies that $X \in \max_{\mathcal{N}}(\mathcal{X})$, and so $\mathcal{X} \subset \max_{\mathcal{N}}(\mathcal{X})$. The same argument applies to the other three basic transformations. ■

If one starts with a single random variable, $\mathcal{X} = \{X\}$, then $N_I \in \mathcal{N}$ is necessary and sufficient for X to be included in $\max_{\mathcal{N}}(X)$ and in $\min_{\mathcal{N}}(X)$. In general though, one can find instances of specific models, $\mathcal{X}$, included in $\max_{\mathcal{N}}(\mathcal{X})$ or in $\min_{\mathcal{N}}(\mathcal{X})$ without N_I belonging to $\mathcal{N}$.

On the other hand, the four combined mechanisms of Definition 3 always work as model extensions, irrespective of whether N_I is in $\mathcal{N}$ or not.

Proposition 6. *The four model transformation in Definition 3 work as a model extension of $\mathcal{X}$, for any $\mathcal{X}$. That is so, even if one of the two new parameters, δ_1 or δ_2, is fixed.*

Proof: $F_{X_\theta} \in \mathcal{X}$ implies that $F_{Y_{\theta,\delta_1,\delta_2}} = h_{N_{\delta_2}} \circ h^{-1}_{N_{\delta_1}}(F_{X_\theta}) \in \max_{\mathcal{N}}(\max^{-1}_{\mathcal{N}}(\mathcal{X}))$ for all $\delta_1, \delta_2 \in \mathcal{D}$, and in particular $F_{X_\theta} = h_{N_{\delta_1}} \circ h^{-1}_{N_{\delta_1}}(F_{X_\theta}) \in \max_{\mathcal{N}}(\max^{-1}_{\mathcal{N}}(\mathcal{X}))$, which means that $\mathcal{X} \subset \max_{\mathcal{N}}(\max^{-1}_{\mathcal{N}}(\mathcal{X}))$. The same argument applies to the other three transformations, and when any of the two new parameters is fixed. ■

Different from the transformations in Definition 3, using $\max_{\mathcal{N}}(\cdot) \cup \min_{\mathcal{N}}(\cdot)$ with a stopping model $\mathcal{N}$ that does not include N_I does not always work as a model extension.

7.3. When are the extensions statistically stable?

Under general uniparametric stopping models, the basic transformations of Definition 2 usually add one dimension to the parameter space, while the combined transformations of Definition 3 usually add two dimensions to it.

Instead, when the stopping model is uniparametric and closed under pgf composition both basic as well as combined transformations add at most a single dimension, and the basic transformations of Definition 2 become restricted versions of the combined transformations of Definition 3 with the extra parameter, η, of the basic transformations taking values on a semi-line and the extra parameter, η, of the combined transformation taking values on the whole real line.

Furthermore, under general stopping models repeated use of these extensions usually keep extending the models. Instead, when the stopping model is closed under pgf composition and the transformation works as an extension, then it is always a statistically stable extension and hence repeated use of that extension leaves the extended model unchanged.

Proposition 7. *If the stopping model $\mathcal{N} = \{N_\eta : h_{N_\eta},\ \eta \in [\eta_0, \infty)\}$ is* "uniparametric and closed under pgf composition" *then the model transformations in Definition 2 are such that:*

1. *if $\eta_0 = 0$, then $\max_{\mathcal{N}}(\cdot)$, $\min_{\mathcal{N}}(\cdot)$, $\max^{-1}_{\mathcal{N}}(\cdot)$, and $\min^{-1}_{\mathcal{N}}(\cdot)$ are statistical model extensions that are statistically stable, and*

2. *if* $\eta_0 > 0$, *then* $\max_{\mathcal{N}}(\mathcal{X})$ *is contracted by* $\max_{\mathcal{N}}(\cdot)$, $\min_{\mathcal{N}}(\mathcal{X})$ *is contracted by* $\min_{\mathcal{N}}(\cdot)$, $\max_{\mathcal{N}}^{-1}(\mathcal{X})$ *is contracted by* $\max_{\mathcal{N}}^{-1}(\cdot)$, *and* $\min_{\mathcal{N}}^{-1}(\mathcal{X})$ *is contracted by* $\min_{\mathcal{N}}^{-1}(\cdot)$, *for all* $\mathcal{X}$.

Proof: By Theorem 2 one has that for any $\mathcal{X}$,

$$\mathcal{Y} = \max_{\mathcal{N}}(\max_{\mathcal{N}}(\mathcal{X})) = \{Y_{\theta,\eta_1,\eta_2} : F_{Y_{\theta,\eta_1,\eta_2}} = h_{N_{\eta_2}} \circ h_{N_{\eta_1}}(F_{X_\theta}),\ \theta \in \Theta, \eta_1, \eta_2 \in [\eta_0, \infty)\} =$$

$$\{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = h_{N_{\eta=\eta_1+\eta_2}}(F_{X_\theta}),\ \theta \in \Theta, \eta \in [2\eta_0, \infty)\} \subset$$

$$\{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = h_{N_\eta}(F_{X_\theta}),\ \theta \in \Theta, \eta \in [\eta_0, \infty)\} = \max_{\mathcal{N}}(\mathcal{X}),$$

and so if $\eta_0 > 0$, then $\max_{\mathcal{N}}(\cdot)$ contracts $\max_{\mathcal{N}}(\mathcal{X})$. When $\eta_0 = 0$,

$$\max_{\mathcal{N}}(\max_{\mathcal{N}}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = h_{N_\eta}(F_{X_\theta}),\ \theta \in \Theta, \eta \in [0, \infty)\} = \max_{\mathcal{N}}(\mathcal{X}),$$

which means that $\max_{\mathcal{N}}(\cdot)$ is a statistically stable extension. The same argument applies to the other three transformations in Definition 2. ■

The next result establishes that under uniparametric stopping models closed under pgf composition, there are only two distinct combined extensions and they are statistically stable.

Proposition 8. *If the stopping model* $\mathcal{N}$ *is "*uniparametric and closed under pgf composition*", then Definition 3 yields only two distinct model extensions which are:*

1. $\mathcal{Y}_1 = \max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\cdot)) = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\cdot))$, *and*
2. $\mathcal{Y}_2 = \min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\cdot)) = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\cdot))$,

and these two extensions are both statistically stable. Furthermore, in that case it holds that:

1. *the model* $\mathcal{Y}_1 = \max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X}))$ *is invariant under* $\max_{\mathcal{N}}(\cdot)$ *and* $\max_{\mathcal{N}}^{-1}(\cdot)$,
2. *the model* $\mathcal{Y}_2 = \min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X}))$ *is invariant under* $\min_{\mathcal{N}}(\cdot)$ *and* $\min_{\mathcal{N}}^{-1}(\cdot)$, *and*
3. *the transformations in Definition 2 are restricted versions of one of these two extensions.*

Proof: By Theorem 3, one has that for any $\mathcal{X}$,

$$\mathcal{Y}_1 = \max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X})) = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = H_{\mathcal{N},\eta}(F_{X_\theta}),\ \theta \in \Theta, \eta \in \mathcal{R}\},$$

and the stability follows from that same theorem, because:

$$\max_{\mathcal{N}}(\mathcal{Y}_1) = \{Y_{\theta,\eta+\eta'} : F_{Y_{\theta,\eta+\eta'}} = H_{\mathcal{N},\eta+\eta'}(F_{X_\theta}),\ \theta \in \Theta, \eta + \eta^{'} \in \mathcal{R}\} = \mathcal{Y}_1,$$

$$\max_{\mathcal{N}}^{-1}(\mathcal{Y}_1) = \{Y_{\theta,\eta-\eta'} : F_{Y_{\theta,\eta-\eta'}} = H_{\mathcal{N},\eta-\eta'}(F_{X_\theta}),\ \theta \in \Theta, \eta - \eta^{'} \in \mathcal{R}\} = \mathcal{Y}_1.$$

By Theorem 3 one also has that:

$$\mathcal{Y}_2 = \min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X})) = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \overline{H}_{\mathcal{N},\eta}(F_{X_\theta}),\ \theta \in \Theta, \eta \in \mathcal{R}\},$$

where $\overline{H}_{\mathcal{N},\eta}(t) = 1 - H_{\mathcal{N},\eta}(1-t)$, and stability follows likewise. ■

Corollary 3. *If $\mathcal{N}$ is "*uniparametric and closed under pgf composition*", then:*

1. $\max_{\mathcal{N}}(\mathcal{X}) \cup \max_{\mathcal{N}}^{-1}(\mathcal{X}) \subset \mathcal{Y}_1 = \max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X})) = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X}))$,
2. $\min_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}^{-1}(\mathcal{X}) \subset \mathcal{Y}_2 = \min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X})) = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X}))$,

and if $N_I \in \mathcal{N}$, then the models on the left and the models on the right are equal.

7.4. What happens with stopping models both closed and extreme reversible?

When two stopping models are closed under pgf composition and extreme reversible, Proposition 8 and the definition of extreme reversibility lead to the next result.

Proposition 9. *If the stopping models $\mathcal{N}$ and $\mathcal{N}^*$ are uniparametric, closed under pgf composition, and extreme reversible, then the two distinct statistically stable model extensions in Definition 3 obtained with $\mathcal{N}$ and the ones obtained with $\mathcal{N}^*$ are the same extensions.*

According to Proposition 8, when a stopping model is closed under pgf composition the four extensions in Definition 3 collapse down into two distinct ones. The next result, stating that when a stopping model is both closed and extreme auto-reversible then these two extensions become a single one, is a straight consequence of the definition of extreme-auto-reversibility.

Proposition 10. *If the stopping model $\mathcal{N}$ is uniparametric, closed under pgf composition, and extreme auto-reversible, then the four statistically stable model extensions in Definition 3 coincide,*

$$\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\cdot)) = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\cdot)) = \min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\cdot)) = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\cdot)),$$

and they coincide with $\min_{\mathcal{N}}(\max_{\mathcal{N}}(\cdot))$ *and with* $\max_{\mathcal{N}}(\min_{\mathcal{N}}(\cdot))$. *If on top of that, $N_I \in \mathcal{N}$ (i.e. $\eta_0 = 0$), this statistically stable model extension also coincides with* $\max_{\mathcal{N}}(\cdot) \cup \min_{\mathcal{N}}(\cdot)$.

The geometric stopping model satisfies all the conditions of Proposition 10. As a consequence, the Marshall-Olkin extension of $\mathcal{X}$, originally defined to be $\max_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}(\mathcal{X})$ when $\mathcal{N}$ is geometric, coincides with the extension of $\mathcal{X}$ obtained through Definition 3 with geometric stopping.

As a consequence, we consider the model extensions in Definition 3 to be the natural way to generalize the Marshall-Olkin extension with stopping models other than geometric. Different from what happens if one generalized Marshall-Olkin through $\max_{\mathcal{N}}(\cdot) \cup \min_{\mathcal{N}}(\cdot)$, by generalizing them through the transformations in Definition 3 one guarantees that these transformations will work as model extensions under any stopping model, $\mathcal{N}$.

8. Examples of statistically stable extensions

When one uses the model extensions of Definition 3 with stopping models that are neither closed under pgf composition nor extreme auto-reversible, one obtains four different extensions that are not statistically stable, and the four basic transformations of Definition 2 are not restricted versions of them. As an example, Appendix 1 presents the four basic and the four combined extensions obtained when $\mathcal{N}$ are the zero-truncated Poisson or the logarithmic models.

Here we present the model extensions in Definition 3 obtained when the stopping models are the ones presented in Section 6.2. Given that these stopping models are all closed under pgf composition, all the extensions obtained here are statistically stable in the sense that applying them twice on any given model leads to the same model as applying them once.

Furthermore, because of Proposition 8 another consequence of all these stopping models being closed under pgf composition is that for them Definition 3 yields at most two distinct extensions, and that the transformations in Definition 2 are restrictions of these two extensions and do not need to be considered apart.

In three of the examples, the stopping models are not auto-reversible, and for them the model extension in Definition 3 based on maxima extends $\mathcal{X} = \{X_\theta : F_{X_\theta},\ \theta \in \Theta\}$ into:

$$\mathcal{Y}_1 = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = H_{\mathcal{N},\eta}(F_{X_\theta}),\ \ \theta \in \Theta, \eta \in (-\infty,\infty)\},$$

and the extension in Definition 3 based on minima extends $\mathcal{X}$ into:

$$\mathcal{Y}_2 = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \overline{H}_{\mathcal{N},\eta}(F_{X_\theta}),\ \ \theta \in \Theta, \eta \in (-\infty,\infty)\},$$

with $H_{\mathcal{N},\eta}(\cdot)$ and $\overline{H}_{\mathcal{N},\eta}(\cdot)$ as in Theorem 3.

In the second and third examples the stopping models are auto-reversible, and hence for them these two extension mechanisms, $\mathcal{Y}_1$ and $\mathcal{Y}_2$, coincide because of Proposition 10. The fourth and fifth families of stopping models considered can be reversible, and

when they are reversible they lead to the same pair of model extensions because of Proposition 9.

Example 8.1: Let the stopping model be the one in Example 6.1,

$$\mathcal{N} = \{N_\eta : h_{N_\eta} = 1-(1-t)^{e^{-\eta}},\ \eta \in [0,\infty)\},$$

which is not extreme auto-reversible but it includes N_I.

The extension of $\mathcal{X}$ obtained through maxima and precursors of maxima is:

$$\mathcal{Y}_1 = \max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = 1-(1-F_{X_\theta})^{e^{-\eta}},\ \theta \in \Theta, \eta \in (-\infty,\infty)\},$$

which is a special case of the extension in Cordeiro and Castro (2009). When one restricts $\eta \geq 0$, here one obtains $\mathcal{Y}_1' = \max_{\mathcal{N}}(\mathcal{X})$, and when one restricts $\eta \leq 0$ one obtains $\mathcal{Y}_1'' = \max_{\mathcal{N}}^{-1}(\mathcal{X})$, and therefore in this case $\mathcal{Y}_1 = \max_{\mathcal{N}}(\mathcal{X}) \cup \max_{\mathcal{N}}^{-1}(\mathcal{X})$.

When $\mathcal{X}$ is for example an exponential random variable, $\mathcal{Y}_1$ becomes the exponential model. Because of the stability of this extension, using it again, now on the exponential model, will leave that model unchanged which means that the exponential model is invariant under this extension. On the other hand, if $\mathcal{X}$ is the logistic model, then $\mathcal{Y}_1$ is the type II generalized logistic model, which will also be invariant under this extension.

In general, when a statistical model is invariant under an extension that is stable, it is because that model can be obtained as the extension of a submodel of it.

The extension of $\mathcal{X}$ obtained through minima and precursors of minima is:

$$\mathcal{Y}_2 = \min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X})) = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = (F_{X_\theta})^{e^{-\eta}},\ \theta \in \Theta, \eta \in (-\infty,\infty)\},$$

which is a special case of the extension in Cordeiro, Ortega and Cunha (2013).

When one restricts $\eta \geq 0$, one obtains $\mathcal{Y}_2' = \min_{\mathcal{N}}(\mathcal{X})$, and when one restricts $\eta \leq 0$ one obtains $\mathcal{Y}_2'' = \min_{\mathcal{N}}^{-1}(\mathcal{X})$, and therefore here $\mathcal{Y}_2 = \min_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}^{-1}(\mathcal{X})$.

In this case, if $\mathcal{X}$ is for example the Gumbel model with the location parameter fixed, then $\mathcal{Y}_2$ is the two parameter Gumbel model, and because of the stability of this extension the two parameter Gumbel model model will be invariant under this extension. On the other hand, when $\mathcal{X}$ is the logistic model then $\mathcal{Y}_2$ becomes the type I generalized logistic model which by stability will also be invariant under this extension.

Example 8.2: Let the stopping model be the zero-truncated geometric in Example 6.2,

$$\mathcal{N} = \{N_\eta : h_{N_\eta} = \frac{t}{(1-t)e^\eta + t},\ \eta \in [0,\infty)\},$$

which is extreme auto-reversible and includes N_I.

As a consequence of this auto-reversibility the extension of $\mathcal{X}$ obtained through maxima and their precursors or through minima and their precursors here coincide, and it is:

$$\mathcal{Y} = \mathcal{Y}_1 = \mathcal{Y}_2 = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \frac{F_{X_\theta}}{(1-F_{X_\theta})e^\eta + F_{X_\theta}},\ \theta \in \Theta, \eta \in (-\infty,\infty)\},$$

which is the Marshall-Olkin extension of $\mathcal{X}$. There is a huge literature using this model extension. Here, when one restricts $\eta \geq 0$ one obtains $\mathcal{Y}' = \max_{\mathcal{N}}(\mathcal{X}) = \min_{\mathcal{N}}^{-1}(\mathcal{X})$, and when one restricts $\eta \leq 0$ one obtains $\mathcal{Y}'' = \min_{\mathcal{N}}(\mathcal{X}) = \max_{\mathcal{N}}^{-1}(\mathcal{X})$. As a consequence, this is the only example considered here where $\mathcal{Y} = \max_{\mathcal{N}}(\mathcal{X}) \cup \min_{\mathcal{N}}(\mathcal{X})$.

When for example $\mathcal{X}$ is the logistic model with the location parameter fixed the extended model, $\mathcal{Y}$, is the two parameter logistic model. Because of statistical stability of this extension, applying it again, now on the two-parameter logistic model, leaves the model unchanged, which means that this two parameter model is invariant under this extension.

Example 8.3: Let the stopping model be the $\mathcal{N}_\alpha$ in Example 6.3 for a given $\alpha \geq 0$. Like the geometric model, this one is also extreme auto-reversible, but it only includes N_I when $\alpha = 0$, which is when it becomes the geometric model.

As a consequence of this auto-reversibility, the extensions of $\mathcal{X}$ obtained through maxima and their precursors and the ones obtained through minima and their precursors coincide and are:

$$\mathcal{Y}_\alpha = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \frac{1}{\alpha} \ln\left(1 + \frac{(e^{\alpha F_{X_\theta}} - 1)(e^\alpha - 1)}{(e^\eta - 1)(e^\alpha - e^{\alpha F_{X_\theta}}) + e^\alpha - 1}\right),\ \theta \in \Theta, \eta \in (-\infty, \infty)\},$$

with $\mathcal{Y}'_\alpha = \max_{\mathcal{N}_\alpha}(\mathcal{X}) = \min_{\mathcal{N}_\alpha}^{-1}(\mathcal{X})$ when one restricts $\eta \geq \alpha$, and with $\mathcal{Y}''_\alpha = \min_{\mathcal{N}_\alpha}(\mathcal{X}) = \max_{\mathcal{N}_\alpha}^{-1}(\mathcal{X})$ when one restricts $\eta \leq -\alpha$. When one restricts $\eta \in (-\alpha, \alpha)$ one obtains $\mathcal{Y}'''_\alpha = \max_{\mathcal{N}_\alpha}(\max_{\mathcal{N}_\alpha}^{-1}(\mathcal{X})) = \max_{\mathcal{N}_\alpha}(\min_{\mathcal{N}_\alpha}(\mathcal{X}))$ with η_1, η_2 such that $\eta_2 - \eta_1 \in (-\alpha, \alpha)$, but this restricted transformation does not coincide with any of the transformations in Definition 2.

Different from what happens under the geometric model with $\alpha = 0$, when $\alpha > 0$ neither $\max_{\mathcal{N}_\alpha}(\mathcal{X})$ nor $\min_{\mathcal{N}_\alpha}(\mathcal{X})$ work as a model extension of $\mathcal{X}$, and $\max_{\mathcal{N}_\alpha}(\mathcal{X}) \cup \min_{\mathcal{N}_\alpha}(\mathcal{X}) \subset \mathcal{Y}_\alpha$ with an inclusion often strict. Hence this is an example where $\max_{\mathcal{N}_\alpha}(\mathcal{X}) \cup \min_{\mathcal{N}_\alpha}(\mathcal{X})$ does not work as a model extension of $\mathcal{X}$, but where using the $\mathcal{Y}_\alpha$ from Definition 3 does.

Example 8.4: Let the stopping model be the $\mathcal{N}_{\alpha,\beta}$ in Example 6.4 for a given $\alpha \geq 0$ and $\beta \geq 1$. Here N_I is not in the model, and the model is not auto-reversible and therefore it yields two different model extension mechanisms.

The extension of $\mathcal{X}$ obtained through maxima and their precursors is:

$$\mathcal{Y}_{1_{\alpha,\beta}} = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \frac{1 - \left(\frac{(1-e^{\alpha+\eta})\left(1-F_{X_\theta}\left(1-e^{-\frac{\alpha}{\beta}}\right)\right)^\beta + e^\eta - 1}{(e^\alpha - e^{\alpha+\eta})\left(1-F_{X_\theta}\left(1-e^{-\frac{\alpha}{\beta}}\right)\right)^\beta + e^\eta - e^\alpha}\right)^{\frac{1}{\beta}}}{1 - e^{-\frac{\alpha}{\beta}}},\ \theta \in \Theta, \eta \in (-\infty, \infty)\},$$

and here one obtains $\mathcal{Y}'_{1_{\alpha,\beta}} = \max_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X})$ when $\eta \geq \alpha$, and $\mathcal{Y}''_{1_{\alpha,\beta}} = \max_{\mathcal{N}_{\alpha,\beta}}^{-1}(\mathcal{X})$ when $\eta \leq -\alpha$. When $\eta \in (-\alpha, \alpha)$ one has that $\mathcal{Y}'''_{1_{\alpha,\beta}} = \max_{\mathcal{N}_{\alpha,\beta}}(\max_{\mathcal{N}_{\alpha,\beta}}^{-1}(\mathcal{X}))$ with η_1, η_2 such

that $\eta_2 - \eta_1 \in (-\alpha, \alpha)$, which can neither be obtained through $\max_{\mathcal{N}_{\alpha,\beta}}(\cdot)$ nor through $\max^{-1}_{\mathcal{N}_{\alpha,\beta}}(\cdot)$.

The model extension of $\mathcal{X}$ obtained through minima and their precursors is:

$$\mathcal{Y}_{2_{\alpha,\beta}} = \{Y_{\theta,\eta} : F_{Y_{\theta,\eta}} = \frac{1 - \left(\frac{(e^{\eta} - e^{\alpha})\left(1 + F_{X_\theta}\left(e^{\frac{\alpha}{\beta}} - 1\right)\right)^{\beta} + e^{\alpha} - e^{\alpha+\eta}}{(e^{\eta} - 1)\left(1 + F_{X_\theta}\left(e^{\frac{\alpha}{\beta}} - 1\right)\right)^{\beta} - e^{\eta+\alpha} + 1} \right)^{\frac{1}{\beta}}}{1 - e^{\frac{\alpha}{\beta}}}, \ \theta \in \Theta, \eta \in (-\infty, \infty)\},$$

and one obtains $\mathcal{Y}'_{2_{\alpha,\beta}} = \min_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X})$ when $\eta \geq \alpha$, and $\mathcal{Y}''_{2_{\alpha,\beta}} = \min^{-1}_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X})$ when $\eta \leq -\alpha$. When $\eta \in (-\alpha, \alpha)$ one has that $\mathcal{Y}'''_{2_{\alpha,\beta}} = \min_{\mathcal{N}_{\alpha,\beta}}(\min^{-1}_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X}))$ with η_1, η_2 such that $\eta_2 - \eta_1 \in (-\alpha, \alpha)$, which can neither be obtained through $\min_{\mathcal{N}_{\alpha,\beta}}(\cdot)$ nor through $\min^{-1}_{\mathcal{N}_{\alpha,\beta}}(\cdot)$.

Here $\max_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X}) \cup \max^{-1}_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X}) \subset \mathcal{Y}_{1\alpha,\beta}$, and $\min_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X}) \cup \min^{-1}_{\mathcal{N}_{\alpha,\beta}}(\mathcal{X}) \subset \mathcal{Y}_{2\alpha,\beta}$ with these inclusions being most often strict.

Example 8.5: Let the stopping model be the $\mathcal{N}_{\alpha,n}$ in Example 6.5 for a given $\alpha \geq 0$ and $n \in \mathbb{N}^+$. This model does not include N_I and it is not extreme auto-reversible, but it is reversible with the $\mathcal{N}_{\alpha,\beta=n}$ in Example 6.4. As a consequence, the model extension obtained with $\mathcal{N}_{\alpha,n}$ through maxima and precursors of maxima, coincide with the model extension obtained with the $\mathcal{N}_{\alpha,\beta=n}$ of Example 6.4 through minima and precursors of minima, and viceversa.

9. Final comments

The main contribution of this article is putting together a set of new concepts needed to define and untangle the properties of a large family of statistical model transformation mechanisms that lead to statistical models useful for the analysis of extreme-value data and in reliability. The concepts introduced are:

1. the notion of N-extreme precursors, which can be understood as the inverse of N-stopped maxima and minima, and the model extension mechanisms derived from them (Definitions 1, 2 and 3), which help generalize Marshall-Olkin extensions beyond geometric stopping,

2. the concept of statistical stability of a statistical model extension (Definition 5), which applies to any statistical model extension and not just to the ones considered in this paper,

3. the idea of extreme reversible and auto-reversible stopping models (Definitions 6 and 7), under which the extensions based on randomly stopped maxima and their

inverses coincide with the extensions based on randomly stopped minima and their inverses,

4. and the idea of stopping models closed under pgf composition (Definition 8), which are the ones leading to statistically stable randomly stopped extreme type of extensions.

All these new concepts are needed for the picture to be complete. In particular, if we touch on methods to generate stopping models that are auto-reversible and/or closed under pgf composition other than the geometric model, it is to help understand that the role played by geometric stopping is not as unique as one might think after reading Marshall Olkin (1997).

A second contribution of this article are a set of theoretical results stating that uniparametric stopping models closed under pgf composition can always be parametrized through $\theta = \Pr(N = 1)$ with a parameter space of the form $(0, \theta_0]$ (Theorem 1), and that the pgfs of these models commute under composition among themselves and with their inverses (Theorems 2 and 3). These results are then used in Section 7 to determine conditions leading to statistically stable extensions.

Only two of the families of statistically stable model extensions presented in Section 8 are based on stopping models that are both closed under pgf composition and extreme auto-reversible. And the geometric model is the only stopping model that we know that shares these two features and includes N_I. Nevertheless, note that in order to obtain statistically stable extensions through Definition 3, one only needs that the stopping model be closed under pgf composition.

The only consequence of using stopping models that, unlike the geometric model, are not extreme auto-reversible is that the extension based on maxima and their inverses does not coincide with the extension based on minima and their inverses, and using stopping models that, unlike geometric, do not include N_I does neither affect the statistical stability nor the fact that the transformations presented in Definition 3 always work as an extension.

Finally, note that our definition of statistical stability is extremely basic and fundamental. A statistical model transformation is statistically stable only if using that transformation twice in a row on any statistical model has the same effect as using that transformation just once. The only reason that we can think for not finding the notion of statistical stability anywhere in the statistical literature is that it might be difficult to prove results of that kind outside the specific context of randomly stopped extreme transformations, and the closely related area of randomly stopped sum transformations; It is easy to check that stopping models closed under pgf composition also lead to randomly stopped sum model extensions that are statistically stable.

We consider statistical stability to be a property that should be central in the study of any type of statistical model extension and not just in the study of the specific extensions considered here, and we intend to keep investigating that.

Acknowledgments

We are grateful for the comments made by the reviewers, which helped us improve this manuscript. This work was supported by the grant PID2021-125380OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe.

References

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N. (1992). *A First Course in Order Statistics*. New York, Wiley.

Consul, P.C. (1984). On the distributions of order statistics for a random sample size. *Statistica Neerlandica*, 38, 249-256.

Cordeiro, G.M., Castro, M. (2009). A new family of generalized distributions. *Journal of Statistical Computation & Simulation*, 81, 883-898.

Cordeiro, G.M., Ortega, E.M.M., Cunha, D.C.C. (2013). The exponential generalized class of distributions. *Journal of Data Science*, 11, 1-27.

Engen, (1974). On species frequency models. *Biometrika*, 61, 263-270.

Fama, E.F., Roll, R. (1968). Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, 63, 817-836.

Gupta, D., Gupta, R.C. (1984). On the distributions of order statistics for a random sample size. *Statistica Neerlandica*, 38, 13-19.

Louzada, F., Beret, E.M.P, Franco, M. A. P. (2012). On the distribution of the minimum or maximum of a random number of i.i.d. lifetime random variables. *Applied Mathematics*, 3, 350-353.

Marshall, A.W., Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84, 641-652.

Rachev, S.T., Resnick, S. (1991). Max-geometric infinite divisibility and stability. *Communications in Statistics. Stochastic Models,* 7, 191-218.

Raghunandanan, K. and Patil, S.A. (1972). On order statistics for random sample size. *Statistica Neerlandica*, 26, 121-126.

Rohatgi, V.K. (1987). Distribution of order statistics with random sample size. *Communications in Statistics. Theory and Methods*, 16, 3739-3743.

Shaked, M. (1975). On the distribution of the minimum and of the maximum of a random number of i.i.d. random variables. In *Statistical Distributions in Scientific Work*. Vol. I. ed. G.P. Patil, S. Kotz and J.K. Ord. Reidel, Dordrecht. pp. 363-380.

Shaked, M., Wong, T. (1997). Stochastic comparisons of random minima and maxima. *Journal of Applied Probability*, 34, 420-425

Appendix 1: Model extensions when $\mathcal{N}$ is the zero truncated Poisson or the logarithmic model

The zero-truncated Poisson(α) model is defined through the set of pgfs:

$$\mathcal{N} = \{N_\alpha : h_{N_\alpha} = \frac{e^{\alpha t} - 1}{e^\alpha - 1},\ \alpha \in [0, \infty)\}.$$

This model includes N_I and therefore both the basic transformations in Definition 2 as well as the combined transformations in Definition 3 are extensions, but this model is neither extreme auto-reversible, because $\Pr[N_\alpha = 1] = 1/E[N_\alpha]$, nor closed under pgf composition, because

$$\Pr(N_\alpha = 1) = \frac{\alpha}{e^\alpha - 1},$$

$$\Pr(N_\alpha = 2) = \frac{1}{2}\frac{\alpha^2}{e^\alpha - 1},$$

and therefore it does not satisfy the necessary condition of Corollary 2 for being closed,

$$\frac{\Pr(N_\alpha = 2)}{\Pr(N_\alpha = 1)(1 - \Pr(N_\alpha = 1))} = \frac{\alpha}{2} + \frac{1}{2}\frac{\alpha^2}{e^\alpha - \alpha - 1} = Constant.$$

The four basic extensions of $\mathcal{X} = \{X_\theta : F_{X_\theta},\ \theta \in \Theta\}$ obtained through Definition 2 are,

$$\max_{\mathcal{N}}(\mathcal{X}) = \{Y_{\theta,\alpha} : F_{Y_{\theta,\alpha}} = \frac{e^{\alpha F_{X_\theta}} - 1}{e^\alpha - 1},\ \theta \in \Theta, \alpha \in [0, \infty)\},$$

$$\max_{\mathcal{N}}^{-1}(\mathcal{X}) = \{Y_{\theta,\alpha} : F_{Y_{\theta,\alpha}} = \frac{\ln(1 + (e^\alpha - 1)F_{X_\theta})}{\alpha},\ \theta \in \Theta, \alpha \in [0, \infty)\},$$

$$\min_{\mathcal{N}}(\mathcal{X}) = \{Y_{\theta,\alpha} : F_{Y_{\theta,\alpha}} = \frac{e^\alpha(1 - e^{-\alpha F_{X_\theta}})}{e^\alpha - 1},\ \theta \in \Theta, \alpha \in [0, \infty)\},$$

$$\min_{\mathcal{N}}^{-1}(\mathcal{X}) = \{Y_{\theta,\alpha} : F_{Y_{\theta,\alpha}} = 1 - \frac{\ln(1 + (e^\alpha - 1)(1 - F_{X_\theta}))}{\alpha},\ \theta \in \Theta, \alpha \in [0, \infty)\},$$

and the four combined extensions of $\mathcal{X}$ obtained through Definition 3 are,

$$\max_{\mathcal{N}}(\max_{\mathcal{N}}^{-1}(\mathcal{X})) =$$

$$\{Y_{\theta,\alpha_1,\alpha_2} : F_{Y_{\theta,\alpha_1,\alpha_2}} = \frac{1}{\mathrm{e}^{\alpha_2} - 1}\left(\left(1 + (\mathrm{e}^{\alpha_1} - 1)\,F_{X_\theta}\right)^{\frac{\alpha_2}{\alpha_1}} - 1\right),\ \theta \in \Theta, \alpha_1, \alpha_2 \in [0, \infty)\},$$

$$\max_{\mathcal{N}}^{-1}(\max_{\mathcal{N}}(\mathcal{X})) =$$

$$\{Y_{\theta,\alpha_1,\alpha_2} : F_{Y_{\theta,\alpha_1,\alpha_2}} = \frac{1}{\alpha_2}\ln\left(1 + \frac{\left(\mathrm{e}^{\alpha_2} - 1\right)\left(\mathrm{e}^{\alpha_1 F_{X_\theta}} - 1\right)}{\mathrm{e}^{\alpha_1} - 1}\right),\ \theta \in \Theta, \alpha_1, \alpha_2 \in [0, \infty)\},$$

$$\min_{\mathcal{N}}(\min_{\mathcal{N}}^{-1}(\mathcal{X})) =$$

$$\{Y_{\theta,\alpha_1,\alpha_2} : F_{Y_{\theta,\alpha_1,\alpha_2}} = \frac{\mathrm{e}^{\alpha_2}}{\mathrm{e}^{\alpha_2}-1}\left(1-\left(1-\left(1-\mathrm{e}^{-\alpha_1}\right)F_{X_\theta}\right)^{\frac{\alpha_2}{\alpha_1}}\right),\ \theta\in\Theta,\alpha_1,\alpha_2\in[0,\infty)\},$$

$$\min_{\mathcal{N}}^{-1}(\min_{\mathcal{N}}(\mathcal{X})) =$$

$$\{Y_{\theta,\alpha_1,\alpha_2} : F_{Y_{\theta,\alpha_1,\alpha_2}} = 1-\frac{1}{\alpha_2}\ln\left(\frac{(\mathrm{e}^{\alpha_2}-1)\left(\mathrm{e}^{\alpha_1\left(1-F_{X_\theta}\right)}-1\right)}{\mathrm{e}^{\alpha_1}-1}+1\right),\ \theta\in\Theta,\alpha_1,\alpha_2\in[0,\infty)\}.$$

Furthermore, according to Example 5.2 the Logarithmic(p) model defined through:

$$\mathcal{N} = \{N_p : h_{N_p} = \frac{\log(1-pt)}{\log(1-p)},\ \ p\in[0,1)\},$$

where $p = 1-e^{-\alpha}$, is extreme reversible with the zero-truncated Poisson model. As a consequence of that property, the set of model extensions obtained through Definitions 2 and 3 using the Logarithmic(p) model coincide with the set of extensions obtained using the zero-truncated Poisson model presented in this Appendix.

The specific extensions for the Logarithmic(p) model are the ones listed above for the truncated Poisson model after replacing α by $-\log(1-p)$, and after switching $\max_{\mathcal{N}}$ and $\min_{\mathcal{N}}^{-1}$ and switching $\min_{\mathcal{N}}$ and $\max_{\mathcal{N}}^{-1}$. For example the $\max_{\mathcal{N}}(\cdot)$ transformation when $\mathcal{N}$ is Logaritmic(p) is the $\min_{\mathcal{N}'}^{-1}(\cdot)$ transformation when $\mathcal{N}'$ is truncated Poisson$(\alpha = -\log(1-p))$, the $\min_{\mathcal{N}}(\cdot)$ transformation is the $\max_{\mathcal{N}'}^{-1}(\cdot)$ transformation, and so on.

SORT 49 (1) January-June 2025, 73-92 DOI: 10.57645/20.8080.02.23

Lattice structures for the stochastic comparison of call ratio backspread derivatives with an application

María Concepción López-Díaz[1], Miguel López-Díaz[2] and Sergio Martínez-Fernández[3]

Abstract

The comparison of investments in financial derivatives is an appealing topic in the optimization of resources. A relevant derivative is the call ratio backspread. Motivated by the need to compare investments in such derivatives, a new family of stochastic orders is introduced. That permits to reach decisions on the allocations of funds in those derivatives under general conditions and without assuming specific probability distributions of the asset prices. Characterizations of the orders are developed. Special emphasis is placed on the existence of infima and suprema in such dominance criteria, which leads to lattice structures on some special spaces and to the reduction of some optimization problems with stochastic dominance constraints. The method is illustrated with an application using real data from financial markets.

MSC: *60E15, 62P05.*

Keywords: *Call ratio backspread derivative, integrated survival function, lattice, stochastic order.*

1. Motivation of the study

The aim of this manuscript is to show how the theory of stochastic orders can be used for reaching decisions on the allocations of funds in call ratio backspread derivatives, entailing some advantages with respect to other methods.

[1] Departamento de Matemáticas, Universidad de Oviedo. C/ Leopoldo Calvo Sotelo 18. E-33007 Oviedo, Spain. cld@uniovi.es

[2] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo. C/ Leopoldo Calvo Sotelo 18. E-33007 Oviedo, Spain. mld@uniovi.es

[3] Unidad de Auditoría de Capital & Impairments, Banco Sabadell. C/ Sena 12, P.I. Can Sant Joan. E-08174 Sant Cugat del Valles, Barcelona, Spain. martinezserg@bancsabadell.com

Received: June 2024.
Accepted: March 2025.

A financial derivative is a financial contract with a value derived from the future price of an underlying asset. Basically, it is an agreement between two parts, a buyer and a seller, that specifies conditions on the dates, resulting values, definitions of the underlying assets, the parties' contractual obligations, and the amount under which payments are to be made between the parties. The assets of derivatives can be of a quite different kind, such as goods and shares, but also indices such as rates of return, rates of interest or exchange rates.

A relevant derivative in financial markets is the (European) call option. That is an agreement that gives to the purchaser of the call option the right, but not the obligation, to buy an agreed quantity of an underlying asset at a specified exercise price, at an exercise date, while paying a premium for this right. If x is the unit price of the underlying asset on the due date, p is the unit exercise price and k is the unit premium, the benefit of the purchaser of the call option per unit of the asset is $(x-p)_+ - k$, where the subscript $+$ denotes the positive part of a real number. The unit profit of the seller of the call option is $k-(x-p)_+$.

Some financial derivatives are formed by means of other derivatives, like the call ratio backspread. Consider the common call ratio backspread, in which two call options are bought with unit exercise price p_2, and a call option is sold with unit exercise price p_1, where $p_1 < p_2$, all of them with the same asset and the same exercise date. From now on, we will refer to it as the call ratio backspread.

The unit benefit of a call ratio backspread is $k_1-(x-p_1)_+ + 2(x-p_2)_+ - 2k_2$, where x is the unit price of the asset at the exercise date, p_1 is the asset unit exercise price of the sale of the call option, k_1 its unit premium and p_2 and k_2 play the same role in the purchases of the call options.

In this manuscript, we propose a model to compare investments in call ratio backspread derivatives in terms of a family of stochastic orders, which does not need specific distributions of the asset prices and makes it possible to detect arbitrage options in financial markets, that is, detecting deals that would lead to a non-zero probability of future profit. Another advantage of the new method is that when the order is satisfied, the expected benefits are ordered whatever price p_1. Thus, an investor does not need to attain some particular values of p_1 to be able to compare investments and find opportunities. When the order is satisfied and the premiums do not follow the same arrangement for a particular value of p_1, there exist arbitrage opportunities. Moreover, we prove that there exist an infinum and a supremum of any two random variables with finite means with respect to any stochastic order of that family. The existence of a supremum and an infimum is useful in optimization problems with stochastic dominance constraints.

The reader is referred, for instance, to the books Dixit and Pindyck (1994), Cohen (2005) and Hull (2015) for an introduction to the field of financial derivatives, and to Müller and Stoyan (2002), Shaked and Shanthikumar (2007), Belzunce, Martínez-Riquelme and Mulero (2016) and Levy (2016) for a comprehensive introduction to the theory and applications of stochastic orderings. Some references which relate stochastic dominance criteria and arbitrage opportunities are Levy (2016), Jarrow (1986) and Ng,

Wong and Xiao (2017), to the best of our knowledge, few manuscripts approach both topics. An article connecting stochastic orders and financial derivatives is López-Díaz, López-Díaz and Martínez-Fernández (2018).

As an application of the proposed model to compare investments in call ratio backspread derivatives, in the present manuscript we compare call ratio backspread derivatives whose assets are the weekly returns of Boeing and Procter & Gamble (P&G), companies in the Dow Jones Industrial Average Index. For that purpose, a result which permits to use statistical inference techniques to test conditions that lead to the call ratio backspread stochastic order is proved. As a consequence, we obtain that the expected benefit of a call ratio backspread derivative with asset the unit weekly revaluation of Boeing is greater (not lower) than the corresponding derivative with the asset unit weekly revaluation of P&G, whatever $p_1 < 1$. Then, if for some $p_1 < 1$, the premium of the derivative associated with Boeing is lower than the premium of P&G, an arbitrage opportunity exists for those derivatives. Moreover, in case of equality of premiums, an investor should choose the Boeing derivative instead of the P&G option.

The structure of the paper is as follows. Section 2 contains the preliminaries of the manuscript. In Section 3, we introduce the mathematical model to analyze the aforementioned problem in terms of a family of stochastic orders and we develop the main characterizations of those families. Section 4 is devoted to the analysis of the existence of infimum and supremum in such orderings. The application described above of the proposed method is developed in Section 5. To conclude, Section 6 contains some final comments and conclusions about the manuscript.

2. Preliminaries

In this section, preliminary concepts and notations are presented.

Throughout the paper, if $a \in \mathbb{R}$, a_+ will stand for $\max\{a,0\}$ and a_- for $\max\{-a,0\}$.

Given a random variable X, F_X will represent its distribution function, EX its expected value and P_X the probability induced by X. Moreover, $\overline{F}_X$ will denote the survival function of X.

The integrated survival function of a random variable X with finite mean is the mapping $\pi_X : \mathbb{R} \to \mathbb{R}$, with $\pi_X(t) = E(X-t)_+$ for any $t \in \mathbb{R}$. It is well-known that $\pi_X(t) = \int_{(t,+\infty)} \overline{F}_X(x)\,dx$.

A stochastic order is a pre-order relation on a set of probabilities. Basically, it aims to order probabilities in accordance with a criterion.

An integral stochastic order $\preceq$ is defined by the comparison of the integrals of real measurable mappings in a certain class. Namely, two probabilities P and Q on $(\mathbb{R},\mathcal{B})$ ($\mathcal{B}$ denotes the usual Borel σ-field) satisfy $P \preceq Q$, when

$$\int_{\mathbb{R}} f\,dP \leq \int_{\mathbb{R}} f\,dQ$$

for any f in that class, such that the integrals exist. That set of mappings is said to be a generator of the order (see Müller (1997) for integral stochastic orders).

If $\preceq$ is a stochastic order on the probabilities on $(\mathbb{R},\mathcal{B})$, and X and Y are two random variables, $X \preceq Y$ will mean $P_X \preceq P_Y$.

The following integral stochastic orders will appear in the manuscript.

Let X and Y be random variables, then

i) X is said to be smaller than Y in the usual stochastic ordering ($X \preceq_{st} Y$) if $E(f(X)) \leq E(f(Y))$ for all increasing mappings $f : \mathbb{R} \to \mathbb{R}$ such that the above expectations exist, equivalently, $F_X \geq F_Y$,

ii) X is said to be smaller than Y in the increasing concave order ($X \preceq_{icv} Y$) if $E(f(X)) \leq E(f(Y))$ for all increasing concave mappings $f : \mathbb{R} \to \mathbb{R}$ such that the above expectations exist.

The notation $X \sim_{st} Y$ will mean that X and Y have the same distribution.

Given P a probability on $(\mathbb{R},\mathcal{B})$ and $T : \mathbb{R} \to \mathbb{R}$ a measurable mapping, $P \circ T^{-1}$ will denote the probability given by $P \circ T^{-1}(B) = P(T^{-1}(B))$ for any $B \in \mathcal{B}$.

If $A \subset \mathbb{R}$, I_A will stand for the indicator mapping of the set A.

3. The call ratio backspread stochastic orders

In this section, we introduce the new family of stochastic orders which arises from the aim to compare call ratio backspread derivatives, providing characterization results of those orders.

From now on, $f_{p_1,p_2,k}$ will stand for the mapping $f_{p_1,p_2,k} : \mathbb{R} \to \mathbb{R}$, with $f_{p_1,p_2,k}(x) = 2(x-p_2)_+ - (x-p_1)_+ + k$ for any $x \in \mathbb{R}$, with $p_1, p_2, k \in \mathbb{R}$, and $p_1 < p_2$. This mapping represents the unit benefit of a call ratio backspread at the expiration date with k equal to $k_1 - 2k_2$.

Given $p_2 \in \mathbb{R}$, we will denote by $\mathcal{F}^{p_2}$ the class of mappings $\mathcal{F}^{p_2} = \{f_{p_1,p_2,k} \mid p_1, k \in \mathbb{R}, p_1 < p_2\}$, that is, the family of mappings which represent the unit benefits of call ratio backspread derivatives where the unit exercise price of the call option purchases is a given value p_2.

Next, the model to compare call ratio backspread derivatives is introduced.

Definition 3.1. *Let X and Y be two random variables. It will be said that X is less than Y in the call ratio backspread stochastic order for the unit exercise price of the call option purchases p_2, if $E(f(X)) \leq E(f(Y))$ for any $f \in \mathcal{F}^{p_2}$ such that the above expectations exist. This relation will be denoted by $X \preceq^{p_2}_{crb} Y$.*

Consider two call ratio backsprad derivatives with common exercise prices and expiration dates. Assume now that their premiums are equal to some value k. Let X and Y stand for the random variables unit prices of the assets of those derivatives at the expiration date. The unit expected benefits of both financial derivatives are $E(f_{p_1,p_2,k}(X))$ and $E(f_{p_1,p_2,k}(Y))$, respectively. The relation $X \preceq^{p_2}_{crb} Y$ means that the expected benefit of the call ratio backspread derivative associated with Y is greater (not lower) than that of X, whatever unit exercise price p_1 of the call option sales and whatever premiums k.

Thus, if $X \preceq_{crb}^{p_2} Y$ is held and the real premium of the second derivative (that of Y) is lower than the premium of the first derivative, an option of arbitrage is being offered in financial markets.

Observe that the model does not assume specific probabilistic distributions of the prices of the underlying assets, such as Brownian movements.

Next, we state different characterization results of the call ratio backspread stochastic orders.

Given $p_2 \in \mathbb{R}$, let $\mathcal{F}_0^{p_2} = \{f_{p_1,p_2,p_2-p_1} \mid p_1 \in \mathbb{R}, p_1 < p_2\}$. Notice that $\mathcal{F}_0^{p_2} \subset \mathcal{F}^{p_2}$. In fact, $\mathcal{F}_0^{p_2}$ is given by the mappings of the class $\mathcal{F}^{p_2}$ whose values at the point p_2 are equal to 0. Both $\mathcal{F}_0^{p_2}$ and $\mathcal{F}^{p_2}$ are generators of the stochastic order $\preceq_{crb}^{p_2}$. Observe that any mapping in $\mathcal{F}^{p_2}$ is a translation of a map in $\mathcal{F}_0^{p_2}$.

The following result says that the analysis of the family of call ratio backspread stochastic orders can be performed for the unit exercise price $p_2 = 0$.

Proposition 3.2. *Let X and Y be random variables. It holds that $X \preceq_{crb}^{p_2} Y$ if and only if $X - p_2 \preceq_{crb}^{0} Y - p_2$.*

Proof. Suppose that $X \preceq_{crb}^{p_2} Y$. Let $f \in \mathcal{F}^0$ such that $E(f(X-p_2))$ and $E(f(Y-p_2))$ exist. Let $T : \mathbb{R} \to \mathbb{R}$ with $T(x) = x - p_2$. We have that

$$\int_{\mathbb{R}} f(x)\, dP_{X-p_2} = \int_{\mathbb{R}} f(x)\, dP_X \circ T^{-1} = \int_{\mathbb{R}} f(T(x))\, dP_X = \int_{\mathbb{R}} f(x-p_2)\, dP_X.$$

Since $f \in \mathcal{F}^0$, $f = f_{p_1,0,k}$ for some $p_1 \in \mathbb{R}$ with $p_1 < 0$ and $k \in \mathbb{R}$. It can be seen that $f \circ T = f_{p_1+p_2,p_2,k}$, which belongs to $\mathcal{F}^{p_2}$. Thus,

$$\int_{\mathbb{R}} f(x-p_2)\, dP_X \leq \int_{\mathbb{R}} f(x-p_2)\, dP_Y = \int_{\mathbb{R}} f(x)\, dP_{Y-p_2},$$

which leads to $X - p_2 \preceq_{crb}^{0} Y - p_2$. The converse can be proved in a similar way. ■

All the results will be developed for the call ratio backspread stochastic order $\preceq_{crb}^{0}$, which we will refer to it as the call ratio backspread order. The counterpart for any unit exercise price p_2 can be immediately derived by Proposition 3.2.

Observe that the translations of the random variables in Proposition 3.2 can lead to variables and prices assuming negative values, even in the case that the original variables were positive.

Proposition 3.3. *Let X and Y be random variables with finite means. Then $X \preceq_{crb}^{0} Y$ if and only if*

$$-tF_X(t) + E(|X| I_{(t,\infty)}(X)) \leq -tF_Y(t) + E(|Y| I_{(t,\infty)}(Y)) \text{ for any } t < 0.$$

Proof. Assume that $X \preceq^0_{crb} Y$. Let $t<0$. Consider the mapping $f_{t,0,-t}$, which belongs to $\mathcal{F}^0_0$. The condition $X \preceq^0_{crb} Y$ implies that

$$\int_{\mathbb{R}} f_{t,0,-t}(x)\,dP_X \leq \int_{\mathbb{R}} f_{t,0,-t}(x)\,dP_Y.$$

Now notice that

$$\int_{\mathbb{R}} f_{t,0,-t}(x)\,dP_X = -tP_X((-\infty,t]) + \int_{(t,0]} -x\,dP_X + \int_{(0,+\infty)} x\,dP_X$$

$$= -tF_X(t) + \int_{(t,+\infty)} |x|\,dP_X = -tF_X(t) + E(|X|I_{(t,\infty)}(X)),$$

which proves one implication.

Conversely, if

$$-tF_X(t) + E(|X|I_{(t,\infty)}(X)) \leq -tF_Y(t) + E(|Y|I_{(t,\infty)}(Y))$$

for any $t<0$, then $E(f_{t,0,-t}(X)) \leq E(f_{t,0,-t}(Y))$ for any $t<0$. Notice that the class of mappings $\{f_{t,0,-t} \mid t<0\}$ is $\mathcal{F}^0_0$, thus, $X \preceq^0_{crb} Y$. ■

The following result provides a characterization of $\preceq^0_{crb}$ in terms of integrated survival functions. It will be key to prove the existence of infimum and supremum in the order.

Proposition 3.4. *Let X and Y be random variables with finite means. We have that $X \preceq^0_{crb} Y$ if and only if*

$$-\pi_X(t) + 2\pi_X(0) \leq -\pi_Y(t) + 2\pi_Y(0) \text{ for any } t<0.$$

Proof. Notice that $X \preceq^0_{crb} Y$ holds if and only if $E(f(X)) \leq E(f(Y))$ for any $f \in \mathcal{F}^0_0$, that is, if and only if $E(f_{p_1,0,-p_1}(X)) \leq E(f_{p_1,0,-p_1}(Y))$ for any $p_1<0$.

Observe that $E(f_{p_1,0,-p_1}(X)) = E(2X_+ - (X-p_1)_+ - p_1)$, and so, $X \preceq^0_{crb} Y$ is equivalent to $2EX_+ - E(X-p_1)_+ \leq 2EY_+ - E(Y-p_1)_+$ for any $p_1<0$, that is, $2\pi_X(0) - \pi_X(t) \leq 2\pi_Y(0) - \pi_Y(t)$ for any $t<0$. ■

Some consequences of the preceding results are developed below.

Proposition 3.5. *Let X and Y be random variables with finite means. If $X \preceq^0_{crb} Y$ and $Y \preceq^0_{crb} X$, then $X_- \sim_{st} Y_-$.*

Proof. By Proposition 3.4, $X \preceq^0_{crb} Y$ and $Y \preceq^0_{crb} X$ are equivalent to $-\pi_X(t) + 2\pi_X(0) = -\pi_Y(t) + 2\pi_Y(0)$ for any $t<0$.

By the Second Fundamental Theorem of Calculus, $\overline{F}_X = \overline{F}_Y$ almost everywhere in $(-\infty, 0)$. Since distribution functions are right continuous, $\overline{F}_X(t) = \overline{F}_Y(t)$ for any $t < 0$, that is, $\overline{F}_X(-t) = \overline{F}_Y(-t)$ for any $t > 0$.

Therefore, $(1 - F_X(-t))I_{(0,+\infty)}(t) = (1 - F_Y(-t))I_{(0,+\infty)}(t)$ for any $t > 0$, hence $(1 - F_X(-t^-))I_{(0,+\infty)}(t) = (1 - F_X(-t^-))I_{(0,+\infty)}(t)$ for any $t > 0$. This is the same as $F_{X_-}(t) = F_{Y_-}(t)$ for any $t > 0$, thus, $X_- \sim_{st} Y_-$. ■

Corollary 3.6. *The relation $\preceq^0_{crb}$ is a partial order on the set of a.s. negative random variables with finite means, where equality is in distribution.*

Proof. The reflexive property is obvious. Transitivity follows from Proposition 3.4. The anti-symmetric property is a consequence of Proposition 3.5. ■

The order $\preceq^0_{crb}$ is not a partial order on the set of random variables with finite mean, but a pre-order. Consider the random variables X and Y with $P(X = 0) = P(X = 2) = 1/2$ and $P(Y = 1) = 1$. It holds that $X \preceq^0_{crb} Y$ and $Y \preceq^0_{crb} X$, but $X \sim_{st} Y$ is false.

4. Lattice structures

Throughout this section, we will prove that there exist an infinum and a supremum of any two random variables with finite means with respect to any call ratio backspread stochastic order. This permits to construct lattice structures on special partially ordered sets.

We prove the following result on integrated survival functions to analyze the case of the infimum.

Proposition 4.1. *Let X and Y be random variables with finite means. Let $h : \mathbb{R} \to \mathbb{R}$ with*

$$h(t) = \max\{\pi_X(t) - 2\pi_X(0), \pi_Y(t) - 2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\}$$

for any $t \in \mathbb{R}$. The mapping h is an integrated survival function of a random variable with finite mean.

Proof. Notice that the mapping h is well-defined since X and Y have finite means.

In accordance with Theorem 1.5.10 in Müller and Stoyan (2002), the mapping h is an integrated survival function of a random variable with finite mean if and only if : *i*) h is decreasing, *ii*) h is convex, *iii*) $\lim_{t\to+\infty} h(t) = 0$, and *iv*) $\lim_{t\to-\infty} h(t) + t \in \mathbb{R}$.

In relation to *i*), since π_X and π_Y are integrated survival functions, both are decreasing, and so is h.

Regarding the second condition, we should prove that for any $t_1, t_2 \in \mathbb{R}$ and any $\lambda \in [0,1]$, it holds that $h(\lambda t_1 + (1-\lambda)t_2) \leq \lambda h(t_1) + (1-\lambda)h(t_2)$.

Notice that π_X and π_Y are convex since they are integrated survival functions. Thus,

$$\begin{aligned}
&\pi_X(\lambda t_1 + (1-\lambda)t_2) - 2\pi_X(0) \\
&\leq \lambda \pi_X(t_1) + (1-\lambda)\pi_X(t_2) - 2(\lambda \pi_X(0) + (1-\lambda)\pi_X(0)) \\
&= \lambda(\pi_X(t_1) - 2\pi_X(0)) + (1-\lambda)(\pi_X(t_2) - 2\pi_X(0)),
\end{aligned}$$

and the same inequality is satisfied with Y. Therefore,

$$\begin{aligned}
&\max\{\pi_X(\lambda t_1 + (1-\lambda)t_2) - 2\pi_X(0),\, \pi_Y(\lambda t_1 + (1-\lambda)t_2) - 2\pi_Y(0)\} \\
&\leq \max\{\lambda(\pi_X(t_1) - 2\pi_X(0)) + (1-\lambda)(\pi_X(t_2) - 2\pi_X(0)), \\
&\quad \lambda(\pi_Y(t_1) - 2\pi_Y(0)) + (1-\lambda)(\pi_Y(t_2) - 2\pi_Y(0))\} \\
&\leq \lambda \max\{\pi_X(t_1) - 2\pi_X(0), \pi_Y(t_1) - 2\pi_Y(0)\} \\
&\quad + (1-\lambda)\max\{\pi_X(t_2) - 2\pi_X(0), \pi_Y(t_2) - 2\pi_Y(0)\},
\end{aligned}$$

which leads to the convexity of h.

With respect to *iii*), $\lim_{t\to+\infty} \pi_X(t) = 0$ and $\lim_{t\to+\infty} \pi_Y(t) = 0$ since π_X and π_Y are integrated survival functions of random variables with finite means, therefore,

$$\lim_{t\to+\infty} h(t) = \max\{-2\pi_X(0), -2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\} = 0.$$

In relation to *iv*), X and Y have finite means, hence $\lim_{t\to-\infty} \pi_X(t) + t = EX$ and $\lim_{t\to-\infty} \pi_Y(t) + t = EY$. Thus,

$$\begin{aligned}
&\lim_{t\to-\infty} h(t) + t \\
&= \lim_{t\to-\infty} \max\{\pi_X(t) - 2\pi_X(0),\, \pi_Y(t) - 2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\} + t \\
&= \lim_{t\to-\infty} \max\{\pi_X(t) + t - 2\pi_X(0),\, \pi_Y(t) + t - 2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\} \\
&= \max\{EX - 2\pi_X(0), EY - 2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\} \in \mathbb{R}.
\end{aligned}$$

Therefore, h is an integrated survival function of a random variable with finite mean.

■

If W is a random variable such that $\pi_W = h$, in accordance with Theorem 1.5.10 in Müller and Stoyan (2002), it holds that $EW = \lim_{t\to-\infty} h(t) + t$. Thus, $EW = \max\{EX - 2\pi_X(0), EY - 2\pi_Y(0)\} + 2\min\{\pi_X(0), \pi_Y(0)\}$.

Proposition 4.2. *Let X and Y be random variables with finite means. Let $h:\mathbb{R}\to\mathbb{R}$, with*

$$h(t)=\max\{\pi_X(t)-2\pi_X(0),\,\pi_Y(t)-2\pi_Y(0)\}+2\min\{\pi_X(0),\pi_Y(0)\}$$

for any $t\in\mathbb{R}$. Let W be a random variable such that $\pi_W=h$. Then, W is an infimum of X and Y in the stochastic order $\preceq^0_{crb}$.

Proof. Proposition 4.1 guarantees the existence of a random variable W with finite mean in the conditions of the statement. Now, notice that

$$h(0)=\max\{-\pi_X(0),-\pi_Y(0)\}+2\min\{\pi_X(0),\pi_Y(0)\}=\min\{\pi_X(0),\pi_Y(0)\}.$$

Therefore, we have that

$$\begin{aligned}-h(t)+2h(0)&=-\max\{\pi_X(t)-2\pi_X(0),\,\pi_Y(t)-2\pi_Y(0)\}\\&\quad-2\min\{\pi_X(0),\pi_Y(0)\}+2\min\{\pi_X(0),\pi_Y(0)\}\\&=-\max\{\pi_X(t)-2\pi_X(0),\,\pi_Y(t)-2\pi_Y(0)\}\\&=\min\{-\pi_X(t)+2\pi_X(0),\,-\pi_Y(t)+2\pi_Y(0)\}.\end{aligned}$$

Thus,

$$-h(t)+2h(0)\le-\pi_X(t)+2\pi_X(0)\ \text{ and }\ -h(t)+2h(0)\le-\pi_Y(t)+2\pi_Y(0)$$

for any $t<0$. By Proposition 3.4, $W\preceq^0_{crb}X$ and $W\preceq^0_{crb}Y$.

Let Z be a random variable with finite mean such that $Z\preceq^0_{crb}X$ and $Z\preceq^0_{crb}Y$. In accordance with Proposition 3.4,

$$-\pi_Z(t)+2\pi_Z(0)\le-\pi_X(t)+2\pi_X(0)\text{ and }-\pi_Z(t)+2\pi_Z(0)\le-\pi_Y(t)+2\pi_Y(0)$$

for all $t<0$. Thus, $-\pi_Z(t)+2\pi_Z(0)\le-h(t)+2h(0)$ for all $t<0$, equivalently, $Z\preceq^0_{crb}W$, which proves the result. ■

Proposition 4.3. *Let X and Y be random variables with finite means. Let Z be an infimum of X and Y in the order $\preceq^0_{crb}$. Then,*

i) $EZ_+=\min\{EX_+,EY_+\}$,

ii) $EZ=\max\{EX-2EX_+,EY-2EY_+\}+2\min\{EX_+,EY_+\}$.

Proof. Let W be the infimum of X and Y in $\preceq^0_{crb}$ given in Proposition 4.2. Since Z and W are infima, then $W \preceq^0_{crb} Z$ and $Z \preceq^0_{crb} W$, that is, $-\pi_Z(t)+2\pi_Z(0) = -\pi_W(t)+2\pi_W(0)$ for any $t<0$.

The Monotone Convergence Theorem implies that $\pi_Z(0) = \pi_W(0)$, which is equal to $\min\{EX_+, EY_+\}$, and so, we derive i).

On the other hand, we have obtained that $\pi_Z(t) = \pi_W(t)$ for any $t<0$, and so,

$$EZ = \lim_{t\to-\infty} \pi_Z(t)+t = \lim_{t\to-\infty} \pi_W(t)+t$$

$$= \max\{EX-2EX_+, EY-2EY_+\}+2\min\{EX_+, EY_+\},$$

which concludes the proof. ■

Proposition 4.4. *Let X and Y be random variables with finite means such that $X \leq 0$ a.s. Then, the infimum of X and Y with respect to the order $\preceq^0_{crb}$ is unique in distribution.*

Proof. Let us suppose that W_1 and W_2 are two infima of X and Y. Since $W_i \preceq^0_{crb} X$, we have that $E(f(W_i)) \leq E(f(X))$ for any $f \in \mathcal{F}^0_0$ and any $i \in \{1,2\}$.

Take the sequence $\{f_{-\frac{1}{n},0,\frac{1}{n}}\}_n \subset \mathcal{F}^0_0$, which is decreasing and whose pointwise convergence is the mapping $g(x) = x_+$ for any $x \in \mathbb{R}$.

The Monotone Convergence Theorem implies that $EW_{i_+} \leq EX_+ = 0$. Therefore, $W_{i_+} = 0$ *a.s.* and so $W_i \leq 0$ *a.s.* for any $i \in \{1,2\}$.

Since W_1 and W_2 are infima of X and Y, we obtain that $W_1 \preceq^0_{crb} W_2$ and $W_2 \preceq^0_{crb} W_1$. Corollary 3.6 leads to $W_1 \sim_{st} W_2$. ■

Next we analyze the case of the supremum of two random variables in the call ratio backspread stochastic order.

Let I be an interval of $\mathbb{R}$ with non-empty interior. Let $f : I \to \mathbb{R}$ be a mapping. We will denote by $vex(f)$ the mapping $vex(f) : I \to \mathbb{R}$, with

$$vex(f)(t) = \sup\{g(t) \mid g \text{ is convex and } g(x) \leq f(x) \text{ for all } x \in I\}$$

for any $t \in I$. This mapping is usually known as the convex hull operator, or the greatest convex minorant.

Proposition 4.5. *Let X and Y be random variables with finite means. Then, there exists a random variable W with finite mean which is a supremum of X and Y in the stochastic order $\preceq^0_{crb}$.*

Proof. Let us consider the mapping $l : (-\infty, 0] \to \mathbb{R}$, with

$$l(t) = \min\{\, \pi_X(t)-2EX_+,\, \pi_Y(t)-2EY_+ \,\}+2\max\{EX_+, EY_+\}$$

for any $t \leq 0$. It holds that

$$l(0) = \min\{-EX_+, -EY_+\} + 2\max\{EX_+, EY_+\} = \max\{EX_+, EY_+\} \geq 0.$$

Define $h : (-\infty, 0] \to \mathbb{R}$, with $h = vex(l)$.

Observe that h is decreasing. Notice that π_X and π_Y are decreasing since they are integrated survival functions, therefore l is decreasing and so is h.

Trivially h is convex.

Let us see that $\lim_{t\to-\infty} h(t) + t \in \mathbb{R}$. Notice that the mapping $t \to h(t) + t$ is convex, which guarantees the existence of that limit.

We have that for any $t \leq 0$ it holds that $h(t) + t \leq l(t) + t$. Now

$$\begin{aligned}\lim_{t\to-\infty} l(t) + t &= \lim_{t\to-\infty} \min\{\pi_X(t) - 2EX_+, \pi_Y(t) - 2EY_+\} \\ &\quad + 2\max\{EX_+, EY_+\} + t \\ &= \lim_{t\to-\infty} \min\{\pi_X(t) + t - 2EX_+, \pi_Y(t) + t - 2EY_+\} + 2\max\{EX_+, EY_+\} \\ &= \min\{EX - 2EX_+, EY - 2EY_+\} + 2\max\{EX_+, EY_+\} \in \mathbb{R}\end{aligned}$$

since X and Y have finite means.

Let $\widetilde{l} : \mathbb{R} \to \mathbb{R}$ given by $\widetilde{l}(t) = \min\{\pi_X(t), \pi_Y(t)\}$ for all $t \in \mathbb{R}$. Define the mapping $\widetilde{h} = vex(\widetilde{l})$.

Clearly $\widetilde{l}(t) \leq l(t)$ when $t \in (-\infty, 0]$. As a consequence $\widetilde{h}(t) \leq h(t)$ for any $t \in (-\infty, 0]$.

In accordance with Müller and Scarsini (2006), the function $\widetilde{h}$ is the integrated survival function of a random variable with finite expectation. Thus, $\lim_{t\to-\infty} \widetilde{h}(t) + t \in \mathbb{R}$.

Since $\lim_{t\to-\infty} \widetilde{h}(t) + t \leq \lim_{t\to-\infty} h(t) + t \leq \lim_{t\to-\infty} l(t) + t$, we conclude that $\lim_{t\to-\infty} h(t) + t \in \mathbb{R}$.

Consider now any mapping $\widehat{h} : \mathbb{R} \to \mathbb{R}$ with $\widehat{h}(t) = h(t)$ for any $t \leq 0$ such that $\widehat{h}$ is continuous, convex, decreasing and with $\lim_{t\to+\infty} \widehat{h}(t) = 0$. Thus, $\widehat{h}$ is an integrated survival function of a random variable with finite mean.

The existence of at least one of such mappings can be guaranteed as follows.

Notice that $l(0) = \max\{EX_+, EY_+\}$. Since the constant mapping $\max\{EX_+, EY_+\}$ is convex, and $l(t) \geq l(0)$ for any $t < 0$, it holds that $h(0) = l(0) = \max\{EX_+, EY_+\} \geq 0$.

If $h(0) = 0$ the extension is trivial by taking $\widehat{h}(t) = 0$ for all $t \geq 0$.

Let $h(0) > 0$. Since h is convex, there exists $h'_-(0)$, the left derivative of h at the point 0 (see, for instance, Roberts and Varberg (1973)). Moreover, $h'_-(0) \leq 0$ due to the decreasing of h.

If $h'_-(0) < 0$, consider the mapping $g : \mathbb{R} \to \mathbb{R}$ with $g(t) = h(0) + h'_-(0)t$ for all $t \geq 0$. This mapping is continuous, convex, strictly decreasing, cuts the x-axis, is tangent to h at the point 0 and $g(0) = h(0)$.

Thus, it is sufficient to take

$$\widehat{h}(t) = \begin{cases} h(t) & \text{if } t \leq 0, \\ g(t) & \text{if } t \in (0, -\frac{h(0)}{h'_-(0)}], \\ 0 & \text{if } t > -\frac{h(0)}{h'_-(0)}. \end{cases}$$

Let us see that $h'_-(0) = 0$ is not possible.

The condition $h(0) > 0$ implies that $\pi_X(0) > 0$ or $\pi_Y(0) > 0$. Let us suppose that $\pi_X(0) \geq \pi_Y(0)$ and so $0 < \pi_X(0)$.

Firstly, consider the case $0 < \pi_Y(0) < \pi_X(0)$.

Since $\pi_X(0) > 0$, π_X is convex and $\lim_{t\to+\infty} \pi_X(t) = 0$, we obtain that π_X is strictly decreasing (when different from 0) and so $\pi'_{X-}(0) < 0$ (left derivative of π_X at the point 0). Moreover,

$$l(t) = \min\{\, \pi_X(t),\, \pi_Y(t) + 2(EX_+ - EY_+)\,\}$$

for any $t \leq 0$.

Let $\widetilde{t} \leq 0$ such that $\pi'_{Y-}(\widetilde{t}) < 0$. Such a point exists since π_Y is an integrated survival function. Take $\alpha < 0$ satisfying that

i) $\pi'_{X-}(0) \leq \alpha$,

ii) $\pi_X(0) + \alpha\widetilde{t} < \pi_Y(0) + 2(\pi_X(0) - \pi_Y(0))$, and

iii) $\pi'_{Y-}(\widetilde{t}) \leq \alpha$.

Such a value α exists since $\pi_X(0) < \pi_Y(0) + 2(\pi_X(0) - \pi_Y(0))$.

Define $g : (-\infty, 0] \to \mathbb{R}$ with $g(t) = \pi_X(0) + \alpha t$ for any $t \leq 0$.

Let us see that $g(t) \leq l(t)$ for any $t \leq 0$.

By condition *i*), $g(t) \leq \pi_X(0) + \pi'_{X-}(0)t \leq \pi_X(t)$ for any $t \leq 0$.

On the other hand, if $t \in [\widetilde{t}, 0)$, condition *ii*) leads to $g(t) \leq \pi_Y(0) + 2(\pi_X(0) - \pi_Y(0)) \leq \pi_Y(t) + 2(\pi_X(0) - \pi_Y(0))$ for any $t \leq 0$.

Moreover, if $t \in (-\infty, \widetilde{t})$, conditions *ii*) and *iii*) imply that

$$g(t) = \pi_X(0) + \alpha\widetilde{t} + \alpha(t - \widetilde{t}) \leq \pi_Y(0) + 2(\pi_X(0) - \pi_Y(0)) + \alpha(t - \widetilde{t})$$

$$\leq \pi_Y(\widetilde{t}) + 2(\pi_X(0) - \pi_Y(0)) + \pi'_{Y-}(\widetilde{t})(t - \widetilde{t}) \leq \pi_Y(t) + 2(\pi_X(0) - \pi_Y(0)).$$

Therefore, $g(t) \leq \min\{\pi_X(t), \pi_Y(t) + 2(\pi_X(0) - \pi_Y(0))\} = l(t)$ for any $t \leq 0$. Since g is convex, $g(t) \leq h(t)$ for any $t \leq 0$. Recall that $g(0) = h(0)$. Thus,

$$0 > \alpha = \lim_{x \to 0^-} \frac{g(x) - g(0)}{x} \geq \lim_{x \to 0^-} \frac{h(x) - h(0)}{x} = h'_-(0).$$

Therefore, $h'_-(0) < 0$.

The case $0 < \pi_X(0) = \pi_Y(0)$ is immediate since $l = \min\{\pi_X, \pi_Y\}$. In this case $\widehat{h} = vex(l)$, where l is defined on the whole real line, satisfies the required conditions.

As a consequence, we have that the mapping $\widehat{h}$ is an integrated survival function of a random variable with finite mean.

Let W be a random variable whose integrated survival function fulfills $\pi_W = \widehat{h}$.

Let us see that W is a supremum of X and Y in the order $\preceq^0_{crb}$.

Let $t < 0$, then

$$\pi_W(t) - 2\pi_W(0)$$

$$\leq \min\{\pi_X(t) - 2EX_+, \pi_Y(t) - 2EY_+\} + 2\max\{EX_+, EY_+\} - 2\pi_W(0)$$

$$\leq \pi_X(t) - 2EX_+ = \pi_X(t) - 2\pi_X(0).$$

Applying Proposition 3.4, we deduce that $X \preceq^0_{crb} W$. In the same way we obtain that $Y \preceq^0_{crb} W$.

Let Z be a random variable with $X \preceq^0_{crb} Z$ and $Y \preceq^0_{crb} Z$. For any $t < 0$

$$\pi_Z(t) - 2\pi_Z(0) \leq \pi_X(t) - 2\pi_X(0) \quad \text{and} \quad \pi_Z(t) - 2\pi_Z(0) \leq \pi_Y(t) - 2\pi_Y(0).$$

Thus, for any $t < 0$

$$\pi_Z(t) - 2\pi_Z(0) + 2\max\{EX_+, EY_+\} \leq l(t).$$

Notice that $\pi_Z(t) - 2\pi_Z(0) + 2\max\{EX_+, EY_+\}$ is convex since π_Z is convex, which implies that for any $t < 0$

$$\pi_Z(t) - 2\pi_Z(0) + 2\max\{EX_+, EY_+\} \leq \widehat{h}(t),$$

which is the same as $\pi_Z(t) - 2\pi_Z(0) \leq \widehat{h}(t) - 2\widehat{h}(0)$ for any $t < 0$, equivalently $\pi_Z(t) - 2\pi_Z(0) \leq \pi_W(t) - 2\pi_W(0)$ for any $t < 0$, that is, $W \preceq^0_{crb} Z$. This concludes the proof of the existence of a supremum in the stochastic order $\preceq^0_{crb}$. ■

Proposition 4.6. *Let X and Y be random variables with finite means. Let Z be a supremum of X and Y in the order $\preceq^0_{crb}$. Then, we have that $EZ_+ = \max\{EX_+, EY_+\}$.*

Proof. Let W be a random variable which is the supremum of X and Y given in Proposition 4.5. Thus, $\pi_W(0) = h(0) = \max\{EX_+, EY_+\}$.

Let Z be another supremum of X and Y in the stochastic order $\preceq^0_{crb}$. Thus, we have that $Z \preceq^0_{crb} W$ and $W \preceq^0_{crb} Z$. By Proposition 3.4 we obtain that $-\pi_Z(t) + 2\pi_Z(0) = -\pi_W(t) + 2\pi_W(0)$ for any $t < 0$. The Monotone Convergence Theorem implies that $\pi_W(0) = \pi_Z(0)$, and so we obtain the result. ■

Proposition 4.7. *Let X and Y be random variables with finite means such that $X \leq 0$ and $Y \leq 0$ a.s. Then, the supremum of X and Y with respect to the order $\preceq^0_{crb}$ is unique in distribution.*

Proof. Let W and Z be two suprema of X and Y in the order $\preceq^0_{crb}$, thus $W \preceq^0_{crb} Z$ and $Z \preceq^0_{crb} W$.

Since $X \leq 0$ and $Y \leq 0$ *a.s.*, it holds that $EX_+ = EY_+ = 0$. Applying Proposition 4.6, we obtain that $\pi_W(0) = 0$ and $\pi_Z(0) = 0$. That is, $EW_+ = 0 = EZ_+$, and so $W \leq 0$ and $Z \leq 0$ *a.s.*

Applying Corollary 3.6 we conclude that $Z \sim_{st} W$. ■

Under the assumption of no arbitrage opportunities, the supremum of two variables corresponds to the price of an asset of the call ratio backspread derivative with greater expected benefit (not lower) than those of the variables and with the smallest possible premium. In a similar way, the infimum is a distribution of the price of an asset of the best call ratio backspread which is cheaper (not more expensive) than those of the variables.

The existence of a supremum and an infimum is useful in optimization problems with stochastic dominance constraints (see, for instance, Dentcheva, Lai and Ruszczyński (2004), Dentcheva and Martínez (2012), Dentcheva and Wolfhagen (2016), Singh and Selvamuthu (2017), Consigli, Dentcheva and Maggioni (2021) and the references therein for stochastic dominance constraints and Müller and Scarsini (2006) for lattice of stochastic orders). In the problem

$$\begin{aligned} &\text{maximize} && h(X) \\ &\text{subject to} && X \preceq^0_{crb} W_i, \;\; i = 1, \ldots, m, \\ &&& X \in \mathcal{C} \end{aligned}$$

where $\mathcal{C}$ is a set of random variables, $h : \mathcal{C} \to \mathbb{R}$ is a real valued functional, and W_i, with $i = 1, \ldots, m$, are random variables, the stochastic constraint is equivalent to $X \preceq^0_{crb} \inf\{W_i, \; i = 1, \ldots, m\}$, where the infimum is in the stochastic order $\preceq^0_{crb}$. Thus, there is only one stochastic constraint instead of m. In a similar way, if the stochastic constraints are subject to $W_i \preceq^0_{crb} X$, $i = 1, \ldots, m$, this is equivalent to $\sup\{W_i, \; i = 1, \ldots, m\} \preceq^0_{crb} X$, the supremum being in the order $\preceq^0_{crb}$.

The order $\preceq^0_{crb}$ is not a partial order but a pre-order and lattice structures are defined on partially ordered sets. Let $\sim^0_{crb}$ be the equivalence relation given by $\preceq^0_{crb}$ in the usual way. The following result follows from Proposition 4.2 and Proposition 4.5.

Proposition 4.8. *Let $\mathcal{M}^1$ be the set of random variables with finite mean and $\mathcal{M}^1/\sim^0_{crb}$ be the set of equivalence classes in $\mathcal{M}^1$ with respect to $\sim^0_{crb}$. Let $\widetilde{\preceq}^0_{crb}$ be the relation on $\mathcal{M}^1/\sim^0_{crb}$ given by $[X]\,\widetilde{\preceq}^0_{crb}[Y]$ when $X \preceq^0_{crb} Y$. The set of equivalence classes $\mathcal{M}^1/\sim^0_{crb}$ endowed with $\widetilde{\preceq}^0_{crb}$ is a lattice.*

An equivalence class is made up of random variables of unit prices of assets on the expiration date of call ratio backspread derivatives whose expected benefits are the same. If $X \in [Y]$, $X + p_2 \sim^{p_2}_{crb} Y + p_2$, for any $p_2 \in \mathbb{R}$, that is, $E(f_{p_1,p_2,k}(X + p_2)) = E(f_{p_1,p_2,k}(Y + p_2))$ whatever $p_1 < p_2$ and $k \in \mathbb{R}$. An equivalence class can be interpreted as those assets of call ratio backspread derivatives whose premiums should be equal if there were not options of arbitrage. If elements of an equivalence class can be bought with different premiums, opportunity of arbitrage are being offered in financial markets.

5. An application of the method

This section illustrates the method developed for the analysis of call ratio backspread derivatives. We will compare call ratio backspread derivatives whose assets are the weekly returns of Boeing and Procter & Gamble (P&G), companies in the Dow Jones Industrial Average Index.

Let x_t stand for the weekly close price of a market value at week t. Notice that $\frac{x_t}{x_{t-1}}$ is the price at the end of week t of a monetary unit invested in such a value at the end of week $t-1$. The weekly return is defined as $\frac{x_t}{x_{t-1}} - 1$, that is, the interest rate during the corresponding week.

We will consider the share prices of Boeing and P&G during the period 2019-23. The data of the prices are public and were taken from https://es.finance.yahoo.com/

In Figure 1, we have depicted the daily evolution of such prices during the above period. Figure 2 shows the evolution of the weekly returns.

Let X be the random variable weekly return of P&G and let Y stand for the variable associated with the weekly returns of Boeing.

To show empirical evidence of an ordering between the corresponding call ratio backspread derivatives, we will make use of Proposition 3.4, which reads that $X \preceq^0_{crb} Y$ if and only if $-\pi_X(t) + 2\pi_X(0) \leq -\pi_Y(t) + 2\pi_Y(0)$ for any $t < 0$.

We have depicted the empirical version of the above quantities, that is, $-\widetilde{\pi}_X(t) + 2\widetilde{\pi}_X(0)$ and $-\widetilde{\pi}_Y(t) + 2\widetilde{\pi}_Y(0)$, where $\widetilde{\pi}_X(t) = \int_t^{+\infty} \overline{\widetilde{F}}_X(x)\,dx$ and $\widetilde{F}_X$ stands for the empirical distribution function of the sample associated with the variable X.

Figure 3 contains that representation. The values of t in the graphic cover all the values of the two samples.

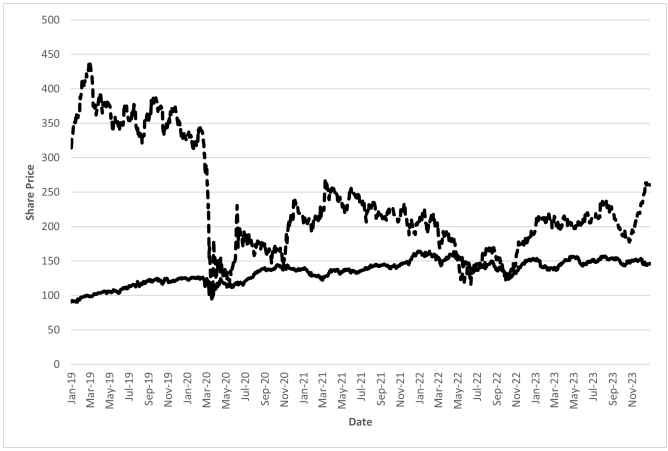


Figure 1. *Evolution of share prices of Boeing and P&G during the years 2019-23. Solid line for P&G, dashed line for Boeing. Dates in horizontal axis, prices in vertical axis.*

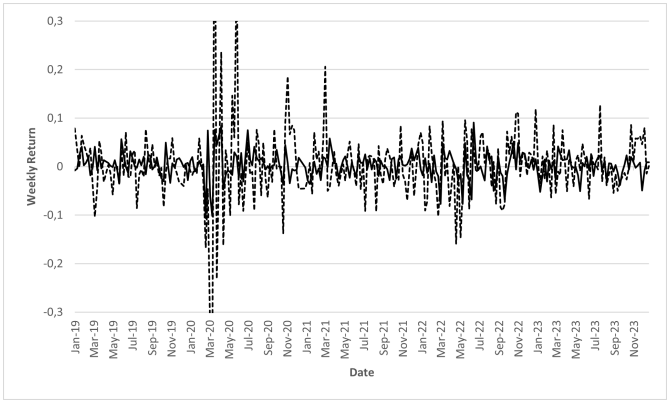


Figure 2. *Evolution of the weekly returns of Boeing and P&G during the years 2019-23. Solid line for P&G, dashed line for Boeing. Dates in horizontal axis, values of the weekly returns in vertical axis.*

Such a representation shows reasonable evidence that X is less that Y in the call ratio backspread stochastic order.

To draw a conclusion on such a relation, we state the following result.

Proposition 5.1. *Let X and Y be random variables with finite means. If $EX_+ \leq EY_+$ and $X_- \preceq_{icv} Y_-$, then $X \preceq^0_{crb} Y$.*

Proof. The condition $X_- \preceq_{icv} Y_-$ is equivalent to

$$\int_{-\infty}^{t} F_{X_-}(x)\,dx \geq \int_{-\infty}^{t} F_{Y_-}(x)\,dx$$

for any $t \in \mathbb{R}$ (see Theorem 4.A.2 in Shaked and Shanthikumar (2007)). Notice that *a.e.* $F_{X_-}(x) = \overline{F}_X(-x)$ when $x \geq 0$, and $F_{X_-}(x) = 0$ if $x < 0$. Thus, when $t < 0$,

$$\int_{-\infty}^{t} F_{X_-}(x)\,dx = 0,$$

and if $t \geq 0$, we conclude that

$$\int_{-\infty}^{t} F_{X_-}(x)\,dx = \int_{0}^{t} F_{X_-}(x)\,dx = \int_{0}^{t} \overline{F}_X(-x)\,dx = \int_{-t}^{0} \overline{F}_X(x)\,dx.$$

Therefore, for any $t \geq 0$,

$$\int_{-t}^{0} \overline{F}_X(x)\,dx \geq \int_{-t}^{0} \overline{F}_Y(x)\,dx,$$

equivalently,

$$-\int_{t}^{0} \overline{F}_X(x)\,dx \leq -\int_{t}^{0} \overline{F}_Y(x)\,dx$$

for any $t < 0$. On the other hand, $EX_+ \leq EY_+$ is $\pi_X(0) \leq \pi_Y(0)$, hence,

$$-\int_{t}^{0} \overline{F}_X(x)\,dx + \pi_X(0) \leq -\int_{t}^{0} \overline{F}_Y(x)\,dx + \pi_Y(0)$$

for any $t < 0$, which is the same as $-\pi_X(t) + 2\pi_X(0) \leq -\pi_Y(t) + 2\pi_Y(0)$ for all $t < 0$, that is, $X \preceq^{0}_{crb} Y$. ■

Notice that the above result permits to use statistical inference techniques to test conditions which lead to the call ratio backspread stochastic order.

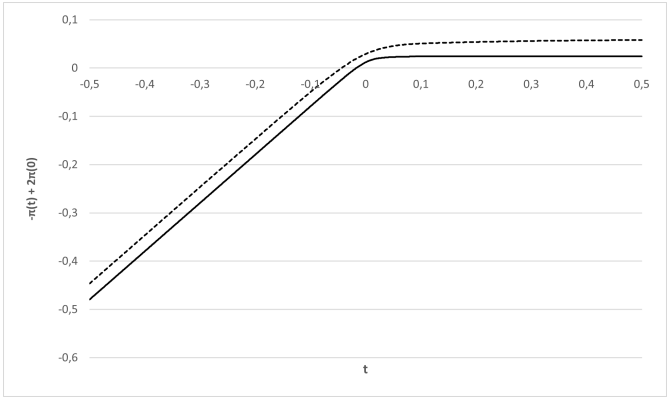


Figure 3. *Representation of the empirical version of the mappings $t \to -\pi(t) + 2\pi(0)$ with $t \in (-0.5, 0.5)$, solid line for P&G, dashed line for Boeing. Values of t in horizontal axis, values of the mappings in vertical axis.*

Some tests have been proposed to infer on the increasing convex order (equivalently, the increasing concave order), see for instance, Zardasht (2015), Berrendero and Cárcamo (2011), Scaillet and Topaloglou (2010) or Baringhaus and Grübel (2009). For our purpose, we took the CX10 test proposed by Berrendero and Cárcamo (2011) since it is quite intuitive, although other tests could be considered.

To apply the corresponding tests, we have divided at random the weekly returns of the period 2019-23 into two disjoint groups, one for Boeing and one for P&G, avoiding paired samples.

The classical run test was applied to all the involved samples (those of X_+, Y_+, X_- and Y_-). The corresponding p-values were greater than the usual level of significance 0.05. Thus, samples can be considered random, assumption needed for the application of the CX10 test and for the test on the comparison of the expectations of the positive parts of the variables. Moreover, the normality assumption was rejected for all the above variables.

Regarding the test with hypothesis

$$H_0 : X_- \preceq_{icv} Y_- \quad \text{versus} \quad H_1 : H_0 \textit{ is false},$$

the p-value of the corresponding samples was higher than 0.99. As a consequence, the null hypothesis is not rejected.

In relation to the test

$$H_0 : EX_+ \leq EY_+ \quad \text{versus} \quad H_1 : EX_+ > EY_+,$$

the p-value of the corresponding samples was 0.9956.

Making use of Proposition 5.1, we conclude that $X \preceq^0_{crb} Y$.

Observe that by Proposition 3.2, the relation $X \preceq^0_{crb} Y$ is equivalent to $X + p_2 \preceq^{p_2}_{crb} Y + p_2$ for any $p_2 \in \mathbb{R}$. Thus, if $p_2 = 1$, we obtain that

$$\frac{X_t}{X_{t-1}} \preceq^1_{crb} \frac{Y_t}{Y_{t-1}},$$

where X_t and Y_t stand for the weekly close prices of P&G and Boeing at week t, respectively.

That is, the expected benefit of a call ratio backspread derivative with asset the unit weekly revaluation of Boeing is greater (not lower) than the corresponding derivative with the asset unit weekly revaluation of P&G, whatever $p_1 < 1$. That shows that if for some $p_1 < 1$, the premium of the derivative associated with Boeing is lower than the premium of P&G, an arbitrage opportunity exists for those derivatives. Moreover, in case of equality of premiums, and investor should choose the Boeing derivative instead of the P&G option.

6. Final comments and conclusions

The present manuscript shows how the theory of stochastic orders can be used for reaching decisions on the allocations of funds in call ratio backspread derivatives. The math-

ematical model proposed in this article permits to compare investments in the above financial derivatives by means of a new family of stochastic orders. That allows to detect possible options of arbitrage.

This procedure entails some advantages with respect to other methods. The proposed technique does not require specific analytical equations or formulas of the prices of the assets, or particular probability distributions of those prices, like Brownian movements, or geometric Brownian movements. Moreover, an advantage of the new method is that when the order is satisfied, the expected benefits are ordered whatever price p_1. Thus, an investor does not need to attain some particular values of p_1 to be able to compare investments and find opportunities. Notice that when the order is satisfied and the premiums do not follow the same arrangement for a particular value of p_1, there exist arbitrage opportunities. On the other hand, we have proved the existence of supremum and infimum of two variables in the new orders, that brings advantages in optimization problems with stochastic constraints.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation under Grant MCI-20-PID2019-104486GB-I00 and Grant MCIU-22-PID2021-123461NB-C22, and Principado de Asturias Government under Grant AYUD/ 2021/50897.

References

Baringhaus, L. and Grübel, R. (2009). Nonparametric two-sample tests for increasing convex order. Bernoulli, 15, 99-123.

Belzunce, F., Martínez-Riquelme, C. and Mulero, J. (2016). An Introduction to Stochastic Orders. Elsevier/Academic Press, Amsterdam.

Berrendero, J.R. and Cárcamo, J. (2011). Tests for the second order stochastic dominance based on L-statistics. Journal of Business & Economic Statistics, 29, 260-270.

Cohen, G. (2005). The Bible of Options Strategies. Pearson Education, Inc., New Jersey.

Consigli, G., Dentcheva, D. and Maggioni, F. (2021). Stochastic optimization: theory and applications. Annals of Operational Research, 292, 575-580.

Dentcheva, D., Lai, B. and Ruszczyński, A. (2004). Dual methods for probabilistic optimization problems. Mathematical Methods of Operations Research, 60, 331-346.

Dentcheva, D. and Martínez, G. (2012). Two-stage stochastic optimization problems with stochastic ordering constraints on the recourse. European Journal of Operational Research, 219, 1-8.

Dentcheva, D. and and Wolfhagen, E. (2016). Two-stage optimization problems with multivariate stochastic order constraints. Mathematics of Operations Research, 41, 1-22.

Dixit, A.K. and Pindyck, R.S. (1994). Investment under Uncertainty. Princeton University Press, Princenton, New Jersey.

Hull, J.C. (2015). Options, futures and other derivatives. Pearson, Boston.

Jarrow, R. (1986). The relationship between arbitrage and first order stochastic dominance. Journal of Finance, 41, 915-921.

Levy, H. (2016). Stochastic dominance. Investment decision making under uncertainty. Third edition. Springer, Cham.

López-Díaz, M.C., López-Díaz, M. and Martínez-Fernández, S. (2018). Stochastic orders to approach investments in condor financial derivatives. Test, 27, 122-146.

Müller, A. (1997). Stochastic orders generated by integrals: a unified study. Advances in Applied Probability, 29, 414-428.

Müller, A. and Scarsini, M. (2006). Stochastic order relations and lattices of probability measures. SIAM Journal on Optimization, 16, 1024-1043.

Müller, A. and Stoyan, D. (2002). Comparison Methods for Stochastic Models and Risks. John Wiley & Sons, Chichester.

Ng, P., Wong, W.-K. and Xiao, Z. (2017). Stochastic dominance via quantile regression with applications to investigate arbitrage opportunity and market efficiency. European Journal of Operational Research, 261, 666-678.

Roberts, A.W. and Varberg, D.E. (1973). Convex functions. Pure and Applied Mathematics, Vol. 57. Academic Press, New York-London.

Scaillet, O. and Topaloglou, N. (2010). Testing for stochastic dominance efficiency. Journal of Business & Economic Statistics, 28, 169-180.

Shaked, M. and Shanthikumar, J.G. (2007). Stochastic Orders. Springer, New York.

Singh, A. and Selvamuthu, D. (2017). Mean-variance optimal trading problem subject to stochastic dominance constraints with second order autoregressive price dynamics. Mathematical Methods of Operations Research, 86, 29-69.

Zardasht, V. (2015). A test for the increasing convex order based on the cumulative residual entropy. Journal of the Korean Statistical Society, 44, 491-497.

SORT 49 (1) January-June 2025, 93-120 DOI: 10.57645/20.8080.02.24

Spatial autoregressive modelling of epidemiological data: geometric mean model proposal

Mabel Morales-Otero[1,2], Christel Faes[3] and Vicente Núñez-Antón[4]

Abstract

We propose the geometric mean spatial conditional model for fitting spatial public health data, assuming that the disease incidence in one region depends on that of neighbouring regions, and incorporating an autoregressive spatial term based on their geometric mean. We explore alternative spatial weights matrices, including those based on contiguity, distance, covariate differences and individuals' mobility. A simulation study assesses the model's performance with mobility-based spatial correlation. We illustrate our proposals by analysing the COVID-19 spread in Flanders, Belgium, and comparing the proposed model with other commonly used spatial models. Our approach demonstrates advantages in interpretability, computational efficiency, and flexibility over the commonly used and previously existing methods.

MSC: *62J05, 62H11, 62M30, 92D30, 62F15.*

Keywords: *Bayesian approaches, COVID-19 incidence, Epidemiology, Spatial modelling.*

1. Introduction

The analysis of spatial data has become widely spread in epidemiology, specially because location can be an important surrogate for lifestyle, environment, as well as genetic

[1] Institute of Data Science and Artificial Intelligence (DATAI), University of Navarra, Calle Universidad 6, 31009, Pamplona, Spain. mmoralesote@unav.es

[2] TECNUN School of Engineering, University of Navarra, Manuel Lardizabal Ibilbidea 13, 20018, Donostia-San Sebastián, Spain.

[3] I-BioStat, Center for Statistics, Hasselt University, Agoralaan gebouw D, 3590 Diepenbeek, Belgium. christel.faes@uhasselt.be

[4] Department of Quantitative Methods, University of the Basque Country (UPV/EHU), Avenida Lehendakari Aguirre 83, 48015 Bilbao, Spain. vicente.nunezanton@ehu.eus

Received: May 2024.
Accepted: March 2025.

and other factors and, therefore, it can provide important insights for public health data analysis. Autoregressive models proposals for analysing spatial data include the Conditional Autoregressive (CAR) model, the auto-Poisson scheme (Besag, 1974) and the Simultaneous Autoregressive (SAR) model (Whittle, 1954), which incorporate the spatial correlation by assuming a conditional covariance structure for an unobservable component included in the regression structure. In addition, the spatial conditional overdispersion models include a spatial lag of the response variable in the regression model specification, which allows to capture the spatial dependence directly observed on neighbouring regions (see Cepeda-Cuervo, Córdoba and Núñez-Antón, 2018; Morales-Otero and Núñez-Antón, 2021). In the case of time series data, Zeger and Qaqish (1988) consider Poisson models that include the logarithm of the past counts in the log-mean regression specification, Knorr-Held and Richardson (2003) propose different autoregressive specifications when including the past counts and Held, Höhle and Hofmann (2005) propose an autoregressive model using an identity link.

An alternative to these models is given by spatial regression models for count data that make use of a spatially structured random effect, which is structured according to a given spatial weights matrix. In this context, two of the most popular models in spatial disease mapping are the Besag-York-Mollié (BYM) model (Besag, York and Mollié, 1991) and the BYM2 model (Riebler et al., 2016). The BYM model incorporates spatial dependence by means of two unobserved latent effects, namely a spatially unstructured random effect and a spatially structured random effect following an Intrinsic Conditional Autoregressive (ICAR) prior (Besag, 1974). In the BYM2 model the latent effect is a weighted average of these two random effects. Another random effects model frequently found in the literature is the Leroux model (Leroux, Lei, and Breslow, 2000). These models are generally estimated using Bayesian inferential methods.

In the aforementioned models, the relationship between two regions is described by a spatial weights matrix, for which several different specifications have been developed (see Anselin, 2002). In most cases, this matrix is fixed and previously specified, a choice that may have an impact on the results of the analysis. Therefore, it is very important for researchers to be able to study how to best describe the spatial structure of the data. Traditionally, spatial weights matrices are based on the adjacency of regions or on the distance among regions. However, there may be situations where the association is not given by the geographical proximity but, instead, it depends on some other connectivity structure or even on the specific characteristics of the regions under study.

In this sense, several authors have explored the use of different weights matrices. Earnest et al. (2007) studied the influence of different specifications of spatial weights matrices on the smoothing properties of the CAR model, obtaining considerable differences in the reported results, which provided a clear evidence about the importance of the proper choice of the spatial structure. In addition, Case, Hines, and Rosen (1993) proposed the use of a similarity matrix based on the inverse of the difference of the values that a given covariate takes in each region, which improved the performance of their fitted models. Ejigu and Wencheko (2020) proposed a weights matrix that took into ac-

count geographical proximity and covariate information simultaneously, which led to a better justification and motivation of the spatial structure present in the data under study.

After the beginning of the pandemic, several authors concentrated their efforts on the different statistical modelling proposals to study COVID-19 data. For example, Sahu and Böhning (2022) proposed a joint spatio-temporal model to analyse the weekly number of cases and deaths related to COVID-19, also presenting different specifications for the spatial and temporal random effects. Konstantinoudis et al. (2022) analysed the weekly number of deaths for several regions in Europe during the period going from 2015 to 2019, fitting a hierarchical Poisson model with a BYM2 specification to these data, thus, being able to evaluate the excess of mortality during the COVID-19 pandemic. Fritz et al. (2022) proposed a Poisson autoregressive model similar to the one in Held et al. (2005), and analysed data from Germany on COVID-19 infections, hospitalizations and intensive care units occupation. Additional references include D'Angelo, Abbruzzo, and Adelfio (2021), Johnson, Ravi, and Braneon (2021) and Natalia et al. (2022), among others. Furthermore, purely spatial approaches have also been used, such as the proposals in Konstantinoudis et al. (2021), where they studied the relationship between COVID-19 related deaths and long-term exposure to air-pollution, fitting a BYM2 model to data concerning the first wave of the disease in England. Other researchers have used the mobility of individuals among regions as a connectivity structure for modelling COVID-19 data. For example, Slater et al. (2022) combined geographical proximity and human mobility data on the BYM specification to spatially model COVID-19 case counts in the regions of Castilla-León and Madrid in Spain from March to June 2020.

In this paper, we propose a geometric mean extension of the spatial conditional models in Cepeda-Cuervo et al. (2018) and Morales-Otero and Núñez-Antón (2021) to account for the spatial autocorrelation that may be present in the data. The spatial conditional model is described in Section 2.1, and the extension is motivated and introduced in Section 2.2. Additionally, we also investigate the use of several spatial weights matrices in the computation of the spatial lag and propose some new possible structures to be implemented, which are discussed in Section 2.3. A simulation study is included in Section 3. The usefulness of our methodological proposals and their comparison with other commonly used spatial models is provided in Section 4. More specifically, a comparison with the BYM2 and Leroux spatial models is included in Section 4.3. In Section 5, we end with a discussion.

2. Methodology

This section reviews the spatial conditional overdispersion models proposed in the literature. Thereafter, we propose an extension of this model and discuss possible weights matrices that could describe the underlying spatial dependency structure.

2.1. Review of the spatial conditional model

The spatial conditional overdispersion models were developed to fit spatial count data, allowing to capture overdispersion and to explain the spatial dependence that may exist in the data, as suggested by Cepeda-Cuervo et al. (2018). These authors assume that the dependent variable Y_i, for regions $i = 1,\ldots,n$, follows a conditional distribution $f(y_i \mid y_{\sim i})$, where y_i represents the observed count in region i and, $y_{\sim i}$, the values in all of the neighbouring regions of the i-th region (without including the i-th region itself). A spatial autoregressive term, more specifically, the lag of the response variable, is incorporated in the regression model specification for the conditional mean $\mathrm{E}(Y_i \mid Y_{\sim i})$. The inclusion of such spatial dependence in the model can explain part of the overdispersion.

In an epidemiological context, interest often goes towards the modelling of the rates of a disease. In this case, Morales-Otero and Núñez-Antón (2021) assumed that the conditioned response variable $(Y_i \mid Y_{\sim i}, \nu_i)$, the total number of cases for $i = 1,\ldots,n$, follows a Poisson distribution, with conditional mean μ_i, so that $\mathrm{E}(Y_i \mid Y_{\sim i}, \nu_i) = \mu_i = \mathrm{P}_i r_i$. Here, P_i represents the population size and r_i the disease rate in the i-th region, for $i = 1,\ldots,n$. They proposed the following regression structure for the conditioned means:

$$\log(\mu_i) = \log(\mathrm{P}_i) + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{Rates} + \nu_i, \tag{1}$$

where an autoregressive component is included for the rates, (i.e., $\mathbf{W}_i\mathbf{Rates} = \sum_{j=1}^{n} w_{ij}\mathrm{Rates}_j$), which is a weighted average of the observed rates $\mathrm{Rates}_j = y_j/\mathrm{P}_j$, with weights specified by the spatial weights matrix $\mathbf{W}$. Here, $\mathbf{x}_i$ is a $1 \times p$ vector of explanatory variables for the i-th observation, $\boldsymbol{\beta} \in \Re^p$ a $p \times 1$ vector of unknown regression parameters that need to be estimated and $\rho \in \Re$ the unknown spatial autoregressive parameter. These parameters and variables belong to the set of all real numbers, as no constraints are imposed. In addition, a normally distributed random effect $\nu_i \sim N(0,\tau)$, with $\tau > 0$, is included to allow for additional unstructured overdispersion in the counts. Note that the assumed spatial structure is given by the matrix $\mathbf{W}$, where its elements, w_{ij}, are weights that represent the strength of the relationship between regions i and j. Section 2.3 includes a detailed description about the different ways these weights can be defined.

2.2. Geometric mean spatial conditional model

Zeger and Qaqish (1988) proposed several models to account for temporal autocorrelation in time series data, including one for count data, where they suggested the use of a Poisson model that incorporates the logarithm of the past counts in the regression model for the logarithm of the mean instead of the past counts. Knorr-Held and Richardson (2003) proposed the use of the term $\log(y_{t-1}+1)$ in order to overcome the issue of the nonexistence of the logarithm, so that it is equal to zero when there are no cases. Held et al. (2005) proposed to regress the mean directly on the past counts instead, but assuming an identity link.

Following the ideas in Zeger and Qaqish (1988) and Knorr-Held and Richardson (2003), we propose the following geometric mean spatial conditional model for count data. As before, we assume a Poisson model for the conditioned response outcomes, that is $(Y_i \mid Y_{\sim i}, \nu_i) \sim \text{Poi}(\mu_i)$, with conditional mean $\text{E}(Y_i \mid Y_{\sim i}, \nu_i) = \mu_i = \text{P}_i \text{r}_i$, following the regression model:

$$\log(\mu_i) = \log(\text{P}_i) + \mathbf{x}_i^\intercal \boldsymbol{\beta} + \rho \mathbf{W}_i \log(\mathbf{Rates}) + \nu_i \tag{2}$$

Here, we believe it is important to mention that, in the presence of zero counts, it would be necessary to use $\log(y_j + 1)$ and $\log(P_j + 1)$ when computing the observed rates (see equation(1)). This model closely resembles the model in equation (1), but here the autoregressive component is a weighted average of the logarithms of the rates, instead of the rates. It can be easily seen that the smoothed estimates of the rates are estimated as:

$$\begin{aligned} \widehat{r_i} &= \exp(\mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}) \exp\left(\frac{1}{n_i} \sum_{j=1}^{n} w_{ij}^* \log(\text{Rates}_j) \right)^{\hat{\rho}} \exp(\nu_i) \\ &= \exp(\mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}) \overline{\overline{\mathbf{Rates}}}_i^{\hat{\rho}} \exp(\nu_i), \end{aligned} \tag{3}$$

with w_{ij}^* representing the non-standardized spatial weights, n_i being the number of neighbours of region i, and $\overline{\overline{\mathbf{Rates}}}_i$ being the geometric mean of the rates included in the vector of the observed rates **Rates**. Note that the geometric mean of a sample $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is defined as $(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$, which can also be expressed as $\exp\left[\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right]$, when $x_i > 0$, for $i = 1, \ldots, n$.

This can also be generalized to the case where the spatial weights matrix is given by some criterion where the weights w_{ij} are not necessarily equal to 0 or 1. This could be the case, for example, in cases where we use criteria based on distance among regions or on the mobility matrix. In these cases, we would have a weighted geometric mean of the rates included in the vector of the observed rates, so that:

$$\overline{\overline{\mathbf{Rates}}}_i = \exp\left(\frac{\sum_{j=1}^{n} w_{ij}^* \log(\text{Rates}_j)}{\sum_{j=1}^{n} w_{ij}^*} \right) \tag{4}$$

Therefore, the estimated value obtained for the spatial parameter ρ would represent how the incidence rate in the regions resembles the (weighted) geometric mean of the rates in their neighbours. Consequently, the use of the logarithm of the rates in the autoregressive component has an important epidemiological interpretation. For a better understanding of this effect, in Section 9 of the supplementary material, we have included a detailed description and better motivation of the effect of the geometric mean of the rates on the estimated disease rate, considering different values of the estimated spatial parameter.

2.3. Spatial weights matrices

As already stated, the models proposed in Section 2 do not impose or need any restrictions when specifying the spatial weights matrix and, therefore, they are very flexible,

allowing the use of a wide range of spatial structures. Moreover, this flexibility makes them a valuable tool for exploring different spatial weights matrices in a specific dataset. This section discusses different possible choices for specifying the weights w_{ij} used in the proposed model in equation (2).

2.3.1. Spatial weights matrices based on contiguity

The spatial structure based on contiguity or adjacency is defined by the spatial weights matrix $\mathbf{W}$, where $w_{ij} = 1$, if region i is adjacent or a neighbour to region j, and $w_{ij} = 0$, otherwise. Different criteria can be assumed to specify whether two regions are adjacent. For example, the Queen contiguity criteria assumes that regions i and j are neighbours if they share at least one point in their boundaries. Most commonly the spatial weights matrix is standardized by rows, so that if region i is adjacent to region j, then $w_{ij} = 1/n_i$, where n_i is the number of neighbours region i has. In this way, the spatial lag $\mathbf{W}_i\mathbf{y}$ can be viewed as a spatial average of the values that the variable takes in all of its neighbouring locations.

First order contiguity is specified when we consider that regions i and j are neighbours if they share at least one point in their boundaries. This specification is also known as Queen contiguity criterion. Extending this criteria by considering that i and j are neighbours if they share a common neighbour, we can define second order contiguity. Third order contiguity can be specified the same way, when it is assumed that regions i and j are adjacent if they share a common neighbour of order two. Contiguity of higher order is also possible to specify by following these ideas.

2.3.2. Spatial weights matrices based on distance

An alternative way to define a spatial structure is to consider a spatial weights matrix where its elements are defined as a function of the distance among the central points of the polygons representing the regions, called the centroids, s_i $(i = 1, \ldots, n)$. Inverse distance weights are specified as $w_{ij} = 1/\|s_i - s_j\|$, with $\|s_i - s_j\|$ being the Euclidean distance between regions i and j. In addition, in the negative exponential criteria the weights are defined so that $w_{ij} = \exp(-\|s_i - s_j\|)$.

Finally, we can also define the distance band weights, with band width given by a critical threshold h. In particular, it is considered that regions i and j are neighbours if their centroid lies within the chosen band. Let s_i be the centroids of the regions under study, for a given threshold h, then $w_{ij} = 1$ if the Euclidean distance between s_i and s_j is smaller than h, that is $\|s_i - s_j\| < h$, and 0 otherwise.

2.3.3. Covariate-based similarity (or difference) matrices

Ejigu and Wencheko (2020) proposed a weights matrix $\mathbf{W}$, which not only takes into account geographical proximity, but also a specific covariate's information. Given an environmental variable e_i $(i = 1, \ldots, n)$ for n regions with centroids s_i, they define the following structure for the weights:

$$w_{ij} = \exp\{-[\alpha|e_i - e_j| + (1-\alpha)\|s_i - s_j\|]\}, \tag{5}$$

where α is a previously selected fixed value between zero and one, $|e_i - e_j|$ is the absolute difference in the value of the environmental covariate between regions i and j and $\|s_i - s_j\|$ is the Euclidean distance between the centroids of regions i and j. The elements in the diagonal of this matrix are zero and it is row standardized. As α approaches zero, the weights give more relevance to the geographical distance, and, when it approaches one, the covariate differences receive more importance.

Following this idea, we also propose an alternative covariate-based similarity matrix, where we will consider both environmental and socio-economic variables to impact the weight amongst regions. Let $\mathbf{W}$ be a traditional weights matrix based on contiguity, distance, or any other criteria, with elements w_{ij}, and $\mathbf{D}$ an $n \times n$ matrix with elements $d_{ij} = 0$ if $i = j$ and:

$$d_{ij} = \exp(-|e_i - e_j|), \text{ for } i \neq j, \tag{6}$$

We then propose the use of the matrix $\mathbf{W} \circ \mathbf{D}$, which is the Hadamard (or element-wise) product of matrices $\mathbf{W}$ and $\mathbf{D}$. In this way, small weights are given to neighbouring regions with large differences in the values of the covariate and to distant regions, while large weights are given to neighbouring regions with similar covariate information and that are geographically close to each other.

A potential concern might arise regarding whether specifying covariate-based similarity matrices in the model described by equation (2), while also including these covariates as independent variables, could lead to endogeneity problems. As discussed by Case et al. (1993), when the weights matrix $\mathbf{W}$ is constructed based on similarities or differences in covariates between municipalities, and the vector of observations for the covariates captures within-municipality variations, this design ensures that the elements of the weights matrix are orthogonal to the explanatory variables. Therefore, by construction, this approach eliminates any induced correlation between the covariates and the error term, thus addressing potential endogeneity issues.

2.3.4. Mobility matrix

The previous proposals presented here for the weights matrices are a representation of how close (in space) and/or how similar (in terms of covariate information) regions are. Another characteristic to define the weights matrix is to assess how much contact there was amongst individuals in the different regions. This is of special interest when considering, for example, an outcome that depends on the contact behaviour, such as is the case in infectious disease incidence. As a proxy for the contact behaviour, and based on mobile phone data, the mobility amongst regions can be used. That is, each element m_{ij} in the mobility matrix $\mathbf{M}$ is defined as the mean proportion of time that people from region i have spent in region j in a given time period. This matrix would then clearly represent a different type of connectivity structure among regions.

2.4. Model estimation and selection

All models considered here are fitted using the integrated nested Laplace approximation (INLA) approach, in the R-INLA package. It should be noted, however, that, in general,

any software methodology that allows for estimation of a generalized linear mixed model can be used to implement this model. This is a great advantage of the proposed method, as one is not restricted to complex estimation tools for fitting spatial models.

In addition, it could be worth addressing the potential risk of spatial confounding in the proposed geometric mean spatial model. Spatial confounding arises when covariates share similar spatial patterns with unobserved spatial processes or random effects. In our model, however, the spatial lag of the logarithm of the observed rates is used as an explanatory variable, directly incorporating the observed spatial structure. Since no additional spatially structured random effects are employed and the spatial structure is assumed to be fully observed, the model theoretically mitigates the issue of spatial confounding. The spatial dependence is captured through the geometric mean of neighbouring observations, minimizing the risk of confounding spatial random effects with covariate effects.

Model comparison is carried out by using the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010), where the smallest values indicate the best fitting model. Additionally, we also use the Conditional Predictive Ordinate (CPO) diagnostic (Pettit, 1990), which is a leave-one-out predictive measure. More specifically, for each observation i, the CPO_i is computed, so that it reflects the posterior probability of observing that value, given the other observations. In this way, we would be able to compute a global value by using the sum of the logarithms for the resulting CPO_i values (i.e., $\text{CPO} = -\sum_{i=1}^{n} \log(\text{CPO}_i)$) As in the case of the WAIC, the model with the smallest CPO value would be considered as the best fitting one.

Furthermore, these model selection criteria ensure a balance between model fit and complexity by penalizing overly complex models, helping in this way to prevent possible overfitting. To further assess the model's generalizability, cross-validation techniques like CPO evaluate predictive performance by measuring how well the model generalizes to unseen data. This approach ensures that the model does not overfit the observed data and is capable of making accurate predictions under new scenarios.

In addition, for all the estimated parameters, noninformative prior distributions are assumed. In particular, for the fixed effects and for the precision parameters, we assume independent normal $N(0, 1 \times 10^5)$ distributions, and gamma $G(1 \times 10^{-4}, 1 \times 10^{-4})$ distributions, respectively.

3. Simulation study

As already mentioned in Section 1, most spatial modelling applications make use of a spatial weights matrix following traditional criteria, such as the ones based on contiguity or distance among the regions. However, in this section, we wish to assess the performance of our proposed models in the selection of the weights matrix, as well as to study the sensitivity of the parameters to a misspecified neighbourhood matrix.

Therefore, we have carried out a simulation study, where we induce correlation in the response variable following the mobility matrix structure. For this purpose, we have

implemented a Gibbs sampling algorithm, which allowed us to generate spatially autocorrelated Poisson data by repeatedly sampling from conditional distributions (see Jackson and Sellers, 2008). In our specific case, we define a set of initial values for the parameters β, ρ and τ and, on each iteration, we draw Poisson samples, where the mean is conditioned on the values of the previous iteration. Additional details on the algorithm described below can be found in Section 2 in the supplementary material. We would like to remark that the data have been simulated using the mobility matrix provided in the dataset corresponding to the COVID-19 cases in the municipalities of Flanders, which will be analysed in Section 4. Moreover, this spatial structure has also been used to obtain the contiguity of order one and the inverse distance spatial weights matrices.

We have defined twelve different scenarios, given the true values for the parameters, which can be consulted in the first column to the left in Table 1. For each case, we have simulated $S^* = 500$ datasets (with the number of regions $n = 300$), and discarded half of them, so that $S = 250$ simulations for each scenario remained. Model (2) has been fitted to each of the simulated dataset, considering three different specifications for the spatial structure, one using the mobility matrix to compute the spatial lag, another one using the contiguity of order one spatial weights matrix, and a third one using the inverse distance spatial weights matrix. In addition, we have also fitted the BYM2 and Leroux models, both using the contiguity of order one spatial weights matrix, which is the standard specification for such models. For further details about these models, refer to Section 8 in the supplementary material.

Table 1 reports the bias, mean squared error (MSE) and the coverage of the estimates obtained from fitting each model to the simulated datasets. This table includes only the results obtained for the geometric mean model using the three different spatial weights structures mentioned above. The BYM2 and Leroux models have different formulations and produce estimates that are not directly comparable to those of the geometric mean model. Therefore, they are excluded from this analysis and will be considered later when evaluating and comparing the models' goodness of fit in the specific dataset under study.

For the scenarios where the parameters' true values were $\beta = -2$ and $\rho = 0.5$ (i.e., first two scenarios), the smallest bias was obtained for the estimations for the model using the mobility matrix, indicating that this is the model where the resulting estimates are closer to the true values of the parameters. In these scenarios the coverage percentages in the models using the mobility matrix are also the largest, indicating that most of the credible intervals of the estimated parameters in these models contain the true values. However, when the true value for β changed to -0.5 (i.e., third and fourth scenarios), the smallest bias and the best coverage were obtained for the model using the contiguity criterion for the weights matrix, which seems to suggests that the value given to the intercept β is having a significant impact on the results. This substantial influence can be attributed to the fact that the model does not include any covariates apart from the offset (i.e., the logarithm of the population in each municipality), the intercept itself, and the spatial lag of the logarithm of the rates. Consequently, the intercept determines the baseline level of expected counts across municipalities, directly influencing the scale of

the predicted counts, the overall variability, and the relative contribution of the spatial lag term.

In the scenarios where the true value for ρ is set to 0.2, the bias of the estimates considerably increases when using the mobility matrix. In fact, the estimations with smallest bias are obtained for the model using the inverse distance criterion for the spatial matrix and, moreover, the coverage is very high for all the models. This can be due to the fact that here we are setting a small value for the spatial parameter and, thus, forcing the mobility connectivity structure to have a smaller relevance in the simulated data.

In addition, for the parameters' true values $\beta = -2$ and $\rho = 0.9$, the smallest bias of the estimates and the best coverage were also obtained for the mobility matrix. Given that, in this case, we are setting a large value for the spatial parameter, more relevance is given to this structure. However, when $\beta = -0.5$ and $\rho = 0.9$, the models using the contiguity and the mobility matrix produce similar values for the bias of the estimations and the coverage for the contiguity matrix highly improves, meaning that, for this specific setting, the spatial structure is not so clearly defined.

The sensitivity of the results to small variations in the parameter ρ is due to the absence of additional covariates or effects in the model, making, therefore, the spatial autoregressive term the primary source of variation. The corresponding result would be that even minor adjustments to ρ can significantly influence the dynamics of the overall model and the resulting estimates.

Finally, from the results included in Table 1, for the precision parameter τ, no significant changes were observed in the bias of the estimations or in the coverage when changing this parameter's value from 5 to 15.

Regarding the predictive accuracy of the models, we can evaluate it by computing the mean squared predictive error (MSPE) of the simulated rates for each simulated dataset $\left[\text{MSPE}_s = \sum_{i=1}^{n}(r_i^{(s)} - \hat{r}_i^{(s)})^2/n\right]$ (Carroll et al., 2015). In this way, we can obtain an average for the model fitted for each of the 250 datasets generated for each scenario, so that $\text{MSPE} = \sum_{s=1}^{\text{S}}(\text{MSPE}_\text{s})/\text{S}$. Note that the models with the lowest values of the MSPE would be considered as the best fitting ones. The results obtained are included in Table 2, where we can see that, in general, the MSPE is small in every scenario, but the smallest values are mostly obtained for the models in which the mobility matrix was used to compute the spatial lag of the log-rates.

Moreover, we have counted the number of times that the information criteria values were smaller in each case so that we can check how many times the "correct" model was selected as the best fitting one. Most of the times, with a very few exceptions, the model where the mobility matrix was used, was selected with the smallest WAIC and CPO values. This indicates that we can indeed, based on the model selection criteria, select the underlying true neighbourhood matrix. These results are included in Table S1 in the supplementary material.

From the results obtained in the simulation study, we can conclude that it is essential to evaluate whether the spatial structure used in a study is the most adequate one. For most of the spatial modelling applications, the spatial weights matrix employed to de-

Table 1. *Results obtained from the models using different weights matrices, fitted to the simulated datasets.*

Fitted model:		Mobility			Contiguity			Inverse distance		
True value		β	ρ	τ	β	ρ	τ	β	ρ	τ
$\beta=-2$,	Bias	0.172	0.044	0.806	-0.718	-0.187	-6.863	12.064	3.049	-6.362
$\rho=0.5$,	MSE	0.031	0.002	0.846	0.524	0.036	47.160	145.754	9.312	40.547
$\tau=15$	Coverage	97%	96%	100%	6%	3%	0%	0%	0%	0%
$\beta=-2$,	Bias	0.347	0.088	0.319	-0.487	-0.130	-1.303	10.892	2.751	-1.217
$\rho=0.5$,	MSE	0.123	0.008	0.113	0.245	0.017	1.705	118.791	7.577	1.489
$\tau=5$	Coverage	84%	82%	100%	68%	67%	0%	0%	0%	0%
$\beta=-0.5$,	Bias	0.232	0.233	0.406	0.037	0.027	-0.219	2.623	2.639	-0.296
$\rho=0.5$,	MSE	0.054	0.054	0.166	0.001	7.769e-04	0.049	6.883	6.966	0.088
$\tau=5$	Coverage	0%	0%	100%	100%	100%	100%	0%	0%	100%
$\beta=-0.5$,	Bias	0.187	0.189	1.181	-0.010	-0.018	-1.285	2.612	2.632	-1.367
$\rho=0.5$,	MSE	0.035	0.036	1.409	2.215e-04	4.489e-04	1.669	6.824	6.932	1.886
$\tau=15$	Coverage	0%	0%	100%	100%	100%	100%	0%	0%	100%
$\beta=-2$,	Bias	0.138	0.055	0.855	-0.539	-0.219	-0.318	0.224	0.088	-0.317
$\rho=0.2$,	MSE	0.020	0.003	0.783	0.293	0.048	0.153	0.282	0.045	0.153
$\tau=15$	Coverage	100%	100%	100%	0%	0%	100%	100%	100%	100%
$\beta=-2$,	Bias	0.229	0.092	0.307	-0.501	-0.203	0.045	-0.066	-0.029	0.045
$\rho=0.2$,	MSE	0.054	0.009	0.097	0.252	0.042	0.005	0.130	0.021	0.005
$\tau=5$	Coverage	100%	100%	100%	68%	60%	100%	100%	100%	100%
$\beta=-0.5$,	Bias	0.083	0.131	0.308	-0.095	-0.156	0.171	-0.071	-0.118	0.169
$\rho=0.2$,	MSE	0.007	0.017	0.095	0.009	0.024	0.030	0.006	0.017	0.029
$\tau=5$	Coverage	100%	100%	100%	100%	100%	100%	100%	100%	100%
$\beta=-0.5$,	Bias	0.082	0.130	0.928	-0.104	-0.171	0.440	-0.058	-0.095	0.438
$\rho=0.2$,	MSE	0.007	0.017	0.869	0.011	0.029	0.203	0.007	0.017	0.201
$\tau=15$	Coverage	100%	100%	100%	100%	100%	100%	100%	100%	100%
$\beta=-2$,	Bias	0.106	0.012	110.710	0.584	0.117	-14.665	63.681	7.041	-14.651
$\rho=0.9$,	MSE	0.013	2.227e-04	38995.408	0.640	0.018	215.050	4085.262	49.935	214.643
$\tau=15$	Coverage	84%	92%	68%	60%	48%	0%	0%	0%	0%
$\beta=-2$,	Bias	0.191	0.020	13.815	0.410	0.093	-4.615	61.910	6.859	-4.664
$\rho=0.9$,	MSE	0.039	4.853e-04	275.350	0.506	0.013	21.701	3860.889	47.386	21.756
$\tau=5$	Coverage	89%	93%	67%	69%	59%	0%	0%	0%	0%
$\beta=-0.5$,	Bias	0.353	0.075	0.327	0.295	0.054	-2.981	15.885	3.387	-3.219
$\rho=0.9$,	MSE	0.125	0.006	0.132	0.089	0.003	8.891	252.426	11.474	10.362
$\tau=5$	Coverage	0%	0%	100%	100%	100%	0%	0%	0%	0%
$\beta=-0.5$,	Bias	0.174	0.037	0.647	0.194	0.033	-11.888	15.331	3.287	-12.256
$\rho=0.9$,	MSE	0.031	0.001	0.819	0.040	0.001	141.331	235.405	10.822	150.208
$\tau=15$	Coverage	43%	45%	100%	100%	100%	0%	0%	0%	0%

Table 2. *Average of the MSPE values obtained from the models using different weights matrices, fitted to the simulated datasets.*

True values	Mobility	Contiguity	Inverse distance	BYM2	Leroux
$\beta = -2, \rho = 0.5, \tau = 15$	2.323e-06	5.173e-05	5.601e-05	1.977e-05	1.973e-05
$\beta = -2, \rho = 0.5, \tau = 5$	9.659e-07	2.994e-05	3.190e-05	1.092e-05	1.042e-05
$\beta = -0.5, \rho = 0.5, \tau = 5$	1.765e-06	1.223e-05	1.212e-05	5.402e-06	5.622e-06
$\beta = -0.5, \rho = 0.5, \tau = 15$	2.292e-06	4.135e-05	4.204e-05	1.696e-05	1.724e-05
$\beta = -2, \rho = 0.2, \tau = 15$	2.764e-06	1.793e-05	1.796e-05	1.555e-05	1.554e-05
$\beta = -2, \rho = 0.2, \tau = 5$	1.683e-06	1.004e-05	1.005e-05	8.350e-05	8.402e-05
$\beta = -0.5, \rho = 0.2, \tau = 5$	3.518e-06	6.521e-06	6.540e-06	5.322e-06	5.516e-06
$\beta = -0.5, \rho = 0.2, \tau = 15$	5.497e-06	1.594e-05	1.598e-05	1.291e-06	1.336e-06
$\beta = -2, \rho = 0.9, \tau = 15$	6.441e-06	9.049e-06	9.917e-06	1.735e-05	1.852e-05
$\beta = -2, \rho = 0.9, \tau = 5$	4.254e-06	9.247e-06	9.594e-06	1.714e-05	2.275e-06
$\beta = -0.5, \rho = 0.9, \tau = 5$	8.752e-07	5.207e-05	4.737e-05	2.345e-05	6.582e-06
$\beta = -0.5, \rho = 0.9, \tau = 15$	2.551e-06	1.071e-04	9.862e-05	1.788e-05	1.482e-05

scribe the spatial structure of the data under study is the one following the contiguity of order one criterion. However, we believe it has been clearly shown that this is not always necessarily the best choice.

In this specific study, it has been shown that when the mobility matrix is the underlying structure, and the model is misspecified, in general, the bias of the estimations is larger than the bias obtained for the model using the mobility matrix. Moreover, information criteria values such as the WAIC and CPO and, also predictive accuracy measures such as the MSPE, have favoured the correctly specified model, selecting it as the best fitting one in almost all cases. Overall, the simulation study illustrates the fact that our proposed model effectively identifies the correct spatial structure when properly specified. However, this does not imply that our model is the correct or best model under a given setting, which was not the original purpose in the simulation study.

4. Illustration of methodology

4.1. Data Exploration

We investigate the spatial distribution of COVID-19 from September 2020 until January 2021 amongst the Flemish municipalities. Figure 1 shows the observed incidence of COVID-19 per 100,000 inhabitants in Flanders' municipalities in the time period con-

sidered, which was the time of the second wave in Belgium. It can be observed that not all municipalities presented the same impact in the second COVID-19 wave. In this study, we wish to assess whether the spatial correlation pattern of the incidence during the second wave of the disease was linked to any social demographics.

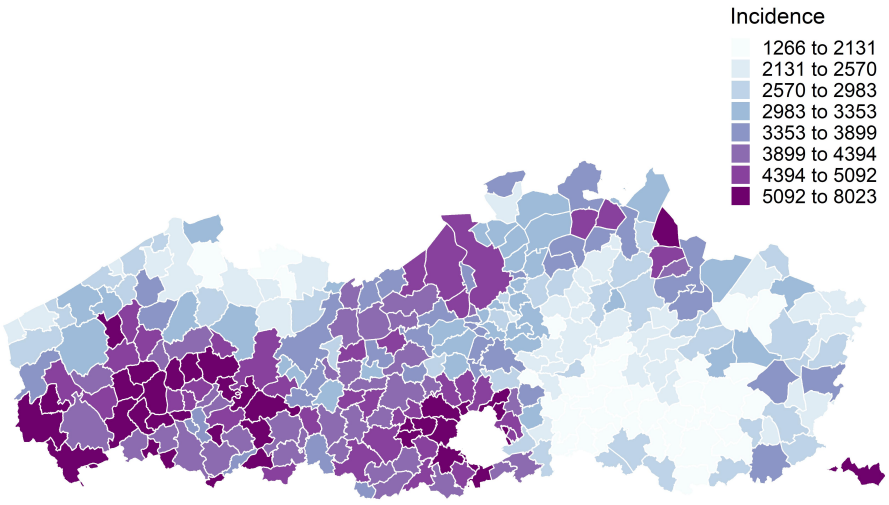


Figure 1. *Spatial distribution of the incidence of COVID-19 per 100,000 inhabitants in Flanders' municipalities from 2020-09-01 to 2021-01-31.*

The data under analysis includes information on the 300 municipalities of Flanders in Belgium, which is available at the website of the Belgian Institute for Public Health (Sciensano) (https://epistat.wiv-isp.be/covid/). Table 3 includes some descriptive statistics for the variables available across municipalities. The outcome of interest is the number of confirmed COVID-cases from 2020-09-01 to 2021-01-11, summarized by the variable `N.cases`. The population size in the municipality is denoted as `P`, and `incidence` is the number of COVID-19 cases in this time period per 100,000 inhabitants. There are also two additional variables available which can be considered as proxies for the socio-economic status and demography of the municipality. These are the percentage of households with a discount on the electricity meter (`budgetmeters`) and the percentage of single-parent households (`single_house`).

4.2. Model Estimates

The COVID-19 incidence map in Figure 1 suggests the presence of spatial autocorrelation in the data, as municipalities with similar values of COVID-19 incidence are grouped together in space. Therefore, in order to be able to characterize the spatial pattern of the second wave of COVID-19 in the data under study, we will further analyse these data by implementing spatial models that account for the spatial dependence. In addition, since we have seen the importance of testing different spatial structures for the

Table 3. *Descriptive statistics for the variables available across municipalities*

	Median	Mean	SD	Min.	Max.
N.cases	544.000	801.243	1570.136	1.000	24387.000
P	15036.500	22097.143	36156.404	79.000	529247.000
incidence	3358.868	3564.184	1235.740	1265.823	8023.070
single_house	7.885	8.021	1.245	5.300	15.300
budgetmeters	1.125	1.229	0.678	0.000	6.860

weights matrix with the simulation study carried out in the previous section, here we will also explore different possible choices for the spatial weights matrix to be included in the fitted models.

We fit both the spatial conditional normal Poisson model in equation (1) (Section 2.1) and the proposed geometric mean spatial conditional normal Poisson model in equation (2) (Section 2.2). As one of our objectives is to be able to select the spatial structure that best accommodates the spatial underlying process in the data, we use the different weights matrices described in Section 2.3, and compare the fitting of the different models by using their WAIC and CPO values. Note that, in this specific application, we do not include any covariates in the linear predictor, as we focus on the spatial modelling by means of the autoregressive terms and on the comparison of the performance of such models.

We believe it is important to mention the fact that, at the beginning of this research, the variables available were included in the model as covariates. However, the results obtained suggested that they did not offer any improvements in models' fitting in terms of information criteria. Therefore, in this specific study, we decided to only employ them when computing the proposed weights matrices based on similarities. It should also be noted that, in this study, we do not aim to identify any risk factor in the spreading of the infection, but to investigate the spatial correlation that may exist in the data and find the structures that best accommodate it.

The results obtained for the fitting of these models are included in Tables 4 and 5, which were fitted by considering ten different options for the weights matrix. First, we have used the spatial weights matrices based on the adjacency among municipalities (contiguity of first and third order). Second, weights were based on the distance among the centroids of the municipalities (inverse distance, negative exponential distance and distance band method). Third, the weights matrices were based on the product between covariate differences and traditional spatial weights, as proposed in Section 2.3. For these similarity matrices, the spatial weights matrices considered are the ones based on contiguity or first order, and that based on the distance band. The variables used to measure whether municipalities have a similar socio-economic status are single_house and budgetmeters. Finally, the mobility matrix was also considered. The heatmaps for these matrices are included in Figure S1 in the supplementary material, where we can clearly see the different structures they represent.

Table 4. *Results obtained after fitting the spatial conditional normal Poisson models to the COVID-19 incidence data in Flanders, for the different weights matrices considered.*

Weights matrix			$\hat{\beta}$	$\hat{\rho}$	$\hat{\tau}$
Contiguity of order 1	WAIC = 2947.7	Mean	-4.378	27.928	27.830
	CPO = 1807.4	SD	(0.041)	(1.121)	(2.440)
		95% CI	(-4.459,-4.297)	(25.728,30.129)	(23.287,32.868)
Contiguity of order 3	WAIC = 2942.7	Mean	-4.425	29.220	18.100
	CPO = 1868.4	SD	(0.060)	(1.637)	(1.542)
		95% CI	(-4.542,-4.307)	(26.006,32.434)	(15.217,21.271)
Inverse distance	WAIC = 2941.2	Mean	-5.649	63.083	19.319
	CPO = 1859.3	SD	(0.120)	(3.333)	(1.646)
		95% CI	(-5.885,-5.413)	(56.538,69.629)	(16.242,22.705)
Negative exponential	WAIC = 2941.4	Mean	-6.006	73.132	12.660
	CPO = 1921.9	SD	(0.221)	(6.152)	(1.062)
		95% CI	(-6.440,-5.573)	(61.049,85.214)	(10.672,14.843)
Distance band	WAIC = 2938.4	Mean	-4.514	31.480	24.663
	CPO = 1822.6	SD	(0.051)	(1.370)	(2.120)
		95% CI	(-4.613,-4.415)	(28.790,34.172)	(20.706,29.031)
$W \circ D$ single_house and	WAIC = 2946.9	Mean	-4.349	27.121	26.956
Contiguity of order 1	CPO = 1813.9	SD	(0.041)	(1.112)	(2.355)
		95% CI	(-4.430,-4.268)	(24.938,29.305)	(22.567,31.814)
$W \circ D$ single_house and	WAIC = 2940.3	Mean	-4.490	30.970	27.609
Distance band	CPO = 1811.3	SD	(0.046)	(1.246)	(2.396)
		95% CI	(-4.580,-4.400)	(28.524,33.418)	(23.142,32.550)
$W \circ D$ budgetmeters and	WAIC = 2949.4	Mean	-4.360	27.500	29.005
Contiguity of order 1	CPO = 1814.6	SD	(0.040)	(1.074)	(2.556)
		95% CI	(-4.438,-4.282)	(25.392,29.609)	(24.247,34.284)
$W \circ D$ budgetmeters and	WAIC = 2938.7	Mean	-4.513	31.657	27.009
Distance band	CPO = 1813.9	SD	(0.047)	(1.293)	(2.336)
		95% CI	(-4.606,-4.420)	(29.120,34.196)	(22.652,31.823)
Mobility	WAIC = 2972.6	Mean	-4.270	25.113	22.010
	CPO = 1849.2	SD	(0.045)	(1.213)	(1.974)
		95% CI	(-4.357,-4.182)	(22.726,27.491)	(18.342,26.093)

When comparing the models' fit related to the different weights matrices included in Table 4, it can be seen that parameter estimates can differ considerably. The estimated value for the autoregressive parameter ρ is large and statistically significant, according to its 95% credible interval, in all models, an indication that there is a clear sign for the existence of spatial autocorrelation. Interpretation of the value of the estimated parameter is difficult, however.

Table 5. *Results obtained after fitting the geometric mean spatial conditional normal Poisson models to the COVID-19 incidence data in Flanders, for the different weights matrices considered.*

Weights matrix			$\hat{\beta}$	$\hat{\rho}$	$\hat{\tau}$
Contiguity of order 1	WAIC = 2945.1	Mean	-0.786	0.770	22.488
	CPO = 1920.6	SD	(0.122)	(0.036)	(1.938)
		95% CI	(-1.027,-0.546)	(0.699,0.840)	(18.871,26.479)
Contiguity of order 3	WAIC = 2942.1	Mean	-0.885	0.740	15.724
	CPO = 1918.5	SD	(0.162)	(0.048)	(1.331)
		95% CI	(-1.202,-0.567)	(0.647,0.834)	(13.235,18.459)
Inverse distance	WAIC = 2941	Mean	3.753	2.108	19.144
	CPO = 1857.8	SD	(0.380)	(0.112)	(1.630)
		95% CI	(3.006,4.500)	(1.887,2.328)	(16.097,22.494)
Negative exponential	WAIC = 2941.4	Mean	4.919	2.451	12.709
	CPO = 1922.7	SD	(0.695)	(0.205)	(1.067)
		95% CI	(3.554,6.283)	(2.049,2.854)	(10.712,14.900)
Distance band	WAIC = 2938.6	Mean	0.241	1.071	25.000
	CPO = 1820.9	SD	(0.157)	(0.046)	(2.151)
		95% CI	(-0.067,0.550)	(0.980,1.161)	(20.985,29.432)
$W \circ D$ single_house and	WAIC = 2945.3	Mean	-0.811	0.762	22.535
Contiguity of order 1	CPO = 1905.7	SD	(0.121)	(0.036)	(1.943)
		95% CI	(-1.048,-0.573)	(0.692,0.832)	(18.909,26.536)
$W \circ D$ single_house and	WAIC = 2940.4	Mean	0.216	1.062	28.129
Distance band	CPO = 1806.3	SD	(0.144)	(0.042)	(2.445)
		95% CI	(-0.066,0.498)	(0.979,1.145)	(23.571,33.170)
$W \circ D$ budgetmeters and	WAIC = 2946.3	Mean	-0.780	0.773	23.622
Contiguity of order 1	CPO = 1928.9	SD	(0.118)	(0.035)	(2.045)
		95% CI	(-1.012,-0.549)	(0.702,0.839)	(19.809,27.838)
$W \circ D$ budgetmeters and	WAIC = 2938.9	Mean	0.268	1.076	27.501
Distance band	CPO = 1812.3	SD	(0.148)	(0.043)	(2.380)
		95% CI	(-0.023,0.558)	(0.991,1.162)	(23.058,32.402)
Mobility	WAIC = 2960.6	Mean	-1.766	0.482	14.842
	CPO = 1915.7	SD	(0.115)	(0.034)	(1.275)
		95% CI	(-1.993,-1.541)	(0.416,0.548)	(12.460,17.466)

The information criteria values obtained (i.e., WAIC) for the fitting of these models indicate that the best fit for the models accounting only for contiguity or distance amongst municipalities is for the distance band spatial weights (WAIC = 2938.4). With regard to the predictive accuracy measure (i.e., CPO), the best fitting model is the one using the contiguity of order one criterion (CPO = 1807.4). As for the models taking into

account the similarity in socio-economic status, the combination of `single_house` or `budgetmeters` and distance bands are the best fitting models (WAIC = 2940.3 and CPO = 1811.3, and WAIC = 2938.7 and CPO = 1813.9, respectively).

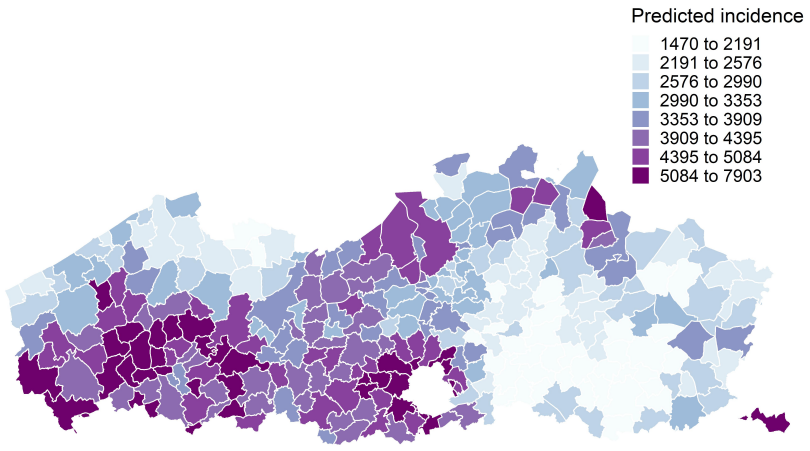
For the model considering the mobility matrix, we can conclude that, according to the information and the predictive criteria, this model did not provide a good fit for the dataset under study. Unlike the simulation study, where the spatial pattern was explicitly constructed based on the mobility matrix and the models accurately identified this structure, in this case, the mobility structure does not seem to be the underlying structure driving the spatial pattern of incidence rates in the dataset under study.

Similar results are observed in Table 5, where the fitting of these models appears to be very similar, according to the WAIC and CPO values, to the ones reported in Table 4. Here, the models with the smallest values were the ones using the distance band weights matrix (WAIC = 2938.6 and CPO = 1820.9) and similarity matrix of the distance band and `single_house` or `budgetmeters` (WAIC = 2940.4 and CPO = 1806.3, and WAIC = 2938.9 and CPO = 1812.3, respectively). In these weights matrices, larger weights are specified for municipalities that lie within the distance band and have similar values of these variables. Therefore, the fitting of these models suggests that this structure could be properly explaining the underlying spatial dependence, assuming that the variables considered represent the socio-economic or demographic characteristics of the population in these municipalities.

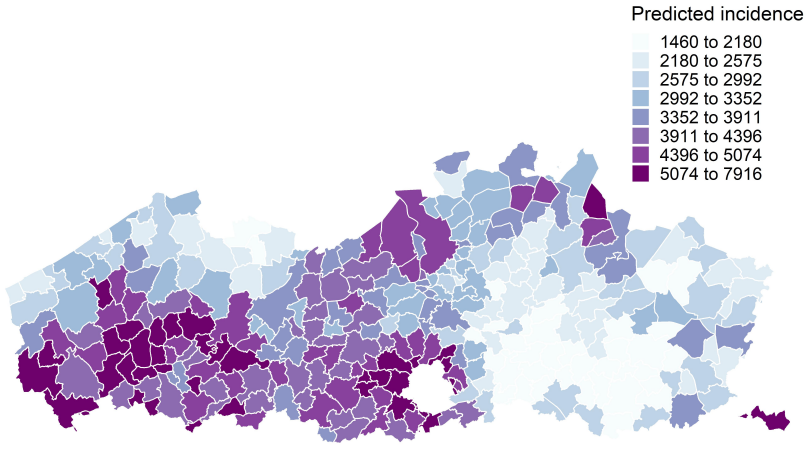
Regarding the spatial autoregressive parameter ρ, here the spatial lag is also significant for all the fitted models, indicating that the spatial autocorrelation is being properly captured. However, we believe it is relevant to mention that, while this finding provides some clear evidence that the autoregressive structure of the model is appropriate for capturing the spatial autocorrelation present in the dataset under study, it does not provide any argument in favor of this being the best or the correct model.

Moreover, the interpretation of parameter ρ can be useful in order to quantify how much the spatial structure considered can influence the resemblance of the incidence rate in a municipality to the geometric mean of the incidence rates of its neighbours. In the models where the distance band matrix was used, the parameter ρ has posterior mean approximately equal to 1, and, thus, in this setting, we find that the rate in a municipality is close to the geometric mean of the rates in the municipalities within the distance band. For the models where the specified weights matrix was either the exponential or the inverse distance, the estimated values of ρ was approximately equal to two, suggesting that the rate in a municipality is the square of the geometric mean of the rates of its neighbours. For the remaining models, this parameter's estimated value was smaller than one. For example, in the model with the mobility matrix, it was $\hat{\rho} = 0.4823$, suggesting that, for this connectivity structure, the rate in a municipality is approximately the squared root of the geometric mean of the rates of its neighbours.

Figure 2 includes the maps of the predicted incidence obtained after fitting the geometric mean spatial conditional normal Poisson models using the spatial weights matrix following the distance band criterion and the similarity spatial matrix combining the dif-



(a) *Predicted incidence obtained from the model using the spatial matrix following the distance band criterion, fitted to the COVID-19 data in Flanders.*



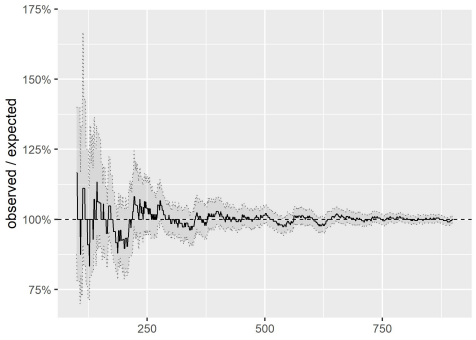
(b) *Predicted incidence obtained from the model using the similarity spatial matrix combining the differences in the variable budgetmeters and the distance bands criterion, fitted to the COVID-19 data in Flanders.*

Figure 2. *Predicted incidence obtained from some of the geometric mean spatial conditional normal Poisson models considered, fitted to the COVID-19 data in Flanders.*

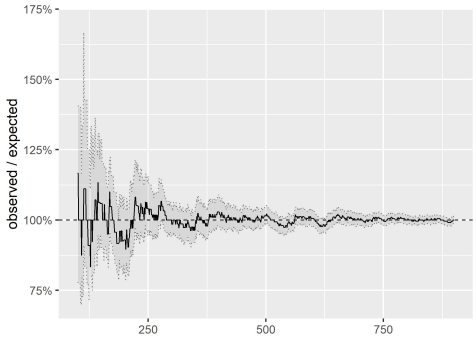
ferences in the variable `budgetmeters` and the distance bands criterion, which were considered as the best fitting ones. Similar maps obtained for some of the other fitted models have been included in Figure S2 in the supplementary material. If we compare these maps with the observed incidence map shown in Figure 1, we can see that, in general, the predictions are quite accurate, as they are very similar to the observed incidence. In addition, when compared to each other, we note that the predictions obtained differ only for a small number of municipalities. In addition, scatterplots of the observed versus the predicted rates, obtained from the fitting of these models are included in Figure S3 in the supplementary material, where it can be seen that the fitted models show high accuracy in the prediction of the incidence rates.

We can also check the distributional assumptions in the fitted models, which is a Poisson distribution, where the overdispersion is accommodated by means of the inclusion of a random effect in the regression for the mean. This can be achieved by using the `distribution_check` function from the R package inlatools (Onkelinx, 2019). Here, simulations are drawn from the model and the empirical cumulative distribution function (eCDF) is computed for the observed response and for the simulated data, so that they can be compared.

Figure 3 includes the plots which illustrate these comparison results for two of the fitted models. In each figure, the black line is the result of dividing the eCDF of the observed data by the median of the eCDF's of the simulated datasets, and the grey bands represent the 95% credible intervals of the simulated data. In addition, the dotted horizontal line placed at 100% indicates where the ratio of the eCDF's is equal to one. If the eCDF is inside the credible intervals, which is the case for all of the models fitted here, the assumed distribution in the model seems to be a plausible one. Moreover, given that the eCDF is quite close to the reference line, these results suggest that the data is well modelled with this distribution.



(a) *Geometric mean model where the spatial matrix follows the distance band criterion.*



(b) *Geometric mean model for the similarity spatial weights matrix combining distance band and the variable* ***budgetmeters****.*

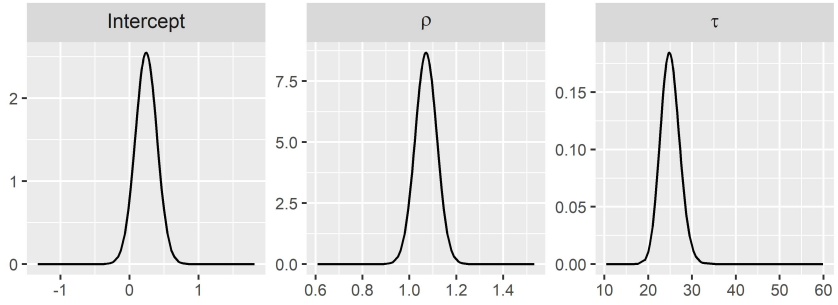
Figure 3. *Distribution check for some of the fitted models.*

Additionally, Figure 4 shows the marginal posterior distribution of the parameters estimated from some of the fitted models, where it can be verified that the normality assumption holds. Distribution checks and the posterior marginals for the estimated parameters in the geometric mean models corresponding to the spatial weights matrix following the contiguity of order one criterion and the mobility matrix have been included in Figures S4 and S5 in the supplementary material.

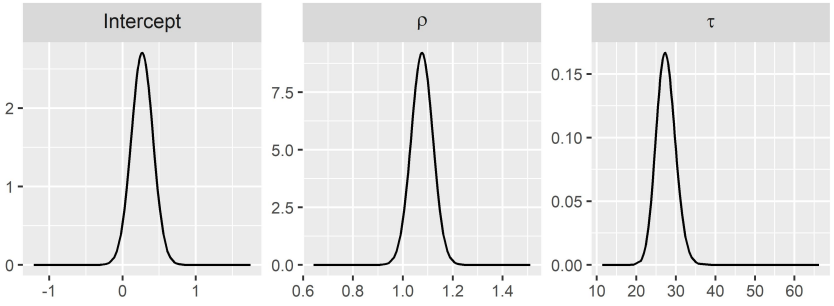
After examining the results obtained in this section, we could conclude that, on the one hand, with the proposed model we present an appealing interpretation of the spatial parameter, given by the geometric mean of the incidence rates. We have shown how this interpretation can change for the different fitted models, indicating how much

the spatial structure considered explains the spatial autocorrelation by means of the geometric mean of the rates in the neighbouring municipalities. On the other hand, by examining different weights matrices, we can have a better idea of the underlying spatial dependence structure of the data. When the similarity matrices based on the distance band were used, the information criteria values were similar to the model considering the traditional distance band matrix. Therefore, taking into account that they provide similar predictions and similar fit, we believe that, for the specific data set considered, this weights matrix could represent a proper choice for modelling the spatial underlying structure of the data.

Finally, we would like to briefly mention that the computation time needed for fitting the models in this section is of approximately one second for each one of the fitted models.



(a) *Geometric mean model where the spatial matrix follows the distance band criterion.*



(b) *Geometric mean model for the similarity spatial weights matrix combining distance band and the variable* ***budgetmeters****.*

Figure 4. *Posterior densities from the parameters estimated for some of the fitted models.*

4.3. Comparison to the BYM2 and Leroux spatial models

In this section, we will fit the BYM2 and Leroux spatial models to the COVID-19 data in Flanders. Additional details about these models have been included in Section 8 in the

supplementary material. We should stress here that one of our main goals in this work is to present the geometric mean proposal as a new extension of the spatial conditional Poisson model in Cepeda-Cuervo et al. (2018). In these models, the interpretation of the spatial parameters is different from that of the BYM2 or the Leroux models. Furthermore, the spatial conditional and the geometric mean models offer the possibility of specifying any weights matrix in a straightforward way, as it is used for computing a spatial lag. In our view, this feature makes these models more appealing for investigating different spatial structures, which is another one of our goals in this work. In the case of the BYM2 or the Leroux models, this is not straightforward due to its limitations, where the assumed spatial structure needs to be symmetric, which is not the case, for example, for the mobility matrix we have employed before. Although it is known that any matrix can be symmetrized, this would include carrying out a previous process, which is not required when fitting our proposed models.

Nevertheless, we believe it can be useful to compare the performance of the proposed methods with that of the BYM2 and Leroux models, often employed in disease mapping applications. Therefore, in order to specify the BYM2, we consider the model in equation (S3) in the supplementary material, where, in order to specify the penalized complexity priors and following Simpson et al. (2017), for the precision parameter τ_s we assume that $\text{Prob}(1/\sqrt{\tau_s} > 0.2/31) = 0.01$ and, for the mixing parameter ϕ_s, $\text{Prob}(\phi_s < 0.5) = 2/3$. Additionally, for the Leroux model, we consider the formulation in equation (S4) in the supplementary material. In this case, the prior for the precision parameter τ_u is a noninformative Gamma distribution (i.e., $\tau_u \sim G(1 \times 10^{-4}, 1 \times 10^{-4})$) and the prior for the spatial parameter ϕ_u is a uniform distribution over the unit interval ($\phi_u \sim U(0,1)$) (Lee, 2013). For the intercept, we assume a noninformative normal prior distribution (i.e., $\beta \sim N(0, 1 \times 10^5)$). Note that, in the BYM2 and Leroux models, the spatial weights matrix is defined based on contiguity of order one. The results obtained after fitting these models to the COVID-19 data in Flanders are included in Table 6 and Table 7.

Table 6. *Results obtained after fitting the BYM2 model to the COVID-19 incidence data in Flanders.*

	Mean	SD	95% CI
$\hat{\beta}$	-3.3924	0.004	(-3.400,-3.385)
$\hat{\tau_s}$	12.057	1.131	(9.885,14.318)
$\hat{\phi_s}$	0.976	0.020	(0.923,0.998)
WAIC = 2932.9 CPO = 1802.6			

Results reported in the previous section indicate that the smallest WAIC resulted for the model using the distance band criterion (WAIC = 2938.6) and the smallest CPO was obtained for the geometric mean model using the similarity matrix of the distance band and `budgetmeter` (CPO = 1806.3). As for the BYM2 model, we can see that the WAIC and CPO values obtained are slightly lower than those obtained for the pre-

Table 7. *Results obtained after fitting the Leroux spatial model to the COVID-19 incidence data in Flanders.*

	Mean	SD	95% CI
$\hat{\beta}$	-3.189	0.258	(-3.590,-2.544)
$\hat{\tau_u}$	6.907	0.603	(5.777,8.150)
$\hat{\phi_u}$	0.989	0.017	(0.942,0.999)
WAIC = 2954.8 CPO = 1958.7			

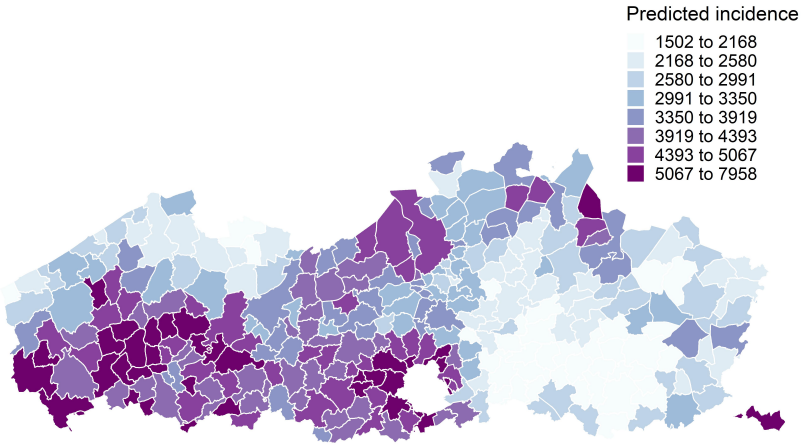
vious models. In contrast, the Leroux model yielded larger information criteria values when compared to both the spatial conditional and BYM2 models, suggesting a weaker goodness of fit to the data under study.

In addition, the value obtained for the mixing parameter in the BYM2 model, $\hat{\phi}_s = 0.976$, suggests that more than 97% of the variability in the data is being explained by the spatially structured effect. Similarly, the estimated spatial parameter in the Leroux model, $\hat{\phi}_u = 0.989$, indicates that most of the variability in the data is explained by the spatial component, which is consistent with the BYM2 estimate of ϕ_s.

Regarding the predictive accuracy of these models, Figure 5 includes the maps of the predicted incidence obtained from their fitting, where we can see that the predictions are very accurate when compared to the map of the observed incidence in Figure 1, and also very similar to the ones obtained in the previous section for our proposed methods (see Figure 2). The scatterplots of the observed versus the predicted incidence rates are also included in Figure S3(i) in the supplementary material, showing some issues in some of the municipalities.

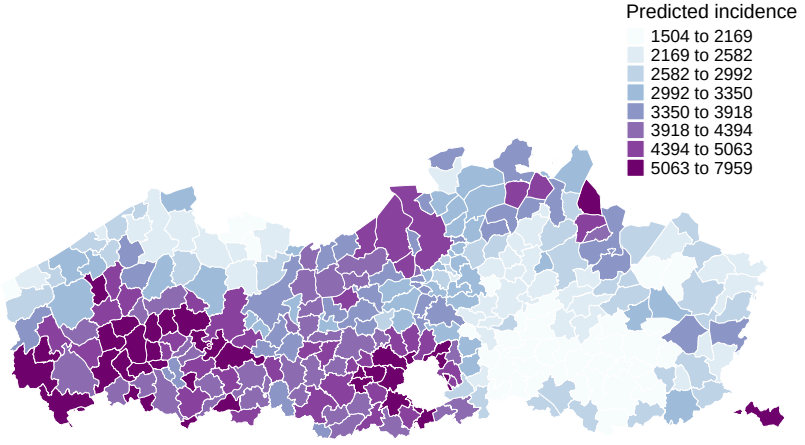
The computation time needed to fit the BYM2 model was of approximately ten seconds, while the Leroux model required five seconds. In contrast, for the spatial conditional and geometric mean models fitted in Section 4.2, the runtime was of about one second for each one of the fitted models. Moreover, in the simulation study carried out in Section 3, we fitted 250 models for each one of the 12 scenarios. When fitting the Leroux and the BYM2 models to these datasets, the runtime increased by a factor of five and ten, respectively, when compared to the geometric mean model. In other words, fitting the Leroux model for the entire simulation study would require more than 4 hours, while the BYM2 model would take about 8 hours, compared to the 50 minutes required for the geometric mean model. Therefore, in our view, this is an important advantage worth mentioning for our proposed models over the BYM2 and the Leroux models, which are commonly used in this area of research.

Despite the fact that the information criteria values favoured the BYM2 model and that its predictive accuracy is similar to the one from the geometric mean model, we restate our goal here of presenting the geometric mean proposal, which can be viewed as an alternative to the Leroux, BYM and BYM2 models, and to investigate the weights matrices which best reflect the spatial underlying process.



(a) *Predicted incidence obtained from the BYM2 model, fitted to the COVID-19 data in Flanders.*

Figure 5. *Predicted incidence obtained from the BYM2 and Leroux models, fitted to the COVID-19 data in Flanders.*



(b) *Predicted incidence obtained from the Leroux model, fitted to the COVID-19 data in Flanders.*

Figure 5. *Predicted incidence obtained from the BYM2 and Leroux models, fitted to the COVID-19 data in Flanders (Continued).*

There are situations where the spatial conditional models may offer a better fit than the Leroux, BYM and BYM2 models, or viceversa. We believe that the choice of the model to fit should depend on the specific objective of the study. For example, Morales-Otero and Núñez-Antón (2021) reported that, given by the information criteria values obtained, the spatial conditional and the BYM and BYM2 models offered a very similar fitting to the infant mortality data they studied. In addition, in Morales-Otero, Gómez-

Rubio and Núñez-Antón (2022), the spatial conditional models were employed in order to illustrate a new fitting approach in INLA.

5. Discussion

In this work, we have studied the geographical spread of COVID-19 cases in the municipalities of Flanders in Belgium during the period going from September 2020 to January 2021. In order to be able to fit these data, we have considered the Bayesian spatial conditional model proposals (Cepeda-Cuervo et al., 2018), which assume the incidence of cases in a municipality is conditional on the incidence of cases in neighbouring municipalities. These models offer a great flexibility and also the possibility that considering different weights matrices can be done in a direct and very simple way.

We have proposed a geometric mean spatial conditional model, where the logarithm of the rates is employed for computing the spatial lag component. This model offers an interpretation of the spatial parameter ρ based on the geometric mean, representing how the incidence rate in one municipality resembles the geometric mean of the rates in its neighbours. For the spatial weights matrix used in these models, we have proposed alternative specifications based on a combination of the similarity of a certain variable in the different locations and the distance between these municipalities. In addition, we have also considered the connectivity structure given by the mobility of individuals among the municipalities under study.

In order to further assess the performance of the proposed methods when the correlation among the different municipalities under study is given by a connectivity pattern, such as, for example, the mobility matrix, we have carried out a simulation study where we induce correlation in the response variable based on this structure. In this study, we have been able to appropriately verify that the models are able to identify the correct spatial structure for most of the cases under study.

In the application to the COVID-19 data in Flanders, we have compared these proposed models with the ones in Cepeda-Cuervo et al. (2018) finding that our proposal provides a similar fit, but offers a particular and straightforward interpretation within the context of the specific dataset under analysis. We have fitted these models by using different definitions for the weights matrices employed to compute the spatial lag, such as the classical ways of accounting for spatial autocorrelation based on contiguity and distance, as well as the similarities weights matrices we proposed as alternatives. In addition, we have also studied the use of the mobility matrix in modelling the COVID-19 incidence data in Flanders, which is given by the proportion of time individuals from one municipality spent in a different one.

In order to provide a comparison of the proposed models with other commonly used models employed in disease mapping applications, we have also fitted the BYM2 and Leroux spatial models to the dataset under study. Results indicate that the BYM2 model provides a similar fit based on information criteria and demonstrates a comparable predictive accuracy to that of our proposed model. However, it may be the case that the

dataset under study may not be the best example to fully justify the need for the geometric mean spatial conditional model. Moreover, we believe it is important to clarify that this study initially began as an investigation into whether the mobility connectivity structure could explain the spatial pattern of COVID-19 incidence across municipalities in Flanders. Addressing this question required flexible spatial models, such as the spatial conditional models proposed by Morales-Otero and Núñez-Antón (2021), which motivated the use of this approach in our analysis. Subsequently, we developed the geometric mean spatial conditional model, adjusting the primary focus of this work to introducing it as a flexible alternative for capturing spatial dependencies and making it possible to specify different spatial structures.

The BYM2 model is well established in this area of research but, at the same time, we also believe that our model provides interpretational advantages, computational simplicity, and the flexibility to easily test for different spatial structures. For example, the computation time required to estimate a geometric mean model is ten and five times shorter, respectively, than the one needed for the BYM2 or Leroux models. We consider this to be a significant advantage of our proposed model, particularly when researchers need to perform simulation studies, such as the one presented in Section 3, where a large number of models must be efficiently fitted. Nonetheless, we recognize that further applications are necessary to fully evaluate the benefits and limitations of the geometric mean model in different contexts, and we intend to continue exploring this model proposal structure in our future research.

In any case, overall results suggest a strong spatial correlation in the dataset under study, which is best explained by the distance band spatial weights matrix. This implies that, for the data under study, the underlying spatial process is well explained and modelled by this spatial structure.

Finally, we believe it it worth mentioning that, in this work, we focus on the analysis of the data corresponding to the time period of the COVID-19 second wave in Flanders, and we have tried to characterize the overall spatial pattern believed to be present in this wave. Our main interest for this specific application does not include transmission. However, for future research we are also interested in performing comparison with the spatial pattern of additional COVID-19 waves in the area under study, by being able to propose a spatio-temporal approach, which is out of the scope of this paper. We have already developed spatio-temporal extensions of the spatial conditional models and specific proposals are in the process of being finalized, so that they are part of a different manuscript to be later submitted for possible publication. Moreover, one of our objectives is to be able to apply these proposals to the comparisons of the different waves in the dataset we have analysed here.

Acknowledgements

This research has been partially funded by Ministerio de Ciencia e Innovación (MCIN, Spain), Agencia Estatal de Investigación (AEI/10.13039/501100011033/) and Fondo Eu-

ropeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa" under the I+D+i research grant PID2020-112951GB-I00 and by the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group) under research grant IT-1508-22.

Supplementary material

Supplementary material has been included in an accompanying document. It includes heatmaps of the weights matrices considered here, scatterplots of the observed versus the predicted rates, additional details about the BYM2 and the Leroux models, among other information that can be useful to the readers.

Code availability

There is an R script available with an example of how to fit the proposed geometric mean spatial conditional model, using a simulated dataset. The link is: https://github.com/mabelmo/Spatial-Autoregressive-Geometric-Mean-Model-.git.

References

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* 27(3), 247–267.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society - Series B* 36, 192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1–20.

Carroll, R., Lawson, A.B., Faes, C., Kirby, R.S., Aregay, M., and Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and Spatio-temporal Epidemiology* 11-15, 45–54.

Case, A., Hines, J.R. Jr., and Rosen, H.S. (1993). Budget spillovers and fiscal policy interdependence: Evidence from the States. *Journal of Public Economics* 52(3), 285–307.

Cepeda-Cuervo, E., Córdoba, M., and Núñez-Antón, V. (2018). Conditional overdispersed models: Application to count area data. *Statistical Methods in Medical Research* 27, 2964–2988.

D'Angelo, N., Abbruzzo, A., and Adelfio, G. (2021). Spatio-temporal spread pattern of COVID-19 in Italy. *Mathematics* 9(19), 2454.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics* 6(54).

Ejigu, B.A. and Wencheko, E. (2020). Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. *Spatial Statistics* 38, 100454.

Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H., and Kauermann, G. (2022). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling* 0(0), 1471082X221124628.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 5(3), 187–199.

Jackson, M.C. and Sellers, K.F. (2008). Simulating discrete spatially correlated Poisson data on a lattice. *International Journal of Pure and Applied Mathematics* 46(1), 137–154.

Johnson, D.P., Ravi, N., and Braneon, C.V. (2021). Spatiotemporal Associations Between Social Vulnerability, Environmental Measurements, and COVID-19 in the Conterminous United States. *GeoHealth* 5(8), e2021GH000423.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics* 52(2), 169–183.

Konstantinoudis, G., Padellini, T., Bennett, J., Davies, B., Ezzati, M., and Blangiardo, M. (2021). Long-term exposure to air pollution and COVID-19 mortality in England: A hierarchical spatial analysis. *Environment International* 146, 106316.

Konstantinoudis, G., Cameletti, M., Gómez-Rubio, V., León Gómez, I., Pirani, M., Baio, G., Larrauri, A., Riou, J., Egger, M., Vineis, P., and Blangiardo, M. (2022). Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. *Nature Communications* 13(482).

Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software* 55(13), 1–24.

Leroux, B.G., Lei, X., and Breslow, N. (2000). Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. *In Statistical Models in Epidemiology, the Environment, and Clinical Trials* (M.E. Halloran and D. Berry, eds.). Springer: New York, USA, 179–191.

Morales-Otero, M., and Núñez-Antón, V. (2021). Comparing Bayesian spatial conditional overdispersion and the Besag-York-Mollié models: Application to infant mortality rates. *Mathematics (Special issue on Spatial Statistics with its Applications)* 9(3), 282.

Morales-Otero, M., Gómez-Rubio, V., and Núñez-Antón, V. (2022). Fitting double hierarchical models with the integrated nested Laplace approximation. *Statistics and Computing* 32(62)

Natalia, Y.A., Faes, C., Neyens, T., and Molenberghs, G. (2022). The COVID-19 wave in Belgium during the Fall of 2020 and its association with higher education. *PLOS ONE* 17(2), e0264516.

Onkelinx, T. (2019). The inlatools package, https://inlatools.netlify.app/. Last accessed 09 August 2022.

Pettit, L.I. (1990). The Conditional Predictive Ordinate for the Normal Distribution. *Journal of the Royal Statistical Society - Series B* 52(1), 175–184.

Riebler, A., Sørbye, S.H., Simpson, D.P., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research* 25(4), 1145–1165.

Sahu, S.K. and Böhning, D. (2022). Bayesian spatio-temporal joint disease mapping of Covid-19 cases and deaths in local authorities of England. *Spatial Statistics* 49, 100519.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., and Sørbye, S.H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* 32(1), 1–28.

Slater, J.J., Brown, P.E., Rosenthal, J.S., and Mateu, J. (2022). Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. *Spatial Statistics* 49, 100540.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11(116), 3571–3594.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* 41(3-4), 434–449.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* 44(4), 1019–1031.

SORT 49 (1) January-June 2025, 121-146 DOI: 10.57645/20.8080.02.25

Leave-group-out cross-validation for latent gaussian models

Zhedong Liu[1], Janet Van Niekerk[2] and Håvard Rue[3]

Abstract

Evaluating the predictive performance of a statistical model is commonly done using cross-validation. Among the various methods, leave-one-out cross-validation (LOOCV) is frequently used. Originally designed for exchangeable observations, LOOCV has since been extended to other cases such as hierarchical models. However, it focuses primarily on short-range prediction and may not fully capture long-range prediction scenarios. For structured hierarchical models, particularly those involving multiple random effects, the concepts of short- and long-range predictions become less clear, which can complicate the interpretation of LOOCV results. In this paper, we propose a complementary cross-validation framework specifically tailored for longer-range prediction in latent Gaussian models, including those with structured random effects. Our approach differs from LOOCV by excluding a carefully constructed set from the training set, which better emulates longer-range prediction conditions. Furthermore, we achieve computational efficiency by adjusting the full joint posterior for this modified cross-validation, thus eliminating the need for model refitting. This method is implemented in the R-INLA package (www.r-inla.org) and can be adapted to a variety of inferential frameworks.

MSC: *62-04 62C10 62F15 62J12.*

Keywords: *Bayesian Cross-Validation, Latent Gaussian Models, R-INLA.*

[1] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. zhedongliu1@gmail.com

[2] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. Department of Statistics, University of Pretoria, South Africa. janet.vanniekerk@kaust.edu.sa

[3] Statistics Program, CEMSE. King Abdullah University of Science and Technology. Kingdom of Saudi Arabia, Thuwal 23955-6900. haavard.rue@kaust.edu.sa

Received: August 2024.
Accepted: March 2025.

1. Introduction

1.1. Rationale and Background

Leave-one-out cross-validation (LOOCV) (Stone, 1974) stands as a popular method to evaluate a statistical model's predictive performance, perform model selections, or estimating some critical parameters in the model. The core concept of LOOCV is elegantly straightforward. Suppose we have data, $\boldsymbol{y} = \{y_i\}$, for $i = 1, \ldots, n$, presumed to be independent and identically distributed (I.I.D.) samples from the true distribution $\pi_T(y)$. Our objective is to determine how well a fitted model can predict a new observation, $\tilde{y}$, sampled from this true distribution. In the Bayesian context, we use the posterior predictive distribution $\pi(y|\boldsymbol{y})$ to predict $\tilde{y}$ sampled from $\pi_T(y)$ as proposed by Geisser and Eddy (1979). Using the logarithmic score (Gneiting and Raftery, 2007), we can compute $E_{\tilde{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ as a metric for prediction ability.

Owing to the lack of $\pi_T(y)$, directly computing the expectation becomes infeasible. Nonetheless, since y_i is an exchangeable sample from $\pi_T(y)$, we can estimate this expectation by evaluating

$$u_{\text{LOOCV}} = \frac{1}{n} \sum_{i=1}^{n} \log \pi(y_i|\boldsymbol{y}_{-i}),$$

where y_i is the testing point and $\boldsymbol{y}_{-i}$ is the training set, and $\boldsymbol{y}_{-i}$ are all data except the ith observation.

The informal interpretation of LOOCV is that it mimics "using $\boldsymbol{y}$ to predict $\tilde{y}$" by "using $\boldsymbol{y}_{-i}$ to predict y_i". This intuitive interpretation is then used to justify, often implicitly, the use of LOOCV as a "default" way to evaluate predictive performance.

However, issues can arise in more complex statistical models where the dependency in the model results in the data not being exchangeable (see Vehtari and Ojanen (2012) for a complete discussion of cross-validation (CV) for several types of exchangeability); we describe these kinds of models as "dependent" cases for the purpose of this paper. An intuitive dependent case is a time series. Burman, Chow and Nolan (1994) proposed a block CV method for dependent data from a stationary process, acknowledging the need for a different approach to CV than LOOCV. McQuarrie and Tsai (1998) propose modified cross-validation (MCV) where dependent data chunks are removed together with the relevant point to account for the dependence in a time series (and other dependent data generating models). Bergmeir and Benítez (2012) investigated the properties of blocked CV and other approaches for robust time series model evaluations (see also Bergmeir, Hyndman and Koo (2018) for a study on k-fold CV), while Bürkner, Gabry and Vehtari (2020) proposed a leave-future-out CV strategy. Cerqueira, Torgo and Mozetič (2020) investigated CV and holdout approaches for time series models and concluded that the out-of-sample holdout procedure is more accurate for non-stationary processes than LOOCV.

Besides time series, spatial dependence models come to mind for which Valavi et al. (2018) proposed a buffering strategy by leaving out specific spatial points or areas and

spatial and environmental blocking. Spatial blocking forms clusters of data points according to spatial effects, and environmental blocking forms clusters using K-means (Hartigan and Wong, 1979) on the covariates. Other examples of dependent cases are longitudinal data for multiple subjects in a study (Saeb et al., 2017) and hierarchical models (see Gelman et al. (1995) and Vehtari and Ojanen (2012, Section 5.1.4). Racine (2000) proposed an hv-block CV approach for dependent data while Merkle, Furr and Rabe-Hesketh (2019) considers a multilevel model and shows that marginal WAIC is akin to LOOCV. Roberts et al. (2017) advocate a block cross-validation, partitioning ecological data based on inherent patterns, when the prediction task is not simply short-range prediction. Rabinowicz and Rosset (2022) offers a modification to LOOCV, ensuring an unbiased measure of predictive performance given the correlation between new and observed data, where the unbiasedness is in the sense of randomized both observed and new data. We should note that an assumed prediction task determines the correlation between new and observed data.

In dependent cases, LOOCV can provide a restricted assessment of the models' predictive performance since LOOCV cannot evaluate longer-range prediction. Even in terms of short-range prediction, it is not clear what is short- or longer-range in dependent models that are not purely temporal or spatial models where the range has a physical interpretation. We use the concepts of short-range and longer-range predictions, acknowledging that these concepts can have overlapping meanings.

We thus propose a framework that emulates longer-range prediction scenarios, for hierarchical models, by constructing non-random leave-out sets based on model-based correlations. This can be viewed as a complementary approach to LOOCV for evaluating predictive performance, providing additional informative insights of the predictive ability for dependent cases.

1.2. The prediction task

The critical observation is that the meaning of "prediction" is not clearly defined when we are far away from exchangeability, so that $\boldsymbol{y}$ are *non-exchangeable* samples of $\pi_T(y)$. $\pi_T(\tilde{y}|\boldsymbol{y})$ lacks a unique definition in dependent cases as without a clear *prediction task*, i.e., how we imagine a new data point, $\tilde{y}$, is generated given observed data $\boldsymbol{y}$. This ambiguity extends to the act of "using $\boldsymbol{y}$ to predict $\tilde{y}$" as it is uncertain what our target, $\tilde{y}$, represents. To illustrate these concepts, let us discuss some more concrete examples.

Time-series model

Assume data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}$ is a time-series, observed sequentially at time $1, 2, \ldots, T$. The inherent prediction task is to predict future values, given the temporal nature of the data. We can predict a new observation at $k \geq 1$ steps into the future by $\pi(y_{T+k}|y_1, \ldots, y_T)$.

In this example, the LOOCV will be computed from

$$\pi(y_t|y_1, \ldots, y_{t-1}, y_{t+1}, \ldots, y_T), \qquad t = 1, \ldots, T,$$

which is often referred to as interpolation or imputation of missing values, rather than a prediction. However, the predictive performance of time series models is often assessed through leave-future-out cross-validation (LFOCV) (Bürkner et al., 2020):

$$\sum_{T'=T_0}^{T-k} \log \pi(y_{T'+k}|y_1, \ldots, y_{T'}),$$

where T' starts from time $T_0 > 1$ as we need some data to estimate the model.

The message from this example is that LOOCV, when applied to such models, is essentially evaluating short-range prediction performance rather than longer-range predictive performance.

We acknowledge two issues. First, the distinction between short and longer-range prediction is not always clear-cut, leading to overlapping concepts. For example, a one-step-ahead forecast leans more towards short range than a two-step-ahead prediction. In contrast, a one-step-ahead forecast leans less towards short-range than a missing value imputation. However, this does not deter our discussion. Secondly, while an ideal model succeeds in all prediction tasks, real-world scenarios require us to settle for the definition of the "best fit". Consequently, our choice of evaluation should align with our specific objectives.

Multilevel model

Figure 1 illustrates an example of a multilevel model. Consider observations of student grades or performance. This data exhibits a hierarchical structure: students belong to classes, classes reside within schools, and schools are nested within regions. This hierarchical arrangement is significant because it introduces correlated random effects attributed to the class, school, and region levels, substantially deviating from the exchangeable case.

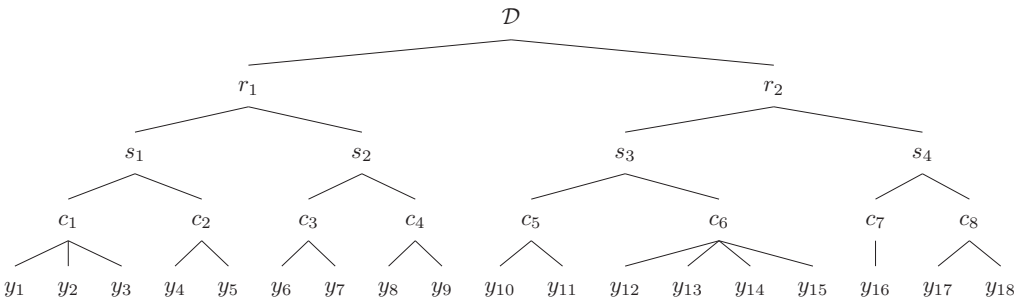


Figure 1. *A nested multilevel model.*

Given such a model, the prediction task becomes ambiguous. Are we aiming to predict the performance of an unobserved student from an observed class? Or are we trying to predict the performance of an unobserved student in an unobserved class, school, or

even region? This difficulty mirrors the challenges in defining asymptotic regimes for these models. As students, classes, schools, and regions can grow indefinitely in various ways, it is unclear whether one of such choices is the most reasonable.

To evaluate predictive performance within this context, users must first explicitly define their prediction task and then evaluate the model according to this definition. It should be noted that applying LOOCV would evaluate the prediction of individual students within observed classes. In our view, this mimics more short-range prediction rather than longer-range prediction, and another framework is needed to quantify the predictive ability for a new student in a new class in a new school in a new region, for example. Our proposal provides some insight into this kind of prediction task.

1.3. LGOCV: Complementing LOOCV for dependent cases

Our discussions illuminate an important insight: when dealing with models that lead to non-exchangeable data, the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions than predictive performance for longer range as is usually implied by "out-of-sample" prediction. This prompts the question: What is a suitable approach moving forward?

One observation is the absence of a "one size fits all" solution. Each model may possess a natural prediction task-or several-based on its intended application. Thus, for a specific assessment of predictive performance, we need to define these prediction tasks explicitly. One can then evaluate distinct predictive performance metrics using our proposed leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n}\sum_{i=1}^{n} \log(\pi(y_i|\boldsymbol{y}_{-I_i})). \tag{1}$$

Here, the *group* (denoted by I_i) is an index set including i. This configuration facilitates that the pair $(y_i, \boldsymbol{y}_{-I_i})$ mimics a specified prediction task, with $\boldsymbol{y}_{-I_i}$ being the data subset excluding the data indexed by I_i. In a multilevel model, as depicted in Figure 1, predicting a student's grade from an unseen class necessitates that I_i includes i and all observations from student i's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. Therefore, even in complex cases, LOOCV is often applied for its simplicity-even if it leans more towards imputation or short-range prediction.

Developing a framework that evaluates a model's longer-range prediction like the proposed LGOCV, necessitates the construction of the leave-out group I_i for each datapoint y_i. Our approach constructs a model-based group, I_i, for each i by using the prior or posterior correlation among the set of linear predictors. Though we will delve into the construction of I_i in Section 3, an initial understanding is that I_i comprises the data points that correspond to the linear predictors that are most informative for predicting the testing linear predictor, and thus the testing point, y_i. This set ensures that

our LGOCV focuses less on short-range prediction (interpolation) and more on longer-range prediction than LOOCV. In other words, LGOCV tests the model on more difficult prediction tasks since the most influential points are removed together with the testing point, instead of some arbitrary (possibly uncorrelated) point(s). The user needs to only provide a number that indicates the "degree of the independence" between the prediction point and the rest of the data", and we compute these groups for each datapoint in an automated way. In various practical examples, we will show how this model-based procedure produces reasonable groups. Advanced spatial examples applying the proposed method are presented by Adin et al. (2024). For a simple time-series example, our new approach will correspond to evaluating $\pi(y_t|y_1,\ldots,y_{t-k},y_{t+k},\ldots,y_T)$, for fixed $k>1$. This corresponds to removing a sequence of data with length $2k-1$, to predict the central one. As we see, this task mimics a longer-range prediction task. Our interpretation is that LGOCV quantifies the model's ability to predict longer-range more appropriately than LOOCV, when $k>1$, and is similar to the cross-validation procedure proposed by Burman et al. (1994) for stationary processes.

There are two key challenges to address to make our proposal practical. Firstly, we must quantify the information contributed by one data point in predicting another; this is crucial for the group construction. Secondly, we face the computational task of evaluating u_{LGOCV} given a set of groups. The naive computation of LGOCV by fitting models across all potential training sets and evaluating their utility against corresponding testing points is computationally infeasible, especially given the resource-demanding nature of modern statistical models. However, these challenges can be handled elegantly within the framework of latent Gaussian models (LGMs) combined with the integrated nested Laplace approximation (INLA) inference, as detailed in Rue et al. (2009, 2017); Van Niekerk and Rue (2024); Van Niekerk et al. (2023). Throughout this paper, we will assume that our model is an LGM. We will discuss how to integrate the automatic group construction and the fast computation of u_{LGOCV} using the INLA framework. Notably, our proposed methodology has been incorporated into the R-INLA package (www.rinla.org), extending its applicability across all LGMs supported by R-INLA.

1.4. Theoretical aspects

Cross-validation (CV), particularly LOOCV, is frequently considered as an estimator of $E_{\tilde{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ or $E_{\tilde{y},\boldsymbol{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$. The first expectation describes the generalized predictive performance given a specific training set, while the second expectation describes the generalized predictive performance averaged over different identically distributed training sets. These expectations can be evaluated when assuming the existence of the joint density $\pi_T(\tilde{y},\boldsymbol{y})$, representing the true data generation process. Under the assumption of exchangeability and some regularity conditions on the model, the Bernstein-Von-Mises theorem states that $\log\pi(\tilde{y}|\boldsymbol{y})$ converges to a random variable irrelevant to $\boldsymbol{y}$. Consequently, $E_{\tilde{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ and $E_{\tilde{y},\boldsymbol{y}}[\log\pi(\tilde{y}|\boldsymbol{y})]$ become equivalent in the limit. If we further assume that $\tilde{y}$ is sampled from the same distribution as all the training data, LOOCV is an asymptotically unbiased estimator of the expectations. Commonly used informa-

tion criteria, such as AIC (Akaike, 1973), WAIC (Watanabe, 2010), are asymptotically equivalent to LOOCV in fully exchangeable cases. This type of analysis is prevalent in the literature with various settings (Stone, 1974, 1977; Yang, 2007; Shao, 1993).

However, a similar analysis does not hold for dependent cases in general. Firstly, the existence of different prediction tasks means that both the model prediction, $\pi(\tilde{y}|\boldsymbol{y})$, and the true data generation process, $\pi_T(\tilde{y}|\boldsymbol{y})$, are not uniquely defined as discussed in Section 1.2. Secondly, the asymptotic scheme is not uniquely defined, even with a specific prediction task. For example, in a temporal model where data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ is a time-series, observed at time $t_1 < t_2 < \cdots < t_n$ and we denote the last time step as T. Several meanings of of $n \to \infty$ can be considered:

- $T \to \infty$ and $t_i - t_{i-1}$ is a constant
- $t_i - t_{i-1} \to 0$ and T is a constant
- $t_i - t_{i-1} \to 0$ and $T \to \infty$ with $T(t_i - t_{i-1})$ fixed

These scenarios correspond to observing more future data and having higher sample rates within a time frame. As mentioned in Section 1.2, multilevel data can also have various asymptotic regimes. Thirdly, if the data generation process is not stationary, the model will not converge under certain asymptotic regimes, which differentiate $E_{\tilde{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ from $E_{\tilde{y},\boldsymbol{y}}[\log \pi(\tilde{y}|\boldsymbol{y})]$ even in asymptotic scenarios. These points highlight that the estimand of CV is not uniquely defined in dependent cases, preventing the establishment of an asymptotic analysis framework.

From the perspective of CV, it is also inappropriate to consider it an estimator since each summand in CV should be viewed as a sample from different distributions due to the relevance of data indexes in dependent cases. For example, if we compute LOOCV in a time series. Each y_t is sampled from a different conditional distribution $\pi_T(y_t|\boldsymbol{y}_{-t})$ and thus the average $\frac{1}{n}\sum_{t=1}^{T} \log \pi(y_t|\boldsymbol{y}_{-t})$ cannot be considered as an estimator in general. Therefore, it is more reasonable to view CV as a predictive measurement rather than an estimator of an expectation. This perspective allows us to interpret the proposed LGOCV as the averaged predictive performance for similar prediction tasks, created systematically by the model.

While the proposal of Merkle et al. (2019) for multilevel model demonstrates that marginal WAIC is akin to LOOCV, we note that conditional WAIC aligns with LGOCV, where a hierarchical level, such as a school, defines the groups. The h-block CV of Burman et al. (1994) is a special case of LGOCV for a stationery model. LFOCV proposed by Bürkner et al. (2020) is similar to LGOCV as shown in Section 5. The spatial buffering proposed by (Valavi et al., 2018) ensures that no test data is spatially next to any training data, and is a special case of LGOCV for model with only spatial effects. LGOCV this provides a framework where no training data is placed next to the test data in terms of the entire model, and not just specific components thereof.

1.5. Plan of paper

We propose the model-based LGOCV to evaluate longer-range prediction performance for latent Gaussian models, as a special case of a hierarchical model. Complementing this, we introduce a computational method to approximate u_{LGOCV} without model refitting, which is crucial for practical implementation of our proposal. Our computational technique also facilitates the calculation of u_{LGOCV} with user-specified groups.

Section 2 introduces LGMs and explains how they can be efficiently inferred using INLA. In Section 3, we discuss the model-based group construction method for LGMs. This method can be implemented in two ways: by using the prior correlation matrix or the posterior correlation matrix of the latent linear predictors. In Section 4, we demonstrate how to approximate the LGOCV predictive density. Finally, in Section 5, we compare the approximated LGOCV with the exact LGOCV computed by Markov chain Monte Carlo (MCMC) and present some applications. We conclude with a general discussion in Section 6.

2. Latent Gaussian models

This section briefly introduces LGMs, as detailed in Rue et al. (2009, 2017); Van Niekerk and Rue (2024); Van Niekerk et al. (2023), since the model-based group construction and fast approximations rely on them. The LGMs can be formulated by

$$\begin{aligned} &y_i|\eta_i,\boldsymbol{\theta} \sim \pi(y_i|\eta_i,\boldsymbol{\theta}),\\ &\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{f}, \quad \boldsymbol{f}|\boldsymbol{\theta} \sim N(0,\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

In LGMs, each y_i is independent conditioned on its corresponding linear predictor η_i, and hyperparameters $\boldsymbol{\theta}$; $\boldsymbol{\eta}$ is a linear combination of $\boldsymbol{f}$, which is assigned with a Gaussian prior with zero mean and a precision matrix parameterized by $\boldsymbol{\theta}$; $\boldsymbol{A}$ is the design matrix mapping $\boldsymbol{f}$ to $\boldsymbol{\eta}$; $\pi(\boldsymbol{\theta})$ is a prior density of hyperparameters. It is worth mentioning that the prior precision matrix $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})$ is very sparse, which is leveraged to speed up the inference.

The model is quite general because $\boldsymbol{f}$ can combine many modeling components, including linear model, spatial components, temporal components, spline components, etc (Wang, Yue and Faraway, 2018; Krainski et al., 2018; Gómez-Rubio, 2020). It is also common with linear constraints on the latent effects $\boldsymbol{f}$ (Rue and Held, 2005).

We can approximate $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ at some configurations, $\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_k$. The configurations are located around the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, denoted by $\boldsymbol{\theta}^*$, for numerical integration. Approximations of $\pi(\boldsymbol{\eta}|\boldsymbol{\theta},\boldsymbol{y})$ are computed using the linear relation, $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{f}$. The Gaussian approximation of $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ plays an essential role, which is outlined as follows.

We have $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ for a given $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{f}^{\mathsf{T}}\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})\boldsymbol{f} + \sum_{i=1}^{n} \log(\pi(y_i|\eta_i,\boldsymbol{\theta}))\right\}, \tag{3}$$

whose mode is $\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})$. The Gaussian approximation of $\pi(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y})$ is

$$\pi_G(\boldsymbol{f}|\boldsymbol{\theta},\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{f}^{\mathsf{T}}(\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})+\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A})\boldsymbol{f}+\boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{f}\right\}. \tag{4}$$

In (4), $b_i(\boldsymbol{\theta},\boldsymbol{y}) = g_i'(\eta_i^*) - g_i''(\eta_i^*)\eta_i^*$, and $\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})$ is a diagonal matrix with $C_{ii}(\boldsymbol{\theta},\boldsymbol{y}) = -g_i''(\eta_i^*)$, where $g_i(\eta_i) = \log(\pi(y_i|\eta_i,\boldsymbol{\theta}))$ and $\eta_i^* = \boldsymbol{A}_i\boldsymbol{\mu}_f(\boldsymbol{\theta},\boldsymbol{y})$ with $\boldsymbol{A}_i$ being ith row of $\boldsymbol{A}$. The Gaussian approximation is denoted by,

$$\boldsymbol{f}|\boldsymbol{y},\boldsymbol{\theta} \approx N(\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}),\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})), \tag{5}$$

where $\boldsymbol{\mu}_f(\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y})^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})$ and $\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})+\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A}$ are the approximated posterior mean and precision matrix of $\boldsymbol{f}$ given $\boldsymbol{\theta}$.

3. Model-based Group Construction

The primary feature of our proposed group construction is that it requires a choice of correlation matrix (prior or posterior) for the linear predictors, and a single mandatory parameter to adjust the difficulty of the prediction task. This parameter is termed "the number of level sets". It can be interpreted as the strength of the non-dependence between the group to leave out and the rest of the data. A higher value would thus ensure that the leave-out group is more independent from the rest of the data, than a lower value. A higher indpeendnece between the leave-out data and the rest of the data simluates a more difficult prediction task for the model. Based on this value and the correlation matrix choice, all other processes are automated. In a multivariate Gaussian distribution, we can quantify the information provided by a data point to predict another data point by the variance reduction of the conditional distribution, and the variance reduction is a function of their correlation coefficient. To elaborate, if X and Y are both Gaussian random variables, the variance reduction resulting from knowing X when predicting Y equates to $\sigma_Y^2\rho^2$, where σ_Y^2 is the marginal variance of Y and ρ is the correlation between X and Y.

In LGMs, the linear predictors, $\boldsymbol{\eta}$, represent the underlining data generation process of data in (2). The linear predictors are designed to have a Gaussian prior and approximated to be Gaussian in posterior given $\boldsymbol{\theta}$ therefore, we can use the absolute value of the correlation matrix of $\boldsymbol{\eta}$ to represent the information provided by one data point to predict another data point. We evaluate those correlation matrices at the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, denoted by $\boldsymbol{\theta}^*$. Then, we have correlation matrices of $\boldsymbol{\eta}$ derived from the prior precision matrix, $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta}^*)$, and the posterior precision matrix, $\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta}^*,\boldsymbol{y})$. We call the former one prior correlation matrix, denoted by $\boldsymbol{R}_{\text{prior}}$, and the latter one posterior correlation matrix, denoted by $\boldsymbol{R}_{\text{post}}$. Note that the correlation matrices are not fully evaluated and stored to avoid computational burden as they are dense and large; thus, care has to be applied to the implementation to make it feasible. The group would vary with $\boldsymbol{\theta}$. We use $\boldsymbol{\theta}^*$ because it has the highest weight in the posterior. This preference arises because we

frequently employ non-informative priors for hyperparameters. This approach ensures that our focus remains on evidence from the data rather than on arbitrary assumptions about hyperparameters.

Manually constructed groups are often based on prior knowledge and some structured effects, represented by $\boldsymbol{f}$. To imitate this process, we can compute the correlation matrix from a submatrix of $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})$. The correlation matrix, $\boldsymbol{R}_{\text{prior}}$, derived from the submatrix of the prior precision matrix, is a correlation matrix conditioning on those unselected effects. The groups constructed using $\boldsymbol{R}_{\text{prior}}$ are viewed to be solely user-defined in the way that it only depends on the priors and not on the data. In some situations this could be motivated, but in general we recommend using $\boldsymbol{R}_{\text{post}}$ to construct groups because the data will be informative to determine the importance of each effect.

When using a correlation matrix $\boldsymbol{R}$, it is natural to select a fixed number of η_j most correlated to η_i and include their index in the group I_i. However, this approach can be problematic as some linear predictors may have identical absolute correlations to η_i, e.g., in a model with only intercept, all the linear predictors are correlated to each other with correlation 1. Instead, we include all indices of η_j's with identical absolute correlations to η_i in I_i if one of them is included. We define a level set as all η_j's with the same absolute correlation to η_i and determine the group size based on the number of level sets, denoted as m. Setting a higher m results in a less dependent training set and testing point. We recommend using a small number of level sets, such as $m = 3$, as a high value of m can result in a large leave-out group size.

The automated group construction process thus involves selecting the number of level sets, m, and the correlation matrix to use, $\boldsymbol{R}_{\text{prior}}$ or $\boldsymbol{R}_{\text{post}}$. For each i, we can associate m level sets with the m largest absolute correlations to η_i, and the union of those level sets forms I_i. As an illustration, we outline the automated group construction procedure in Algorithm 1.

4. Approximation of LGOCV predictive density

In this section, we will explore the process of approximating $\pi(y_i|\boldsymbol{y}_{-I_i})$. The results are straightforward but tedious in implementation; thus, it is crucial to exercise caution to ensure that all potential numerical instabilities are accounted for. Through empirical testing, this new method has shown to be both more accurate and stable compared to the approach outlined in Rue et al. (2009), when $I_i = i$.

We start by writing $\pi(y_i|\boldsymbol{y}_{-I_i})$ as nested integrals,

$$\pi(y_i|\boldsymbol{y}_{-I_i}) = \int_{\boldsymbol{\theta}} \pi(y_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})d\boldsymbol{\theta} \tag{6}$$

$$\pi(y_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}) = \int \pi(y_i|\eta_i, \boldsymbol{\theta})\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})d\eta_i. \tag{7}$$

The integral (6) is computed by the numerical integration (Rue et al., 2009), and the integral (7) is computed by Gauss-Hermite quadratures (Liu and Pierce, 1994) as the

conditional posterior density $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ will be approximated by a Gaussian distribution. The key to the accuracy of (7) is that the likelihood, $\pi(y_i|\eta_i, \boldsymbol{\theta})$, is known such that small approximation errors of $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ diminish due to the integration. The accuracy of (6) relies on the accuracy of (7) and the assumption that the removal of $\boldsymbol{y}_{I_i}$ does not have a dramatic impact on the posterior.

Algorithm 1: Find groups for all data points

```
Input: A correlation matrix choice R (posterior correlation is the default),
  Number of level sets m;
Output: A list containing the groups for all data points;
Calculate R from the model;
N ← number of rows in R;
groups ← initialize N empty lists;
for i = 1 to N do
    r ← absolute values of the i-th row of R;
    ordered indices ← indices of r sorted by value in decreasing order;
    current absolute correlation ← 1;
    k ← 1 ;
    for j = 1 to m do
        while current absolute correlation == r[ordered indices[k]] do
            groups[i].append(ordered indices[k]);
            k ← k + 1;
        end
        current absolute correlation ← r[ordered indices[k]];
    end
end
return groups;
```

The computation of the nested integrals reduces to the computation of $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$. We will approximate $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ by a Gaussian distribution, denoted by $\pi_G(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$, and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$ by correcting the approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ in Rue et al. (2009). We further improve the mean of $\pi_G(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ using variational Bayes (Van Niekerk and Rue, 2024) in the implementation. In this section, we focus on the explanation of computing $\pi_G(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ and an approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$.

Computing $\pi_G(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$

The mean and variance of $\pi_G(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ can be obtained by

$$\begin{aligned} \mu_{\eta_i}(\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}) &= \boldsymbol{A}_i \boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}), \\ \sigma^2_{\eta_i}(\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}) &= \boldsymbol{A}_i \boldsymbol{Q}_{\boldsymbol{f}}^{-1}(\boldsymbol{\theta}, \boldsymbol{y}_{-I_i}) \boldsymbol{A}_i^{\top}. \end{aligned} \tag{8}$$

The computation of $\pi_G(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ requires the mode of $\pi(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$ for each i at each configuration of $\boldsymbol{\theta}$, which is computationally expensive. With the mode at full data, we

use an approximation to avoid the optimization step,

$$\boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \boldsymbol{Q}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}) - \boldsymbol{A}_{I_i}^{\mathsf{T}}\boldsymbol{C}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{A}_{I_i}, \tag{9}$$

$$\boldsymbol{\mu}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \tilde{\boldsymbol{\mu}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})^{-1}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y}) - \boldsymbol{A}_{I_i}^{\mathsf{T}}\boldsymbol{b}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})), \tag{10}$$

where $\boldsymbol{A}_{I_i}$ is a submatrix of $\boldsymbol{A}$ formed by rows of $\boldsymbol{A}$, $\boldsymbol{b}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})$ is a subvector of $\boldsymbol{b}(\boldsymbol{\theta},\boldsymbol{y})$, and $\boldsymbol{C}_{I_i}(\boldsymbol{\theta},\boldsymbol{y})$ is a principal submatrix of $\boldsymbol{C}(\boldsymbol{\theta},\boldsymbol{y})$. When the posterior is Gaussian, the approximation is exact as (9) and (10) define the precision matrix and the mean of the posterior. It seems easy to obtain the moments using (8), but the decomposition of $\tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is too expensive. To avoid the decomposition of $\tilde{\boldsymbol{Q}}_{\boldsymbol{f}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$, we use the linear relation $\boldsymbol{\eta}_{I_i} = \boldsymbol{A}_{I_i}\boldsymbol{f}$ to map all the computation on $\boldsymbol{f}$ to $\boldsymbol{\eta}_{I_i}$. We compute $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ through $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ as shown in the Appendix A using a low rank representation, where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ is the posterior covariance matrix of $\boldsymbol{\eta}_{I_i}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is the covariance matrix of $\boldsymbol{\eta}_{I_i}$ with $\boldsymbol{y}_{I_i}$ left out. The computation of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{I_i}}(\boldsymbol{\theta},\boldsymbol{y})$ is non-trivial, especially when linear constraints are applied, which is demonstrated in Appendix B.

The approximation is more accurate when $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is close to Gaussian. The Gaussianity of $\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}$ comes from three sources. Firstly, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is nearly Gaussian, when η_i is connected to large amount of data (Rue et al., 2009). Secondly, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is dominated by the Gaussian prior, which happens when η_i is connected to very few data. Thirdly, the log-likelihood can be close to the log-likelihood of a Gaussian distribution, resulting in the Gaussianity of $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ due to the conjugacy. Thus, $\pi(\eta_i|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ is rarely far away from a Gaussian distribution.

Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$

To approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i})$, we use the relation,$\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-I_i}) \propto \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})}$, where we can approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ at configurations as in Rue et al. (2009). We need to compute $\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) \approx \int \pi(\boldsymbol{y}_{I_i}|\boldsymbol{\eta}_{I_i},\boldsymbol{\theta})\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})d\boldsymbol{\eta}_{I_i}$. A Laplace approximation can be applied to this integral,

$$\pi_{\mathrm{LA}}(\boldsymbol{y}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i}) = \frac{\pi(\boldsymbol{y}_{I_i}|\boldsymbol{\eta}^*_{I_i},\boldsymbol{\theta})\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})}{\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y})}, \tag{11}$$

where $\boldsymbol{\eta}^*_{I_i}$ is the mode of $\pi_G(\boldsymbol{\eta}^*_{I_i}|\boldsymbol{\theta},\boldsymbol{y})$. Note that the correction of the hyperparameter reuses $\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y}_{-I_i})$ and $\pi_G(\boldsymbol{\eta}_{I_i}|\boldsymbol{\theta},\boldsymbol{y})$.

5. Simulations and Applications

This section showcases two simulated examples and two real data applications. The code is available at https://github.com/zhedongliu/LGOCV.

We start with a simulation that tests the approximation accuracy in a multilevel model with various response types. Following this, a time series forecasting simulation is presented. This allows LGOCV results with automatically constructed groups to be compared to the LFOCV. We then delve into disease mapping, contrasting group constructions derived from various strategies. Finally, we apply our methodology to intricate models using a large dataset, as documented by Lowe et al. (2021). For the construction of the leave-out group for LGOCV, we used Algorithm 1 in Section 3. The procedures detailed in Sections 3 and 4 have been integrated into R-INLA, ensuring that all computational tasks in this section are executed through R-INLA.

Simulated Multilevel Model with Various Responses

This example is a simulation that demonstrates the accuracy of the approximation described in Section 4. The main purpose is to compare $\pi(y_i|\boldsymbol{y}_{-I_i})$ computed using an approximation in Section 4 and the same quantity computed using MCMC. Furthermore, we use automatic group construction with the number of level sets equal to 1, corresponding to predicting a data point from a new class.

We simulate data according to the following process. Initially, we simulate 10 class means, denoted as $\boldsymbol{s}$, from a standard normal distribution. Next, we compute 100 linear predictors, $\eta_i = \mu + s_{j(i)}$, where $\mu = \log(10)$ and $j(i)$ is a function mapping data index i to the group index j. For this function, we set $j(i) = \lceil \frac{i}{10} \rceil$, where the ceiling function, $\lceil x \rceil$, rounds a number up to the nearest integer. We generate responses according to the linear predictor and one of three response types; Gaussian, binomial and exponential. The mean of the Gaussian response is η_i, and the standard deviation is 0.1. We generate binomial responses with a success probability of $\frac{1}{1+e^{-\eta_j}}$ for 20 trials. The exponential responses are generated with a mean of e^{η_j}.

We consider the model,

$$\begin{aligned} &\log(\tau_s) \sim N(0, 10^{-4}), \quad \mu \sim N(0, 10^{-4}), \quad s_j|\tau_s \sim N(0, \tau_s), \\ &\eta_i = \mu + s_j, \quad y_i|\eta_i \sim \text{response model}(\eta_i), \end{aligned} \tag{12}$$

where the second parameter of the Gaussian distribution is the precision, and the likelihood is specified according to the data generation process with the given response model.

As a reference, we let the MCMC runs for 10^8 iterations, which makes the Monte Carlo errors negligible. The large size of MCMC samples is required because the predictive distributions are influenced by the tails of $\pi(\eta_i|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$. In Figure 2, (a), (c), and (d) show the data against its group index, which presents a clear group structure; (b), (d), and (e) show the comparison of $\pi(y_i|\boldsymbol{y}_{-I_i})$ obtained from the approximations and MCMC. We use Rstan (Stan Development Team, 2022) for the MCMC.

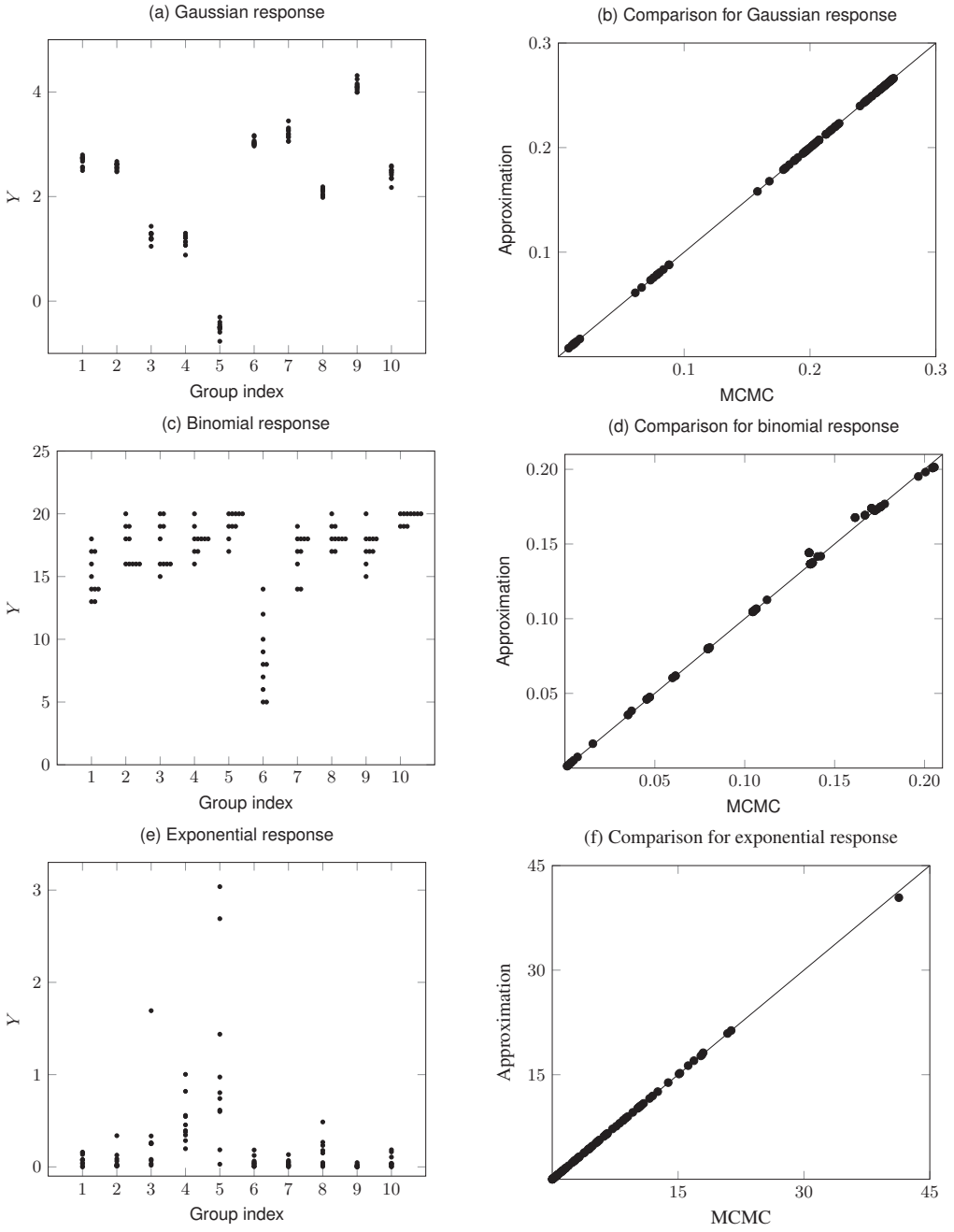


Figure 2. *Comparison of $\pi(y_i|\mathbf{y}_{-I_i})$ from approximations and MCMC. First column: y-axis shows response value, x-axis shows group index. Second column: y-axis shows LGOCV from proposed approximation, x-axis shows LGOCV from MCMC.*

This example shows that the approximations are highly accurate. When the response is Gaussian, the approximation almost equals the MCMC results, where the main difference is due to MCMC sampling errors, as our approach is exact up to numerical integration in this case. Also, under both non-Gaussian cases, the results are close to the long-run MCMC results.

Time Series Forecasting

In this example, we will demonstrate how the automatic LGOCV method can measure the forecasting performance of a time series model, while LOOCV is not effective in doing so.

We will first simulate 2000 data points using the following procedure: We will simulate an AR(1) time series by using $u_i = 0.9u_{i-1} + \varepsilon_{u_i}$, where ε_{u_i} follows a standard Gaussian distribution. Next, we will compute linear predictors by calculating $\eta_i = \mu + u_i$, with μ set to 2. Finally, the Gaussian responses have mean η_i, and a standard deviation of 0.1.

We fit a time series model on the simulated data:

$$\begin{aligned} &\mu \sim N(0, 10^{-4}), \quad \boldsymbol{u} \sim N(0, \boldsymbol{Q}_u), \\ &\eta_i = \mu + u_i, \quad y_i|\eta_i \sim N(\eta_i, 100), \end{aligned}$$

where $\boldsymbol{Q}_u$ is determined by an AR(1) model with the true parameters.

The prediction task is k steps forward forecasting for $k = \{1, 2, \ldots, 10\}$ using the true model. The natural cross-validation for these prediction tasks is LFOCV. To replicate the LFOCV, the group in LGOCV for testing point y_i and k steps forward prediction includes $\boldsymbol{y}_{(i-k+1):n}$. We can compute LFOCV for every k, denoted by LFOCV(k). To make the training set similar to the data set, the last 500 data points will be used as testing points, which means $i = \{1501, \ldots 2000\}$ in (1), and the quantity is averaged over 500 data points. We can also compute LGOCV using automatically constructed groups with the number of level sets, $m = \{1, 2, \ldots, 10\}$, denoted by LGOCV(m). In this setting, the automatically constructed group for a testing point y_i with a number of level sets equal to m includes $\boldsymbol{y}_{\max(1,i-m+1):\min(n,i+m-1)}$. Also, LGOCV(1) is equivalent to LOOCV in this model.

To compare LGOCV and LFOCV, we will fit a natural spline to have LFOCV(t) for t as a real number (see Figure 3 (a)) and map the number of level sets in LGOCV to the steps ahead in LFOCV (see Figure 3 (b)). We can see that LOOCV measures approximately 0.4 steps forward forecasting when the simplest prediction task is one step forward forecasting. LGOCV(2) represents roughly a one-step forward forecasting performance of the model. As the number of level sets increases in LGOCV, it represents more steps forward forecasting performance. Note that the specific translation between the automatic LGOCV and LFOCV is only valid in this model and may not be applicable in other models.

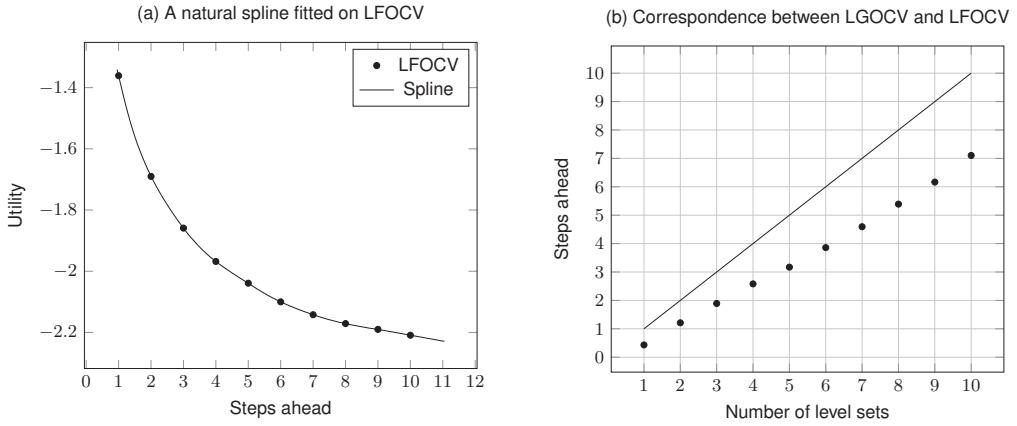


Figure 3. *Comparison of Automatic LGOCV and LFOCV. LOOCV measures approximately 0.4 steps forward forecasting. LGOCV(2) roughly represents a one-step forward forecasting performance.*

Disease Mapping

In this example, we will present groups constructed by different automatic group construction strategies. We will see the differences between those groups and get an idea to choose a proper group construction strategy.

We applied a disease mapping model to data detailing cancer incidence by location (Besag, York and Mollié, 1991; Wakefield, Best and Waller, 2000; Held et al., 2005). This dataset captures oral cavity cancer cases in Germany from 1986-1990 (Held et al., 2005). The response y_i indicates the cases in area i over five years. The case count in each region is influenced by its population and age distribution. The expected case count E_i in the region i is derived from its age distribution and population, ensuring $\sum_i y_i = \sum_i E_i$. Additionally, the covariate x_i represents tobacco consumption in area i.

We fit the following model on the data set:

$$\begin{aligned} y_i|\eta_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= \mu + f_{\text{rw}}(x_i) + u_i + v_i, \end{aligned} \tag{13}$$

where μ is an intercept, $\boldsymbol{u}$ is a spatially structured component, $\boldsymbol{v}$ is an unstructured component (Krainski et al., 2018), and $\boldsymbol{f}_{rw}$ is an intrinsic second-order random-walk model of the covariate x_i (Rue and Held, 2005).

In Figure 4, we illustrate groups formed through various automatic group construction strategies. The testing point is located in the black region, while the data in the group are located in grey areas. As seen in Figure 4 (a) and (b), Groups from $\boldsymbol{R}_{\text{prior}}$ focus solely on spatial effects. Groups from $\boldsymbol{R}_{\text{post}}$ exhibit mostly strong spatial patterns, such as Figure 4 (c). Yet, some points, like in Figure 4 (d), indicate non-spatial patterns. This arises as all model components, including fixed and random effects, priors, and the response variable, are considered.

The spatial patterns in posterior groups may justify incorporating spatial effects into the model, given that data retains this pattern in correlation. In practice, groups from $\boldsymbol{R}_{\text{post}}$ offer a more balanced representation. However, $\boldsymbol{R}_{\text{prior}}$ with selective effects resemble those from manually defined groups.

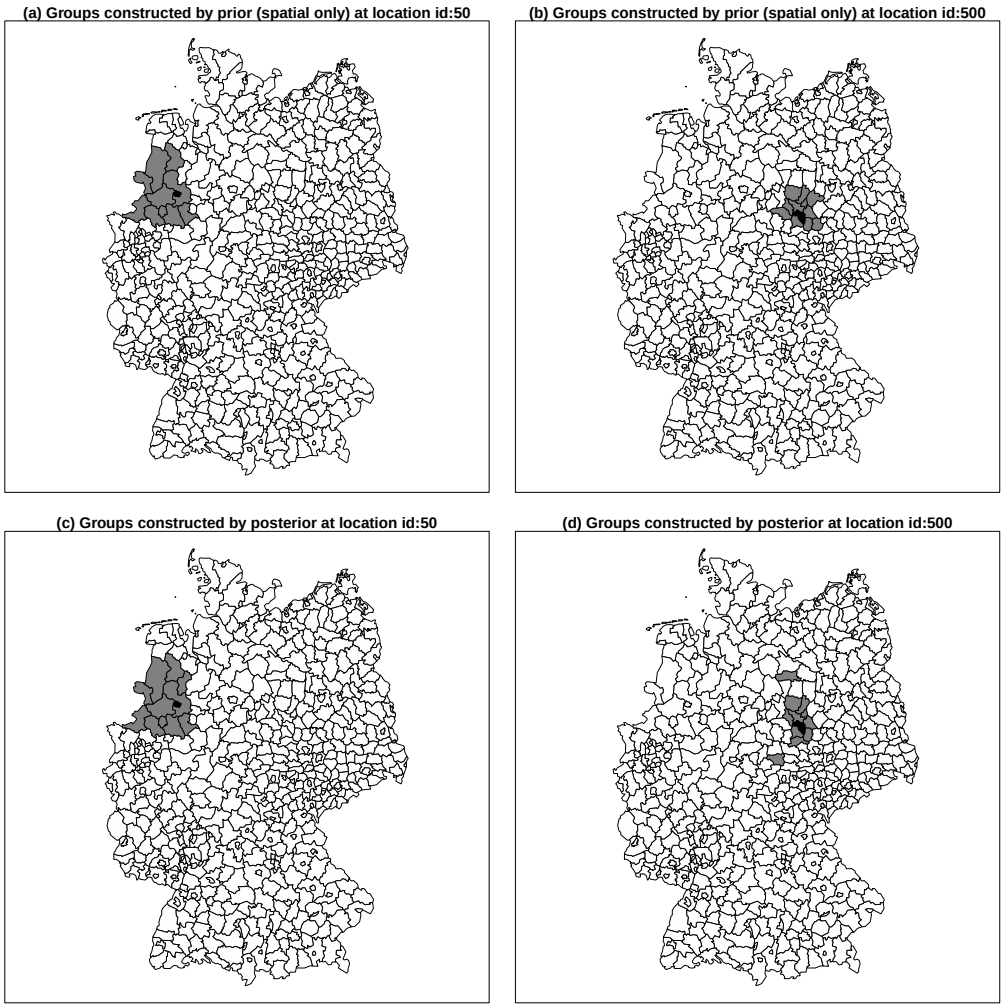


Figure 4. *Groups by different automatic construction strategies. The testing point is located in the black region, and the data in the group are located in the grey regions. In (d), the group constructed by posteriors contains some non-spatial patterns.*

Dengue Risk in Brazil

In this real-world example, we will demonstrate the scalability and adaptability of the automatic LGOCV method in a complex model structure and a large sample size. The automatically constructed model-based groups are consistent with the domain knowledge that dengue disease is prevalent in summer.

We will repeat the variable selection process as shown in Lowe et al. (2021) using the automatic LGOCV. The model chosen by LGOCV is considered to have better predictive power for longer-range predictions, than those selected based on other criteria because the most informative data points for predicting the target are excluded from the training set.

The models study the influence of extreme hydrometeorological hazards on dengue risk, factoring in Brazil's urbanization levels. Our dataset, with $127,224$ samples representing $12,895,293$ dengue cases, covers Brazil's 558 microregions from January 2001 to December 2019. Given the dataset's magnitude and the model's intricacy, LGOCV or LOOCV calculations require the approximation method detailed in Section 4.

Data points include month, year, microregion, and state. The candidate covariates encompass the monthly average of daily minimum (T_{min}) and maximum temperatures (T_{max}), the palmer drought severity index (PDSI), the urbanization levels: overall (u), centered at high (u_1), intermediate (u_2), and more rural levels (u_3) and the access to water supply: overall (w) and centered at high-frequency shortages (w_1), intermediate (w_2), and low-frequency shortages (w_3). To preprocess these covariates' specifics, refer to Lowe (2021).

The data generating model is chosen to be negative binomial, to account for overdispersion. The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component (Riebler et al., 2016). The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

```
y ~ 1 + covariates + f(month, model = "rw1", replicate = state, cyclic = TRUE)
    +f(microregion, model = "bym2", replicate = year).
```

In short, we write this model as $\mathrm{y} \sim 1 + \mathrm{covariates} + f_t + f_s$. The number of parameters in this model is $21,567$ with $127,224$ observations for the full model. The appendix of Lowe et al. (2021) and its repository Lowe (2021) provide full details about the models and data.

The model accounts for temporal effects with spatial replicates and spatial effects with temporal replicates, complicated by various constraints. Given its intricacy and the lack of a clear prediction task, crafting groups for LGOCV manually is challenging. Hence, utilizing our automatic group construction through posterior correlation is beneficial. For model comparisons, using the same groups across different models is recommended. The base model, which only incorporates structured components, is chosen for group building. Most automatic groups cluster data from the same year, location, and nearby months to the testing points. Figure 5 displays the relative month frequencies in the group, given the testing points correspond to a specific month. The chart suggests the first half-year data better informs predictions. Even in July and November testing

points, the group frequently includes that data, aligning with the known prevalence of dengue during summer. See Figure 5 (c) and (d) for details.

The results of model selection using deviance information criterion (DIC), LOOCV, and LGOCV ($m = 2,3,4$) are presented in Table 1. The candidate models are those referenced in Lowe (2021). To transform equation 1 into a loss function, we calculated its negative value.

From Table 1 we note that LOOCV prefers the spatio-temporal model that incorporates access to water while the spatio-temporal model with urbanization as a covariate, is preferred by LGOCV. This result is interesting since we can conclude that the same model might not necessarily perform well for short- and longer-range prediction. The practitioner thus needs to decide what the goal of the modeling is, and then choose the model to be used accordingly. If we want to predict dengue risk for a new unobserved area or time point, it seems that urbanization has a better prediction ability than access to water. Note also that as we increase the number of level sets, we are defining a prediction target with an increased range, thus moving further away from LOOCV.

Table 1. *Comparative evaluation of models for predicting variable y based on various environmental factors. This table presents the model selection results, including each model's Deviance Information Criterion (DIC), LOOCV, and LGOCV scores.*
Note: We offset DIC by 826841.66, LOOCV by 3.2721, LGOCV ($m = 2$) by 3.314, LGOCV ($m = 3$) by 3.3763 and LGOCV ($m = 4$) by 3.4372.

Index	Model	DIC	LOOCV	LGOCV		
				($m = 2$)	($m = 3$)	($m = 4$)
1	$y \sim 1 + f_t + f_s$	3615.38	0.0151	0.0158	0.0206	0.0270
2	$y \sim 1 + T_{min} + f_t + f_s$	1562.96	0.0064	0.0067	0.0088	0.0098
3	$y \sim 1 + T_{max} + f_t + f_s$	2228.73	0.0091	0.0098	0.0133	0.0163
4	$y \sim 1 + PDSI + f_t + f_s$	2167.12	0.0092	0.0095	0.0126	0.0184
5	$y \sim 1 + PDSI + T_{min} + f_t + f_s$	160.43	0.0006	0.0006	0.0012	0.0023
6	$y \sim 1 + PDSI + T_{max} + f_t + f_s$	900.65	0.0038	0.0038	0.0057	0.0084
7	$y \sim 1 + PDSI + T_{min} + PDSI * u_1 + u + f_t + f_s$	38.21	0.0002	0*	0*	0*
8	$y \sim 1 + PDSI + T_{min} + PDSI * u_2 + u + f_t + f_s$	39.13	0.0002	0*	0*	0*
9	$y \sim 1 + PDSI + T_{min} + PDSI * u_3 + u + f_t + f_s$	28.64	0.0002	0*	0*	0*
10	$y \sim 1 + PDSI + T_{min} + PDSI * w_1 + w + f_t + f_s$	6.68	0*	0.0005	0*	0.0014
11	$y \sim 1 + PDSI + T_{min} + PDSI * w_2 + w + f_t + f_s$	0*	0*	0.0005	0*	0.0015
12	$y \sim 1 + PDSI + T_{min} + PDSI * w_3 + w + f_t + f_s$	4.55	0*	0.0006	0*	0.0014

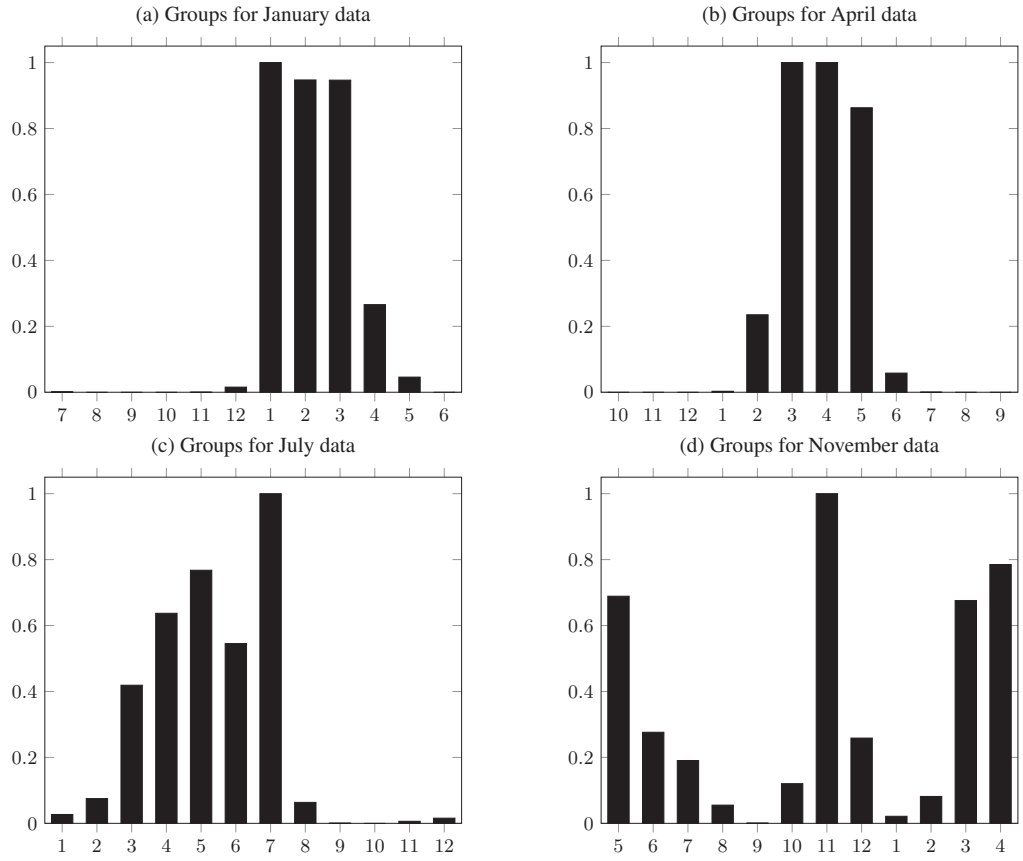


Figure 5. *Groups for testing points from a specific month. y-axis: relative frequency, x-axis: month of data measurement in groups. The first half-year data are more informative for prediction. As shown in (c) and (d), even in July and November, the group often includes data consistent with the known summer prevalence of dengue. Note that dengue is prevalent in the summer months which are approximately November to February.*

6. Discussion

An over-reliance on LOOCV to evaluate predictive capacity in general persists in statistical practice, despite concerns raised in studies such as Roberts et al. (2017); Vehtari et al. (2019). LOOCV can provide an evaluation of short-range predictive ability with well-established asymptotics for some models. On the other hand, what can we do to evaluate the longer-range predictive ability of complex models? Various approaches for specific models, such as time series or spatial models have been proposed, where custom CV procedures are designed to mimic a longer-range prediction task than that of LOOCV. We have introduced an automated approach for evaluating the longer-range prediction ability of any latent Gaussian model, namely LGOCV. LGOCV is designed to be applicable to all models that are latent Gaussian models, and thus provides a framework

for general longer-range predictive ability evaluation without the need for case-by-case considerations. Moreover, we propose a computationally efficient approach to calculate LGOCV scores and metrics based on the INLA methodology. We have shown that our approximate LGOCV implementation is almost exact when compared with the results from MCMC, albeit at a much lower computational cost. This enables practitioners to use the LGOCV approach for complex models and large data.

Our approach is designed for latent Gaussian models and some ideas can thus be extended to the case of non-latent Gausian models with careful consideration of the computational cost associated with this endeavor. LGOCV for LGMs is computationally efficient in INLA, since it is fully parallelizable by computing the necessary quantities only at the mode of the hyperparameters. For huge data ($n > 10^6$) however, the cost will be high since the cost increases linearly in n, albeit much lower than other available approaches. For huge data, performing CV on a subset or constructing the groups manually could be considered. Nonetheless, for LGMs, the proposed LGOCV could be considered as the most feasible approach for longer-range predictive ability evaluation.

The choice of the number of level sets determine the prediction task and thus the degree of independence between the leave-out group and the rest of the data. There is not a one-to-one correspondence between the number of level sets and the number of points to leave out as shown in the simulations and applications, although a higher number of level sets would imply a longer range for the prediction task, than a lower number. The choice of the number of level sets remains arbitrary since it is a user-defined parameter, we recommend a low number like $m = 2$ or $m = 3$ if there is no clear indication of what else m should be. There exists no optimal value of m in general, since it would imply different prediction tasks for different levels of dependency. In our applications, and those of others who have applied the LGOCV framework, it is shown that LGOCV provides the information we need to evaluate longer-range prediction ability, and complements the information from LOOCV.

It is pertinent to note that the proposed LGOCV do not replace a custom CV strategy designed by modelers, tailored for specific applications. We pose it as an alternative default strategy for longer-range prediction ability evaluation, that complements LOOCV in assessing the predictive ability of an LGM, while being computationally efficient and practical for real-world scenarios.

Acknowledgments

The authors thank D. Castro-Camilo, D. Rustand, and E. Krainski for valuable discussions and suggestions.

A. On the computation of $\Sigma_{\eta_{I_i}}(\theta, y_{-I_i})$ and $\mu_{\eta_{I_i}}(\theta, y_{-I_i})$

In this section, we let I_i be I and drop θ to simplify the notation. We have a random vector $\eta_I|y \sim N(\mu_{\eta_I}(y), \Sigma_{\eta_I}(y))$, which can be viewed as a posterior distribution with prior $\eta_I|y_{-I} \sim N(\mu_\eta(y_{-I}), \Sigma_\eta(y_{-I}))$ and likelihood $\pi_G(y_I|\eta_I) \propto \exp\left\{-\frac{1}{2}\eta_I^T C(y_I)\eta_I + b(y_I)\eta_I\right\}$. Now, we need to use the posterior and the likelihood to obtain the prior.

If $\Sigma_{\eta_I}(y)$ is full rank, we have $Q_{\eta_I}(y) = \Sigma_{\eta_I}(y)^{-1}$ and $b_{\eta_I}(y) = Q_{\eta_I}(y)\mu_{\eta_I}(y)$. By conjugacy of Gaussian prior and Gaussian likelihood, $Q_{\eta_I}(y_{-I}) = Q_{\eta_I}(y) - C(y_I)$ and $b_{\eta_I}(y_{-I}) = Q_{\eta_I}(y)\mu_{\eta_I}(y) - b(y_I)$. Then we have desired $\mu_{\eta_I}(y_{-I})$ and $\Sigma_{\eta_I}(y_{-I})$.

If $\Sigma_{\eta_I}(y)$ is singular, we let $\eta|y = Bz|y$, where $B = V\Lambda$ with V containing eigenvectors corresponding to non-zero eigenvalues, Λ containing square root of non-zero eigenvalues on its diagonal, and $z|y \sim N(\mu_z(y), \mathcal{I})$, where $\mathcal{I}$ is an identity matrix and $\mu_z(y) = B^\intercal \mu_{\eta_I}(y)$. By conjugacy, we have $Q_z(y_{-I}) = \mathcal{I} - B^\intercal C(y_I)B$ and $b_z(y_{-I}) = \mu_z(y) - B^\intercal b(y_I)$. It is followed by $\mu_z(y_{-I}) = Q_z(y_{-I})^{-1} b_z(y_{-I})$. Then mean and covariance of $z|y_{-I}$ is $\mu_\eta(y_{-I}) = B\mu_z(y_{-I})$, $\Sigma_\eta(y_{-I}) = B\Sigma_z(y_{-I})B^\intercal$.

B. On the computation of $\Sigma_{\eta_{I_i}}(\theta, y)$ and $\mu_{\eta_{I_i}}(\theta, y)$ with Linear Constraints

We start by illustrating how to compute $\Sigma_{\eta_{I_i}}(\theta, y)$ and $\mu_{\eta_{I_i}}(\theta, y)$ without linear constraints. $\mu_{\eta_{I_i}}(\theta, y)$ is simply obtained by $\mu_{\eta_{I_i}}(\theta, y) = A_{I_i}\mu_f(\theta, y)$. However, we never store large dense matrix like $Q_f(\theta, y)^{-1}$. Thus, $\Sigma_{\eta_{I_i}}(\theta, y)$ cannot be obtained by using matrix multiplication $A_{I_i} Q_f(\theta, y)^{-1} A_{I_i}^\intercal$. Instead, we compute $\Sigma_\eta(\theta, y)$ entry by entry and use the result to fill in entries of $\Sigma_{\eta_{I_i}}(\theta, y)$. We compute $\Sigma_\eta(\theta, y)_{i,j}$ by solving

$$Q_f(\theta, y)x = A_i$$

and $\Sigma_\eta(\theta, y)_{i,j} = A_j x$. The computation is fast because A and $Q_f(\theta, y)$ are sparse, and the factorization of $Q_f(\theta, y)$ is reused.

When linear constraints $\mathcal{C}f = e$ are applied on f, we have

$$\Sigma_f(\theta, y)^* = Q_f(\theta, y)^{-1} - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}\mathcal{C}Q_f(\theta, y)^{-1},$$
$$\mu_f(\theta, y)^* = \mu_f(\theta, y) - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}(\mathcal{C}\mu_f - e),$$

where $\Sigma_f(\theta, y)^*$ and $\mu_f(\theta, y)^*$ are the mean and the covariance matrix after applying constraints (Rue and Held, 2005). Because $\mu_f(\theta, y)^*$, is always stored, the computation of $\mu_{\eta_{I_i}}(\theta, y)$ is simple. We need to propagate the effects of linear constraints to $\Sigma_\eta(\theta, y)_{i,j}$. This is achieved by computing (Rue and Held, 2005)

$$x^* = x - Q_f(\theta, y)^{-1}\mathcal{C}^\intercal(\mathcal{C}Q_f(\theta, y)^{-1}\mathcal{C}^\intercal)^{-1}\mathcal{C}x,$$

where x solves $Q_f(\theta, y)x = A_i$. Then $\Sigma_\eta(\theta, y)^*_{i,j} = A_j x^*$.

References

Adin, A., Krainski, E. T., Lenzi, A., Liu, Z., Martínez-Minaya, J., and Rue, H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, page 100843.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó Location Budapest, Hungary.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for bayesian time series models. *Journal of Statistical Computation and Simulation*, 90(14):2499–2523.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied statistics*, 28(1):100–108.

Held, L., Natário, I., Fenton, S. E., Rue, H., and Becker, N. (2005). Towards joint disease mapping. *Statistical methods in medical research*, 14(1):61–82.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

Lowe, R. (2021). Data and R code to accompany 'Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study'.

Lowe, R., Lee, S. A., O'Reilly, K. M., Brady, O. J., Bastos, L., Carrasco-Escobar, G., de Castro Catão, R., Colón-González, F. J., Barcellos, C., Carvalho, M. S., et al. (2021). Combined effects of hydrometeorological hazards and urbanisation on dengue risk in brazil: a spatiotemporal modelling study. *The Lancet Planetary Health*, 5(4):e209–e219.

McQuarrie, A. D. and Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific.

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84:802–829.

Rabinowicz, A. and Rosset, S. (2022). Cross-validation for correlated data. *Journal of the American Statistical Association*, 117(538):718–731.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of econometrics*, 99(1):39–61.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., and Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience*, 6(5):gix019.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

Stan Development Team (2022). RStan: the R interface to Stan. R package version 2.21.5.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Biorxiv*, page 357798.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for bayesian inference with inla. *Computational Statistics & Data Analysis*, 181:107692.

Van Niekerk, J. and Rue, H. (2024). Low-rank variational bayes correction to the laplace method. *Journal of Machine Learning Research*, 25(62):1–25.

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Vehtari, A., Simpson, D. P., Yao, Y., and Gelman, A. (2019). Limitations of "limitations of bayesian leave-one-out cross-validation for model selection". *Computational Brain & Behavior*, 2(1):22–27.

Wakefield, J. C., Best, N., and Waller, L. (2000). Bayesian approaches to disease mapping. *Spatial epidemiology: methods and applications*, pages 104–127.

Wang, X., Yue, Y. R., and Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. CRC Press.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.

Information for authors

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.

Authors must indicate the **source of funding** for the articles.

Authors can **preprint** their manuscripts during the submission process and showcase their work to the global research community, before it is accepted or published. This can be done in any non-commercial preprint server such as **archiv**.

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.

The Library of Catalonia (depending on the Government of Catalonia) periodically records SORT website and stores it indefinitely in the repository **PADICAT** (Patrimoni Digital de Catalunya: https://www.padicat.cat/)

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

Creative Commons License

Copyright notice and author opinions

Authors transfer the exploitation rights of their works to the journal. The Institut d'Estadística de Catalunya holds the copyright ownership of the contents published in the journal. Authors may deposit a copy of their works in repositories, as specified in the self-archiving policy. Published articles represent the author's opinions; the journal SORT does not necessarily agree with the opinions expressed in the published articles.

Self-archiving policy

The journal SORT allows the deposit and dissemination of the published version of the contributions in any institutional, subject and/or multidisciplinary repository. The repository must contain the information about the publication in the journal and the corresponding link. The journal allows the deposit and dissemination of article preprints, but recommends being linked to the published version.

Statement of ethics and good practices

As a journal co-edited by the Universitat de Barcelona, SORT declares that follows its "Declaration of Ethics and Good Practices for Scientific Journals" (https://diposit.ub.edu/dspace/handle/2445/97665)

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76
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.